УДК 681.321

Г.Ф. КРИВУЛЯ, А.А. ДАВЫДОВ

Харьковский национальный университет радиоэлектроники, Украина

ОПТИМИЗАЦИЯ БИНАРНЫХ РЕШАЮЩИХ ДЕРЕВЬЕВ ПРИ ИНТЕЛЛЕКТУАЛЬНОЙ ДИАГНОСТИКЕ КОМПЬЮТЕРНЫХ СИСТЕМ

В данной статье рассматривается оптимизация бинарных деревьев решений путем уменьшения размерности дерева и определения несущественных атрибутов (диагностических признаков), которые не влияют на процесс принятия решения о состоянии компьютерной системы. Оптимизированное дерево решений позволяет решать задачи классификации состояния объекта в системах диагностирования с меньшими временными и аппаратурными затратами. Применение теории ПФ для оптимизации БДР позволяет минимизировать число атрибутов (диагностических признаков) за счет исключения несущественных атрибутов, которые не влияют на принятие решений при классификации состояний компьютерной системы.

Ключевые слова: *бинарные деревья решений, минимизация, диагностические признаки, компьютерные системы, переключательная функция, оптимизация.*

Введение

В настоящее время интеллектуальные системы, использующие знания экспертов, стали неотъемлемыми компонентами современных автоматизированных систем различного назначения. При этом за последние десятилетия получили широкое распространение экспертные диагностические системы для различных сфер применения – от медицинских объектов до атомных электростанций [1, 2]. Эффективность таких систем, трудоемкость их проектирования, эксплуатации и развития, их устойчивость к изменению предметной области зависят от средств, использованных для представления знаний и методов обработки этих знаний. В качестве основных моделей представления знаний в интеллектуальных системах используются системы продукций, фреймовые структуры, семантические сети и логические системы.

Основные трудности при проектировании интеллектуальных диагностических систем связаны с тем, что такие системы разрабатываются для плохо формализованных предметных областей, в которых знания неточны, неполны, противоречивы и изменчивы. Это делает необходимым разработку эффективных методов представления и обработки таких знаний.

В частности, возникает необходимость пополнения, обобщения и классификация диагностической информации. При этом наиболее удобной формой представления знаний для компьютерной обработки являются логические системы в виде деревьев решений [3].

Деревья решений (Decision Trees) являются удобным инструментом в системах поддержки принятия решений интеллектуального анализа данных (Data Mining) и являются одним из наиболее популярных средств для задач классификации в диагностических системах Они создают иерархическую структуру классифицирующих правил типа "если … то …" (If -Then), имеющую вид дерева. Конечными узлами дерева являются "листья", соответствующие найденным решениям и объединяющие некоторое количество объектов классифицируемой выборки.

Область применения деревьев решений в интеллектуальных диагностических системах довольно значительна, но все задачи, решаемые этим методом, могут быть объединены в три следующие группы:

1. Описание данных: деревья решений позволяют хранить информацию о данных в компактной форме, т.е. вместо обширных таблиц данных мы можем хранить дерево решений, которое содержит в концентрированной форме точное описание объектов.

2. Классификация: деревья решений отлично справляются с задачами классификации, т.е. отнесения объектов к одному из заранее известных классов; при этом целевая переменная должна быть измерена в порядковой шкале.

3. Регрессия: если целевая переменная имеет непрерывные значения, деревья решений позволяют установить зависимость целевой переменной от независимых (входных) переменных. Например, к этому классу относятся задачи численного прогнозирования (предсказания значений целевой переменной).

Для диагностирования компьютерной системы (КС), которая представляет собой сложный технический объект, успешно используются бинарные деревья решений (БДР, Binary Decision Trees) [4]. При проектировании диагностического обеспечения КС первоначально составляется диагностическая матрица, которая содержит двоичные диагностические признаки и возможные состояния объекта. По данной матрице конструируется БДР, затем на основе анализа дерева решается задача классификации, т.е. принимается решение о работоспособности отдельных компонентов КС.

В связи с большой размерностью БДР для реальных технических объектов актуальной задачей являются методы оптимизации построенных деревьев решений. В [5] для такой оптимизации используется генетический алгоритм, с использованием которого значительно уменьшается время оптимизации.

Целью настоящей работы является оптимизация БДР за счет уменьшении размерности дерева и определения несущественных атрибутов (диагностических признаков), которые не влияют на процесс принятия решения о состоянии КС.

Оптимизация деревьев решений с использованием теории ПФ

Представим БДР в виде переключательной функции (ПФ) $f(x_1, x_2, ..., x_n)$ n аргументов, где n – количество диагностических признаков, а функция f(X),которая описывает состояния КС. Как правило, диагностическая матрица, по которой строится БДР, задана не на всех 2^n двоичных наборах и ПФ в этом случае является неполностью определенной. Данное обстоятельство позволяет выполнить минимизацию ПФ (оптимизацию БДР).

Очевидно, что если ПФ $f(x_1, x_2, ..., x_n)$ n аргументов имеет m определенных наборов, на которых функция принимает значения, равные нулю или единице, то при $m < 2^n$ существуют $p = 2^n - m$ неопределенных наборов значений аргументов [6]. Доопределение этих наборов двоичными значениями имеет 2^p вариантов, из которых целёсообразно выбрать такое, которое обеспечивает минимальное значение ПФ по сравнению с другими вариантами.

Для сложных ПФ большой размерности одним из перспективных методов получения минимальной ПФ является скобочная форма (СФ) ПФ, которая по своей сути совпадает с БДР.

Получение скобочной формы для полностью определенных ПФ основано, как правило, на принципе функциональной декомпозиции. Примером декомпозиции является представление функции $f(x_1, x_2, ..., x_n)$ в виде разложения по i-му аргументу в следующей форме:

$$f(X) = \overline{x}_i f(x_1, x_2, ..., x_{i-1}, 0, x_{i+1}, ..., x_n) \vee \vee x_i f(x_1, x_2, ..., x_{i-1}, 1, x_{i+1}, ..., x_n). \quad (1)$$

Определение 1. Функция $f(x_1, x_2, ..., x_{i-1}, 0, x_{i+1}, ..., x_n)$ называется $\overline{x}_i$ -компонентой исходной функции $f(X)$. Обозначим $\overline{x}_i$ – компоненту как $\overline{f^{x_i}}(x_1, x_2, ..., x_n)$. Соответственно функцию

$$f(x_1, x_2, ..., x_{i-1}, 1, x_{i+1}, ..., x_n)$$

назовем x_i -компонентой $f(X)$ и обозначим ее

$$f^{x_i}{}_i(x_1, x_2, ..., x_n).$$

Определение 2. Переменная x_i называется существенной для функции $f(X)$, если $\overline{f^{x_i}} \neq f^{x_i}$. Если же $\overline{f^{x_i}} = f^{x_i}$ то переменная x_i называется фиктивной.

В графическом виде разложение исходной ПФ по ее переменным осуществляется следующим образом. Внутри каждого узла БДР записывается переменная, по которой производится разложение функции. Нижнее левое ребро каждого узла реализует $\overline{x}_i$ -компоненту входной функции узла, а правое ребро – x_i -компоненту этой функции.

Определение 3. БДР функции $f(x)$ называется регулярной, если внутри каждого ряда узлы имеют одинаковые номера аргументов ПФ.

Ограничимся рассмотрением только регулярных БДР. Входными для узлов нижнего ряда являются значения функции $f(x)$ на соответствующих наборах переменных. Для неполностью определенной ПФ это могут быть 0,1 или x, при доопределении x равен 0 или 1.

Рассмотрим алгоритм получения минимальной формы ПФ на примере диагностической матрицы, содержащей 4 диагностических признака. В процессе обучения матрица определена на 7 двоичных наборах, остальные 9 наборов неопределенны (табл. 1).

Таблица 1

Обучающая выборка

№	X_1	X_2	X_3	X_4	f
0	0	0	0	0	0
4	0	1	0	0	1
5	0	1	0	1	1
11	1	0	1	1	0
12	1	1	0	0	1
14	1	1	1	0	1
15	1	1	1	1	0

БДР, построенное без процедур оптимизации и соответствующая данной матрице, приведено на рис. 1.

Используя данное бинарное дерево решений, получим:

$$f(x_1, x_2, x_3, x_4) = \bar{x}_1\bar{x}_2 \vee x_1x_2\bar{x}_3 \vee x_1x_2x_3\bar{x}_4 . \quad (2)$$

Рассмотрим обучающую выборку в виде карты Карно представленной в табл. 2.

Таблица 2

Карта Карно

X_1X_2	X_3X_4			
	00	01	11	10
00	0	x	x	x
01	1	1	x	x
11	1	x	0	1
10	x	x	0	x

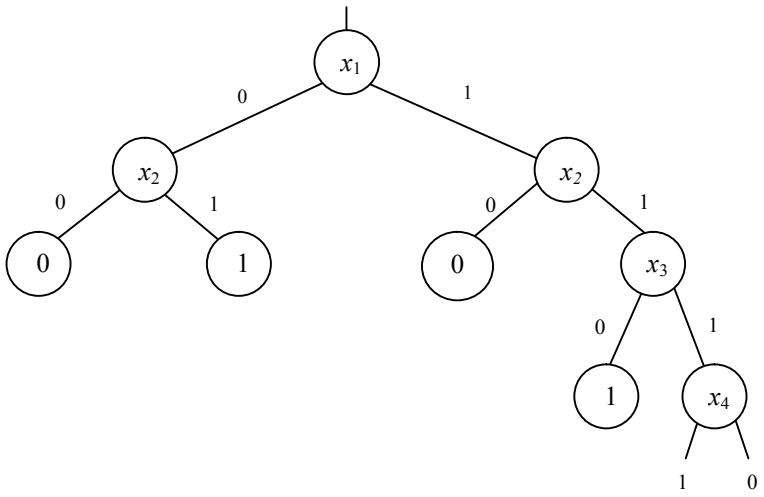

Рис. 1. БДР ПФ

Данное представление дает возможность получения минимальной формы представления. Доопределим наборы 1101 и 0110 единицами и получим:

$$f_{min} = X_2\overline{X_3} \vee X_2\overline{X_4} = X_2(\overline{X_3} \vee \overline{X_4}) \quad (3)$$

Рассмотрим представление БДР в виде ПФ четырех аргументов. Пусть ПФ задана таблицей истинности (табл. 3) соответствующей обучающей выборке представленной в табл. 2.

На рис. 2 приведена БДР данной ПФ: входами нижнего яруса являются значения ПФ в порядке возрастания наборов. Входным значением для набора $\bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4$ является $f^{\bar{x}_4}$ — компонента функции, равная нулю. Второй путь соответствует первому набору — $\bar{x}_1\bar{x}_2\bar{x}_3x_4$. Крайний правый путь на БДР соответствует пятнадцатому набору ПФ $x_1x_2x_3x_4$. Для него значение функции равно нулю.

Возможные входы ячеек первого ряда — это различные паросочетания из трехэлементного множества входных значений ПФ {0, 1, x}.

При этом выходные функции узлов нижнего ряда обозначим следующим образом:

Входная функция	0	x_i	$\bar{x}_i$	1	a	b	c	d	x
Входное значение	00	01	10	11	0x	x0	1x	x1	xx

Отметим, что для совпадающих входных значений 00, 11, xx переменная x фиктивна и выходные функции соответственно равны 0, 1, x.

Запишем по БДР скобочную форму ПФ четырех аргументов, доопределяя все неопределенные наборы нижнего ряда нулями: $f_0 = x_2(\bar{x}_1\bar{x}_3 \vee x_1\bar{x}_4)$.

Данная форма содержит 5 букв и не является минимальной. Доопределим теперь наборы таким образом, чтобы обеспечить максимально простую форму ПФ (рис. 2). Получим функцию $f_0 = x_2(\bar{x}_1 \vee \bar{x}_4)$, которая содержит уже 3 буквы.

Очевидно, чем больше в БДР ПФ фиктивных переменных, тем проще конечный вид скобочной формы ПФ. Добиться увеличения фиктивных переменных в БДР можно двумя путями: оптимальным доопределением значений функции на неопределенных входных наборах и перестановкой переменных в рядах БДР.

Целесообразно иметь алгоритм перенаправленного перебора переменных в ярусах $x_2(\bar{x}_1 \vee \bar{x}_4)$ БДР, чтобы исключить перебор всех возможных вариантов расположения переменных при решении данной задачи.

Введем количественную оценку оптимальности БДР с помощью понятия сходности БДР.

Таблица 3

Таблица истинности

Переменные, функция	№ наборов															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
x_2	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
x_3	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
x_4	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
$f(x_1, x_2, x_3, x_4)$	0	x	x	x	1	1	x	x	x	x	x	0	1	x	1	0

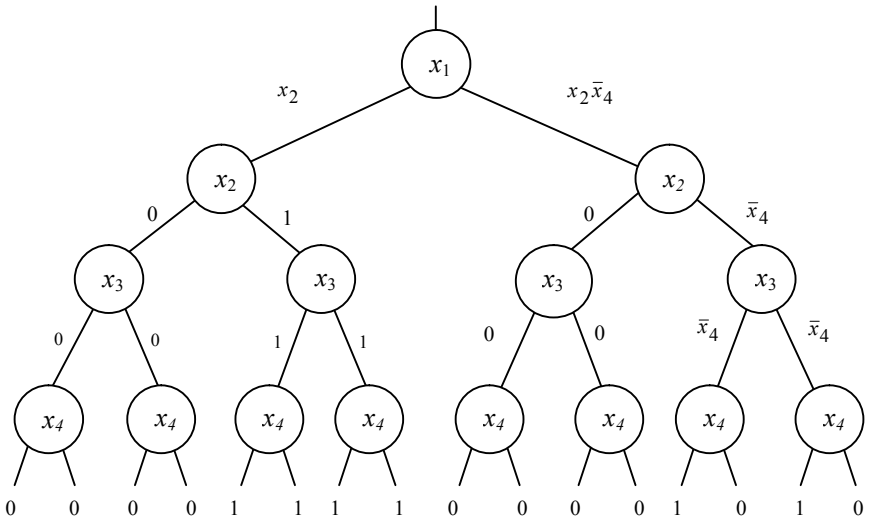

Рис. 2. Дерево решений

Определение 4. Сходностью БДР переключательной функции n аргументов назовем следующую сумму:

$$S = \sum_{i=1}^{n} k_i \cdot 2^{i-1}, \qquad (4)$$

где k_i, – число сходных узлов в i-м ярусе.

Сходными являются узлы, для которых обе входные функции одинаковы.

Если для i-го яруса выполняется условие $k_i = 2^{n-i}$, то переменная x_i фиктивна, так как все узлы i-го яруса сходны..

БДР на рис. 2 имеет сходность:

$$S = 6 \cdot 2^0 + 4 \cdot 2^1 = 14.$$

Для определения сходных узлов БДР используем операцию сравнения возможных входных двоичных значений узла: 0, 1 x, y в соответствии с правилами, приведенными в табл. 4.

Таблица 4

Правила сравнения
взаимно дополнительных ячеек

~	0	1	x	y
0	1	0	1	0
1	0	1	1	0
x	1	1	1	1
y	0	0	1	1

Рассмотрим алгоритм нахождения БДР с максимальной сходностью.

Для БДР ПФ четырех аргументов, приведенной на рис. 2, таблица Венна имеет вид, показанный в табл. 5. Ячейки таблицы Венна, двоичные номера наборов которых отличаются в одном разряде, соответствующего переменной x_i назовем взаимно

дополнительными по x_i. Две взаимно дополнительные ячейки сходны, если в них записано одинаковое значение булевой переменной.

Для вычисления сходности в каждом ярусе БДР по каждой переменной необходимо осуществлять сравнение всех взаимно дополнительных ячеек таблицы Венна:

$$k_i(x_i) = \sum_{j=1}^{2^{n-i}} z_j (i = 1,...,n), \qquad (5)$$

где z_j — результат сравнения взаимно дополнительных ячеек по x_j.

Для ПФ (табл. 5) получим БДР, представленную на рис. 3.

Таблица 5

Таблица Венна для БДР ПФ

x_1x_2	x_3x_4			
	00	01	10	11
00	0	x	x	x
01	1	1	x	x
10	x	x	x	0
11	1	x	1	0

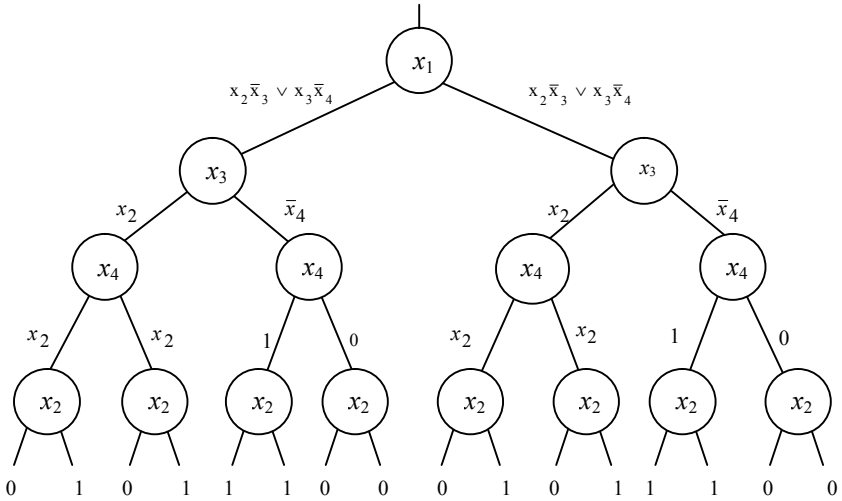

Рис. 3. БДР ПФ

Доопределим входные значения ПФ при выбранном распределении переменных по ярусам БДР.

Доопределение реализуется путем использования правила покоординатного пересечения булевых переменных [6].

После доопределения получаем полный вектор входных значений: 01 01 11 00 01 01 11 00, который используется для получения минимальной формы ПФ. Запишем для БДР, представленной на рисунке 3, дизъюнктивную форму ПФ четырех аргументов:

$$f = x_2\bar{x}_3 \vee x_3\bar{x}_4 .$$

Выводы

Применение теории ПФ для оптимизации БДР позволяет минимизировать число атрибутов (диагностических признаков) за счет исключения несущественных атрибутов, которые не влияют на принятие решений при классификации состояний компьютерной системы.

Если все атрибуты дерева существенны и их число невозможно уменьшить, то минимизация осуществляется путем оптимального расположения атрибутов дерева по его ярусам, что дает более простую форму ПФ и более простое дерево.

Оптимизированное дерево решений позволяет решать задачи классификации состояния объекта в системах диагностирования с меньшими временными и аппаратурными затратами.

Литература

1. Krivoulya G. Fuzzy expert system for diagnosis of computer failures / G.Krivoulya, Z.Dudar, D.Kucherenko // Proceeding of the 10th International Conference CADSM'2009, Ukraine – 2009. – P. 225-230.

2. Кривуля Г.Ф., Кучеренко Д.Е. Интеллектуальные средства диагностирования состояний компьютерных систем управления / Г.Ф. кривуля, Д.Е. Кучеренко // Інфомаційно-керуючі системи на залізничному транспорт. – 2009. – № 4. – С. 23-28.

3. Breiman L. Classification and Regression Trees / L. Breiman, J.H. Friedman, R.A. Olshen, C.T. Stone. – Wadsworth, Belmont, California, 1984.

4. Morris Rosenthal. Computer Repair with Diagnostic Flowcharts / Rosenthal Morris // Foner Books. – 2004. – P. 1-113.

5. Sung-Hyuk Cha. Genetic algorithm for constructing compact binary decision trees / Cha Sung-Hyuk, Tappert Charles // Journal of pattern recognition research. – 2009. – 1. – P. 1-13.

6. Кривуля Г.Ф. Минимизация неполностью определенных переключательных функций с помощью граф-схем / Г.Ф. Кривуля // АСУ и приборы автоматики. – 1981. – № 57. – С. 87-96.

Поступила в редакцию 10.02.2010

Рецензент: д-р техн. наук, проф., проф. кафедры В.И. Хаханов, Харьковский национальный университет радиоэлектроники, Харьков.

ОПТИМІЗАЦІЯ БІНАРНИХ ВИРІШАЛЬНИХ ДЕРЕВ ПРИ ІНТЕЛЕКТУАЛЬНІЙ ДІАГНОСТИЦІ КОМП'ЮТЕРНИХ СИСТЕМ

Г.Ф. Кривуля, А.А. Давидов

У статті розглядається оптимізація бінарних дерев рішень за рахунок зменшення розмірності дерева та визначення неістотних атрибутів (діагностичних ознак), які не впливають на процес вибору рішення про стан комп'ютерної системи. Оптимізоване дерево рішень дозволяє вирішувати завдання класифікації стану об'єкту в системах діагностування з меншими часовими і апаратурними витратами.

Ключові слова: бінарні дерева рішень, мінімізація, діагностичні ознаки, комп'ютерні системи, двійкова функція, оптимізація.

OPTIMIZATION OF BINARY DECISION TREES OF THE INTELLECTUAL DIAGNOSTICS COMPUTER SYSTEMS

G.F. Krivoulya, A.A. Davidov

In given article are considered optimizations binary decision trees at the expense of diminishing of dimension of tree and determination of unimportant attributes (diagnostic signs) which do not influence on the process of decision-making about the state computer system. The optimized decision tree allows to decide the tasks of classification of the objects state in the diagnosis systems with less temporal and apparatus expenses.

Key words: binary decision trees, minimization, diagnostic signs, computer systems, switch function, optimization.

Кривуля Геннадий Федорович – д.т.н., профессор, зав. кафедрой автоматизации проектирования вычислительной техники Харьковского национального университета радиоэлектроники, Харьков, Украина, e-mail: krivoulya@i.ua.

Давыдов Андрей Андреевич – аспирант кафедры АПВТ, Харьковского национального университета радиоэлектроники, Харьков, Украина, e-mail: AndreyHnure@yandex.ru.