

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NGUYỄN THỊ THU HẰNG

**NGHIÊN CỨU MỘT SỐ THUẬT TOÁN
RÚT GỌN THUỘC TÍNH TRONG
BẢNG QUYẾT ĐỊNH TẬP GIÁ TRỊ**

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: GS.TS VŨ ĐỨC THI

Thái Nguyên – năm 2015

LỜI CẢM ƠN

Trước hết, tôi xin gửi lời cảm ơn sâu sắc đến thầy hướng dẫn khoa học GS.TS Vũ Đức Thi về những chỉ dẫn khoa học, định hướng nghiên cứu và tận tình hướng dẫn tôi trong suốt quá trình làm luận văn.

Tôi cũng xin cảm ơn các Thầy trong viện Công Nghệ Thông Tin, các Thầy Cô trong trường Đại học Công Nghệ Thông Tin và Truyền Thông - Đại học Thái Nguyên đã quan tâm chỉ bảo và trực tiếp giảng dạy, giúp đỡ trong suốt quá trình học tập và nghiên cứu.

Cuối cùng, tôi xin cảm ơn gia đình và bạn bè, những người đã luôn ủng hộ và động viên tôi để tôi yên tâm nghiên cứu luận văn này.

Học viên

Nguyễn Thị Thu Hằng

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình của riêng tôi, dưới sự hướng dẫn của GS.TS Vũ Đức Thi. Các số liệu và kết quả nghiên cứu trong luận văn này là trung thực.

Mọi tham khảo trong luận văn đều được trích dẫn rõ ràng tên tác giả, tên công trình, thời gian, địa điểm công bố.

Mọi sao chép không hợp lệ, vi phạm quy chế đào tạo hay gian trá tôi xin hoàn toàn chịu trách nhiệm.

Học viên

Nguyễn Thị Thu Hằng

MỤC LỤC

LỜI CẢM ƠN	i
DANH MỤC CÁC THUẬT NGỮ	vi
BẢNG KÝ HIỆU, TỪ VIẾT TẮT.....	vii
DANH MỤC BẢNG.....	ix
DANH MỤC HÌNH	x
MỞ ĐẦU	1
CHƯƠNG 1: KHÁI QUÁT VỀ HỆ THÔNG TIN TẬP GIÁ TRỊ VÀ.....	4
BÀI TOÁN RÚT GỌN THUỘC TÍNH.....	4
1.1. Hệ thông tin và mô hình tập thô truyền thống	4
1.1.1. Hệ thông tin.....	4
1.1.2. Bảng quyết định	6
1.1.3. Tập rút gọn và tập lõi	7
1.1.4. Mô hình tập thô truyền thống	9
1.1.5. Ma trận phân biệt được và hàm phân biệt được	13
1.2. Hệ thông tin tập giá trị và mô hình tập thô dung sai.....	15
1.2.1. Hệ thông tin tập giá trị	15
1.2.2. Quan hệ dung sai.....	17
1.2.3. Bảng quyết định tập giá trị.....	18
1.2.4. Tập thô dựa trên quan hệ dung sai.....	19
1.2.5. Ma trận dung sai.....	20
1.2.6. Rút gọn thuộc tính trong bảng quyết định tập giá trị.....	21
CHƯƠNG 2: RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH TẬP GIÁ TRỊ.....	26
2.1. Đặt vấn đề.....	26
2.2. Cơ sở lý thuyết.....	26
2.2.1. Hàm phân biệt ngẫu nhiên.....	26

2.2.2. Bảng ngẫu nhiên CT và bảng ngẫu nhiên dựa trên quan hệ dung sai TCT.....	27
2.2.3. Giá trị thuộc tính biểu diễn qua mô hình lưới	37
2.3. Thuật toán tìm tập rút gọn thuộc tính.....	40
2.3.1. Thuật toán 2.1 - Tìm tập rút gọn thuộc tính trong bảng quyết định tập giá trị	40
2.3.2. Ví dụ minh họa thuật toán 2.1.....	41
2.4. Thuật toán tìm xấp xỉ trong hệ thông tin tập giá trị	44
2.4.1. Thuật toán 2.2- Thuật toán tìm xấp xỉ trên, xấp xỉ dưới sử dụng hàm phân biệt.....	44
2.4.2. Ví dụ minh họa thuật toán 2.2.....	45
CHƯƠNG 3: PHẦN MỀM THỬ NGHIỆM THUẬT TOÁN TÌM TẬP RÚT GỌN TRONG BẢNG QUYẾT ĐỊNH TẬP GIÁ TRỊ VÀ ỨNG DỤNG TRONG BÀI TOÁN CHẨN ĐOÁN BỆNH VIÊM GAN B	48
3.1. Phát biểu bài toán.....	48
3.2. Mô tả và xử lý dữ liệu	48
3.2.1. Mô tả dữ liệu	48
3.2.2. Xử lý dữ liệu	50
3.3. Thử nghiệm chương trình	52
3.4. Đánh giá kết quả.....	54
3.5. Kết luận chương	55
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	57
TÀI LIỆU THAM KHẢO.....	58

DANH MỤC CÁC THUẬT NGỮ

Thuật ngữ tiếng Việt	Thuật ngữ tiếng Anh
<i>Tập thô</i>	<i>Rough Set</i>
<i>Hệ thống tin đơn trị</i>	<i>Information System</i>
<i>Hệ thống tin đơn trị đầy đủ</i>	<i>Complete Information System</i>
<i>Hệ thống tin đơn trị không nhất quán</i>	<i>Inconsistent Information System</i>
<i>Bảng quyết định</i>	<i>Decision Table</i>
<i>Hệ thống tin tập giá trị</i>	<i>Set valued Information System</i>
<i>Bảng quyết định tập giá trị</i>	<i>Set Valued Decision Information System</i>
<i>Quan hệ không phân biệt được</i>	<i>Indiscernibility Relation</i>
<i>Quan hệ dung sai</i>	<i>Tolerance Relation</i>
<i>Xấp xỉ dưới</i>	<i>Lower Approximation</i>
<i>Xấp xỉ trên</i>	<i>Upper Approximation</i>
<i>Rút gọn thuộc tính</i>	<i>Attribute Reduction</i>
<i>Tập rút gọn</i>	<i>Reduct</i>
<i>Tập lõi</i>	<i>Core</i>
<i>Ma trận phân biệt</i>	<i>Indiscernibility Matrix</i>
<i>Hàm phân biệt</i>	<i>Indiscernibility Function</i>
<i>Bảng ngẫu nhiên</i>	<i>Contingency Table</i>
<i>Bảng ngẫu nhiên dựa trên quan hệ dung sai</i>	<i>Tolerance Based Contingency Table</i>

BẢNG KÝ HIỆU, TỪ VIẾT TẮT

Ký hiệu, từ viết tắt	Diễn giải
$S = U, A, V, f$	Hệ thông tin
$T = U, C \cup D, V, f$	Bảng quyết định
$IS = U, A, V, f$	Hệ thông tin tập giá trị
$DS = (U, C \cup d, V, f)$	Bảng quyết định tập giá trị
$u \ a$	Giá trị của đối tượng u tại thuộc tính a
$IND \ B$	Quan hệ B – không phân biệt
$u \ B$	Lớp tương đương chứa u của quan hệ $IND \ B$
U / B	Phân hoạch của U sinh bởi tập thuộc tính B
$COVER \ U$	Tập tất cả các phủ của U
$\partial_B(u)$	Hàm quyết định suy rộng của đối tượng u đối với B
\underline{BX}	B – xấp xỉ dưới của X trong hệ thông tin
\overline{BX}	B – xấp xỉ trên của X trong hệ thông tin
$BN_B \ X$	B – miền biên của X trong hệ thông tin
$POS_B \ D$	B – miền dương của D trong hệ thông tin
T_B	Quan hệ dung sai của tập thuộc tính B
$\overline{T_B}(X)$	Xấp xỉ trên của X trong hệ thông tin tập giá trị
$\underline{T_B}(X)$	Xấp xỉ dưới của X trong hệ thông tin tập giá trị
$BND_{T_B}(X)$	Miền biên của X trong hệ thông tin tập giá trị
$NEG_{T_B}(X)$	Miền ngoài của X trong hệ thông tin tập giá trị
$POS_{T_B}(X)$	Miền dương của X trong hệ thông tin tập giá trị
CT_B	Bảng ngẫu nhiên của tập thuộc tính B
TCT_B	Bảng ngẫu nhiên dựa trên quan hệ dung sai

	<i>của tập thuộc tính B</i>
M_{DT}	<i>Ma trận phân biệt</i>
$discern(A)$	<i>Hàm phân biệt</i>
IS_p	<i>Hệ thông tin giá trị tập đại diện</i>
DS_p	<i>Bảng quyết định giá trị tập đại diện</i>
U_p	<i>Tập đối tượng đại diện của hệ thông tin tập giá trị</i>

DANH MỤC BẢNG

Bảng 1. 1: Ví dụ về hệ thông tin	5
Bảng 1. 2. Bảng quyết định về bệnh cúm	7
Bảng 1. 3. Bảng rút gọn thứ nhất của hệ thống bệnh cúm R_1	8
Bảng 1. 4. Bảng rút gọn thứ hai của hệ thống bệnh cúm R_2	9
Bảng 1. 5. Thông tin về bệnh cúm	10
Bảng 1. 6. Ma trận phân biệt được xây dựng từ Bảng 1.2	14
Bảng 1. 7. Hệ thông tin tập giá trị	16
Bảng 1. 8. Bảng quyết định tập giá trị	18
Bảng 1. 9. Ma trận phân biệt theo hướng quyết định.....	21
Bảng 1. 10. Bảng quyết định về các xe hơi.....	23
Bảng 1. 11. Bảng quyết định tập giá trị	24
Bảng 2. 1. Bảng phân biệt ngẫu nhiên biểu diễn giá trị tập thuộc tính và hàm phân biệt	32
Bảng 2. 2. Minh họa giá trị của hàm phân biệt	36
Bảng 2. 3. Bảng quyết định tập giá trị bao gồm 4 cột thuộc tính	41
Bảng 2. 4. Bảng quyết định tập giá trị bao gồm 4 cột thuộc tính điều kiện và cột d_x	45

DANH MỤC HÌNH

Hình 2. 1. Cấu trúc của bảng quyết định tập giá trị	39
Hình 3. 1. Bảng dữ liệu đầu vào.....	49
Hình 3. 2. Tập dữ liệu sau khi xử lý.....	52
Hình 3. 3. Giao diện nhập dữ liệu	52
Hình 3. 4. Màn hình hiển thị thông tin các thuộc tính	53
Hình 3. 5. Kết quả thực hiện với bộ dữ liệu thử nghiệm	53
Hình 3. 6. Tập dữ liệu sau khi rút gọn	55

MỞ ĐẦU

Trong những năm gần đây chứng kiến sự phát triển mạnh mẽ và sôi động của công nghệ thông tin trong lĩnh vực khai phá dữ liệu và phát triển tri thức trong cơ sở dữ liệu. Nhiều nhóm nhà khoa học trên thế giới quan tâm nghiên cứu và phát triển lý thuyết tập thô vào các lĩnh vực khoa học nổi bật

Lý thuyết tập thô - do Zdzislaw Pawlak [11] đề xuất vào những năm đầu thập niên tám mươi của thế kỷ hai mươi - được xem là công cụ hữu hiệu để giải quyết các bài toán phân lớp, phát hiện luật...chứa dữ liệu không đầy đủ, không chắc chắn. Từ khi xuất hiện, lý thuyết tập thô đã được sử dụng hiệu quả trong các bước của quá trình khai phá dữ liệu và khám phá tri thức, bao gồm tiền xử lý số liệu, khai phá dữ liệu và đánh giá kết quả thu được. Rút gọn thuộc tính và trích lọc luật quyết định (luật phân lớp) là hai ứng dụng chính của lý thuyết tập thô trong khai phá dữ liệu.

Rút gọn thuộc tính thuộc giai đoạn tiền xử lý dữ liệu còn trích lọc luật thuộc giai đoạn khai phá dữ liệu. Rút gọn thuộc tính là ứng dụng quan trọng nhất trong lý thuyết tập thô. Mục tiêu của rút gọn thuộc tính là loại bỏ các thuộc tính dư thừa để tìm ra các thuộc tính cốt yếu và cần thiết trong cơ sở dữ liệu. Với bảng quyết định tập giá trị, rút gọn thuộc tính là tìm tập con nhỏ nhất của tập thuộc tính với điều kiện bảo toàn thông tin phân lớp của bảng quyết định. Đối với một bảng quyết định tập giá trị có thể có nhiều tập rút gọn khác nhau. Tuy nhiên, trong thực tế thường không đòi hỏi tìm tất cả các tập rút gọn mà chỉ cần tìm được một tập rút gọn tốt nhất theo một tiêu chuẩn đánh giá nào đó là đủ.

Vì vậy, mỗi phương pháp rút gọn thuộc tính đều đề xuất một thuật toán Heuristic tìm tập rút gọn. Các thuật toán này giảm thiểu đáng kể khối lượng tính toán, nên có thể áp dụng với bài toán có khối lượng dữ liệu lớn.

Trong các bài toán thực tế, các bảng quyết định thường thiếu giá trị trên miền giá trị thuộc tính, gọi là các bảng quyết định không đầy đủ. Trên bảng quyết định không đầy đủ, Kryszkiewicz [10] đã mở rộng quan hệ tương đương trong lý thuyết tập thô truyền thống thành quan hệ dung sai và đề xuất mô hình tập thô dung sai nhằm trích lọc luật trực tiếp không qua bước xử lý giá trị thiếu. Trên xu thế đó, có rất nhiều tài liệu nghiên cứu các phương pháp rút gọn thuộc tính trong hệ thông tin đơn trị. Tuy nhiên đó mới là hệ đơn trị, luận văn này tôi đi vào “NGHIÊN CỨU MỘT SỐ THUẬT TOÁN RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH TẬP GIÁ TRỊ”.

Mục tiêu của luận văn trình bày có chọn lọc về các khái niệm cơ bản nhất trong lý thuyết tập thô trong phạm vi xem xét bài toán rút gọn thuộc tính. Khảo sát một số thuật toán liên quan đến bảng quyết định tập giá trị, thuật toán giải quyết bài toán rút gọn thuộc tính trong tập thô truyền thống và tập thô dung sai trong hệ thông tin tập giá trị. Phần tiếp theo của luận văn là khai thác thuật toán tìm tập rút gọn thuộc tính trong bảng quyết định tập giá trị.

Đối tượng nghiên cứu là bài toán rút gọn thuộc tính trong bảng quyết định tập giá trị như đã trình bày ở phần mục tiêu luận văn.

Phạm vi nghiên cứu Giới hạn ở bài toán rút gọn thuộc tính trong bước tiền xử lý số liệu và một phần bước trích lọc tri thức từ bảng dữ liệu tập giá trị của khai phá dữ liệu. Ứng dụng thuật toán rút gọn thuộc tính trong bảng quyết định tập giá trị vào bài toán chẩn đoán bệnh viêm gan B.

Phương pháp nghiên cứu là nghiên cứu lý thuyết có sử dụng phương pháp nghiên cứu thực nghiệm.

* Cấu trúc của luận văn gồm 3 chương như sau:

- **Chương 1: khái quát về hệ thông tin tập giá trị và bài toán rút gọn thuộc tính:** chương này trình bày có chọn lọc về các khái niệm cơ bản nhất về tập thô truyền thống, tập thô dung sai.

- **Chương 2: Rút gọn thuộc tính trong bảng quyết định tập giá trị:** chương này khai thác các thuật toán trong hệ thông tin tập giá trị: thuật toán tìm tập rút gọn thuộc tính trong bảng quyết định tập giá trị và thuật toán tìm xấp xỉ trên- xấp xỉ dưới sử dụng hàm phân biệt trong bảng quyết định tập giá trị.

- **Chương 3: Phần mềm thử nghiệm thuật toán rút gọn thuộc tính trong bảng quyết định tập giá trị và ứng dụng trong bài toán chẩn đoán bệnh Viêm gan B:** trên cơ sở lý thuyết đã nghiên cứu, toàn bộ chương này đi vào xây dựng phần mềm thực nghiệm, phân tích thiết kế chương trình và đưa ra kết quả của chương trình.

- **Kết luận.**

CHƯƠNG 1: KHÁI QUÁT VỀ HỆ THÔNG TIN TẬP GIÁ TRỊ VÀ BÀI TOÁN RÚT GỌN THUỘC TÍNH

1.1. Hệ thông tin và mô hình tập thô truyền thống [1]

1.1.1. Hệ thông tin

Hệ thông tin là một bảng dữ liệu gồm p cột ứng với p thuộc tính và n hàng ứng với n đối tượng. Một cách hình thức, hệ thông tin được định nghĩa như sau:

Định nghĩa 1.1. Hệ thông tin là một bộ tứ $IS = (U, A, V, f)$ trong đó U là tập hữu hạn, khác rỗng các đối tượng; A là tập hữu hạn, khác rỗng các thuộc tính; $V = \bigcup_{a \in A} V_a$ với V_a là tập giá trị của thuộc tính $a \in A$; $f : U \times A \rightarrow V_a$ là hàm thông tin, $\forall a \in A, u \in U, f(u, a) \in V_a$.

Với mọi $u \in U, a \in A$, ta ký hiệu giá trị thuộc tính a tại đối tượng u là $a(u)$ thay vì $f(u, a)$. Nếu $B = \{b_1, b_2, \dots, b_k\} \subseteq A$ là một tập con các thuộc tính thì ta ký hiệu bộ các giá trị $b_i(u)$ bởi $B(u)$. Như vậy, nếu u và v là hai đối tượng, thì ta viết $B(u) = B(v)$ nếu $b_i(u) = b_i(v)$ với mọi $i = 1, \dots, k$.

Ví dụ 1.1. Cho hệ thông tin trong Bảng 1.1 khi đó ta có:

Tập các đối tượng $U = \{u_1, \dots, u_7\}$

Tập các thuộc tính $A = \{\text{Độ tuổi}, \text{Số buổi}, \text{Thi đậu}\}$

Tập giá trị của thuộc tính độ tuổi, số buổi và thi đậu là:

$V_{\text{độ tuổi}} = \{16 - 30, 31 - 45, 46 - 60, 16 - 30\}$

$V_{\text{số buổi}} = \{0, 50, 1 - 25, 26 - 49\}$

$V_{\text{thi đậu}} = \{\text{có}, \text{không}\}$

$f(u_1, \text{độ tuổi}) = (16 - 30)$, $f(u_2, \text{số buổi}) = 0 \dots$ tương ứng là các giá trị của các đối tượng u_1, u_2 trên các thuộc tính *độ tuổi*, *số buổi*.

Bảng 1. 1: Ví dụ về hệ thông tin

U	Độ tuổi	Số buổi	Thi đậu
u_1	16 - 30	50	Có
u_2	16 - 30	0	Không
u_3	31 - 45	1 - 25	Không
u_4	31 - 45	1 - 25	Có
u_5	46 - 60	26 - 49	Không
u_6	16 - 30	26 - 49	Có
u_7	46 - 60	26 - 49	Không

Xét hệ thông tin $IS = U, A, V, f$, mỗi tập con các thuộc tính $P \subseteq A$ xác định một quan hệ hai ngôi trên U , ký hiệu là $IND\ P$, xác định bởi

$$IND\ P = \{u, v \in U \times U \mid \forall a \in P, a(u) = a(v)\}.$$

$IND\ P$ là quan hệ P - không phân biệt được. Dễ thấy rằng $IND\ P$ là một quan hệ tương đương trên U . Nếu $u, v \in IND\ P$ thì hai đối tượng u và v không phân biệt được bởi các thuộc tính trong P . Quan hệ tương đương $IND\ P$ xác định một phân hoạch trên U , ký hiệu là $U / IND\ P$ hay U / P . Ký hiệu lớp tương đương trong phân hoạch U / P chứa đối tượng u là u_p , khi đó $u_p = \{v \in U \mid u, v \in IND\ P\}$.

Ví dụ 1.2. Xét hệ thông tin đơn trị với các thuộc tính: Độ tuổi, Số buổi, Thi đậu được cho trong Bảng 1.1 ta có:

$$U / \{\text{Độ tuổi}\} = \{u_1, u_2, u_6\}, \{u_3, u_4\}, \{u_5, u_7\}$$

$$U / \{\text{Số buổi}\} = \{u_1, u_2\}, \{u_3, u_4\}, \{u_5, u_6, u_7\}$$

$$U / \{\text{Thi đậu}\} = \{u_1, u_4, u_6\}, \{u_2, u_3, u_5, u_7\}$$

Giả sử chọn $P = \{\text{Độ tuổi, Số buổi, Thi đậu}\}$ ta dễ dàng thu được một phân hoạch của U được sinh bởi P là:

$$U/P = \{u_1, u_2, u_3\}, \{u_4, u_5, u_7, u_6\}$$

Như vậy, các đối tượng u_1, u_2 không phân biệt được về độ tuổi, nhưng phân biệt được về số buổi và thi đậu. Tương tự u_3, u_4 không phân biệt được về độ tuổi và số buổi, nhưng phân biệt được về thi đậu, ...

1.1.2. Bảng quyết định

Một lớp đặc biệt của các hệ thông tin có vai trò quan trọng trong nhiều ứng dụng là bảng quyết định đầy đủ, gọi tắt là *bảng quyết định- decision table*. Bảng quyết định là một hệ thông tin DS với tập thuộc tính A được chia thành hai tập khác rỗng rời nhau C và D , lần lượt được gọi là tập thuộc tính điều kiện và tập thuộc tính quyết định. Tức là $DS = U, C \cup D, V, f$ với $C \cap D = \emptyset$.

Bảng quyết định DS được gọi là nhất quán- *consistent* nếu D phụ thuộc hàm vào C , tức là với mọi $u, v \in U, C(u) = C(v)$ kéo theo $D(u) = D(v)$. Ngược lại thì gọi là không nhất quán- *inconsistent* hay mâu thuẫn. Theo định nghĩa miền dương, bảng quyết định là nhất quán khi và chỉ khi $POS_C D = U$. Trong trường hợp bảng không nhất quán thì $POS_C D$ chính là tập con cực đại của U sao cho phụ thuộc hàm $C \rightarrow D$ đúng.

Ví dụ 1.3. Cho bảng quyết định về bệnh cúm (Bảng 1.2) trong đó tập thuộc tính điều kiện $C = \{\text{Mệt mỏi, Đau đầu, Đau cơ, Thân nhiệt}\}$ và tập thuộc tính quyết định $D = \{\text{Cảm cúm}\}$.

Bảng 1. 2. Bảng quyết định về bệnh cúm

U	Mệt mỏi	Đau đầu	Đau cơ	Thân nhiệt	Cảm cúm
u_1	Có	Có	Có	Bình thường	Không
u_2	Có	Có	Có	Cao	Có
u_3	Có	Có	Có	Rất cao	Có
u_4	Có	Không	Có	Bình thường	Không
u_5	Có	Không	Không	Cao	Không
u_6	Có	Không	Có	Rất cao	Có

Ta có $U / C = \{C_1, C_2, C_3, C_4, C_5, C_6\}$ với

$C_1 = \{u_1\}$, $C_2 = \{u_2\}$, $C_3 = \{u_3\}$, $C_4 = \{u_4\}$, $C_5 = \{u_5\}$, $C_6 = \{u_6\}$.

$U / D = \{D_1, D_2\}$ với $D_1 = \{u_1, u_4, u_5\}$, $D_2 = \{u_2, u_3, u_6\}$;

Trong trường hợp này, Bảng 1.2 là một bảng quyết định nhất quán.

1.1.3. Tập rút gọn và tập lõi

Trong bảng quyết định, các thuộc tính điều kiện được phân thành ba nhóm: *thuộc tính lõi (core attribute)*, *thuộc tính rút gọn (reductive attribute)* và *thuộc tính dư thừa (redundant attribute)*. *Thuộc tính lõi* là thuộc tính không thể thiếu trong việc phân lớp chính xác tập dữ liệu. Thuộc tính lõi xuất hiện trong tất cả các tập rút gọn của bảng quyết định. *Thuộc tính dư thừa* là những thuộc tính mà việc loại bỏ chúng không ảnh hưởng đến việc phân lớp tập dữ liệu, thuộc tính dư thừa không xuất hiện trong bất kỳ tập rút gọn nào của bảng quyết định. *Thuộc tính rút gọn* là thuộc tính xuất hiện trong một tập rút gọn nào đó của bảng quyết định.

Với bảng quyết định $DS = U, C \cup D, V, f$. Thuộc tính $c \in C$ được gọi là *không cần thiết* (*dispensable*) trong DS nếu $POS_C D = POS_{(C-c)} D$; Ngược lại, c được gọi là *cần thiết* (*indispensable*). Tập tất cả các thuộc tính cần thiết trong DS được gọi là tập lõi và được ký hiệu là $PCORE C$. Khi đó, thuộc tính cần thiết chính là thuộc tính lõi. Như vậy, thuộc tính không cần thiết là *thuộc tính dư thừa* hoặc *thuộc tính rút gọn*.

Nếu tập thuộc tính $R \subseteq C$ thỏa mãn:

$$1) POS_R(D) = POS_C(D)$$

$$2) \forall r \in R, POS_{R-r}(D) \neq POS_C(D)$$

thì R là một tập rút gọn của C . R được gọi là tập rút gọn dựa trên miền dương còn được gọi là tập rút gọn Pawlak.

Từ lý thuyết nêu trên, R là tập rút gọn nếu nó là tập tối thiểu thỏa mãn $POS_R D = POS_C D$. Rõ ràng là có thể có nhiều tập rút gọn của C . Ta ký hiệu $RED(C)$ là tập tất cả các rút gọn của C . Khi đó $CORE C = \bigcap_{R \in RED C} R$

Ví dụ 1.4. Xét bảng quyết định đơn trị về bệnh cúm cho ở Bảng 1.2.

Bảng này có hai tập rút gọn là $R_1 = \{\text{Đau cơ, Thân nhiệt}\}$ (xem bảng 1.3) và $R_2 = \{\text{Đau đầu, Thân nhiệt}\}$ (xem bảng 1.4). Như vậy tập lõi là $CORE(C) = \{\text{Thân nhiệt}\}$ và Thân nhiệt là thuộc tính cần thiết duy nhất. Các thuộc tính Đau đầu, Đau cơ đều không cần thiết theo nghĩa là, từ bảng dữ liệu có thể loại bỏ một trong hai thuộc tính này mà vẫn chẩn đoán đúng bệnh. Tức là: $POS_{\{\text{Đau cơ, Thân nhiệt}\}}(\{\text{Cảm cúm}\}) = POS_C(\{\text{Cảm cúm}\})$

$$POS_{\{\text{Đau đầu, Thân nhiệt}\}}(\{\text{Cảm cúm}\}) = POS_C(\{\text{Cảm cúm}\}).$$

Bảng 1. 3. Bảng rút gọn thứ nhất của hệ thống bệnh cúm R_1

U	Đau cơ	Thân nhiệt	Cảm cúm
-----	--------	------------	---------

u_1, u_4	Có	Bình thường	Không
u_2	Có	Cao	Có
u_3, u_6	Có	Rất cao	Có
u_5	Không	Cao	Không

Bảng 1. 4. Bảng rút gọn thứ hai của hệ thống bệnh cúm R_2

U	Đau đầu	Thân nhiệt	Cảm cúm
u_1	Có	Bình thường	Không
u_2	Có	Cao	Có
u_3	Có	Rất cao	Có
u_4	Không	Bình thường	Không
u_5	Không	Cao	Không
u_6	Không	Rất cao	Có

1.1.4. Mô hình tập thô truyền thống

a. Định nghĩa xấp xỉ trên- xấp xỉ dưới

Cho hệ thống tin $IS = U, A, V, f$, tập thuộc tính $B \subseteq A$ và tập đối tượng $X \subseteq U$. Trong lý thuyết tập thô truyền thống của Pawlak [10], để biểu diễn tập X thông qua các lớp tương đương của U/B (còn gọi là biểu diễn X bằng tri thức có sẵn B), người ta xấp xỉ X bởi hợp của một số hữu hạn các lớp tương đương của U/B . Có hai cách xấp xỉ tập đối tượng X thông qua tập thuộc tính B , được gọi là B -xấp xỉ dưới và B -xấp xỉ trên của X , ký hiệu là lượt là $\underline{B}X$ và $\overline{B}X$, được xác định như sau:

$$\underline{B}X = \{u \in U \mid u_B \subseteq X\}, \quad \overline{B}X = \{u \in U \mid u_B \cap X \neq \emptyset\}.$$

Tập $\underline{B}X$ bao gồm tất cả các phần tử của U chắc chắn thuộc vào X , còn tập $\overline{B}X$ bao gồm các phần tử của U có thể thuộc vào X dựa trên tập thuộc tính B . Từ hai tập xấp xỉ nêu trên, ta định nghĩa các tập $BN_B X = \overline{B}X - \underline{B}X$: B -miền biên của X , $U - \overline{B}X$: B -miền ngoài của X .

B -miền biên của X là tập chứa các đối tượng có thể thuộc hoặc không thuộc X , còn B -miền ngoài của X chứa các đối tượng chắc chắn không thuộc X . Sử dụng các lớp của phân hoạch U/B , các xấp xỉ dưới và trên của X có thể viết lại:

$$\underline{B}X = \bigcup \{Y \in U/B \mid Y \subseteq X\}, \quad \overline{B}X = \bigcup \{Y \in U/B \mid Y \cap X \neq \emptyset\}.$$

Trong trường hợp $BN_B X = \emptyset$ thì X được gọi là *tập chính xác (exact set)*, ngược lại X được gọi là *tập thô (rough set)*.

Với $B, D \subseteq A$, ta gọi B -miền dương của D là tập được xác định như sau

$$POS_B(D) = \bigcup_{X \in U/D} \underline{B}X$$

Rõ ràng $POS_B(D)$ là tập tất cả các đối tượng u sao cho với mọi $v \in U$ mà $u B = v B$ ta đều có $u D = v D$. Nói cách khác $POS_B(D) = \{u \in U \mid u_B \subseteq u_D\}$.

Ví dụ 1.5. Xét hệ thông tin biểu diễn các triệu chứng cúm của bệnh nhân

Bảng 1. 5. Thông tin về bệnh cúm

U	Đau đầu	Thân nhiệt	Cảm cúm
u_1	Có	Bình thường	Không
u_2	Có	Cao	Có
u_3	Có	Rất cao	Có
u_4	Không	Bình thường	Không
u_5	Không	Cao	Không
u_6	Không	Rất cao	Có

u_7	Không	Cao	Có
u_8	Không	Rất cao	Không

Ta có: $U / \{\text{Đau đầu}\} = u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8$

$U / \{\text{Thân nhiệt}\} = u_1, u_4, u_2, u_5, u_7, u_3, u_6, u_8$

$U / \{\text{Cảm cúm}\} = u_1, u_4, u_5, u_8, u_2, u_3, u_6, u_7$

$U / \{\text{Đau đầu, Cảm cúm}\} = u_1, u_2, u_3, u_4, u_5, u_8, u_6, u_7$

Như vậy, các bệnh nhân u_2, u_3 không phân biệt được về đau đầu (a_1) và cảm cúm (a_3), nhưng phân biệt được về thân nhiệt (a_2).

Các lớp không phân biệt được bởi $B = \{\text{Đau đầu, Thân nhiệt}\}$ là:

$u_1, u_2, u_3, u_4, u_5, u_7, u_6, u_8$.

Đặt $X = \{u \mid u(\text{Cảm cúm}) = \text{Có}\} = u_2, u_3, u_6, u_7$. Khi đó:

$\underline{B}X = u_2, u_3$ và $\overline{B}X = u_2, u_3, u_5, u_6, u_7, u_8$. Như vậy, B -miền biên của X là tập hợp $BN_B X = u_5, u_6, u_7, u_8$. Nếu đặt $D = \{\text{Cảm cúm}\}$ thì

$U / D = X_1 = u_1, u_4, u_5, u_8; X_2 = u_2, u_3, u_6, u_7$, $\underline{B}X_1 = u_1, u_4$; $\underline{B}X_2 = u_2, u_3$,

$POS_B(D) = \bigcup_{X \in U/D} \underline{B}X = u_1, u_2, u_3, u_4$.

Từ định nghĩa trên ta đưa ra các tính chất của tập xấp xỉ:

b. Tính chất của tập xấp xỉ

Cho $X \subseteq U$ và $B \subseteq A$. Khi đó:

$$1) \quad \underline{B}(X) \subseteq X \subseteq \overline{B}(X).$$

$$2) \quad \underline{B}(\emptyset) = \overline{B}(\emptyset) = \emptyset, \underline{B}(U) = \overline{B}(U) = U$$

$$3) \quad \overline{B}(X \cup Y) = \overline{B}(X) \cup \overline{B}(Y)$$

$$4) \quad \underline{B}(X \cup Y) = \underline{B}(X) \cup \underline{B}(Y)$$

$$5) \quad X \subseteq Y \Rightarrow \underline{B}(X) \subseteq \underline{B}(Y) \text{ và } \overline{B}(X) \subseteq \overline{B}(Y)$$

$$6) \quad \underline{B}(X \cup Y) \supseteq \underline{B}(X) \cup \underline{B}(Y)$$

$$7) \quad \overline{B}(X \cap Y) \subseteq \overline{B}(X) \cap \overline{B}(Y)$$

$$8) \quad \underline{B}(U - X) = U - \overline{B}(X)$$

$$9) \quad \overline{B}(U - X) = U - \underline{B}(X)$$

$$10) \quad \underline{B}(\underline{B}(X)) = \overline{B}(\underline{B}(X)) = \underline{B}(X)$$

$$11) \quad \overline{B}(\overline{B}(X)) = \underline{B}(\overline{B}(X)) = \overline{B}(X)$$

Với các khái niệm của tập xấp xỉ đối với phân hoạch U/B , các tập thô được chia thành 4 lớp cơ bản như sau:

- a) Tập X là B -xác định thô nếu $\underline{B}X \neq \emptyset$ và $\overline{B}X \neq U$.
- b) Tập X là B -không xác định trong nếu $\underline{B}X = \emptyset$ và $\overline{B}X \neq U$.
- c) Tập X là B -không xác định ngoài nếu $\underline{B}X \neq \emptyset$ và $\overline{B}X = U$.
- d) Tập X là B -không xác định hoàn toàn nếu $\underline{B}X = \emptyset$ và $\overline{B}X = U$.

Sau đây là ý nghĩa trực quan của việc phân lớp này:

- 1) X là có thể xác định thô theo B nghĩa là với tập B ta có thể quyết định phần tử nào của U thuộc X , và phần tử nào thuộc $U - X$.
- 2) X là không thể xác định phía trong theo B nghĩa là với tập B ta có thể quyết định phần tử nào của U thuộc về $U - X$, nhưng không thể quyết định phần tử nào đó là thuộc X hay không.
- 3) X là không thể xác định phía ngoài theo B nghĩa là với tập B ta có thể quyết định phần tử nào của U thuộc về X , nhưng không thể quyết định phần tử nào đó của U là thuộc $U - X$ hay không.

- 4) X hoàn toàn không thể xác định theo B nghĩa là với tập B ta có thể quyết định phần tử nào đó của U thuộc về X hoặc thuộc về $U - X$ hay không.

c. Độ chính xác của xấp xỉ

Tập thô còn có thể đặc trưng hóa dưới hình thức hình thức số bằng hệ số phản ánh độ chính xác của xấp xỉ:

$$\alpha_B(X) = \frac{\text{Card}|\underline{B}(X)|}{\text{Card}|\overline{B}(X)|}$$

Trong đó $|X|$ biểu diễn số phần tử của tập $X \neq \emptyset$

Rõ ràng ta có $0 \leq \alpha_B(X) \leq 1$

Nếu $\alpha_B(X) = 1$. X là rõ theo B (X là chính xác theo B), ngược lại, nếu $\alpha_B(X) < 1$, X là thô theo B (X là gần đúng theo B).

1.1.5. Ma trận phân biệt được và hàm phân biệt được

Xét bảng quyết định $DS = U, C \cup D, V, f$ với $U = u_1, u_2, \dots, u_n$. Ma trận phân biệt của DS , ký hiệu $M = (m_{ij})_{n \times m}$, là một ma trận đối xứng mà mỗi phần tử của nó là một tập hợp các thuộc tính được xác định như sau:

$$m_{ij} = \begin{cases} \emptyset & \text{if } u_i(D) = u_j(D) \\ c \in C \mid u_i(c) \neq u_j(c) & \text{if } u_i(D) \neq u_j(D) \end{cases}$$

Như vậy, m_{ij} là tập hợp gồm tất cả các thuộc tính điều kiện có thể xếp đối tượng u_i và u_j vào các lớp tương đương khác nhau đối với quan hệ không biệt được trên mỗi thuộc tính của tập thuộc tính này. Hay nói cách khác hai đối tượng u_i và u_j mà $u_i(D) \neq u_j(D)$ có thể phân biệt với nhau bởi một thuộc tính bất kỳ trong tập m_{ij} . Nếu $m_{ij} = \emptyset$ thì u_i và u_j bằng nhau trên tập thuộc tính D hoặc trong trường hợp bảng quyết định đã cho là không nhất quán, hai đối

tượng u_i và u_j có cùng giá trị trên tập thuộc tính điều kiện nhưng khác nhau trên tập thuộc tính quyết định.

Ví dụ 1.6. Xét bảng quyết định như trong Bảng 1.2 ta có,

c_1, c_2, c_3, c_4 : tương ứng cho tập các thuộc tính điều kiện $\{Mệt mỏi, Đau đầu, Đau cơ, Thân nhiệt\}$.

d : ký hiệu cho thuộc tính quyết định $\{Cảm cúm\}$.

Ta có ma trận phân biệt được tương ứng cho trong Bảng 1.6. Đây là ma trận đối xứng nên chúng ta chỉ trình bày ma trận tam giác dưới.

Bảng 1. 6. Ma trận phân biệt được xây dựng từ Bảng 1.2

U	u_1	u_2	u_3	u_4	u_5	u_6
u_1	\emptyset					
u_2	c_4	\emptyset				
u_3	c_4	\emptyset	\emptyset			
u_4	\emptyset	c_2, c_4	c_2, c_4	\emptyset		
u_5	\emptyset	c_2, c_4	c_2, c_3, c_4	\emptyset	\emptyset	
u_6	c_2, c_4	\emptyset	\emptyset	c_4	c_3, c_4	\emptyset

Do bảng quyết định trong ví dụ này không nhất quán nên $m_{23} = \emptyset$.

Trong Bảng 1.3 cho thấy hai đối tượng u_3 và u_2 có cùng giá trị quyết định ($u_3(d) = u_2(d) = \text{"có"}$) hay nói cách khác u_3 và u_2 cùng thuộc một lớp tương đương của phân hoạch $IND(D)$. Trong khi đó $m_{42} = \{c_2, c_4\}$ điều này nói lên rằng hai đối tượng u_2 và u_4 có giá trị quyết định khác nhau và chúng có thể phân biệt được với nhau bởi các thuộc tính c_2 và c_4 nhưng không phân biệt được bởi các thuộc tính c_1 và c_3 .

Để tìm tập rút gọn dựa vào ma trận phân biệt được, người ta đưa vào khái niệm hàm phân biệt được f_r xác định như sau: $f_r(u_j) = \bigwedge_{j \neq i} (\vee m_{ij})$ với mỗi $u_i \in U$, trong đó mỗi thuộc tính cho tương ứng một biến logic cùng tên và:

- 1) $\vee m_{ij}$ là biểu thức tuyển của tất cả các biến $c \in m_{ij}$, nếu $m_{ij} \neq \emptyset$
- 2) $\vee m_{ij} = true$, nếu $m_{ij} = \phi$ và $u_i(D) = u_j(D)$.
- 3) $\vee m_{ij} = false$, nếu $m_{ij} = \phi$ và $u_i(D) \neq u_j(D)$.

Như vậy $f_r(u_i)$ chứa những bộ thuộc tính có thể phân biệt u_i với các đối tượng khác trong DS . Do đó $\bigwedge f_r(u_i)$ sẽ xác định tất cả các rút gọn trong bảng quyết định.

1.2. Hệ thông tin tập giá trị và mô hình tập thô dung sai [1]

1.2.1. Hệ thông tin tập giá trị

Lý thuyết tập thô truyền thống do Pawlak [12] đề xuất là công cụ hiệu quả để giải quyết các bài toán rút gọn thuộc tính và trích lọc luật trên các hệ thông tin đơn trị. Với các hệ thông tin trong thực tế, giá trị một đối tượng tại một thuộc tính có thể là một tập giá trị. Ta hiểu như sau: ví dụ xét hệ thông tin có đối tượng “Nguyễn Văn A” tại thuộc tính “Ngoại ngữ” là “Anh, Pháp, Nga”, nghĩa là Nguyễn Văn A biết ngoại ngữ tiếng Anh, hoặc tiếng Pháp, hoặc tiếng Nga. Hệ thông tin như vậy được gọi là hệ thông tin tập giá trị.

Dưới đây là cách tiếp cận của hệ thông tin tập giá trị:

Loại thứ nhất: Với $x \in X$, $a \in A$, $a(x)$ dùng theo nghĩa “và”. Giả sử, a là thuộc tính làm quen với các ngôn ngữ lập trình thì giá trị thuộc tính $a(u) = \{C++, Java, Pascal\}$ được hiểu theo cách: u biết được cả 3 ngôn ngữ lập trình $C++, Java, Pascal$.

Loại thứ hai: Với $x \in U$, $a \in A$, $a(x)$ dùng theo nghĩa “hoặc”. Giả sử, a là thuộc tính làm quen với các ngôn ngữ lập trình thì giá trị thuộc tính $a(u) =$

$\{C++, Java, Pascal\}$ được hiểu theo cách: u biết được một trong 3 ngôn ngữ hoặc $C++, Java, Pascal$ với giá trị thuộc kiểu số. Ví dụ thuộc tính “tuổi” có $b(x) = [20, 25]$ được hiểu là: đối tượng u trong độ tuổi 20 và 25. Hệ thông tin không đầy đủ với một số giá trị thuộc tính bị thiếu đều thuộc hệ thông tin tập giá trị.

Loại thứ ba: Kết hợp của hai mô hình trên, một số thuộc tính trong hệ thông tin được hiểu theo nghĩa “và” như ví dụ thuộc tính “làm quen ngôn ngữ lập trình” và một số thuộc tính hiểu theo nghĩa “hoặc” như thuộc tính “tuổi”.

Qua 3 cách tiếp cận của hệ thông tin tập giá trị trên, luận văn xây dựng theo hướng tiếp cận thứ hai. Với $x \in U$, $a \in A$, $a(x)$ dùng theo nghĩa “hoặc”.

Định nghĩa 1.2.[9].

Hệ thông tin tập giá trị là một bộ tứ $IS = (U, A, V, f)$ trong đó:

U : là tập hữu hạn khác rỗng, được gọi là tập vũ trụ các đối tượng

A : là tập hữu hạn khác rỗng các thuộc tính

$V = \bigcup_{a \in A} V_a$ với V_a là tập giá trị của thuộc tính $a \in A$

f : là hàm thông tin, $f: U \times A \rightarrow 2^V$ là ánh xạ tập giá trị

Ví dụ 1.7. Bảng 1.7 minh họa một hệ thông tin tập giá trị gồm:

Đối tượng $U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$

Các tập thuộc tính $A = \{Nghe, Ngôn ngữ nói, Đọc, Viết\}$

$V = \{E, F, G\}$

Bảng 1. 7. Hệ thông tin tập giá trị

U	Nghe (A)	Ngôn ngữ nói (S)	Đọc (R)	Viết (W)
u_1	{E}	{E}	{F, G}	{F, G}
u_2	{E, F, G}	{E, F, G}	{F, G}	{E, F, G}
u_3	{E, G}	{E, F}	{F, G}	{F, G}

u_4	{E, F }	{E, G}	{F, G}	{F}
u_5	{F, G}	{F, G}	{F, G}	{F}
u_6	{ F }	{F}	{E, F}	{E, F}
u_7	{E, F, G}	{E, F, G}	{E, G}	{E, F, G}
u_8	{E, F}	{F, G}	{E, F, G}	{E, G}
u_9	{F, G}	{G}	{F, G}	{F, G}
u_{10}	{E, F}	{E, G}	{F, G}	{E, F}

1.2.2. Quan hệ dung sai

Xét hệ thông tin $IS = (U, A, V, f)$, với mỗi tập con thuộc tính $B \subseteq A$, một quan hệ nhị phân trên U được định nghĩa như sau:

$$T_B = \{ u, v \in U \times U \mid \forall b \in B, b(u) \cap b(v) \neq \emptyset \}$$

Dễ thấy rằng T_B không phải là quan hệ tương đương vì có tính phản xạ, đối xứng nhưng không có tính bắc cầu. T_B được gọi là *quan hệ dung sai* và rõ ràng. Đặt $T_B(u) = \{ v \in U \mid (u, v) \in T_B \}$, $T_B(u)$ được gọi là một lớp dung sai.

Ký hiệu: $U / T_B = \{ T_B(u) \mid u \in U \}$ biểu diễn tập tất cả các lớp dung sai sinh bởi quan hệ T_B , khi đó U / T_B hình thành một “phủ” của U vì các lớp dung sai trong U / T_B có thể giao nhau và $\bigcup_{u \in U} T_B(u) = U$. Dễ thấy rằng nếu $C \subseteq B$ thì $T_B(u) \subseteq T_C(u)$ với mọi $u \in U$.

Định nghĩa 1.3[14]. Cho $IS = (U, A, V, f)$ là hệ thông tin tập giá trị. Với thuộc tính $b \in A$ ta có ký hiệu: $[x]_{T_B} = \{ v \in U \mid (x, v) \in T_B \}$ là lớp dung sai của $x \in U$. Chúng ta ký hiệu: $U / T_B = \{ [x]_{T_B} \mid x \in U \}$ là họ của tất cả các lớp dung sai của T_B .

1.2.3. Bảng quyết định tập giá trị

Bảng quyết định tập giá trị $DS = (U, C \cup d, V, f)$ trong đó:

U : là tập đối tượng khác rỗng

C : là tập thuộc tính điều kiện khác rỗng

d : là thuộc tính quyết định với $C \cap d = \emptyset$; $V = V_C \cup V_d$, V_C là tập giá trị của các thuộc tính điều kiện và V_d là tập giá trị của thuộc tính quyết định.

$f: U \times C \rightarrow 2^{V_C}$ là ánh xạ tập giá trị, còn $f: U \times d \rightarrow V_d$ là ánh xạ đơn trị.

Ví dụ 1.8. Biểu diễn bảng quyết định tập giá trị. Có 10 đối tượng và 4 thuộc tính điều kiện. Các đối tượng trong bảng thuộc vào 1 trong 2 lớp quyết định:

Bảng 1. 8. Bảng quyết định tập giá trị

U	Nghe (A)	Ngôn ngữ nói (S)	Đọc (R)	Viết (W)	Quyết định(D)
x ₁	{E}	{E}	{F, G}	{F, G}	No
x ₂	{E, F, G}	{E, F, G}	{F, G}	{E, F, G}	No
x ₃	{E, G}	{E, F}	{F, G}	{F, G}	No
x ₄	{E, F }	{E, G}	{F, G}	{F}	No
x ₅	{F, G}	{F, G}	{F, G}	{F}	No
x ₆	{ F }	{F}	{E, F}	{E, F}	Yes
x ₇	{E, F, G}	{E, F, G}	{E, G}	{E, F, G}	Yes
x ₈	{E, F}	{F, G}	{E, F, G}	{E, G}	Yes
x ₉	{F, G}	{G}	{F, G}	{F, G}	Yes

x_{10}	$\{E, F\}$	$\{E, G\}$	$\{F, G\}$	$\{E, F\}$	Yes
----------	------------	------------	------------	------------	-----

Đặt $B = \{\text{Nghe, Ngôn ngữ nói}\}$

Ta tìm được các lớp dung sai của $x_1, x_2, x_3, x_4, x_7, x_{10}$ như sau:

$$x_1, x_2, x_3, x_4, x_7, x_{10}$$

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} = U$$

$$x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_{10}$$

1.2.4. Tập thô dựa trên quan hệ dung sai

Trong lý thuyết tập thô, xấp xỉ trên và xấp xỉ dưới của đối tượng được định nghĩa dựa trên lớp quan hệ bất khả phân biệt. Còn trong hệ tin tập giá trị thì hai khái niệm này được định nghĩa dựa trên quan hệ dung sai.

Định nghĩa 1.4. Xấp xỉ trên và xấp xỉ dưới [14]

Cho hệ thông tin tập giá trị $IS = (U, A, V, f)$. Đặt T_B là quan hệ dung sai với $B \subseteq A$. Xấp xỉ trên và xấp xỉ dưới của tập bất kỳ $X \subseteq U$ được định nghĩa như sau:

$$\underline{T}_B(X) = \{x \in U : [x]_{T_B} \subseteq X\};$$

$$\overline{T}_B(X) = \{x \in U : [x]_{T_B} \cap X \neq \emptyset\}$$

Tập $\underline{T}_B(X)$ gồm tất cả các phần tử của U chắc chắn thuộc vào X , còn tập $\overline{T}_B(X)$ bao gồm các phần tử của U có khả năng được phân loại vào X dựa vào tập thuộc tính B .

Với các xấp xỉ trên, ta gọi $BND_{T_B}(X)$ miền biên của X là tập $BND_{T_B}(X) = \overline{T}_B(X) - \underline{T}_B(X)$, miền ngoài của X là tập $NEG_{T_B}(X) = U \setminus \overline{T}_B(X)$ và miền dương (miền khẳng định) của X là $POS_{T_B}(X)$ được định nghĩa bởi xấp xỉ dưới, miền biên $BND_{T_B}(X)$ là vùng khác nhau giữa xấp xỉ trên và xấp xỉ dưới, miền ngoài $NEG_{T_B}(X)$ là phần bù của xấp xỉ trên.

Ví dụ 1.9. Xét Bảng 1.8 giả sử $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$, $B = \{\text{Nghe, Ngôn ngữ nói, Đọc, Viết}\}$ ta tìm các xấp xỉ của X như sau:

Trước tiên ta tìm các phân lớp dung sai của từng đối tượng

$$[x_1]_{T_B} = \{x_1, x_2, x_3, x_4, x_7, x_{10}\}, [x_2]_{T_B} = [x_7]_{T_B} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}.$$

$$[x_3]_{T_B} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_{10}\}, [x_4]_{T_B} = \{x_1, x_2, x_3, x_4, x_5, x_7, x_9, x_{10}\}.$$

$$[x_5]_{T_B} = \{x_2, x_3, x_5, x_6, x_7, x_9, x_{10}\}, [x_6]_{T_B} = \{x_2, x_5, x_6, x_7, x_8\}.$$

$$[x_8]_{T_B} = \{x_2, x_3, x_6, x_7, x_8, x_9, x_{10}\}, [x_9]_{T_B} = \{x_2, x_4, x_5, x_7, x_8, x_9, x_{10}\}.$$

$$[x_{10}]_{T_B} = \{x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{10}\}.$$

Chúng ta có:

$$[x_6]_{T_B} \subseteq X, [x_1]_{T_B} \not\subseteq X, [x_2]_{T_B} \not\subseteq X, [x_3]_{T_B} \not\subseteq X, [x_4]_{T_B} \not\subseteq X, [x_5]_{T_B} \not\subseteq X, [x_7]_{T_B} \not\subseteq X, [x_8]_{T_B} \not\subseteq X, [x_9]_{T_B} \not\subseteq X \text{ và } [x_{10}]_{T_B} \not\subseteq X.$$

Nên $\underline{T_B}(X) = \{x_6\}$, và vì $\forall i \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, $[x_i]_{T_B} \cap X \neq \emptyset$ do đó

$$\overline{T_B}(X) = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$$

Vùng dương của X ta tính được $POS_{T_B}(X) = \{x_6\}$.

Miền biên $BND_{T_B}(X) = \{x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{10}\}$.

Miền ngoài $NEG_{T_B}(X) = \emptyset$.

1.2.5. Ma trận dung sai

Định nghĩa 1.5. Ma trận dung sai

Cho $IS = (U, A, V, f)$ là hệ thông tin tập giá trị. Ma trận $M_{IS} = [m_{ij}]_{i,j=1}^n$ của IS là ma trận $n \times n$ với $m_{ij} = a \in A : (i, j) \in T_a$ và cặp đối tượng $(u_i, u_j) \in U^2$.

Theo tiếp cận mô hình tập thô dung sai trên hệ thông tin tập giá trị, khai thác được khái niệm rút gọn dựa trên ma trận dung sai. Ma trận dung sai của hệ quyết định tập giá trị DS là $M_{DS} = [m_{ij}]_{n \times n}$, các phần tử m_{ij} được xác định như sau:

$$m_{ij} = \begin{cases} a \mid a \in AT, (u_i, u_j) \in T_{AT} & d(u_i) \neq d(u_j) \\ \emptyset & otherwise \end{cases}$$

Định nghĩa 1.6. Cho bảng quyết định tập giá trị $DS = U, C \cup d$ và ma trận dung sai. $M_{DS} = [m_{ij}]_{n \times n}$. Nếu $R \subseteq C$ thỏa mãn:

$$1) \quad R \cap m_{ij} \neq \emptyset \text{ với mọi } m_{ij} \neq \emptyset$$

$$2) \quad \text{Với mọi } r \in R, R - r \text{ không thỏa mãn (1) thì } R \text{ được gọi là một}$$

tập rút gọn của DS dựa trên ma trận phân biệt.

Chúng ta biểu diễn ma trận phân biệt từ Bảng 1.6 như sau

Bảng 1. 9. Ma trận phân biệt theo hướng quyết định

M_{IS}	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1						A, S		S	A, S	
x_2										
x_3						A			S	
x_4						S			W	
x_5									W	
x_6	A, S		A	S						
x_7										
x_8	S			W	W					
x_9	A,S		S							
x_{10}										

Trong bảng trên phần màu xám biểu diễn các đối tượng có lớp quyết định giống nhau.

1.2.6. Rút gọn thuộc tính trong bảng quyết định tập giá trị

Sử dụng khái niệm miền dương mở rộng trong quan hệ quyết định tập giá trị, luận án đưa ra khái niệm tập rút gọn dựa trên miền dương.

Định nghĩa 1.7. Cho bảng quyết định tập giá trị $DS = U, C \cup d$. Nếu $R \subseteq C$ thỏa mãn:

- 1) $POS_R d = POS_C d$
- 2) $\forall R' \subset R, POS_{R'} d \neq POS_C d$

thì R được gọi là một *tập rút gọn của DS dựa trên miền dương*.

Định nghĩa 1.8. Cho hệ quyết định giá trị tập $DS = U, C \cup d$ Với $u \in U$, $\partial_C(u) = \{v \mid v \in T_C(u)\}$ được gọi là *hàm quyết định suy rộng* của đối tượng u trên tập thuộc tính C .

Nếu $|\partial_C(u)|=1$ với mọi $u \in U$ thì DS là *nhất quán*, trái lại DS là *không nhất quán*. Từ $T_C = \bigcap_{a \in C} T_a$, theo định nghĩa hàm quyết định suy rộng ta dễ dàng suy ra $\partial_C u = \bigcap_{a \in C} \partial_a u$ với mọi $u \in U$. Nếu $B \subseteq C$ thì từ $T_C u \subseteq T_B u$ ta dễ dàng suy ra $\partial_C u \subseteq \partial_B u$ với mọi $u \in U$.

Ví dụ 1.7. Xét bảng quyết định đầy đủ $DS = U, C \cup d$ cho ở Bảng 1.10 với thuộc tính quyết định d (*Gia tốc*), với $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ Và các tập thuộc tính $C = \{\text{Đơn giá, KM đã đi, Kích thước, Tốc độ, Gia tốc}\}$

Bảng 1. 10. Bảng quyết định về các xe hơi

Ô tô	Đơn giá	Km đã đi	Kích thước	Tốc độ	Gia tốc
u_1	Cao	Cao	Đầy đủ	Thấp	Tốt
u_2	Thấp	Cao	Đầy đủ	Thấp	Tốt
u_3	Cao	Cao	Gọn nhẹ	Cao	Xấu
u_4	Cao	Cao	Đầy đủ	Cao	Tốt
u_5	Thấp	Cao	Đầy đủ	Cao	Tuyệt hảo
u_6	Thấp	Cao	Đầy đủ	Cao	Tốt

Ta có $U/\{d\} = \{X_1, X_2, X_3\}$ với $X_1 = \{u_1, u_2, u_4, u_6\}$, $X_2 = \{u_3\}$, $X_3 = \{u_5\}$.

Các tập xấp xỉ dưới đối với C là $\underline{C}X_1 = \{u_1, u_2\}$, $\underline{C}X_2 = \{u_3\}$, $\underline{C}X_3 = \{\emptyset\}$.

Do đó, $POS_C(\{d\}) = \{u_1, u_2, u_3\}$

Hàm quyết định suy rộng của các đối tượng trên tập thuộc tính A là:

$$\partial_C(u_1) = \{Tốt\}$$

$$\partial_C(u_2) = \{Tốt\}$$

$$\partial_C(u_3) = \{Xấu\}$$

$$\partial_C(u_4) = \{Tốt, Tuyệt hảo\}$$

$$\partial_C(u_5) = \{Tốt, Tuyệt hảo\}$$

$$\partial_C(u_6) = \{Tốt, Tuyệt hảo\}.$$

Do đó, DT là bảng quyết định không nhất quán.

Định nghĩa 1.9. Cho bảng quyết định giá trị tập $DS = U, C \cup d$. Nếu

$R \subseteq C$ thỏa mãn:

$$(1) \partial_R u = \partial_C u \text{ với mọi } u \in U$$

$$(2) \text{ Với mọi } \forall R' \subset R, \text{ tồn tại } u \in U \text{ sao cho } \partial_{R'} u \neq \partial_C u$$

thì R được gọi là một tập rút gọn của DS dựa trên hàm quyết định suy rộng.

Ví dụ 1.11. Xét bảng quyết định tập giá trị $DS = U, C \cup d$ cho ở Bảng 1.11 với $U = u_1, u_2, u_3, u_4, u_5, u_6$ và $C = a_1, a_2, a_3, a_4$ và cột thuộc tính d .

Bảng 1. 11. Bảng quyết định tập giá trị

U	a_1	a_2	a_3	a_4	d
u_1	$\{0\}$	$\{0\}$	$\{1, 2\}$	$\{1, 2\}$	0
u_2	$\{0, 1, 2\}$	$\{0, 1, 2\}$	$\{1, 2\}$	$\{0, 1, 2\}$	1
u_3	$\{1, 2\}$	$\{0, 1\}$	$\{1, 2\}$	$\{1, 2\}$	0
u_4	$\{0, 1\}$	$\{0, 2\}$	$\{1, 2\}$	$\{1\}$	1
u_5	$\{1, 2\}$	$\{1, 2\}$	$\{1, 2\}$	$\{1\}$	0
u_6	$\{1\}$	$\{1, 2\}$	$\{0, 1\}$	$\{0, 1\}$	1
u_7	$\{0\}$	$\{0\}$	$\{1, 2\}$	$\{1, 2\}$	0
u_8	$\{1\}$	$\{1, 2\}$	$\{0, 1\}$	$\{0, 1\}$	1

Với đối tượng $u_1 \in U$ ta có

$$T_{a_1} u_1 = u_1, u_3, u_4, u_5$$

$$T_{a_3} u_1 = u_1, u_2, u_4, u_5, u_6$$

$$T_{a_2} u_1 = U$$

$$T_{a_4} u_1 = u_1, u_2, u_6$$

Do đó $T_C u_1 = T_{a_1} u_1 \cap T_{a_2} u_1 \cap T_{a_3} u_1 \cap T_{a_4} u_1 = u_1$.

Tương tự, ta tìm các quan hệ dung sai với đối tượng $u_2 \in U$ như sau:

$$T_{a_1} u_2 = u_2, u_3, u_5, u_6$$

$$T_{a_3} u_2 = u_1, u_2, u_4, u_5, u_6$$

$$T_{a_2} u_2 = U$$

$$T_{a_4} u_2 = u_1, u_2, u_6$$

Do đó $T_C u_2 = T_{a_1} u_2 \cap T_{a_2} u_2 \cap T_{a_3} u_2 \cap T_{a_4} u_2 = u_2, u_6$.

Các quan hệ dung sai với đối tượng $u_3 \in U$

$$T_{a_1} u_3 = U$$

$$T_{a_3} u_3 = u_3$$

$$T_{a_2} u_3 = U$$

$$T_{a_4} u_3 = u_3, u_4, u_5, u_6$$

$$T_C u_3 = T_{a_1} u_3 \cap T_{a_2} u_3 \cap T_{a_3} u_3 \cap T_{a_4} u_3 = u_3$$

Các quan hệ dung sai với đối tượng $u_4 \in U$

$$T_{a_1} u_4 = u_1, u_3, u_4, u_5 \quad T_{a_3} u_4 = u_1, u_2, u_4, u_5, u_6$$

$$T_{a_2} u_4 = U \quad T_{a_4} u_4 = u_3, u_4, u_5, u_6$$

$$T_C u_4 = T_{a_1} u_4 \cap T_{a_2} u_4 \cap T_{a_3} u_4 \cap T_{a_4} u_4 = u_4, u_5$$

Các quan hệ dung sai với đối tượng $u_5 \in U$

$$T_{a_1} u_5 = U \quad T_{a_3} u_5 = u_1, u_2, u_4, u_5, u_6$$

$$T_{a_2} u_5 = U \quad T_{a_4} u_5 = u_3, u_4, u_5, u_6$$

$$T_C u_5 = T_{a_1} u_5 \cap T_{a_2} u_5 \cap T_{a_3} u_5 \cap T_{a_4} u_5 = u_4, u_5, u_6$$

Các quan hệ dung sai với đối tượng $u_6 \in U$

$$T_{a_1} u_6 = \{u_2, u_3, u_5, u_6\} \quad T_{a_3} u_6 = u_1, u_2, u_4, u_5, u_6$$

$$T_{a_2} u_6 = U \quad T_{a_4} u_6 = U$$

$$T_C u_6 = T_{a_1} u_6 \cap T_{a_2} u_6 \cap T_{a_3} u_6 \cap T_{a_4} u_6 = u_2, u_5, u_6$$

Vậy ta có:

$$T_C u_1 = u_1, T_C u_2 = u_2, u_6, T_C u_3 = u_3,$$

$$T_C u_4 = u_4, u_5, T_C u_5 = u_4, u_5, u_6, T_C u_6 = u_2, u_5, u_6.$$

Hơn nữa, $\partial_C(u_1) = \partial_C(u_2) = 1$, $\partial_C(u_3) = 0$, $\partial_C(u_4) = \partial_C(u_5) = \partial_C(u_6) = 1, 2$. Do đó, DS không nhất quán.

CHƯƠNG 2: RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH TẬP GIÁ TRỊ

2.1. Đặt vấn đề

Rút gọn thuộc tính trong hệ quyết định là quá trình lựa chọn tập con nhỏ nhất của tập thuộc tính điều kiện mà bảo toàn thông tin phân lớp của hệ quyết định. Trong các hệ quyết định đơn trị có rất nhiều phương pháp rút gọn thuộc tính dựa trên mô hình tập thô truyền thống được đề xuất. Trong hệ thông tin tập giá trị, mở rộng quan hệ tương đương trong lý thuyết tập thô truyền thống thành quan hệ dung sai và xây dựng mô hình tập thô dung sai bằng cách mở rộng các định nghĩa xấp xỉ trên, xấp xỉ dưới, miền dương... dựa trên quan hệ dung sai.

Trong chương này, luận văn đi khai thác một số thuật toán liên quan đến bảng quyết định tập giá trị. Đầu tiên, khai thác thuật toán tìm xấp xỉ trên- xấp xỉ dưới của một tập dữ liệu trong bảng quyết định tập giá trị của hệ thông tin tập giá trị. Phần tiếp theo tôi khai thác thuật toán tìm tập rút gọn thuộc tính trong bảng quyết định tập giá trị dựa trên cơ sở của các hàm phân biệt và bảng ngẫu nhiên.

2.2. Cơ sở lý thuyết

Phần này tập trung khai thác tới một dạng đặc biệt của bảng quyết định, trong đó tập thuộc tính quyết định chỉ bao gồm một thuộc tính quyết định d , vì vậy bảng quyết định được biểu diễn dưới dạng $DS = (U, C \cup d)$. Trong trường hợp bảng quyết định tập giá trị, giá trị tại thuộc tính quyết định d là đơn trị.

2.2.1. Hàm phân biệt ngẫu nhiên

Xây dựng hàm phân biệt ngẫu nhiên là công cụ để giải quyết bài toán rút gọn thuộc tính được định nghĩa dựa trên hai đại lượng đó là vùng dương và

vùng biên. Đây là cách tiếp cận để đánh giá thuộc tính (hoặc tập thuộc tính) dựa trên số lượng các cặp đối tượng có nhãn khác nhau, chúng được phân biệt bởi một thuộc tính (tập thuộc tính) và được gọi là hàm phân biệt ngẫu nhiên.

Định nghĩa 2.1. Hàm phân biệt ngẫu nhiên

Cho $DS = U, C \cup d$ là bảng quyết định tập giá trị với quan hệ dung sai T_B trên 2^C . Hàm phân biệt ngẫu nhiên của bảng quyết định tập giá trị $discern: 2^C \rightarrow Z^+$ được định nghĩa như sau:

$$\forall B \subseteq C : discern(B) = | \{ (x, y) \in U \times U \mid (d(x) \neq d(y)) \wedge \exists_{b \in B} (x, y) \in T_b \} |$$

Dưới đây là một số tính chất cơ bản của hàm phân biệt ngẫu nhiên :

1). Trong bảng quyết định tập giá trị $DS = U, C \cup d$ và $a \in C$, giá trị của hàm phân biệt ngẫu nhiên $discern(a)$ chính là số lần xuất hiện của thuộc tính a trong ma trận phân biệt M_{DS} : $discern(a) = | \{ a \mid a \in M_{DS} \} |$.

2). Hàm phân biệt ngẫu nhiên có tính đơn điệu tăng theo nghĩa sau đây:

$$\forall B_1, B_2 \subseteq C \text{ nếu } B_1 \subseteq B_2 \text{ thì } discern(B_1) \leq discern(B_2).$$

3). Cho $DS = U, C \cup d$ là bảng quyết định tập giá trị, $\forall B_1, B_2 \subseteq C, B_1 \subseteq B_2$.

$$\text{Khi đó } discern(B_1) \leq discern(B_2) \text{ nếu } \forall (x, y) \in U^2 : (x, y) \in T_{B_1} \Rightarrow (x, y) \in T_{B_2}.$$

2.2.2. Bảng ngẫu nhiên CT và bảng ngẫu nhiên dựa trên quan hệ dung sai TCT

Cho $DS = (U, C \cup d)$ là bảng quyết định tập giá trị. Đặt $|U| = n$ và $U = \{u_1, \dots, u_n\}$. Ta có thể thấy với thuộc tính $a \in A$, số lần xuất hiện của a trong ma trận M_{DT} được tính bằng $DTIME(n^2)$ khi quét tất cả các ô trong ma trận. Thời gian này có thể được cải thiện khi người ta sử dụng hàm phân biệt đã đề xuất và bảng cấu trúc điều kiện gọi là bảng ngẫu nhiên. Khái niệm bảng ngẫu nhiên được định nghĩa dựa trên bảng quyết định hệ thông tin đơn trị. Rõ

ràng bảng ngẫu nhiên là một cấu trúc, nó biểu diễn các thông tin về phân lớp liên quan đến thuộc tính quyết định dựa trên quan hệ không phân biệt được.

Việc sử dụng cấu trúc này giúp chúng ta xác định một cách nhanh chóng tần suất xuất hiện của các thuộc tính trong ma trận mà không cần kiểm tra sự xuất hiện của các thuộc tính trong các ô. Phần này luận văn sẽ giới thiệu lần lượt các khái niệm liên quan đến bảng ngẫu nhiên.

2.2.2.1. Bảng CT_B được xây dựng dựa trên quan hệ tương đương $IND(B)$

Cho $DS = U, C \cup d$ là bảng quyết định tập giá trị, V_d là giá trị của tập quyết định (hay còn gọi là nhãn) của bảng quyết định DS , $U / IND(B) = x_{1B}, x_{2B}, \dots, x_{n_B B}$ là phân hoạch của U được xác định theo quan hệ không phân biệt được $IND(B)$ với $B \subseteq A$. Khi đó, *bảng phân biệt ngẫu nhiên (contingency table) liên quan tới B (CT_B)* là một bảng $CT_B = [CT_B[i, j]]_{i \in \{1, \dots, n_B\}}^{j \in \{1, \dots, |V_d|\}}$ với các giá trị phần tử được định nghĩa như sau:

$$CT_B[i, j] = |\{y \in U : y \in [x_i]_B \wedge d(y) = j\}|$$

Trong ngữ cảnh cố định B, gọi ngắn gọn *bảng phân biệt ngẫu nhiên liên quan tới B* là *bảng phân biệt ngẫu nhiên*.

Với các điều kiện trên, độ phân biệt cục bộ liên quan tới lớp không phân biệt được x_{iB} , được ký hiệu là $\delta_B x_{iB}$, được xác định như sau:

$$\begin{aligned} \delta_B x_{iB} &= \left| (y, z) \in x_{iB} \times U \setminus x_{iB} : dec(y) \neq dec(z) \right| \\ &= \sum_{j_1 \neq j_2, x_k \notin x_{iB}} CT_{i, j_1} . CT_{k, j_2} \\ &= \sum_{j_1 \neq j_2} CT_{i, j_1} \cdot |D_{j_2}| - CT_{i, j_2} \end{aligned}$$

trong đó $|D_j|$ là ký hiệu đề chỉ lực lượng của tập hợp các đối tượng thuộc lớp quyết định D_j với $j = 1, \dots, |V_d|$ $D_j = \{u \in U \mid d(u) = v_{d_j}\}$. Vì vậy độ phân biệt cơ sở (*basic discernibility measure*, hay *hàm phân biệt ngẫu nhiên*) của tập

thuộc tính B, ký hiệu là $discern(B)$, được định nghĩa là số cặp các đối tượng phân biệt:

$$discern(B) = \sum_i \delta_B([x_i]_B) = \frac{1}{2} \sum_{i=1}^{n_B} \sum_{j_1 \neq j_2} CT[i, j_1] (|D_{j_2}| - CT[i, j_2]) \quad (2.1)$$

Ví dụ 2.1. Tạo bảng phân biệt ngẫu nhiên CT và tính hàm $discern$ của các thuộc tính $\{Nghe(A), Ng\ddot{u}n\ ng\ddot{u}i\ n\ddot{o}i\ (S), \acute{D}o\c{c}\ (R), Vi\acute{e}t\ (W)\}$ từ Bảng 1.8 đã được cho trong Ví dụ 1.8

Tính hàm $discern$ lần lượt với các thuộc tính $\{A, S, R, W\}$

* Xét thuộc tính A (*Nghe*), tính phân hoạch U/A , từ Bảng 1.8, có:

$$x_{1\ A} = x_1 \quad (\text{Value là } \{E\})$$

$$x_{2\ A} = x_{7\ A} = x_2, x_7 \quad (\text{Value là } \{E, F, G\})$$

$$x_{3\ A} = x_3 \quad (\text{Value là } \{E, G\})$$

$$x_{4\ A} = x_{8\ A} = x_{10\ A} = x_4, x_8, x_{10} \quad (\text{Value là } \{E, F\})$$

$$x_{5\ A} = x_{9\ A} = x_5, x_9 \quad (\text{Value là } \{F, G\})$$

$$x_{6\ A} = x_6 \quad (\text{Value là } \{F\})$$

Như vậy phân hoạch U/A có 6 lớp tương đương.

Chúng ta tính hàm $discern(A)$ theo công thức (2.1):

$$\sum_{j_1 \neq j_2} CT[i, j_1] \cdot |D_{j_2}| - CT[i, j_2] \quad \text{và}$$

$$discern(A) = \sum_i \delta_A([x_i]_A) = \frac{1}{2} \sum_{i=1}^{n_A} \sum_{j_1 \neq j_2} CT[i, j_1] \cdot (|D_{j_2}| - CT[i, j_2])$$

với $n_A = 6$

1. Ta tính $\delta_A u_{1\ A}$: Ta có $u_{1\ A}$ là lớp nhất quán, $j_1 = NO, j_2 = YES$ và $CT[1, No] = 1, |D_{YES}| = 5, CT[1, YES] = 0$. Do đó $\delta_A u_{1\ A} = 1 \cdot 5 - 0 = 5$

2. Tính $\delta_A x_{2\ A}$: $x_{2\ A}$ là lớp không nhất quán.

- Xét $j_1 = NO, j_2 = YES, CT_{2,NO} = 1, |D_{YES}| = 5, CT_{2,YES} = 1$. Do đó

$$\delta_A x_{2A}^{NO} = 1 \cdot 5 - 1 = 4.$$

- Xét $j_1 = YES, j_2 = NO, CT_{2,YES} = 1, |D_{NO}| = 5, CT_{2,NO} = 1$. Do đó

$$\delta_A x_{2A}^{YES} = 1 \cdot 5 - 1 = 4$$

$$\text{Vậy } \delta_A x_{2A} = \delta_A x_{2A}^{NO} + \delta_A x_{2A}^{YES} = 4 + 4 = 8.$$

1) Làm tương tự với x_{3A} (một lớp nhất quán), nhận được:

$$\delta_A x_{3A} = 1 \cdot (5 - 0) = 5$$

2) Làm tương tự với x_{4A} (một lớp không nhất quán), nhận được:

$$\delta_A x_{4A} = \delta_A x_{4A}^{NO} + \delta_A x_{4A}^{YES} = 3 + 8 = 11.$$

3) Làm tương tự với x_{5A} (một lớp không nhất quán), nhận được:

$$\delta_A x_{5A} = \delta_A x_{5A}^{NO} + \delta_A x_{5A}^{YES} = 4 + 4 = 8.$$

4) Làm tương tự với x_{6A} (một lớp nhất quán), nhận được:

$$\delta_A x_{6A} = 1 \cdot (5 - 0) = 5.$$

Áp dụng công thức tính hàm $discern(A)$, chúng ta có:

$$discern(A) = \sum_i \delta_A([x_i]_A) = \frac{1}{2} \sum_{i=1}^6 \sum_{j_1 \neq j_2} CT[i, j_1].CT[i, j_2] = \frac{1}{2} (5 + 8 + 5 + 11 + 8 + 5) = 21.$$

• Xét thuộc tính S (Ngôn ngữ nói), tính phân hoạch U/S , từ Bảng 1.8 ta có:

$$x_{1S} = x_1 \quad (\text{Value là } \{E\})$$

$$x_{2S} = x_{7S} = x_2, x_7 \quad (\text{Value là } \{E, F, G\})$$

$$x_{3S} = x_3 \quad (\text{Value là } \{E, F\})$$

$$x_{4S} = x_{10S} = x_4, x_{10} \quad (\text{Value là } \{E, G\})$$

$$x_{5_s} = x_{8_s} = x_5, x_8 \quad (\text{Value là } \{F, G\})$$

$$x_{6_s} = x_6 \quad (\text{Value là } \{F\})$$

$$x_{9_s} = x_9 \quad (\text{Value là } \{G\})$$

Như vậy phân hoạch U/S có 7 lớp tương đương.

Với cách làm tương tự, chúng ta tính được giá trị hàm $discern(S)$ theo công thức (2.1) như dưới đây :

$$1) \quad \text{Với lớp nhất quán } x_{1_s} : \text{nhận được : } \delta_s x_{1_s} = 1.(5-0) = 5.$$

$$2) \quad \text{Với lớp không nhất quán } x_{2_s} \text{ nhận được :}$$

$$\delta_s x_{2_s} = \delta_s x_{2_s}^{NO} + \delta_s x_{2_s}^{YES} = 4 + 4 = 8.$$

$$3) \quad x_{3_s} \text{ (một lớp nhất quán) } \delta_s x_{3_s} = 1.(5-0) = 5$$

$$4) \quad x_{4_s} \text{ (lớp không nhất quán) có :}$$

$$\delta_s x_{4_s} = \delta_s x_{4_s}^{NO} + \delta_s x_{4_s}^{YES} = 4 + 4 = 8,$$

$$5) \quad x_{5_s} \text{ (lớp không nhất quán) có:}$$

$$\delta_s x_{5_s} = \delta_s x_{5_s}^{NO} + \delta_s x_{5_s}^{YES} = 4 + 4 = 8,$$

$$6) \quad x_{6_s} \text{ (lớp nhất quán) có } \delta_s x_{6_s} = 1.(5-0) = 5$$

$$7) \quad x_{9_s} \text{ (lớp nhất quán) có } \delta_s x_{9_s} = 1.(5-0) = 5.$$

Áp dụng công thức tính hàm $discern(S)$, nhận được:

$$discern(S) = \frac{1}{2} \sum_{i=1}^7 \sum_{j_1 \neq j_2} CT[i, j_1] \cdot |D_{j_2}| \cdot -CT[i, j_2] = \frac{1}{2} (5+8+5+8+8+5+5) = 22.$$

• Xét thuộc tính R (Độc), tính phân hoạch U/R , từ Bảng 1.8 ta có:

$$x_{1_R} = x_{2_R} = x_{3_R} = x_{4_R} = x_{5_R} = x_{9_R} = x_{10_R} \quad (\text{Value là } \{F, G\})$$

$$= x_1, x_2, x_3, x_4, x_5, x_9, x_{10}$$

$$x_{6_R} = x_6 \quad (\text{Value là } \{E, F\})$$

$$x_{7_R} = x_7 \quad (\text{Value là } \{E, G\})$$

$$x_{8 \ R} = x_8$$

(Value là {E, F, G})

Như vậy phân hoạch U/R có 4 lớp tương đương.

Tính tương tự, chúng ta có:

$$\text{discern}(R) = \frac{1}{2} \sum_{i=1}^4 \sum_{j_1 \neq j_2} CT[i, j_1] \cdot |D_{j_2}| - CT[i, j_2] = \frac{1}{2}(15+5+5+5) = 15.$$

- Xét thuộc tính W (Viết), tính phân hoạch U/W , từ Bảng 1.8 ta có:

$$x_{1 \ W} = x_{3 \ W} = x_{9 \ W} = x_1, x_3, x_9 \quad (\text{Value là } \{F, G\})$$

$$x_{2 \ W} = x_{7 \ W} = x_2, x_7 \quad (\text{Value là } \{E, F, G\})$$

$$x_{4 \ W} = x_{5 \ W} = x_4, x_5 \quad (\text{Value là } \{F\})$$

$$x_{6 \ W} = x_{10 \ W} = x_6, x_{10} \quad (\text{Value là } \{E, F\})$$

$$x_{8 \ W} = x_8 \quad (\text{Value là } \{E, G\})$$

Như vậy phân hoạch U/W có 5 lớp tương đương. Bằng cách tính tương tự, nhận được

$$\text{discern}(W) = \frac{1}{2} \sum_{i=1}^5 \sum_{j_1 \neq j_2} CT[i, j_1] \cdot |D_{j_2}| - CT[i, j_2] = \frac{1}{2}(11+8+10+10+5) = 22.$$

Dưới đây là Bảng 2.1 biểu diễn giá trị của hàm phân biệt mở rộng trong bảng CT với các kết quả được tính trong Ví dụ 2.1 liên quan đến tập thuộc tính đơn giá trị của tập dữ liệu cho trong Bảng 1.8 :

Bảng 2. 1. Bảng phân biệt ngẫu nhiên biểu diễn giá trị tập thuộc tính và hàm phân biệt

Nghe (A)			Ngôn ngữ nói (S)			Đọc (R)			Viết (W)		
Value	No	Yes	Values	No	Yes	Values	No	Yes	Value	No	Yes
s									s		
<i>E</i>	1	0	<i>E</i>	1	0	<i>E, F</i>	0	1	<i>F</i>	2	0
<i>F</i>	0	1	<i>F</i>	0	1	<i>F, G</i>	0	1	<i>E, F</i>	0	2
<i>E, F</i>	1	2	<i>G</i>	0	1	<i>E, G</i>	5	2	<i>E, G</i>	0	1

F, G	1	1	E, F	1	0	E, F, G	0	1	F, G	2	1
E, G	1	0	E, G	1	1				E, F, G	1	1
E, F, G	1	1	F, G	1	1						
			E, F, G	1	1						
discern (A) = 21			discern (S) = 22			discern (R) = 15			discern (W) = 22		

Nhận xét 1: Cho $DS = U, C \cup d$ là bảng quyết định giá trị tập. Đặt $IND(B)$ là quan hệ không phân biệt được với $B \subseteq C$. Giả sử n_B là ký hiệu số lớp không phân biệt được được định nghĩa dựa trên $IND(B)$. Với bảng phân biệt ngẫu nhiên CT_B , giá trị của hàm phân biệt (B) có thể xác định độ phức tạp theo thời gian là $O(dn_B)$; và đường biên được tính bằng $O(dn)$, trong đó $n = card(U)$ hay $n = |U|$ (số đối tượng) và d là số lớp quyết định.

2.2.2.2. Bảng TCT_B được xây dựng dựa trên quan hệ dung sai T

Cho $DS = U, C \cup d$ là bảng quyết định tập giá trị.

Cho $U / IND(B) = x_{1_B}, x_{2_B}, \dots, [x_{n_B}]_B$ là phân hoạch của U được định nghĩa dựa trên quan hệ bất khả phân biệt $IND(B)$, T_B tương ứng là quan hệ không phân biệt được và quan hệ dung sai liên quan tới tập thuộc tính điều kiện $B \subseteq C$. Khi đó, *bảng ngẫu nhiên dung sai liên quan tới B* là một bảng hai chiều $TCT_B = [TCT_{i,j}]_{i \in 1, \dots, n_B}^{j \in 1, \dots, |V_d|}$ với các giá trị phần tử được định nghĩa như sau:

$$TCT_B_{i,j} = \left| y \in x_{i_{T_B}} \wedge dec(y) = j \right|$$

Bảng ngẫu nhiên dung sai là bảng biểu diễn sự phân biệt của các lớp dung sai liên quan tới thuộc tính quyết định. Khó có thể phân hoạch được vì các lớp dung sai trong U / TB có thể giao nhau. Vì vậy, luận văn xin trình bày ra khái

niệm độ phân biệt ngẫu nhiên dung sai (độ ngẫu nhiên dung sai) và độ phân biệt tổng quát của một tập thuộc tính.

Độ ngẫu nhiên dung sai (hay hàm phân biệt mở rộng dựa trên quan hệ dung sai) được tính toán theo công thức sau:

$$\begin{aligned}\delta_B([x_i]_{T_B}) &= |\{(y, z) \in [x_i]_{T_B} \times (U \setminus [x_i]_{T_B}) : dec(y) \neq dec(z)\}| \\ &= \sum_{j_1 \neq j_2, x_k \in [x_i]_{T_B}} CT_B[i, j_1] \times CT_B[k, j_2] \\ &= \sum_{j_1 \neq j_2} CT_B[i, j_1] (|D_{j_2}| - TCT_B[i, j_2])\end{aligned}$$

Độ phân biệt tổng (hay hàm ngẫu nhiên dung sai) của tập thuộc tính B, ký hiệu là $discern(B)$, được định nghĩa như sau:

$$discern(B) = \sum_i \delta_B([x_i]_{T_B}) = \frac{1}{2} \sum_{i=1}^{n_B} \sum_{j_1 \neq j_2} CT_B[i, j_1] (|D_{j_2}| - TCT_B[i, j_2]) \quad (2.2)$$

Chúng ta ký hiệu $CT(B) \otimes TCT(B)$ là phép toán tính theo công thức 2.2. Được tính bằng tổng các phân hoạch rời nhau dựa trên quan hệ dung sai với $j_1, j_2 \in \{1, \dots, |V_d|\}$, $j_1 \neq j_2$. Rõ ràng là hàm phân biệt $discern(B)$ phụ thuộc vào 2 bảng ngẫu nhiên đó là bảng CT_B và bảng ngẫu nhiên tổng quát hóa TCT_B .

Nhận xét 2: Cho $DS = U, C \cup d$ là bảng quyết định tập giá trị, và hai bảng CT_B và bảng TCT_B , với $B \subseteq A$. Đặt $x_1_B, x_2_B, \dots, [x_{n_B}]_B$ là các lớp không phân biệt được được định nghĩa dựa trên quan hệ $IND(B)$. Các phần tử của bảng phân biệt ngẫu nhiên $TCT_B = [TCT_B[i, j]]_{i \in \{1, \dots, n_B\}}^{j \in \{1, \dots, |V_d|\}}$ có thể tính thông qua bảng CT_B lấy trên các đối tượng thuộc các lớp dung sai:

$$TCT_B[i, j] = \sum_{k \in 1, \dots, n_B, x_k \in [x_i]_{T_B}} CT[k, j]$$

Ví dụ 2.2. Tạo bảng phân biệt ngẫu nhiên dựa trên quan hệ dung sai để tính hàm $discern$ từ Bảng dữ liệu 1.8.

- Xét thuộc tính A (Nghe), tính phủ U / T_A , từ Bảng 1.8 ta có:

$$x_{1 \ T_A} = x_1, x_2, x_3, x_4, x_7, x_8, x_{10}$$

$$x_{2 \ T_A} = x_{4 \ T_A} = x_{7 \ T_A} = x_{8 \ T_A} = x_{10 \ T_A} = U$$

$$x_{3 \ T_A} = x_1, x_2, x_3, x_4, x_5, x_7, x_8, x_9, x_{10}$$

$$x_{5 \ T_A} = x_{9 \ T_A} = x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$$

$$x_{6 \ T_A} = x_2, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$$

Tính hàm *discern* A bởi công thức (2.2):

$$\delta_A([x_i]_{T_A}) = \sum_{j_1 \neq j_2} CT_A[i, j_1](|D_{j_2}| - TCT[i, j_2]) \text{ và}$$

$$Discern(A) = \sum_i \delta_A([x_i]_{T_A}) = \frac{1}{2} \sum_{i=1}^{n_A} \sum_{j_1 \neq j_2} CT_A[i, j_1](|D_{j_2}| - TCT_A[i, j_2]) \text{ với } n_A = 5.$$

1. Tính $\delta_A \ x_{1 \ T_A}$: Theo tính toán ở trên, $x_{1 \ T_A}$ là lớp nhất quán, $j_1 = NO, j_2 = YES$ và $CT \ 1, No = 1, |D_{YES}| = 5, TCT \ 1, YES = 3$.

$$\text{Do đó } \delta_A \ x_{1 \ T_A} = 1.(5 - 3) = 2.$$

2. Đối với $x_{2 \ T_A}$ (một lớp không nhất quán):

- Xét $j_1 = NO, j_2 = YES$, $CT \ 2, No = 1, |D_{YES}| = 5, TCT \ 2, YES = 5$. Do đó

$$\delta_A \ x_{2 \ T_A}^{NO} = 1.(5 - 5) = 0.$$

- Xét $j_1 = YES, j_2 = NO$, $CT \ 2, YES = 1, |D_{NO}| = 5, TCT \ 2, NO = 5$. Do đó

$$\delta_A \ x_{2 \ T_A}^{YES} = 1.(5 - 5) = 0.$$

$$\text{Vậy } \delta_A \ x_{2 \ T_A} = \delta_A \ x_{2 \ T_A}^{NO} + \delta_A \ x_{2 \ T_A}^{YES} = 0 + 0 = 0.$$

3. Tương tự với $x_{3 \ T_A}$ (lớp nhất quán) nhận được $\delta_A \ x_{3 \ T_A} = 1.(5 - 4) = 1$

4. Với $x_{5 \ T_A}$ là (lớp không nhất quán) nhận được

$$\delta_A x_{5 T_A} = \delta_A x_{5 T_A}^{NO} + \delta_A x_{5 T_A}^{YES} = 0 + 1 = 1$$

5. với $x_{6 A}$ là (lớp nhất quán), nhận được $\delta_A x_{6 T_A} = 1.(5-3) = 2$.

Áp dụng công thức tính hàm *discern* A , có:

$$discern A = \frac{1}{2} \sum_{i=1}^5 \sum_{j_1 \neq j_2} CT_{A i, j_1} \cdot |D_{j_2}| - TCT_{A i, j_2} = \frac{1}{2} 2 + 0 + 1 + 1 + 2 = 3$$

• *Tương tự với thuộc tính S (Ngôn ngữ nói):*

$$discern S = \frac{1}{2} \sum_{i=1}^7 \sum_{j_1 \neq j_2} CT_{S i, j_1} \cdot |D_{j_2}| - TCT_{S i, j_2} = \frac{1}{2} 2 + 0 + 1 + 1 + 1 + 2 + 1 = 4$$

• *Tương tự với thuộc tính R (Đọc):*

$$discern R = \frac{1}{2} \sum_{i=1}^2 \sum_{j_1 \neq j_2} CT_{R i, j_1} \cdot |D_{j_2}| - TCT_{R i, j_2} = \frac{1}{2} 0 + 0 = 0$$

• *Tương tự với thuộc tính W (Viết):*

$$discern W = \frac{1}{2} \sum_{i=1}^3 \sum_{j_1 \neq j_2} CT_{W i, j_1} \cdot |D_{j_2}| - TCT_{W i, j_2} = \frac{1}{2} 2 + 2 = 2$$

Bảng 2.2. Biểu diễn giá trị của hàm phân biệt mở rộng trong bảng phân biệt ngẫu nhiên CT_B dựa trên quan hệ dung sai với các kết quả được tính từ Ví dụ 2.2.

Bảng 2. 2. Minh họa giá trị của hàm phân biệt

$a_1 = \text{Audition}$					
Set	CT_{a_1}		TCT_{a_1}		δ_{a_1}
values	No	Yes	No	Yes	
E	1	0	4	3	$1.(5-3) + 0.(5-4) = 2$
F	0	1	3	5	$0.(5-5) + 1.(5-3) = 2$
E, F	1	2	5	5	$1.(5-5) + 2.(5-5) = 0$
F, G	1	1	4	5	$1.(5-5) + 1.(5-4) = 1$
E, G	1	0	4	4	$1.(5-4) + 0.(5-4) = 1$

E, F, G	1	1	5	5	$1.(5-5) + 1.(5-5) = 0$
$\text{discern}(a_1) = \frac{1}{2}(2+2+0+1+1+0)$					

Nhận xét 3: Cho tập thuộc tính $B \subseteq C$ bằng CT_B có thể nhận được nếu sử dụng câu lệnh *SQL* theo mẫu *SELECT COUNT .. GROUP BY B*.

Nhận xét 4: Bảng phân biệt ngẫu nhiên dựa trên quan hệ dung sai có thể tính được độ phức tạp theo thời gian trong trường hợp xấu nhất là $O(n_B^2 d)$ với n_B số các bản ghi dựa trên bảng CT_B , đó chính là miền biên được tính bằng lực lượng lớn nhất của các giá trị thuộc tính và d là số lớp quyết định.

Đối với bảng dữ liệu mà số lượng giá trị của các thuộc tính quan trọng nhỏ hơn số các đối tượng thì thuật toán tính TCT_B rất hiệu quả. Trong trường hợp khi số các bản ghi trong bảng phân biệt ngẫu nhiên CT lớn thì việc tạo bảng TCT từ bảng CT có thể tốn nhiều thời gian. Vì vậy trong phần tiếp theo luận án đưa ra một kỹ thuật tăng tốc độ thuật toán khi tính TCT . Mục đích của kỹ thuật mới là biểu diễn thông tin liên quan giữa giá trị của các tập thuộc tính đó là giá trị của thuộc tính thể hiện dạng lưới.

2.2.3. Giá trị thuộc tính biểu diễn qua mô hình lưới

Lưới trong toán học được biểu diễn bởi một tập đối tượng có thứ tự trong đó có hai yếu tố là cận trên và cận dưới. Lưới có thể biểu diễn bằng đồ thị có hướng $G = (V, E)$, với V là tập các đỉnh và E là tập các cạnh. Các đỉnh biểu diễn các phần tử của tập có thứ tự. Hướng của cạnh $(v_i, v_j) \in E$ nếu các phần tử của v_i nhỏ nhất thì v_j và khoảng cách v_i và v_j tạo thành một cạnh. Cấu trúc này được áp dụng để lưu trữ các giá trị của thuộc tính có giá trị đặc biệt

Định nghĩa 2.3. Lưới của giá trị thuộc tính

Cho $DS = U, C \cup d$ là bảng quyết định tập giá trị. Với V_a là miền giá

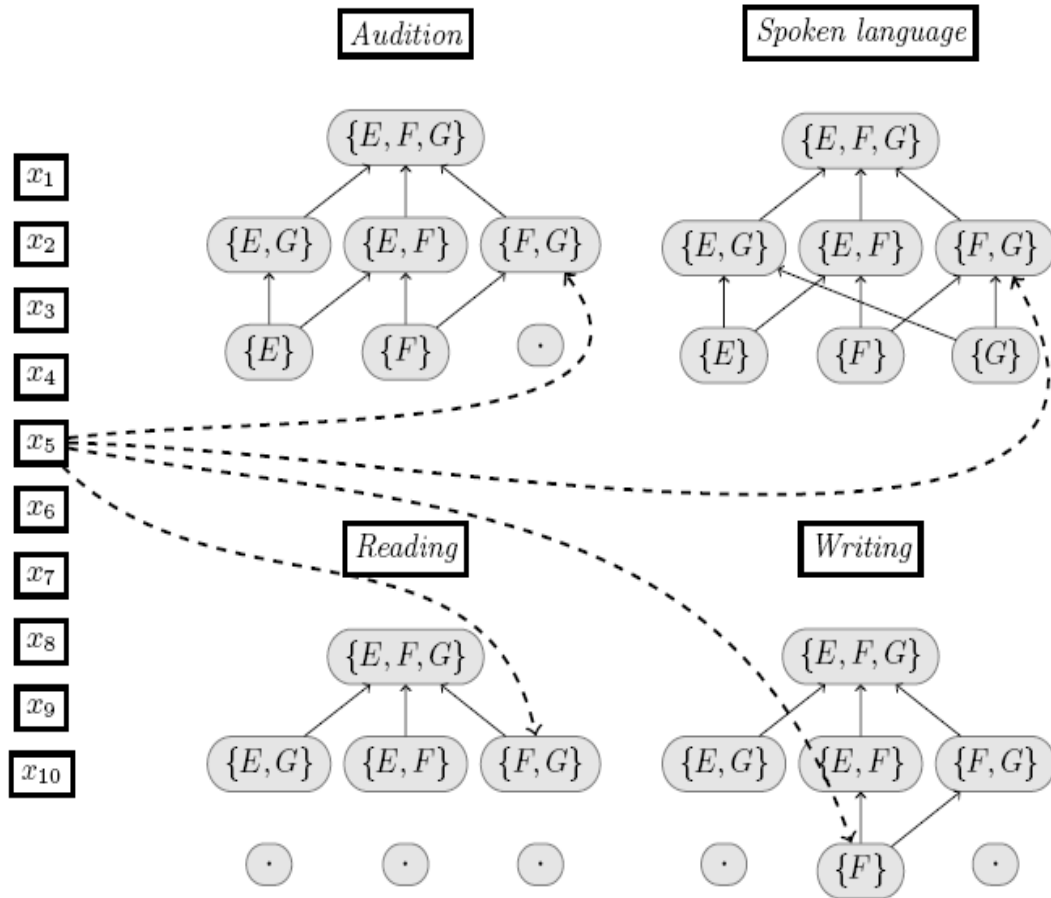
trị của thuộc tính $a \in C$. Lưới của thuộc tính a là một đồ thị có hướng được định nghĩa như là tập đối tượng có thứ tự $Latt(a) = (2^{V_a}, r)$, với $r \subseteq 2^{V_a} \times 2^{V_a}$ là một phần có thứ tự được định nghĩa bởi $r = (X, Y) \in 2^{V_a} \times 2^{V_a} : X \subseteq Y$.

Lưới liên quan tới thuộc tính a được ký hiệu là $Latt(a)$.

Tập của tất cả các lưới liên quan tới tất cả các thuộc tính ký hiệu $Latt(C)$.

Nhận xét 5: Cho $DS = U, C \cup d$ là bảng quyết định tập giá trị có thể chuyển sang thành cấu trúc bảng quyết định $SDTS = (U, Latt(C) \cup d)$.

Biểu diễn giá trị của thuộc tính qua mô hình lưới được minh họa bởi hình 2.1 (ở dưới) ta có thể thấy rằng lớp dung sai của tất cả các nút của $Latt(a)$ có thể xác định qua hai hướng của lưới quét: đó là từ dưới lên và từ trên xuống. Như vậy một lớp dung sai của một nút ở mức thấp nhất chứa ít nhất là một phần tử. Xét mô hình lưới từ dưới lên, tại mỗi nút biểu diễn tập hợp danh sách của lớp dung sai là những nút con. Xét mô hình lưới từ trên xuống những nút cha sẽ chia thành các nhánh chứa tập hợp danh sách các lớp dung sai con.



Hình 2. 1. Cấu trúc của bảng quyết định tập giá trị

Nhận xét 6: Cho thuộc tính $a \in C$, độ phức tạp của thuật toán tính lớp dung sai của các nút trong $Latt(a)$ được tính dựa vào đường biên và bằng $s_a * n_a$, với s_a là số lần xuất hiện giá trị của thuộc tính a và n_a là số các nút của lưới ($(n_a = |U / IND(a)|)$).

Khi có $Latt(a)$, với $a \in C$ ta có thể tính được bảng TCT_B .

Nhận xét 7: Cho $Latt(C)$ là tập dàn định nghĩa bởi tập thuộc tính điều kiện C . Giả sử mỗi một lớp dung sai $Latt(a)$ thì mỗi nút có thể tính toán được như sau. Cho tập thuộc tính $B \subseteq C$, độ phức tạp thời gian của thuật toán tính bảng ngẫu nhiên dung sai TCT_B từ bảng CT_B dựa vào $Latt(C)$ là $O(n_B)$, với n_B là số bản ghi trong bảng ngẫu nhiên cơ bản CT_B .

2.3. Thuật toán tìm tập rút gọn thuộc tính

2.3.1. Thuật toán 2.1 - Tìm tập rút gọn thuộc tính trong bảng quyết định tập giá trị

Đầu vào: Bảng quyết định tập giá trị $DS = U, C \cup d$.

Đầu ra: Tập rút gọn R .

Phương pháp:

1. Sinh tập các dàn $Latt(C)$;
2. $R = \emptyset$;
3. $discern(R) \leftarrow 0$;
4. **while** ($discern(R) < \max_disc(C)$) **do**
5. $\max_discern \leftarrow 0$;
6. **for** ($a_i \in C$) **do**
7. $B \leftarrow R \cup a_i$
8. Tạo bảng CT_B ;
9. Tạo bảng TCT_B sử dụng CT_B và $Latt(C)$;
10. Xác định $discern(B) = CT_B \otimes TCT_B$; theo công thức 2.2
11. **if** ($discern(B) > \max_discern$) **then**
12. $\max_discern \leftarrow discern(B)$;
13. $best_attribute \leftarrow a_i$;
14. **end if**
15. **end for**
16. $C \leftarrow C \setminus \{best_attribute\}$;
17. $R \leftarrow R \cup \{best_attribute\}$;
18. **end while**

* **Kết luận:** Giả sử k là số thuộc tính điều kiện và n là số đối tượng. Độ phức tạp để tính hàm phân biệt qua bảng CT là $O(kn^2)$.

Xét vòng lặp *While*, trước tiên xét độ phức tạp để tính hàm phân biệt qua hai bảng *CT* và *TCT* là $O(kn^2)$. Như vậy độ phức tạp của vòng *For* là $O(k^2n^2)$. Độ phức tạp thời gian để chọn thuộc tính rút gọn tốt nhất của lệnh *if* là $O(k)$. Do đó độ phức tạp của vòng lặp *While* là $O(k^3n^3)$. Vì vậy, độ phức tạp của thuật toán 2.1 là $O(k^3n^3)$

2.3.2. Ví dụ minh họa thuật toán 2.1

Ví dụ 2.3. Xét bảng quyết định tập giá trị $DS = (U, C \cup d)$ cho ở Bảng 2.3 với $U = u_1, u_2, u_3, u_4, u_5, u_6$ và $C = a_1, a_2, a_3, a_4$.

Bảng 2. 3. Bảng quyết định tập giá trị bao gồm 4 cột thuộc tính
(a_1, a_2, a_3, a_4)

U	a_1	a_2	a_3	a_4	d
u_1	{1}	{1}	{1}	{0}	1
u_2	{0}	{0, 1}	{1}	{0}	1
u_3	{0, 1}	{0, 1}	{0}	{1}	0
u_4	{1}	{0, 1}	{1}	{1}	1
u_5	{0, 1}	{0, 1}	{1}	{1}	2
u_6	{0}	{1}	{1}	{0, 1}	1

Minh họa các bước thực hiện của thuật toán rút gọn thuộc tính trong bảng quyết định tập giá trị

■ Chuẩn bị

Đặt $B = C$, từ bảng quyết định tập giá trị ta có:

$$u_{1 \ T_B} = u_1$$

$$u_{2 \ T_B} = u_2, u_6$$

$$u_{3 \ T_B} = u_3$$

$$u_{4 \text{ } T_B} = u_4, u_5$$

$$u_{5 \text{ } T_B} = u_4, u_5, u_6$$

$$u_{6 \text{ } T_B} = u_2, u_5, u_6$$

$$U / T_B = \{u_1, u_2, u_3\}, \{u_4, u_5\}, \{u_6\}$$

► Mô tả chi tiết cách tính $\text{discern } B = \text{discern } C$

Như vậy cần tính hàm $\text{discern } B$ cho 6 lớp tương đương theo công thức

$$\delta_B u_{i \text{ } T_B} = \sum_{j_1 \neq j_2} CT_B(i, j_1) \cdot |D_{j_2}| - TCT_B(i, j_2) \quad \text{và}$$

$$\text{discern } B = \sum_i \delta_B u_{i \text{ } T_B} = \frac{1}{2} \sum_{i=1}^{n_B} \sum_{j_1 \neq j_2} CT_B(i, j_1) \cdot |D_{j_2}| - TCT_B(i, j_2) \quad \text{với } n_B = 6.$$

1) Tính $\delta_B u_{1 \text{ } T_B}$: Theo tính toán trên nhận được $u_{1 \text{ } B}$ là lớp nhất quán.

- Xét $j_1 = 1, j_2 = 0$ và $CT_{1,1} = 2, |D_0| = 1, TCT_{1,0} = 1$.

Do đó $\delta_B u_{1 \text{ } B} = 1 \cdot 1 - 0 = 1$.

- Xét $j_1 = 1, j_2 = 2$ và $CT_{1,1} = 2, |D_2| = 1, TCT_{1,2} = 0$.

Do đó $\delta_B u_{1 \text{ } B} = 1 \cdot 1 - 0 = 1$.

Vậy: $\delta_B u_{1 \text{ } B} = 1 + 1 = 2$.

2) Tương tự, với $u_{2 \text{ } B}$ (lớp nhất quán) nhận được $\delta_B u_{2 \text{ } B} = 2$, với $u_{3 \text{ } B}$ là (lớp nhất quán) nhận được $\delta_B u_{3 \text{ } B} = 4 + 1 = 5$, với $u_{4 \text{ } B}$ (lớp nhất quán) nhận được $\delta_B u_{4 \text{ } B} = 1 + 1 = 2$, với $u_{5 \text{ } B}$ (lớp nhất quán) nhận được $\delta_B u_{5 \text{ } B} = 2 + 0 = 2$, với $u_{6 \text{ } B}$ (lớp nhất quán) nhận được $\delta_B u_{6 \text{ } B} = 2 + 0 = 2$.

Tổng hợp các kết quả tính toán trên, ta có: $\text{Discern}(B) = \text{Discern}(C) = 7$.

Tính tương tự tính các giá trị $discern a_1, discern a_2, discern a_3, discern a_4$,
nhận được:

$$Discern B = Discern a_1 = 0$$

$$Discern B = Discern a_2 = 0$$

$$Discern B = Discern a_3 = 5$$

$$Discern B = Discern a_4 = 4$$

Minh họa thuật toán trên qua Ví dụ 2.3.

1) *Tạo Latt(C)*

2) *Gán giá trị ban đầu:*

$$R = \emptyset;$$

$$discern R = 0$$

$$Max_disc = discern(C) = 7$$

3) *Thực hiện vòng lặp WHILE lần thứ 1*

$$max_discern = 0$$

$$best_attribute = 0$$

Thực hiện vòng lặp FOR chọn thuộc tính tốt nhất:

▪ Xét thuộc tính a_1 ta có $B = a_1$, $discern B = 0$

▪ Xét thuộc tính a_2 ta có $B = a_2$, $discern B = 0$

▪ Xét thuộc tính a_3 ta có $B = a_3$, $discern B = 5$, do đó $max_discern = 5$,

$$best_attribute = a_3$$

▪ Xét thuộc tính a_4 ta có $B = a_4$, $discern B = 4$

Do đó, kết thúc vòng lặp FOR, $C = C \setminus a_3 = a_1, a_2, a_4$, $R = R \cup a_3 = a_3$.

Kiểm tra điều kiện $discern(R) = 5 \neq discern(C) = 7$.

4) *Thực hiện vòng lặp WHILE lần thứ 2.*

Kiểm tra $\text{discern}(R) = 5 \neq \text{discern}(C) = 7$

$\text{max_discern} = 0$

$\text{best_attribute} = 0$

▪ Xét thuộc tính a_1 ta có $B = R \cup a_1 = a_1, a_3$, $\text{discern } B = 5$ do đó $\text{max_discern} = 5$, $\text{best_attribute} = a_1$

▪ Xét thuộc tính a_2 ta có $B = R \cup a_2 = a_2, a_3$, $\text{discern } B = 5$

▪ Xét thuộc tính a_4 ta có $B = R \cup a_4 = a_3, a_4$, $\text{discern } B = 7$ do đó $\text{max_discern} = 7$, $\text{best_attribute} = a_4$.

Kết thúc vòng lặp FOR, $C = C \setminus a_4 = a_1, a_2$, $R = R \cup a_4 = a_3, a_4$.

Kiểm tra điều kiện $\text{discern}(R) = 7 = \text{discern}(C) = 7$, kết thúc vòng lặp WHILE.

5) Thuật toán dừng, tập rút gọn tốt nhất của bảng quyết định tập giá trị tìm được theo thuật toán 2.1 là $R = a_3, a_4$.

2.4. Thuật toán tìm xấp xỉ trong hệ thông tin tập giá trị

2.4.1. Thuật toán 2.2- Thuật toán tìm xấp xỉ trên, xấp xỉ dưới sử dụng hàm phân biệt

Đầu vào: Hệ thông tin giá trị tập $IS = (U, A)$ với $X \subseteq U$, $B \subseteq A$

Quan hệ dung sai T_B , $U / \text{IND}(B) = \{1, 2, \dots, n_B\}$.

Đầu ra: Xấp xỉ trên của X , xấp xỉ dưới của X .

Phương pháp:

1. Tạo bảng quyết định $DS = (U, C \cup \{d_X\})$
2. Tạo bảng CT_B ;
3. Tạo bảng TCT_B từ bảng CT_B ;
4. **for** $i \in \{1, 2, \dots, n_B\}$ **do**

5. Tính hàm $v_i = \frac{TCT[i,1]}{TCT[i,1] + TCT[i,0]}$
6. **if** ($v_i = 1$) **then**
7. LowerAppr $\leftarrow \{i\}$
8. **else**
9. **if** ($v_i > 0$) **then**
10. Upper Appr $\leftarrow \{i\}$
11. **end if**
12. **end if**
13. **end for**

* **Kết luận:** Giả sử n là số đối tượng. Độ phức tạp để tạo bảng DS và bảng TCT là $O(kn^2)$.

Xét vòng lặp for độ phức tạp để tính hàm v là $O(kn^2)$. Độ phức tạp thời gian để xấp xỉ trên và xấp xỉ dưới của từng vòng lặp if là $O(kn)$ (bỏ qua d phân lớp quyết định). Vì vậy, độ phức tạp của Thuật toán 2.2 là $O(kn^2)$

2.4.2. Ví dụ minh họa thuật toán 2.2

Ví dụ 2.4. Xét hệ thông tin tập giá trị $IS = (U, A)$ cho ở Bảng 1.5 (bỏ đi thuộc tính quyết định d).

Giả sử $B = A$, $X = u_3, u_4, u_5, u_6$,

$U / IND B = u_1, u_2, u_7, u_3, u_4, u_8, u_{10}, u_5, u_9, u_6$

Tính xấp xỉ trên và xấp xỉ dưới của X theo Thuật toán 2.2

1) Tạo bảng quyết định như sau:

Bảng 2. 4. Bảng quyết định tập giá trị bao gồm 4 cột thuộc tính điều kiện và cột d_x

U	Audition(A)	Spoken	Reading(R)	Writing(W)	d_x
-----	-------------	--------	------------	------------	-------

		Language(S)			
u_1	$\{E\}$	$\{E\}$	$\{F, G\}$	$\{F, G\}$	0
u_2	$\{E, F, G\}$	$\{E, F, G\}$	$\{F, G\}$	$\{E, F, G\}$	0
u_3	$\{E, G\}$	$\{E, F\}$	$\{F, G\}$	$\{F, G\}$	1
u_4	$\{E, F\}$	$\{E, G\}$	$\{F, G\}$	$\{F\}$	1
u_5	$\{F, G\}$	$\{F, G\}$	$\{F, G\}$	$\{F\}$	1
u_6	$\{F\}$	$\{F\}$	$\{E, F\}$	$\{E, F\}$	1
u_7	$\{E, F, G\}$	$\{E, F, G\}$	$\{E, G\}$	$\{E, F, G\}$	0
u_8	$\{E, F\}$	$\{F, G\}$	$\{E, F, G\}$	$\{E, G\}$	0
u_9	$\{F, G\}$	$\{G\}$	$\{F, G\}$	$\{F, G\}$	0
u_{10}	$\{E, F\}$	$\{E, G\}$	$\{F, G\}$	$\{E, F\}$	0

2) Tạo $TCT_B = [TCT_B \ i, j]_{i \in 1, \dots, n_B}^{j \in 0, 1}$. Ta có:

Tính phủ U / T_B , từ bảng trên ta có:

$$u_{1 \ T_B} = u_1, u_2, u_3, u_4, u_7, u_8, u_{10}$$

$$u_{2 \ T_B} = u_{4 \ T_B} = u_{7 \ T_B} = u_{8 \ T_B} = u_{10 \ T_B} = U$$

$$u_{3 \ T_B} = u_1, u_2, u_3, u_4, u_5, u_7, u_8, u_9, u_{10}$$

$$u_{5 \ T_B} = u_{9 \ T_B} = u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}$$

$$u_{6 \ T_B} = u_2, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}$$

$$TCT \ 1,0 = 5, \ TCT \ 1,1 = 2$$

$$TCT \ 2,0 = 6, \ TCT \ 2,1 = 4$$

$$TCT \ 3,0 = 6, \ TCT \ 3,1 = 3$$

$$TCT \ 4,0 = 5, \ TCT \ 4,1 = 4$$

$$TCT_{5,0} = 5, TCT_{5,1} = 3$$

3) Thực hiện vòng lặp For với $i \in 1, 2, \dots, 5$

- Với $i=1$, $v_1 = \frac{TCT_{1,1}}{TCT_{1,1} + TCT_{1,0}} = \frac{2}{2+5} \approx 0.28$, do đó $UpperAppr = 1$;
- Với $i=2$, $v_2 = \frac{TCT_{2,1}}{TCT_{2,1} + TCT_{2,0}} = \frac{4}{4+6} = 0.4$, do đó $UpperAppr = 1, 2$;
- Với $i=3$, $v_3 = \frac{TCT_{3,1}}{TCT_{3,1} + TCT_{3,0}} = \frac{3}{3+6} \approx 0.33$, do đó $UpperAppr = 1, 2, 3$;
- Với $i=4$, $v_4 = \frac{TCT_{4,1}}{TCT_{4,1} + TCT_{4,0}} = \frac{4}{4+5} \approx 0.44$, do đó $UpperAppr = 1, 2, 3, 4$;
- Với $i=5$, $v_5 = \frac{TCT_{5,1}}{TCT_{5,1} + TCT_{5,0}} = \frac{3}{3+5} \approx 0.37$, do đó

$$UpperAppr = 1, 2, 3, 4, 5 ;$$

4) Kết luận:

- ◆ Xấp xỉ trên của tập X đã cho là U .
- ◆ Xấp xỉ dưới của tập X đã cho là rỗng.

CHƯƠNG 3: PHẦN MỀM THỬ NGHIỆM THUẬT TOÁN TÌM TẬP RÚT GỌN TRONG BẢNG QUYẾT ĐỊNH TẬP GIÁ TRỊ VÀ ỨNG DỤNG TRONG BÀI TOÁN CHẨN ĐOÁN BỆNH VIÊM GAN B

3.1. Phát biểu bài toán

Viêm gan siêu vi B là một loại virus tấn công lá gan, gây ra bệnh viêm gan. Tổ chức Y tế Thế giới thông kê có khoảng 350 triệu người nhiễm virus viêm gan B và tại Việt Nam có khoảng 20% dân số nhiễm virus viêm gan B. Những người nhiễm virus viêm gan B nếu không được kiểm soát và điều trị tốt sẽ gây ra viêm gan, xơ gan và ung thư gan.

Như đã biết, từ một bảng quyết định có nhiều đối tượng, tập luật quyết định rút trích được là rất lớn. Để thu gọn tập luật quyết định mà không làm mất đi tính đặc trưng của bảng quyết định ta đi thu gọn tập thuộc tính.

Dựa trên cơ sở về “*Thuật toán 2.1 - Tìm tập rút gọn thuộc tính trong bảng quyết định tập giá trị*” luận văn xây dựng phần mềm thử nghiệm thuật toán căn cứ vào các triệu chứng của bệnh Viêm gan B.

3.2. Mô tả và xử lý dữ liệu

3.2.1. Mô tả dữ liệu

Chương trình tìm luật rút gọn cho các thuộc tính điều kiện từ đó đưa ra luật quyết định để dùng vào cơ sở tri thức của các hệ chuyên gia nhằm mục đích chẩn đoán bệnh. Số thuộc tính rút gọn phải nhỏ hơn số thuộc tính ban đầu và có giá trị như nhau trong việc đưa ra luật quyết định. Luật mới tạo ra có số thuộc tính nhỏ hơn và không ảnh hưởng đến việc đưa ra quyết định. Số liệu thực nghiệm được lấy từ kho dữ liệu UCI với bộ dữ liệu viêm gan **Hepatitis.data[15]** để sinh luật quyết định phục vụ cho các bác sĩ chuyên ngành chẩn đoán bệnh viêm gan cho bệnh nhân.

* Thông tin về các thuộc tính

Attribute information:

1. Class: DIE, LIVE
2. AGE: 10, 20, 30, 40, 50, 60, 70, 80
3. SEX: male, female
4. STERIOD: no, yes
5. ANTIVIRALS: no, yes
6. FATIGUE: no, yes
7. MALAISE: no, yes
8. ANOREXIA: no, yes
9. LIVER BIG: no, yes
10. LIVER FIRM: no, yes
11. SPLEEN PALPABLE: no, yes
12. SPIDERS: no, yes
13. ASCITES: no, yes
14. VARICES: no, yes
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
-- see the note below
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
17. SGOT: 13, 100, 200, 300, 400, 500,
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. HISTOLOGY: no, yes

Class Distribution:

DIE: 32

LIVE: 123

* Bảng dữ liệu đầu vào theo kho UCI

2	30	2	1	2	2	2	2	1	2	2	2	2	2	2	1.00	85	18	4.0	?	1
2	50	1	1	2	1	2	2	1	2	2	2	2	2	2	0.90	135	42	3.5	?	1
2	78	1	2	2	1	2	2	2	2	2	2	2	2	2	0.70	96	32	4.0	?	1
2	31	1	?	1	2	2	2	2	2	2	2	2	2	2	0.70	46	52	4.0	80	1
2	34	1	2	2	2	2	2	2	2	2	2	2	2	2	1.00	?	200	4.0	?	1
2	34	1	2	2	2	2	2	2	2	2	2	2	2	2	0.90	95	28	4.0	75	1
1	51	1	1	2	1	2	1	2	2	1	1	2	2	?	?	?	?	?	?	1
2	23	1	2	2	2	2	2	2	2	2	2	2	2	2	1.00	?	?	?	?	1
2	39	1	2	2	1	2	2	2	1	2	2	2	2	2	0.70	?	48	4.4	?	1
2	30	1	2	2	2	2	2	2	2	2	2	2	2	2	1.00	?	120	3.9	?	1

Hình 3. 1. Bảng dữ liệu đầu vào

3.2.2. Xử lý dữ liệu

Chương trình tiến hành khai phá dữ liệu trong cơ sở dữ liệu bệnh nhân bị viêm gan. Dữ liệu đầu vào là một file text có cấu trúc như sau:

- Các thuộc tính điều kiện tương ứng với 19 triệu chứng thu thập được từ bệnh nhân có biểu hiện viêm gan, được kí hiệu: $\{a_1, a_2, a_3, \dots, a_{19}\}$.
- Mỗi dòng là thông tin về một bệnh nhân, trên mỗi dòng bệnh nhân là thể hiện các thuộc tính, giữa hai thuộc tính là “dấu cách”. Với các thuộc tính có giá trị tập được cách nhau bởi dấu “,”.
- Với mỗi bộ số liệu thiếu giá trị được chọn, chúng tôi tiến hành chuyển đổi sang bộ số liệu tập giá trị bằng cách thay thế các giá trị thiếu (kí hiệu bởi “?”) thành một tập giá trị ngẫu nhiên. Các giá trị trong tập giá trị ngẫu nhiên đó nằm trong miền giá trị của thuộc tính đó có các giá trị $\{0, 1, 2\}$.
- Thuộc tính cuối cùng là thuộc tính quyết định mang giá trị $\{1\}$ - Có bệnh, $\{0\}$ - Không có bệnh.

** Các thuộc tính tương ứng với một số đại lượng dùng để xác định tình trạng bệnh của bệnh nhân:*

- a_1 : Age- số tuổi của bệnh nhân 10 đến 80

Lớp 0: [10-23], Lớp 1: [23- 46], Lớp 2: [46-80]

- a_2 : Sex- Giới tính: 1 = Male, 2 = Female
- a_3 : Steroid- Thuốc kháng sinh: 1 = No, 2 = Yes
- a_4 : Antiviral- Thuốc kháng viruts: 1 = No, 2 = Yes
- a_5 : Fatigue- Mệt mỏi: 1 = No, 2 = Yes
- a_6 : Malaise- Khó chịu: 1 = No, 2 = Yes
- a_7 : Anorexia- Chán ăn: 1 = No, 2 = Yes

- a₈: Liver big- Gan sưng to: 1 = No, 2 = Yes
- a₉: Liver firm- Viêm gan: 1 = No, 2 = Yes
- a₁₀: Spleen palpable- Viêm lá lách: 1 = No, 2 = Yes
- a₁₁: Spiders- Mạch máu hình nhện trên da: 1 = No, 2 = Yes
- a₁₂: Ascites- Hạch ở ổ bụng: 1 = No, 2 = Yes
- a₁₃: Varices- Giãn tĩnh mạch: 1 = No, 2 = Yes
- a₁₄: Bilirubin- Sắc tố da (vàng da):
Lớp 0: [0.39-1.20], Lớp 1: [1.20- 2.40], Lớp 2: [2.40- 4.00]
- a₁₅: Alk phosphate- Huyết tương:
Lớp 0: [33-72.3], Lớp 1: [72.3- 144.6], Lớp 2: [144.6- 250]
- a₁₆: SGOT- Enzym ở men gan:
Lớp 0: [13- 162.3], Lớp 1: [162.3- 324.6], Lớp 2: [324.6- 500]
- a₁₇: Albumin- Nồng độ albumin:
Lớp 0: [2.1- 3.0], Lớp 1: [3.8- 4.5], Lớp 2: [5.0- 6.0]
- a₁₈: Protine- Tình trạng đông máu của người bệnh: từ 10 đến 90
Lớp 0: [10-26.7], Lớp 1: [26.7- 53.4], Lớp 2: [53.4- 90]
- a₁₉: Histology- Tiểu sử mắc bệnh: 0= No, 1 = Yes
- d: Class- Lớp quyết định: 0 = Live, 1 = Die

Sau khi tiến hành xử lý dữ liệu, ta thu được bảng dựa vào Dữ liệu đầu vào ở hình 3.1 như sau:

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	d
0	2	1	2	2	2	2	1	2	2	2	2	2	0	0	0	1	0,1,2	1	1
1	1	1	2	1	2	2	1	2	2	2	2	2	0	1	0	0	0,1,2	1	1
2	1	2	2	1	2	2	2	2	2	2	2	2	0	0	0	1	0,1,2	1	1
1	1	1	2	1	2	2	2	2	2	2	2	2	2	0	0	0	1	2	1
1	1	2	2	2	2	2	2	2	2	2	2	2	0	0,1,2	0	1	0,1,2	1	1
1	1	2	2	2	2	2	2	2	2	2	2	2	0	0	0	1	2	1	1
1	1	1	2	1	2	1	2	2	1	1	2	2	0,1,2	0,1,2	0,1,2	0,1,2	0,1,2	1	0
0	1	2	2	2	2	2	2	2	2	2	2	0	0,1,2	0,1,2	0,1,2	0,1,2	0,1,2	1	1
1	1	2	2	1	2	2	2	1	2	2	2	2	0	0,1,2	0	1	0,1,2	1	1
0	1	2	2	2	2	2	2	2	2	2	2	0	0,1,2	0	1	0,1,2	1	1	1

Hình 3. 2. Tập dữ liệu sau khi xử lý

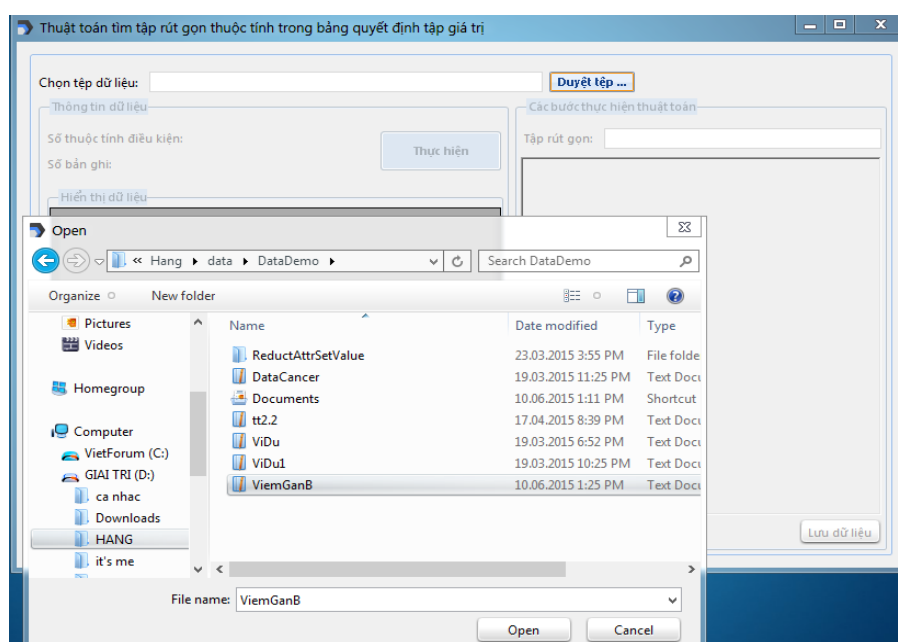
3.3. Thử nghiệm chương trình

❖ Công nghệ và công cụ phát triển ứng dụng

Ứng dụng được xây dựng trên bộ công cụ Microsoft Visual Studio 2012, trên nền tảng .Net Framework 4.0. Sử dụng hệ quản trị CSDL Microsoft SQL Server 2008.

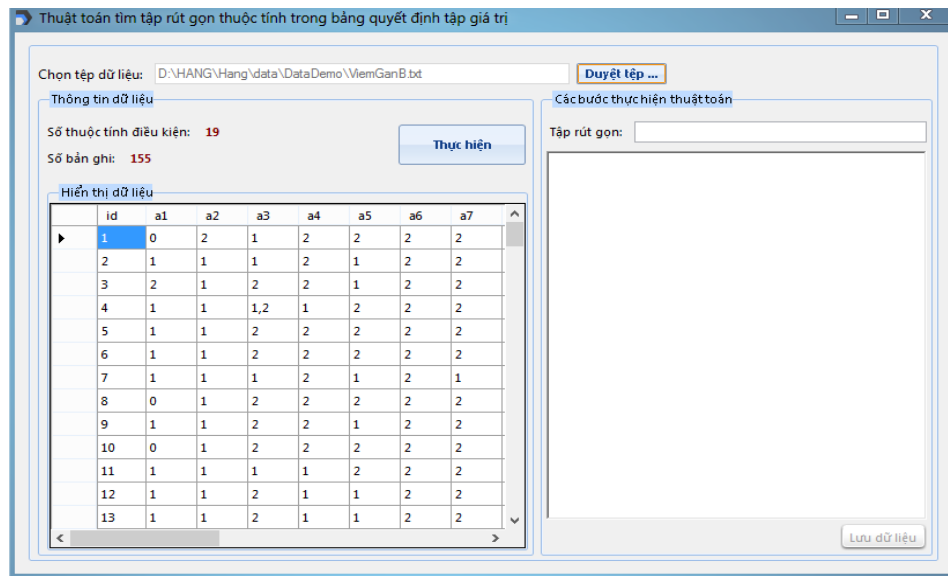
❖ Chức năng nhập dữ liệu

Người sử dụng sẽ tiến hành nhập dữ liệu thông qua nút **Duyệt tệp ...** của tab “Chọn tệp dữ liệu”. Lúc này người sử dụng sẽ chọn tệp văn bản được xây dựng sẵn trên máy tính theo định dạng tệp text.



Hình 3. 3. Giao diện nhập dữ liệu

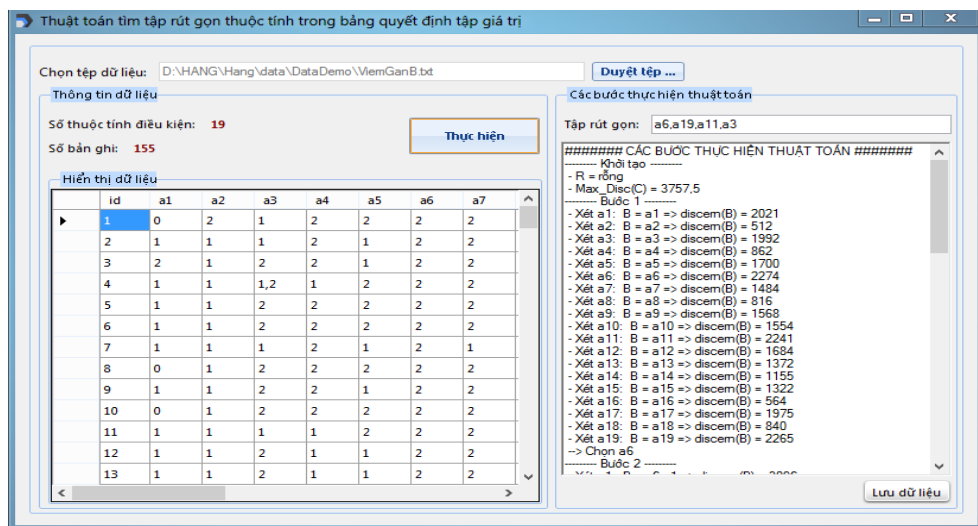
Sau khi chọn tệp dữ liệu, chương trình sẽ xử lý và hiển thị thông tin về số cột thuộc tính, số bản ghi dữ liệu và toàn bộ dữ liệu trong tệp dữ liệu nguồn.



Hình 3. 4. Màn hình hiển thị thông tin các thuộc tính

❖ Chức năng thực hiện thuật toán

Sau khi nhập dữ liệu, người sử dụng chọn nút **Thực hiện** để chương trình thực hiện thuật toán. Chương trình sẽ đưa ra hai kết quả: tập rút gọn thu được và các bước xử lý của thuật toán.



Hình 3. 5. Kết quả thực hiện với bộ dữ liệu thử nghiệm

Sau khi có kết quả rút gọn, người sử dụng sử dụng nút **Lưu dữ liệu** để lưu lại kết quả trên theo định dạng file text.

Dữ liệu thực hiện thuật toán gồm 155 bản ghi. Các giá trị trong tập giá trị ngẫu nhiên đó nằm trong miền giá trị của thuộc tính đó. Việc chuyển đổi được thực hiện bằng công cụ chuyển đổi do tác giả xây dựng. Cách tiếp cận rút gọn theo thuật toán đã trình bày là theo hướng top-down, có nghĩa là việc xây dựng tập rút gọn sẽ bắt đầu từ tập rỗng, sau đó bổ sung lần lượt các thuộc tính có độ quan trọng lớn nhất tính theo giá trị hàm phân biệt. Quá trình bổ sung các thuộc tính kết thúc khi giá trị hàm phân biệt của tập thuộc tính thu được bằng với giá trị của toàn bộ các thuộc tính quyết định.

3.4. Đánh giá kết quả

Thuật toán khởi tạo tập rút gọn là rỗng và tính giá trị hàm phân biệt với toàn bộ thuộc tính quyết định. Tiếp đó, tính lần lượt giá trị hàm phân biệt với từng thuộc tính và lựa chọn thuộc tính nào có giá trị lớn nhất (thuộc tính a_6) để bổ sung vào tập rút gọn. Tiếp tục quá trình, thuộc tính này sẽ được ghép cặp với các thuộc tính còn lại và sẽ được tính giá trị hàm phân biệt. Cặp nào có giá trị lớn nhất thì thuộc tính thuộc cặp đó sẽ bổ sung vào tập rút gọn (cặp a_6 - a_{19} nên thuộc tính tiếp theo được bổ sung vào tập rút gọn là a_{19}). Cứ tiếp tục ghép tập rút gọn với các thuộc tính còn lại, thao tác này dừng cho đến khi giá trị hàm phân biệt bằng với giá trị hàm phân biệt của toàn bộ thuộc tính.

Sau khi kết thúc chương trình dựa vào thuật toán “*Tìm tập rút gọn thuộc tính trong bảng quyết định tập giá trị*”, chương trình thu được kết quả như sau: Từ các tập thuộc tính điều kiện chương trình thu được tập rút gọn có số thuộc tính điều kiện nhỏ hơn số thuộc tính điều kiện ban đầu. Như vậy, thay vì phải dựa vào 19 thuộc tính ban đầu thì ta có thể chỉ dựa vào 4 thuộc tính (a_3 : Steroid- Thuốc kháng viêm, a_6 : Malaise- Khó ở, a_{11} : Spiders- Mạch máu hình nhện trên da, a_{19} : Histology- Tiền sử mắc bệnh) đã rút gọn trong bảng quyết định để đưa ra kết luận bệnh nhân có mắc bệnh hay không.

a3	a6	a11	a19	d
1	2	2	1	1
1	2	2	1	1
2	2	2	1	1
1, 2	2	2	1	1
2	2	2	1	1
2	2	2	1	1
1	2	1	1	0
2	2	2	1	1
2	2	2	1	1
2	2	2	1	1

Hình 3. 6. Tập dữ liệu sau khi rút gọn

- *Kết quả rút gọn*

Dòng 1: - *Nếu* bệnh nhân không sử dụng thuốc kháng sinh - cảm thấy khó chịu trong người- xuất hiện mạch máu hình nhện dưới da- tiểu sử mắc bệnh là không *Thì* mắc bệnh.

Dòng 3: - *Nếu* bệnh nhân có sử dụng thuốc kháng sinh - cảm thấy khó chịu trong người- xuất hiện mạch máu hình nhện dưới da- tiểu sử mắc bệnh là không *Thì* mắc bệnh.

Dòng 4:- *Nếu* bệnh nhân có thể dùng hay không dùng thuốc kháng sinh- cảm thấy khó chịu trong người- xuất hiện mạch máu hình nhện dưới da- tiểu sử mắc bệnh là không *Thì* mắc bệnh.

Dòng 7: *Nếu* bệnh nhân không sử dụng thuốc kháng sinh- cảm thấy khó chịu trong người- không xuất hiện mạch máu hình nhện dưới da- tiểu sử mắc bệnh là không *Thì* không mắc bệnh.

Trên cơ sở nghiên cứu lý thuyết, đã xây dựng một chương trình rút gọn thuộc tính, tạo ra một tập luật hỗ trợ trong việc phát hiện bệnh viêm gan B.

3.5. Kết luận chương

Chương này, tác giả cài đặt thành công thuật toán rút gọn thuộc tính thuộc tính cho bảng quyết định các triệu chứng viêm gan B. Nội dung chủ yếu trình bày về bước tiền xử lý dữ liệu và áp dụng thuật toán rút gọn trên bộ dữ liệu sau khi đã xử lý để thu được tập rút gọn.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

I. Những kết quả chính của luận văn

1. Luận văn trình bày những khái niệm cơ bản về hệ thông tin cùng các khái niệm có liên quan. Trình bày những nội dung về hệ thông tin tập giá trị cùng các khái niệm làm nền tảng cho bài toán rút gọn thuộc tính.

2. Trình bày các khái niệm về tập rút gọn trên hệ thông tin và hệ thông tin tập giá trị.

3. Khai thác hai thuật toán đối với bảng quyết định tập giá trị, thuật toán rút gọn thuộc tính trong bảng quyết định tập giá trị và thuật toán tính xấp xỉ trên- xấp xỉ dưới của một tập trong hệ thông tin tập giá trị.

II. Hướng phát triển tiếp theo của luận văn

1. Trên bảng quyết định tập giá trị, tiếp tục đi sâu vào nghiên cứu rút gọn thuộc tính trong trường hợp khi bổ sung tập đối tượng.

2. Tiếp tục nghiên cứu các hàm phân biệt khác trên hệ thông tin giá trị tập. Trên cơ sở đó, khai thác và tìm hiểu các phương pháp mới hiệu quả hơn các phương pháp đã có.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

- [1]. Nguyễn Long Giang (2012). Nghiên cứu một số phương pháp khai phá dữ liệu theo tiếp cận lý thuyết tập thô, *Luận án Tiến sĩ*, Viện Công Nghệ Thông Tin.
- [2]. Hoàng Thị Lan Giao (2007). Khía cạnh đại số và lôgic phát hiện luật theo tiếp cận tập thô, *Luận án Tiến sĩ*, Viện Công Nghệ Thông Tin.
- [3]. Phùng Thị Thu Hiền, Lê Quang Hào, Nguyễn Quang Khanh, Nguyễn Bá Tường (2010). Định nghĩa tập thô theo hàm thuộc thô, *Tạp chí nghiên cứu Khoa học kỹ thuật và công nghệ quân sự (2010)*, tr. 50 - 54.
- [4]. Phùng Thị Thu Hiền, Lê Quang Hào, Nguyễn Bá Tường (2011). Những vấn đề của trích chọn đặc trưng trong hệ tin, *Tạp chí nghiên cứu Khoa học kỹ thuật và công nghệ quân sự (2011)*, tr. 60 - 63.
- [5]. Nguyễn Đức Thuận (2010). Phủ tập thô và độ đo đánh giá hiệu năng tập luật quyết định, *Luận án Tiến sĩ*, Viện Công Nghệ Thông Tin.

Tài liệu tiếng Anh

- [6]. B. Kolman, R.C. Busby, S.C. Ross, Discrete Mathematical Structures, fifth ed., Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003.
- [7]. G. Liu, The axiomatization of the rough set upper approximation operations, *Fundamenta Informaticae* 69 (3) (2006) 331-342.
- [8]. G. Liu, Axiomatic systems for rough sets and fuzzy rough sets, *International Journal of Approximate Reasoning* 48 (3) (2008) 857-867.
- [9]. Y.Guan, H. Wang, Set-valued information systems, *Information Sciences* 176 (17) (2006) 2507-2525.
- [10]. Nguyen Sinh Hoa, Nguyen H. Son (1996), "Some Efficient Algorithms for Rough Set Methods", *Proceedings of the sixth International Conference on*

Information Processing Management of Uncertainty in Knowledge-Based Systems, pp. 1451-1456.

[11]. Pawlak Z. (1982), “Rough sets”, *International Journal of Computer and Information Science*, 11, pp. 341-356.

[12]. Z. Pawlak, A. Skowron, Rough sets: some extensions, *Information Sciences* 177 (1) (2007) 2840.

[13]. Pawlak Z., *Rough sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, 1991.

[14]. Junbo Zhang, Tianrui Li, Da Ruan, Dun Liu, Rough sets based matrix approaches with dynamic attribute variation in set-valued information systems, *International Journal of Approximate Reasoning*, Volume 53, Issue 4, June 2012, Pages 620-635.

[15]. The UCI machine learning repository,
<https://archive.ics.uci.edu/ml/datasets/Hepatitis>