ĐÁNH GIÁ CHẤT LƯỢNG LUẬT QUYẾT ĐỊNH ĐA TRỊ DỰA TRÊN TIẾP CẬN HÀM Ý THỐNG KÊ

Phan Tấn Tài[1], Lê Đức Thắng[1], Huỳnh Xuân Hiệp[1,2]

[1] Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ
[2] Nhóm nghiên cứu liên ngành DREAM-CTU/IRD, Trường Đại học Cần Thơ

pttai@ctu.edu.vn, ldthang@ctu.edu.vn, hxhiep@ctu.edu.vn

TÓM TẮT *- Trong bài viết này chúng tôi giới thiệu một tiếp cận mới trong việc đánh giá chất lượng luật quyết định đa trị dựa trên phân tích hàm ý thống kê (statistical implicative analysis). Nghiên cứu được xem xét bắt đầu từ một hệ thống thông tin quyết định đa trị (set-valued decision information system) cùng với tập luật quyết định đa trị được sinh ra tương ứng. Từ đây, tập luật quyết định đa trị sẽ được phân tích và đánh giá chất lượng trên cơ sở phân tích hàm ý thống kê. Với kết quả đánh giá chất lượng tập luật quyết định đa trị, các luật quyết định đa trị sẽ được sắp xếp theo các mức độ ưu tiên khác nhau dựa trên các độ đo như chỉ số hàm ý (implication indice) và cường độ hàm ý (implication intensity). Các kịch bản thực nghiệm chỉ ra các luật quyết định đa trị tốt, các luật quyết định đa trị chưa tốt và vai trò của các luật quyết định đa trị. Đây chính là vấn đề mà mô hình luật quyết định đa trị trước đây chưa thể hiện được này.*

Từ khóa *- Hệ thống thông tin đa trị, lớp tương đồng tối đại, luật quyết định đa trị, hàm ý thống kê.*

I. GIỚI THIỆU

Lý thuyết tập thô (rough sets theory) [2][3][15] là một công cụ phân tích dữ liệu hiệu quả, được sử dụng trong mô hình đại diện thuộc tính-giá trị để mô tả sự phụ thuộc giữa các thuộc tính và đánh giá ý nghĩa của các thuộc tính cùng với các luật quyết định. Ngoài ra, sinh luật trong các hệ thống thông tin không đầy đủ (rules in incomplete information systems) [11], kỹ thuật khối thích hợp tối đại cho việc sinh luật trong các hệ thống thông tin không đầy đủ (maximal consistent block technique for rule acquisition in incomplete information systems) [12], tập thô được nghiên cứu để phân tích quyết định đa tiêu chí (rough sets theory for multicriteria decision analysis) [6] đã được nghiên cứu và có nhiều ứng dụng đem lại hiệu quả nhất định. Đặc biệt là có nhiều tiếp cận mới trong việc sinh luật quyết định đa trị trong các hệ thống thông tin quyết định đa trị [7][10] đã xuất hiện trong thời gian gần đây.

Lý thuyết tập thô cổ điển dựa trên các mối quan hệ không phân biệt được và các nghiên cứu chủ yếu dựa trên các hệ thống thông tin đầy đủ. Tuy nhiên, một số đặc điểm của các thuộc tính trong một hệ thống thông tin có thể không biết hoặc đa giá trị. Hơn nữa, các thuộc tính này đôi khi có miền trị có một thứ tự và thứ tự của các đặc điểm của các thuộc tính này đóng một vai trò rất quan trọng [17]. Ngoài ra, trong mối quan hệ giữa ngữ cảnh hình thức và hệ thống thông tin đa trị đã chỉ ra rằng các ngữ cảnh có thể chuyển đổi thành một hệ thống thông tin đa trị giá trị [15]. Ở một khía cạnh khác, một quan hệ mờ trong hệ thống thông tin đa trị (Fuzzy Set-valued Information Systems (FSVISs)) cũng được đề cập đến [18]. Điều này, cho ta thấy khá nhiều các khía cạnh khác nhau của lý thuyết tập thô, cũng như các vấn đề về hệ thống thông tin đa trị và các luật quyết định đa trị đã được nghiên cứu mạnh mẽ. Tuy nhiên, *Hiện nay các luật quyết định đa trị được sinh từ hệ thống thông tin quyết định đa trị chưa được đánh giá chất lượng, "Vai trò" các luật quyết định đa trị là như nhau*. Nói một cách khác hơn là chưa xem xét các độ đo "hấp dẫn" cần thiết cho các luật quyết định đa trị, để từ đó đánh giá chất lượng, chỉ ra vai trò và các khuynh hướng khách quan của tập luật quyết định đa trị.

Trong bài viết này, chúng tôi đề xuất một tiếp cận mới trong việc đánh giá chất lượng các luật quyết định đa trị dựa trên tiếp cận hàm ý thống kê [4][5]. Phân tích hàm ý thống kê (Statistical Implicative Analysis – SIA) được đề xuất bởi Gras [4][5], nhằm phát hiện những khuynh hướng trong một tập hợp các thuộc tính. SIA cung cấp một phương pháp để đánh giá độ hấp dẫn của các luật và cấu trúc của chúng trong việc khám phá mối quan hệ của luật ở các mức độ hàm ý khác nhau. Để đánh giá chất lượng các luật quyết định đa trị, hệ thống thông tin quyết định đa trị và tập luật quyết định đa trị được chuyển về dạng thức có thể phân tích và sắp xếp theo mức độ ưu tiên trong phân tích hàm ý thống kê.

Bài viết gồm 5 phần: phần thứ nhất giới thiệu tổng quan, phần thứ hai trình bày luật quyết định đa trị, phần thứ ba giới thiệu về hàm ý thống kê, phần thứ tư trình bày mô hình đánh giá chất lượng luật quyết định đa trị, phần thứ năm giới thiệu thực nghiệm và sau cùng là phần kết luận.

II. LUẬT QUYẾT ĐỊNH ĐA TRỊ

A. Các hệ thống thông tin

1. Hệ thống thông tin đơn trị và hệ thống thông tin đa trị

Một hệ thống thông tin được định nghĩa như gồm một bộ bốn (O, AT, V, f), trong đó, O là một tập hữu hạn không rỗng gồm N đối tượng $\{x_1, x_2, \dots, x_N\}$, AT là một tập hữu hạn không rỗng gồm n thuộc tính $\{a_1, a_2, \dots, a_n\}$, $V = \cup_{a \in AT} V_a$ với V_a là miền giá trị của thuộc tính a, $f: OxAT \rightarrow V$ là hàm thông tin, nghĩa là $f(x,a) \epsilon V_a$, với $x \in O$.

Nếu ứng với mỗi đối tượng $x_{i\ (i=1,2,\dots,N)}$, mỗi thuộc tính $a \in AT$ tương ứng trong hệ thống thông tin (O, AT, V, f) có một giá trị duy nhất (a unique attribute value) thì (O, AT, V, f) được gọi là hệ thống thông tin đơn trị (a single-valued information system) và ngược lại (O, AT, V, f) được gọi là hệ thống thông tin đa trị (Set-valued (multi-valued) information system) [7][13][14].

Ví dụ xem xét một tập dữ liệu đơn giản cho trong bảng 1, các tập *O, AT, V, f* được xác định như sau: $O=\{x_1, x_2, x_3, x_4, x_5\}$, $AT=\{a_1, a_2, a_3, a_4\}$, $V_{a_1} = \{0, 1, 2\}$ (các giá trị trên cột a_1của bảng 1), $V_{a_2} = \{1, 0, 2\}$ (các giá trị trên thuộc tính a_2 của bảng 1),…và $f(x_1, a_1) = 0$, $f(x_1, a_2) = 1$.

Bảng 1. Hệ thống thông tin đơn trị $\boldsymbol{S = (O, AT, V, f)}$, gồm 5 đối tượng $\{\boldsymbol{x_1, x_2, x_3, x_4, x_5}\}$ và 4 thuộc tính $\{\boldsymbol{a_1, a_2, a_3, a_4}\}$

O	a_1	a_2	a_3	a_4
x_1	0	1	0	1
x_2	0	0	2	1
x_3	1	2	0	0
x_4	2	1	1	3
x_5	2	1	1	2

2. Hệ thống thông tin đầy đủ và hệ thống thông tin không đầy đủ

Xét hệ thống thông tin (O, AT, V, f), khi đó miền giá trị của thuộc tính V có thể chứa một ký hiệu đặc biệt * để chỉ rằng giá trị thuộc tính là không biết. Miền giá trị thuộc tính nào khác với ký hiệu đặc biệt * thì được gọi là miền giá trị thuộc tính chính quy. Một hệ thống thông tin mà trong đó mọi miền giá trị thuộc tính đều là chính quy thì được gọi là hệ thống thông tin đầy đủ (complete information system), ngược lại thì được gọi là hệ thống thông tin không đầy đủ (incomplete information system) [11].

Ví dụ xét hệ thống thông tin $S = (O, AT, V, f)$ cho trong bảng 2, $V_{a_1} = \{0, 1, 3, *\}$ nên $S = (O, AT, V, f)$ trong trường hợp này là hệ thống thông tin không đầy đủ.

Bảng 2. Hệ thống thông tin không đầy đủ $\boldsymbol{S = (O, AT, V, f)}$, gồm 5 đối tượng $\{\boldsymbol{x_1, x_2, x_3, x_4, x_5}\}$ và 4 thuộc tính $\{\boldsymbol{a_1, a_2, a_3, a_4}\}$

O	a_1	a_2	a_3	a_4
x_1	*	4	2	1
x_2	3	4	2	2
x_3	1	2	*	0
x_4	2	1	1	5
x_5	*	1	1	0

3. Quan hệ không phân biệt

Xét hệ thống thông tin đơn trị đầy đủ $S = (O, AT, V, f)$ và $A \subseteq AT$, khi đó một quan hệ không phân biệt (indiscernibility relation) [11], ký hiệu là $IND(A)$ được định nghĩa: $IND(A) = \{(x, y) \in OxO / \forall a \in A, f(x, a) = f(y, a)\}$. Ví dụ xét bảng 1, nếu $A = \{a_1, a_2, a_3\}$ ta có $IND(A) = \{(x_4, x_5)\}$. Với $A \subseteq AT$, $IND(A)$ là quan hệ tương đương và là một bộ phận của O. Nếu gọi $I_A(x)$ là tập các đối tượng có quan hệ không phân biệt với x thì $I_A(x) = \{y \in O / (x, y) \in IND(A)\}$, hiển nhiên $x \in I_A(x)$. Theo bảng 1, với $A = \{a_1, a_2, a_3\}$, ta có $I_A(x_4) = I_A(x_5) = \{x_4, x_5\}$.

Xét hệ thống thông tin không đầy đủ $S = (O, AT, V, f)$ và $A \subseteq AT$, khi đó một quan hệ tương tự (similarity relation) [11], ký hiệu là $SIM(A)$ được định nghĩa: $SIM(A) = \{(x, y) \in OxO / \forall a \in A, f(x, a) = f(y, a) | f(x, a) = * | f(y, a) = *\}$. Ví dụ từ bảng 2, nếu $A = \{a_1, a_2, a_3\}$ ta có $SIM(A) = \{(x_1, x_2), (x_4, x_5)\}$. Nếu gọi $S_A(x)$ là tập các đối tượng có quan hệ tương tự với x thì $S_A(x) = \{y \in O / (x, y) \in SIM(A)\}$. Theo bảng 2, với $A = \{a_1, a_2, a_3\}$, ta có $S_A(x_4) = S_A(x_5) = \{x_4, x_5\}$ vì $(x_4, x_5) \in SIM(A)$.

B. Hệ thống thông tin quyết định đa trị

Hệ thống thông tin quyết định đa trị là một bộ 4: $(O, C \cup \{d\}, V, f)$ [7]. Trong đó: O là một tập hợp hữu hạn khác rỗng các đối tượng, C là một tập hợp hữu hạn khác rỗng các thuộc tính điều kiện, d là thuộc tính quyết định, $C \cap \{d\} = \emptyset$, $V = V_C \cup V_d$, với V_C là hợp miền giá trị các thuộc tính điều kiện, V_d là miền giá trị thuộc tính quyết định,

f là ánh xạ từ $Ox(C \cup \{d\})$ đến V sao cho: $f: OxC \to 2^{|V_c|}$ là một ánh xạ đa trị, $\forall x \in O, c \in C : f(x,c) = c(x)$, $f: Ox\{d\} \to V_d$ là một ánh xạ đơn trị, $\forall x \in O: f(x,d) = d(x)$.

Một hệ thống thông tin quyết định đa trị được trình bày dưới dạng bảng, còn được gọi là bảng quyết định đa trị. Ví dụ về hệ thống thông tin quyết định đa trị được minh họa như bảng 3 bên dưới.

Bảng 3. Hệ thống thông tin quyết định đa trị gồm 10 đối tượng $\{\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_{10}}\}$, 5 thuộc tính điều kiện $\{\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_5}\}$ và 1 thuộc tính quyết định $\boldsymbol{d}$.

O	c_1	c_2	c_3	c_4	c_5	d
x_1	{1}	{0,1}	{0}	{1,2}	{2}	3
x_2	{0,1}	{2}	{1,2}	{0}	{0}	1
x_3	{0}	{1,2}	{1}	{0,1}	{0}	1
x_4	{0}	{1}	{1}	{1}	{0, 2}	2
x_5	{2}	{1}	{0,1}	{0}	{1}	2
x_6	{0,2}	{1}	{0,1}	{0}	{1}	2
x_7	{1}	{0,2}	{0,1}	{1}	{2}	3
x_8	{0}	{2}	{1}	{0}	{0,1}	1
x_9	{1}	{0,1}	{0,2}	{1}	{2}	3
x_{10}	{1}	{1}	{2}	{0,1}	{2}	2

C. Lớp tương đồng

Trong hệ thống thông tin quyết định đa trị $(O, C \cup \{d\}, V, f)$, với mỗi thuộc tính $b \in C$ thì quan hệ tương đồng theo b được ký hiệu và định nghĩa như sau: $T_b = \{(x,y)/\, x,y \in O: b(x) \cap b(y) \neq \emptyset\}$. Với mỗi tập hợp các thuộc tính điều kiện $B \subseteq C$ quan hệ tương đồng theo B được định nghĩa: $T_B = \{(x,y)/x,y \in O \wedge \forall b \in B: b(x) \cap b(y) \neq \emptyset\} = \bigcap_{b \in B} T_b$. Khi $(x,y) \in T_B$ người ta nói x tương đồng với y theo B, ký hiệu là xT_By [7].

Trong một hệ thống thông tin quyết định đa trị $(O, C \cup \{d\}, V, f)$, ta định nghĩa [7] $T_B(x) = \{y \in O \,/\, yT_Bx\} = \{y \,/ y \in O, \forall b \in B: b(x) \cap b(y) \neq \emptyset\}$ là một lớp tương đồng của $x \in O$ dựa trên tập thuộc tính điều kiện $B \subseteq C$.

Theo bảng 1, nếu gọi $B = C$ thì ta có $T_B(x_1) = T_B(x_7) = \{x_1, x_7, x_9\}$, $T_B(x_2) = T_B(x_8) = \{x_2, x_3, x_8\}$, $T_B(x_3) = \{x_2, x_3, x_4, x_8\}$, $T_B(x_4) = \{x_3, x_4\}$, $T_B(x_5) = T_B(x_6) = \{x_5, x_6\}$, $T_B(x_9) = \{x_1, x_7, x_9, x_{10}\}$, $T_B(x_{10}) = \{x_9, x_{10}\}$.

D. Lớp tương đồng tối đại

1. Các khái niệm

Khi N là một lớp tương đồng theo B và $\forall x \in O \backslash N$ và tồn tại $y \in N$ sao cho y không tương đồng với x theo B thì N được gọi là lớp tương đồng tối đại (maximal tolerance class) theo B [7]. Nếu gọi tập hợp tất cả các lớp tương đồng tối đại theo B trong O là λ_B thì khi đó λ_B phủ O: $O = \bigcup_{N \in \lambda_B} N$.

Lớp tương đồng tối đại có đặc trưng sau: xét λ_B tập hợp các lớp tương đồng tối đại. Giả sử rằng B có các thuộc tính $B = \{b_1, b_2, \dots, b_m\}$. Đặc trưng của lớp tương đồng tối đại $N \in \lambda_B$ được định nghĩa như sau: $des(N) = \{\bigcap_{x \in N} b_1(x), \bigcap_{x \in N} b_2(x), .., \bigcap_{x \in N} b_m(x)\}$. Đặc trưng của lớp tương đồng tối đại mô tả giá trị chung của các thuộc tính điều kiện của các đối tượng thuộc về lớp tương đồng tối đại đó.

2. Giải thuật phân lớp tương đồng tối đại

Chúng tôi đề xuất một giải thuật phân lớp tương đồng tối đại như sau:

- *Dữ liệu đầu vào:* hệ thống thông tin quyết định đa trị gồm một bộ 4: $S = (O, C \cup \{d\}, V, f)$. Trong đó: O là một tập hợp hữu hạn khác rỗng gồm N đối tượng, C là một tập hợp hữu hạn khác rỗng các thuộc tính điều kiện, d là thuộc tính quyết định, $C \cap \{d\} = \emptyset$, $V = V_C \cup V_d$, với V_C là hợp miền giá trị các thuộc tính điều kiện, V_d là miền giá trị thuộc tính quyết định, f là ánh xạ từ $Ox(C \cup \{d\})$ đến V sao cho: $f: OxC \to 2^{|V_c|}$ là một ánh xạ đa trị và $f: Ox\{d\} \to V_d$ là một ánh xạ đơn trị.
- *Dữ liệu đầu ra:* các lớp tương đồng tối đại N_k $(k = 1, 2, \dots)$
- Giải thuật:
  - $k = 1$;
  - Với mỗi $x \in O$, ta thực hiện như sau:
    - Khởi tạo: $\{N_k = \{x\}$ // là lớp tương đồng tối đại chứa x; $desN_K = \{$Giá trị tương đồng của $N_k\}$= $\{f(x,c) \,/\, \forall c \in C\}$ }
    - Với mỗi $x_i \in O$ $(i = 1, 2, .., N) \wedge x_i \notin N_k$
      Nếu$(\forall c \in C, \forall f(x,c) \in desN_K: f(x_i,c) \cap f(x,c) \neq \emptyset)$ thì
      $\{ N_k = N_k + \{x_i\}$; $des = \{f(x_i,c) \cap f(x,c) \,/\, \forall c \in C, \forall f(x,c) \in desN_k\}$; $desN_k = des$ }

Ngược lại bỏ qua

* Nếu ($N_k \neq N_j, \forall j: 1 \leq j < k$ (N_j là lớp tương đồng tối đại đã xác định trước đó)) thì

{ Ghi nhận N_k là một lớp tương đồng đối đại; $k = k + 1$}

3. Ví dụ minh họa giải thuật phân lớp tương đồng

Xét hệ thống thông tin quyết định đa trị được cho trong bảng 3, kết quả áp dụng giải thuật phân lớp tương đồng cho 5 lớp tương đồng tối đại như bảng 4.

Bảng 4. Kết quả phân lớp tương đồng tối đại và các đặc trưng tương ứng.

Lớp tương đồng tối đại N_i	Đặc trưng $des(N_i)$					d
	c_1	c_2	c_3	c_4	c_5	
$N_1 = \{x_1, x_7, x_9\}$	1	0	0	1	2	3
$N_2 = \{x_2, x_3, x_8\}$	0	2	1	0	0	1
$N_3 = \{x_3, x_4\}$	0	1	1	1	0	1∨2
$N_4 = \{x_5, x_6\}$	2	1	{0, 1}	0	1	2
$N_5 = \{x_9, x_{10}\}$	1	1	2	1	2	2∨3

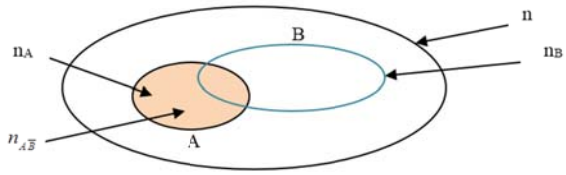
III. HÀM Ý THỐNG KÊ

A. Khái niệm về hàm ý thống kê

Phân tích hàm ý thống kê cung cấp một phương pháp để đánh giá độ hấp dẫn của các luật và cấu trúc của chúng trong việc khám phá mối quan hệ của luật ở các mức độ khác nhau. Gọi O là tập gồm n đối tượng được mô tả bởi một tập V hữu hạn các biến (thuộc tính). Với mỗi $x \in O$, $a \in V$, giá trị của đối tượng x đối với biến a ký hiệu là a(x). Trường hợp $a(x) = 1$ ta nói đối tượng x thỏa biến a.

Vấn đề đặt ra là: "có thể tin đến mức độ nào để biến b là đúng khi biến a là đúng"? . Nói một cách khác, làm thế nào để biết các đối tượng $x \in O$ có thỏa biến b hay không khi biết rằng các đối tượng này thỏa biến a.

Một cách trực quan, chúng ta xét A ⊂ O sao cho $A = \{x/a(x) = true\}$, B ⊂ O sao cho $B = \{x/b(x) = true\}$. Gọi $n = card(O)$, $n_a = card(A)$; gọi $n_b = card(B)$, $n_{a\wedge\overline{b}} = card(A \cap \overline{B})$. Khi đó, Luật $a \to b$ được định lượng sự ngẫu nhiên theo n, n_a, n_b và $n_{a\wedge\overline{b}}$.



Hình 1. Biểu diễn tập các đối tượng dựa trên tiếp cận hàm ý thống kê

Trong trường hợp A ⊂ B thì $a \to b$ là đúng. Tuy nhiên trong thực tế khá phổ biến là một vài đối tượng thỏa biến a nhưng không thỏa biến b thì luật $a \to b$ cần được xem xét.

Luật $a \to b$ được cho là có thể chấp nhận với một ngưỡng α cho trước nếu xác suất xuất hiện trường hợp $card(A \cap \overline{B})$ lớn hơn $card(X \cap \overline{Y})$ nhỏ hơn hoặc bằng α [4][5], với hai tập con X và Y của O, lần lượt có số phần tử là n_a, n_b tương ứng. Nghĩa là: $\Pr\left(card(X \cap \overline{Y}) \leq card(A \cap \overline{B})\right) \leq \alpha$.

B. Chỉ số hàm ý và cường độ hàm ý

Chỉ số hàm ý (implication indice) [4][5] của $a \to b$ được định nghĩa như sau: $q(a, \bar{b}) = \frac{n_{a\wedge\bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}}$

Cường độ hàm ý (implication intensity) [4][5] của luật $a \to b$ được định nghĩa là: $\varphi(a, b) = \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$

IV. MÔ HÌNH ĐÁNH GIÁ CHẤT LƯỢNG LUẬT QUYẾT ĐỊNH ĐA TRỊ

A. Biểu diễn luật quyết định đa trị

Trong một hệ thống thông tin quyết định đa trị $(O, C \cup \{d\}, V, f)$, với $N \in \lambda_B$ là một lớp tương đồng tối đại theo $B \subset C$ và ta đặt $d(N) = \{i/\exists x \in N, d(x) = i\}$. Khi đó: $des(N) \to \bigvee_{i \in d(N)}(d, i)$ hay $\bigwedge_{b_i \in B}(b_i, \bigcap_{x \in N} b_i(x)) \to \bigvee_{i \in d(N)}(d, i)$ là một luật quyết định được xác định bởi N [7].

Theo bảng 4, tập luật quyết định đa trị sinh được (R_S) như sau: $(1,0,0,1,2) \to (d,3)$, $(0,2,1,0,0) \to (d,1)$, $(0,1,1,1,0) \to (d,1) \vee (d,2)$, $(2,1,(0,1),0,1) \to (d,2)$, $(1,1,2,1,2) \to (d,2) \vee (d,3)$. Sau khi tách vế phải, tập luật R_S được biểu diễn như sau: $(1,0,0,1,2) \to (d,3)$, $(0,2,1,0,0) \to (d,1)$, $(0,1,1,1,0) \to (d,1)$, $(0,1,1,1,0) \to (d,2)$, $(2,1,(0,1),0,1) \to (d,2)$, $(1,1,2,1,2) \to (d,2)$, $(1,1,2,1,2) \to (d,3)$.

B. Giải thuật chuyển hệ thống thông tin quyết định đa trị sang dạng nhị phân

Để chuyển hệ thống thông tin quyết định đa trị sang dạng nhị phân, chúng tôi đề xuất một giải thuật làm tương tự Apriori như sau:

Dữ liệu đầu vào: hệ thống thông tin quyết định đa trị gồm một bộ 4: $S = (O, AT, V, f)$. Trong đó: O là một tập hợp hữu hạn khác rỗng các đối tượng, $AT = C \cup \{d\}$, C là một tập hợp hữu hạn khác rỗng các thuộc tính điều kiện, d là thuộc tính quyết định, $C \cap \{d\} = \emptyset$, $V = V_C \cup V_d$, với V_C là hợp miền giá trị các thuộc tính điều kiện, V_d là miền giá trị thuộc tính quyết định, f là ánh xạ từ $Ox(C \cup \{d\})$ đến V sao cho: $f: OxC \to 2^{|V_C|}$ là một ánh xạ đa trị, $f: Ox\{d\} \to V_d$ là một ánh xạ đơn trị.

Dữ liệu đầu ra: hệ thống thông tin quyết định đa trị nhị phân gồm một bộ 4: $S_B = (O, AT_B, V_B, f_B)$. Trong đó: O là một tập hợp hữu hạn khác rỗng các đối tượng, $AT_B = \{"a = v" / \forall\, a \in AT, \forall v \in V_a\}$, $V_B = \{0,1\}$, f_B là ánh xạ từ $OxAT_B$ đến V_B sao cho: $f_B: OxAT_B \to V_B$ là một ánh xạ nhị phân.

Ta gọi hàm $g(x,a,v)$ là hàm kiểm tra xem tập giá trị của đối tượng $x \in O$ tại thuộc tính $a \in AT$ có chứa giá trị v ($v \in V_a$) hay không? Khi đó, $g(x,a,v)$ *được định nghĩa như sau*: $g(x,a,v) = \begin{cases} true, & \text{nếu } v \in f(x,a) \\ false, & \text{nếu } v \notin f(x,a) \end{cases}$

- Giải thuật:
  - Với mỗi đối tượng $x \in O$ trong S, thực hiện:
    - với mỗi thuộc tính $a \in AT$, thực hiện:

{ với mỗi $v \in V_a$, cần xét:

{ nếu $g(x,a,v) = true$ thì gán $f_B(x, "a = v") = 1$, ngược lại gán $f_B(x, "a = v") = 0$ }

}

C. Đánh giá chất lượng các luật quyết định đa trị

Với mỗi luật quyết định đa trị có dạng $a \to b$ *sẽ* được định lượng sự ngẫu nhiên theo n, n_a, n_b và $n_{a \wedge \bar{b}}$ bằng độ đo chỉ số hàm ý hay cường độ hàm ý. Sau khi tập luật R_S được định lượng ngẫu nhiên bằng một độ đo và các luật được sắp xếp theo một thứ tự giá trị độ đo từ cao đến thấp, chúng tôi đề xuất chọn các luật tốt nhất theo có hai hướng: *(i) chọn các luật tốt nhất dựa vào một ngưỡng β cho trước, (ii) chọn m luật có tốt nhất (có giá trị độ đo cao nhất)*. Ngoài ra, ta có thể xem xét thêm các luật được cho là "xấu nhất", theo hai hướng*: (i) dựa vào một ngưỡng θ cho trước, (ii) lấy k luật có giá trị độ đo thấp nhất.*

D. Giải thuật tổng quát đánh giá chất lượng tập luật quyết định đa trị

Giải thuật tổng thể đánh giá chất lượng tập luật quyết định đa trị được tiến hành như sau:

i. Thực hiện tiền xử lý dữ liệu gốc đưa về hệ thống thông tin quyết định đa trị (SDIS)
ii. Sinh tập luật quyết định đa trị (R) từ SDIS
iii. Chuyển R sang tập luật quyết định đa trị (R_s) sao cho vế phải chỉ có một giá trị
(tách các luật quyết định đa trị có vế phải nhiều hơn một giá trị quyết định thành nhiều luật quyết định đa trị mà vế phải chỉ có một giá trị quyết định tương ứng)
iv. Chuyển SDIS sang dạng nhị phân ($SDIS_B$)
v. Chuyển tập luật quyết định đa trị R_s sang dạng nhị phân (R_B)
vi. Với mỗi luật quyết định đa trị dạng nhị phân: (□ → b) ∈ R_B
- Dựa vào $SDIS_B$: thống kê tính: n, n_a, n_b và $n_{a \wedge \bar{b}}$
- Tính giá trị độ đo chỉ số hàm ý và giá trị độ đo cường độ hàm ý dựa trên các giá trị n, n_a, n_b và $n_{a \wedge \bar{b}}$

vii. Sắp xếp thứ tự tập luật R_s dựa trên giá trị chỉ số hàm ý hoặc giá trị cường độ hàm ý
viii. Chọn lọc các luật quyết định đa trị tốt dựa trên một ngưỡng β cho trước hay chọn m luật đầu tiên

V. THỰC NGHIỆM

A. Bài toán thực nghiệm

Vai trò chính của cố vấn học tập (CVHT) trong một trường đại học là tư vấn học tập, nghiên cứu khoa học và rèn luyện cho sinh viên. Tuy nhiên làm thế nào để CVHT có thể tư vấn cho tất cả sinh viên trong lớp chuyên ngành có hiệu quả và ít mất thời gian nhất, đó là vấn đề cấp thiết được đặt ra hiện nay. Một giải pháp được đề xuất là tổ chức mô hình tư vấn học tập nhóm dựa trên cơ sở tiếp cận lớp tương đồng tối đại trong hệ thống thông tin quyết định đa trị [13]. Theo đó, CVHT sẽ tổ chức tư vấn theo nhóm sinh viên tương đồng về kết quả học tập ở 3 mức độ: "yếu-kém", "trung

bình-khá" và "giỏi-xuất sắc". Nghĩa là sinh viên có kết quả xếp loại yếu hay kém sẽ được tư vấn ở cấp "độ yếu kém", sinh viên có kết quả xếp loại trung bình hay khá sẽ được tư vấn ở cấp độ "trung bình-khá" và sinh viên có kết quả xếp loại giỏi hay xuất sắc sẽ được tư vấn ở cấp độ "giỏi-xuất sắc".

Để có thể tổ chức tư vấn theo nhóm sinh viên, ta cần thực hiện phân nhóm sinh viên tương đồng. Thông thường phân nhóm sinh viên tương đồng dựa trên xếp loại học tập cả năm, nhưng cách làm này có những hạn chế nhất định. Ví dụ một sinh viên có xếp loại học tập học kỳ thứ nhất "khá", học kỳ thứ hai "yếu", và cả năm là "trung bình". Khi phân nhóm sinh viên tương đồng dựa vào xếp loại cả năm, sinh viên này thuộc nhóm sinh viên tương đồng "trung bình" (thực tế sinh viên này đang có kết quả học tập giảm). Trong khi đó, với tiếp cận hệ thống thông tin đa trị, nếu "kết quả học tập" được xem là một thuộc tính đa trị qua nhiều học ký thì khi phân nhóm sinh viên tương đồng, sinh viên này thuộc ít nhất hai nhóm sinh viên: tương đồng "yếu" và tương đồng "khá". Điều này dẫn đến khi sinh luật quyết định đa trị sẽ cho kết quả khác hơn so với cách giải quyết vấn đề dựa trên tiếp cận hệ thống thông tin đơn trị.

Từ phân tích trên, vấn đề thực nghiệm được đặt ra là cần đưa vào một độ đo hấp dẫn (interestingness measure) để đánh giá chất lượng các luật quyết định đa trị được sinh từ tập dữ liệu kết quả học tập của một số lớp chuyên ngành hệ thống thông tin của Trường Đại học Cần Thơ trong một năm học. Kết quả thực nghiệm sẽ là cơ sở khách quan giúp chúng tôi phát hiện các đặc điểm có tính quy luật phổ biến và các đặc điểm có tính bất thường. Từ đó, chỉ ra các hướng tư vấn tốt nhất cho sinh viên dựa vào các luật có độ đo hấp dẫn tốt nhất.

B. Dữ liệu thực nghiệm

1. Dữ liệu gốc

Dữ liệu gốc là điểm tổng kết học kỳ 1 và học kỳ 2, năm 2014-2015 của các lớp ngành Hệ thống thông tin các khóa 37, 38, 39, 40 được kết xuất từ Hệ thống quản lý đào tạo của Trường Đại học Cần Thơ [8]. Bảng điểm tổng kết gồm có 297 dòng (tương ứng với 297 sinh viên) và 6 cột: mã số sinh viên (giá trị mssv đã được thay đổi), trung bình học kỳ 1, rèn luyện học kỳ 1, trung bình học kỳ 2, rèn luyện học kỳ 2 và xếp loại cả năm (bảng 9).

Bảng 5. Bảng điểm tổng kết năm học 2014-2015 gồm 10 sinh viên của các lớp ngành hệ thống thông tin khóa 37-40

mssv	tbhk1	rlhk1	tbhk2	rlhk2	xlcn
s1	3.16	84	0	72	Khá
s2	2.28	82	2.71	78	Trung bình
s3	3.41	93	4	95	Giỏi
s4	2.56	90	1.95	83	Trung bình
s5	1.75	67	2	80	Yếu
s6	2.31	78	2	76	Trung bình
s7	2.36	76	2	74	Trung bình
s8	2.91	83	3.32	85	Khá
s9	2.58	78	3.08	76	Khá
s10	3.57	97	4	92	Giỏi

2. Tiền xử lý dữ liệu gốc

Dữ liệu gốc được xử lý đưa về hệ thống thông tin quyết định đa trị gồm gồm 297 đối tượng $\{s_1, s_2, s_3, \ldots., s_{297}\}$, với 02 thuộc tính điều kiện $\{\text{kqht}\ (c_1),\ \text{kqrl}\ (c_2)\}$ và 01 thuộc tính quyết định $tuvan\ (d)$ (có dạng như bảng 10).

Bảng 6. Hệ thống thông tin quyết định đa trị gồm 297 đối tượng $\{\boldsymbol{s_1, s_2, s_3, \ldots., s_{297}}\}$, 02 thuộc tính điều kiện $\{\boldsymbol{c_1, c_2}\}$ và 01 thuộc tính quyết định $\boldsymbol{d}$.

mssv	*kqht*	*kqrl*	*tuvan*
s1	{4, 1}	{6, 5}	2
s2	{3, 4}	{6, 5}	2
s3	{5, 6}	{7, 7}	3
s4	{4, 2}	{7, 6}	2
s5	{2, 3}	{4, 6}	1
s6	{3, 3}	{5, 5}	2
s7	{3, 3}	{5, 5}	2
s8	{4, 5}	{6, 6}	2
s9	{4, 4}	{5, 5}	2
s10	{5, 6}	{7, 7}	3

Miền giá trị của thuộc tính kết quả học tập "c_1" là $V_{c1} = \{1, 2, 3, 4, 5, 6\}$, các giá trị có ý nghĩa như sau [7]: 1="kém nếu điểm trung bình <1,00", 2="trung bình yếu nếu điểm trung bình từ 1.00 đến 1,99", 3= "trung bình nếu điểm trung bình từ 2.00 đến 2,49", 4="khá nếu điểm trung bình từ 2,50 đến 3,19", 5="giỏi nếu điểm trung bình từ 3,20 đến 3,59" và 6="xuất sắc nếu điểm trung bình từ 3,60 đến 4,00".

Miền giá trị của thuộc tính kết quả rèn luyện "c_2" là $V_{c2} = \{1, 2, 3, 4, 5, 6, 7\}$, các giá trị có ý nghĩa như sau [7]: 1="kém nếu điểm rèn luyện <30", 2="yếu nếu điểm rèn luyện từ 30 đến dưới 50", 3= "trung bình nếu điểm rèn luyện từ 50 đến dưới 60 ", 4="trung bình khá nếu điểm rèn luyện từ 60 đến dưới 70", 5="khái nếu điểm rèn luyện từ 70 đến dưới 80", 6="giỏi nếu điểm rèn luyện từ 80 đến dưới 90" và 7="xuất sắc nếu điểm rèn luyện >90".

Miền giá trị của thuộc tính quyết định "d" là $V_d = \{1, 2, 3\}$, các giá trị quyết định có ý nghĩa như sau: 1="Tư vấn học tập ở mức độ yếu-kém", 2="Tư vấn học tập ở mức độ trung bình-khá", 3=" Tư vấn học tập ở mức độ giỏi-xuất sắc".

C. Công cụ sử dụng

Chúng tôi phát triển công cụ EQSIA (Evaluating the quality of set-valued decision rules based on statistical implicative approach) bằng ngôn ngữ R [1][16], để đánh giá chất lượng các luật quyết định đa trị dựa trên tiếp cận hàm ý thống kê. Mọi kết quả thực nghiệm đều được kết xuất từ công cụ EQSIA.

D. Sinh tập luật quyết định đa trị "tư vấn học tập"

1. Sinh tập các nhóm sinh viên tương đồng tối đại

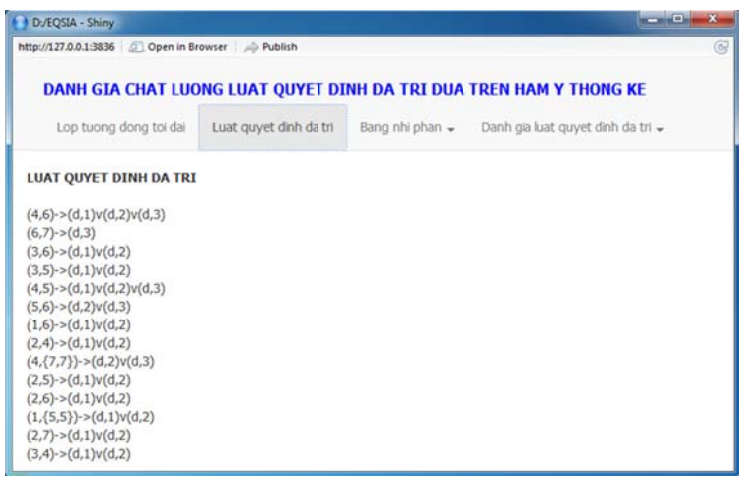
Kết quả phân nhóm sinh viên tương đồng tối đại cho 297 sinh viên ($s_1, s_2, s_3, \ldots, s_{297}$) được 14 nhóm sinh viên tương đồng về kết quả học tập. Mỗi nhóm tương đồng tối đại được diễn tả bằng một tập các giá trị đặc trưng của nó và cùng với giá trị quyết định đa trị tương ứng (hình 2).

	Dac trung	Quyet dinh
,s276,s280,s283,s288,s289,s290,s293,s296,s31,s34,s35,s37,s4,s44,s46,s48,s52,s56,s64,s66,s69,s79,s8,s92}	(4,6)	1 ∨ 2 ∨ 3
{s10,s11,s27,s28,s3,s36,s38,s39,s49}	(6,7)	3
6,s270,s271,s275,s277,s278,s280,s282,s289,s292,s295,s40,s45,s5,s56,s60,s64,s67,s69,s71,s75,s84,s91,s94}	(3,6)	1 ∨ 2
,s32,s33,s40,s42,s45,s53,s59,s6,s61,s62,s65,s67,s68,s69,s7,s70,s73,s74,s77,s84,s87,s91,s93,s94,s97,s99}	(3,5)	1 ∨ 2
55,s260,s266,s267,s271,s275,s276,s280,s283,s289,s31,s33,s34,s37,s41,s47,s48,s52,s63,s69,s70,s89,s9,s99}	(4,5)	1 ∨ 2 ∨ 3
{s109,s110,s12,s138,s14,s145,s151,s153,s23,s283,s30,s35,s43,s45,s49,s60,s75,s8}	(5,6)	2 ∨ 3
{s1,s136,s144,s146,s179,s18,s22,s254,s255,s260,s270,s274,s278,s286,s288,s71,s95}	(1,6)	1 ∨ 2
126,s147,s160,s169,s178,s19,s194,s209,s221,s228,s234,s253,s26,s272,s273,s32,s42,s5,s54,s55,s76,s86,s97}	(2,4)	1 ∨ 2
{s118,s119,s120,s143,s170,s20,s200,s222,s239,s256,s28,s297,s35,s4,s70}	(4, {7,7})	2 ∨ 3
61,s62,s63,s65,s67,s68,s72,s73,s74,s76,s78,s80,s81,s82,s83,s84,s86,s87,s88,s90,s91,s94,s95,s96,s97,s98}	(2,5)	1 ∨ 2
26,s261,s262,s265,s267,s272,s276,s277,s282,s286,s295,s4,s5,s51,s54,s67,s78,s79,s84,s85,s91,s94,s95,s98}	(2,6)	1 ∨ 2
2,s178,s188,s198,s201,s203,s218,s220,s231,s234,s247,s254,s255,s260,s269,s278,s286,s294,s58,s76,s86,s95}	(1, {5,5})	1 ∨ 2
{s116,s119,s170,s207,s227,s239,s256,s282,s4}	(2,7)	1 ∨ 2
1,s105,s167,s169,s194,s201,s209,s218,s219,s221,s228,s258,s26,s269,s273,s279,s32,s42,s5,s56,s93,s97,s99}	(3,4)	1 ∨ 2

Hình 2. Kết quả phân nhóm tương đồng kết quả học tập cho 297 sinh viên.

2. Sinh tập luật quyết định đa trị "tư vấn học tập"

Tập các luật quyết định đa trị sinh được gồm 14 luật quyết định đa trị "tư vấn học tập" tương ứng với 14 nhóm sinh viên tương đồng về kết quả học tập (hình 3).



Hình 3. Kết quả sinh tập luật quyết định đa trị "tư vấn học tập".

3. Tách các luật quyết định đa trị "tư vấn học tập" có vế phải nhiều hơn một quyết định

Kết quả tách các luật quyết định đa trị "tư vấn học tập" có vế phải nhiều hơn một quyết định là một tập luật quyết định đa trị "tư vấn học tập" gồm 29 luật (hình 4).

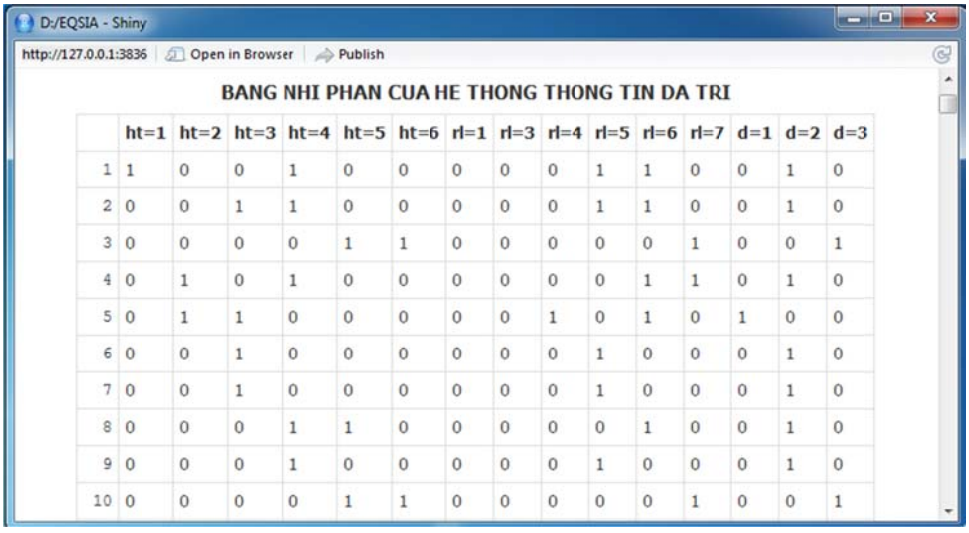
Bảng 7. Kết quả tách tập 14 luật quyết định đa trị "tư vấn học tập"

Luat quyet dinh									
1	(4,6) -> (d, 1)	7	(3,5) -> (d, 1)	13	(5,6) -> (d, 3)	19	(4,{7,7}) -> (d, 3)	25	(1,{5,5}) -> (d, 2)
2	(4,6) -> (d, 2)	8	(3,5) -> (d, 2)	14	(1,6) -> (d, 1)	20	(2,5) -> (d, 1)	26	(2,7) -> (d, 1)
3	(4,6) -> (d, 3)	9	(4,5) -> (d, 1)	15	(1,6) -> (d, 2)	21	(2,5) -> (d, 2)	27	(2,7) -> (d, 2)
4	(6,7) -> (d, 3)	10	(4,5) -> (d, 2)	16	(2,4) -> (d, 1)	22	(2,6) -> (d, 1)	28	(3,4) -> (d, 1)
5	(3,6) -> (d, 1)	11	(4,5) -> (d, 3)	17	(2,4) -> (d, 2)	23	(2,6) -> (d, 2)	29	(3,4) -> (d, 2)
6	(3,6) -> (d, 2)	12	(5,6) -> (d, 2)	18	(4,{7,7}) -> (d, 2)	24	(1,{5,5}) -> (d, 1)		

E. Xác định các giá trị độ đo hàm ý cho các luật quyết định đa trị "tư vấn học tập"

1. Chuyển hệ thống thông tin quyết định đa trị sang dạng nhị phân

Kết quả chuyển hệ thống thông tin quyết định đa trị "kết quả học tập" về dạng nhị phân là một bảng nhị phân gồm 297 dòng (tương ứng với 297 sinh viên) và 16 cột (hình 4 minh họa 10 dòng trong số 297 dòng).

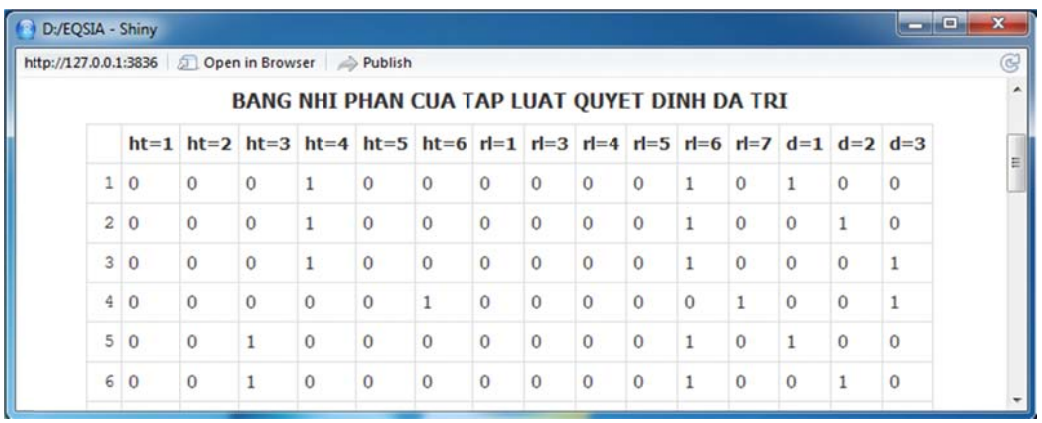


	ht=1	ht=2	ht=3	ht=4	ht=5	ht=6	rl=1	rl=3	rl=4	rl=5	rl=6	rl=7	d=1	d=2	d=3
1	1	0	0	1	0	0	0	0	0	1	1	0	0	1	0
2	0	0	1	1	0	0	0	0	0	1	1	0	0	1	0
3	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1
4	0	1	0	1	0	0	0	0	0	0	1	1	0	1	0
5	0	1	1	0	0	0	0	0	1	0	1	0	1	0	0
6	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
7	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0
8	0	0	0	1	1	0	0	0	0	0	1	0	0	1	0
9	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0
10	0	0	0	0	1	1	0	0	0	0	0	1	0	0	1

Hình 4. Kết quả chuyển hệ thống thông tin quyết định đa trị "kết quả học tập" sang dạng nhị phân.

2. Chuyển tập luật quyết định đa trị "tư vấn học tập" sang dạng nhị phân

Kết quả chuyển tập luật quyết định đa trị "tư vấn học tập" về dạng nhị phân là một bảng nhị phân gồm 29 dòng (tương ứng với 29 luật quyết định đa trị) và 16 cột (hình 5 minh họa 8 dòng trong số 29 dòng).



	ht=1	ht=2	ht=3	ht=4	ht=5	ht=6	rl=1	rl=3	rl=4	rl=5	rl=6	rl=7	d=1	d=2	d=3
1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0
2	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0
3	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1
4	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1
5	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0
6	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0

Hình 5. Kết quả chuyển tập luật quyết định đa trị "tư vấn học tập" sang dạng nhị phân.

3. Xác định các giá trị độ đo hàm ý thống kê cho tập luật quyết định đa trị "tư vấn học tập"

Mỗi luật quyết định đa trị "tư vấn học tập" (bảng 11) được đánh giá chất lượng bằng hai độ đo chỉ số hàm ý và cường độ hàm ý. Kết quả thực nghiệm đánh giá chất lượng 29 luật dựa trên các độ do hàm ý, các luật quyết định này chưa được xếp thứ tự (hình 6 minh họa 10 luật quyết định đầu tiên cùng với các độ đo hàm ý).

DANH GIA LUAT QUYET DINH DA TRI DUA TREN HAM Y THONG KE

	Luat quyet dinh	n	na	nb	nab_	I_indice	I_intensity
1	(4,6) -> (d, 1)	297	81	110	74	3.2206	6e-04
2	(4,6) -> (d, 2)	297	81	170	10	-4.1861	1
3	(4,6) -> (d, 3)	297	81	17	78	0.1873	0.4257
4	(6,7) -> (d, 3)	297	9	17	0	-2.9129	0.9982
5	(3,6) -> (d, 1)	297	72	110	49	0.5446	0.293
6	(3,6) -> (d, 2)	297	72	170	23	-1.4036	0.9198
7	(3,5) -> (d, 1)	297	131	110	77	-0.6036	0.7269
8	(3,5) -> (d, 2)	297	131	170	54	-0.2695	0.6062
9	(4,5) -> (d, 1)	297	78	110	72	3.2661	5e-04
10	(4,5) -> (d, 2)	297	78	170	7	-4.5632	1

Hình 6. Giá trị chỉ số hàm ý, cường độ hàm ý của 10 luật quyết định đa trị trong số 29 luật.

F. Đánh giá chất lượng các luật quyết định đa trị "tư vấn học tập"

1. Đánh giá chất lượng luật quyết định "tư vấn học tập" dựa trên độ đo cường độ hàm ý

Tập 29 luật quyết định đa trị được xếp thứ tự giảm dần dựa trên độ đo cường độ hàm ý (bảng 12). Nếu chọn tập $m = 15$ luật quyết định đa trị đầu tiên có giá trị cường độ hàm ý cao nhất (hay giá trị ngưỡng β= 0.7269): {(4,6) -> (d, 2), (4,5) -> (d,2), (1,{5,5}) -> (d,1), (2,5) -> (d,1), (6,7) -> (d,3), (1,6) -> (d,1), (4,{7,7}) -> (d,2), (2,4) -> (d,1), (3,6) -> (d,2), (2,6) -> (d,1), (5,6) -> (d,3), (2,7) -> (d, 2), (3,4) -> (d, 1), (5,6) -> (d,2), (5,6) -> (d,2), (3,5) -> (d,1)} thì tập $m = 15$ luật quyết định đa trị này chỉ ra các khuynh hướng sau:

(i) Khuynh hướng tư vấn ở mức phù hợp năng lực học tập của sinh viên, có tập 11 luật sau: { (4,6) -> (d,2), (4,5) -> (d,2), (1,{5,5}) -> (d,1), (2,5) -> (d,1), (6,7) -> (d,3), (1,6) -> (d,1), (4,{7,7}) -> (d,2), (2,4) -> (d,1), (3,6) -> (d,2), (2,6) -> (d,1), (5,6) -> (d,3)}. Chẳng hạn: Luật (4,6) -> (d, 2), có nghĩa là (học tập khá ($c_1 = 4$), rèn luyện giỏi ($c_2 = 6$)) → (tư vấn học tập ở mức độ "trung bình-khá", $d = 2$); Luật (4,5) -> (d,2) có nghĩa là (học tập khá ($c_1 = 4$), rèn luyện khá ($c_2 = 5$)) → (tư vấn học tập ở mức độ "trung bình-khá", $d = 2$); Luật (1,{5,5}) -> (d, 1) có nghĩa là (học tập kém ($c_1 = 4$), rèn luyện khá ($c_2 = 5$)) → (tư vấn học tập ở mức độ "yếu-kém", $d = 1$).

(ii) Khuynh hướng tư vấn ở mức cao hơn năng lực học tập của sinh viên, có một luật (2,7) -> (d, 2) ≈ (học tập yếu ($c_1 = 2$), rèn luyện xuất sắc ($c_2 = 7$)) → (tư vấn học tập ở mức độ "trung bình-khá" , $d = 2$).

(iii) Khuynh hướng tư vấn ở mức thấp hơn năng lực học tập của sinh viên, có tập 3 luật sau: {(3,4) -> (d,1), (5,6) -> (d,2), (5,6) -> (d,2), (3,5) -> (d,1)}. Chẳng hạn: Luật (3,4) -> (d,1), có nghĩa là (học tập trung bình ($c_1 = 3$), rèn luyện trung bình khá ($c_2 = 4$)) → (tư vấn học tập ở mức độ "yếu-kém", $d = 1$); Luật (5,6) -> (d,2), có nghĩa là (học tập giỏi ($c_1 = 5$), rèn luyện giỏi ($c_2 = 6$)) → (tư vấn học tập ở mức độ "trung bình-khá", $d = 2$).

Nếu chọn tập $m = 11$ luật quyết định đa trị đầu tiên có giá trị cường độ hàm ý cao nhất là {(4,6) -> (d,2), (4,5) -> (d,2), (1,{5,5}) -> (d,1), (2,5) -> (d,1), (6,7) -> (d,3), (1,6) -> (d,1), (4,{7,7}) -> (d,2), (2,4) -> (d,1), (3,6) -> (d,2), (2,6) -> (d,1), (5,6) -> (d,3) thì tập luật quyết định đa trị này chỉ có một khuynh hướng tư vấn ở mức phù hợp năng lực học tập của sinh viên.

Bảng 8. Kết quả sắp xếp các luật quyết định đa trị dựa trên giá trị cường độ hàm ý giảm dần.

Stt	Luat quyet dinh	n	na	nb	nab_	I_intensity	Stt	Luat quyet dinh	n	na	nb	nab_	I_intensity
1	(4,6) -> (d, 2)	297	81	170	10	1.0000	16	(3,5) -> (d, 2)	297	131	170	54	0.6062
2	(4,5) -> (d, 2)	297	78	170	7	1.0000	17	(4,{7,7})->(d,3)	297	15	17	14	0.5150
3	(1,{5,5})->(d,1)	297	27	110	1	0.9999	18	(4,6) -> (d, 3)	297	81	17	78	0.4257
4	(2,5) -> (d, 1)	297	118	110	47	0.9992	19	(4,5) -> (d, 3)	297	78	17	77	0.3431
5	(6,7) -> (d, 3)	297	9	17	0	0.9982	20	(3,6) -> (d, 1)	297	72	110	49	0.2930
6	(1,6) -> (d, 1)	297	17	110	3	0.9907	21	(2,7) -> (d, 1)	297	9	110	7	0.2877
7	(4,{7,7})->(d,2)	297	15	170	1	0.9837	22	(3,4) -> (d, 2)	297	23	170	12	0.2450
8	(2,4) -> (d, 1)	297	25	110	9	0.9553	23	(2,6) -> (d, 2)	297	58	170	29	0.1996
9	(3,6) -> (d, 2)	297	72	170	23	0.9198	24	(2,4) -> (d, 2)	297	25	170	16	0.0522
10	(2,6) -> (d, 1)	297	58	110	29	0.8933	25	(1,6) -> (d, 2)	297	17	170	14	0.0063
11	(5,6) -> (d, 3)	297	18	17	12	0.8862	26	(2,5) -> (d, 2)	297	118	170	71	0.0019
12	(2,7) -> (d, 2)	297	9	170	2	0.8270	27	(4,6) -> (d, 1)	297	81	110	74	0.0006
13	(3,4) -> (d, 1)	297	23	110	11	0.8199	28	(4,5) -> (d, 1)	297	78	110	72	0.0005
14	(5,6) -> (d, 2)	297	18	170	6	0.7296	29	(1,{5,5})->(d,2)	297	27	170	26	0.0000
15	(3,5) -> (d, 1)	297	131	110	77	0.7269							

2. Đánh giá chất lượng luật quyết định "tư vấn học tập" dựa trên giá trị chỉ số hàm ý

Tập 29 luật quyết định đa trị được xếp thứ tự giảm dần dựa trên độ đo chỉ số hàm ý (bảng 13). Nếu chọn tập $m = 15$ luật quyết định đa trị đầu tiên có giá trị chỉ số hàm ý cao nhất {(1,{5,5}) -> (d,2), (4,5) -> (d, 1), (4,6) -> (d,1), (2,5) -> (d,2), (1,6) -> (d,2), (2,4) -> (d,2), (2,6) -> (d,2), (3,4) -> (d,2), (2,7) -> (d,1), (3,6) -> (d,1), (4,5) -> (d,3), (4,6) -> (d,3), (4,{7,7}) -> (d,3), (3,5) -> (d,2), (3,5) -> (d,1)} thì tập 15 luật quyết định đa trị này chỉ ra các khuynh hướng sau:

(i) Khuynh hướng tư vấn ở mức phù hợp năng lực học tập của sinh viên, có tập 3 luật sau: { (3,4) -> (d,2), (2,7) -> (d,1), (3,5) -> (d,2)}

(ii) Khuynh hướng tư vấn ở mức cao hơn năng lực học tập của sinh viên, có tập 8 luật sau: { (1,{5,5}) -> (d,2), (2,5) -> (d,2), (1,6) -> (d,2), (2,4) -> (d,2), (2,6) -> (d,2), (4,5) -> (d,3), (4,6) -> (d,3), (4,{7,7}) -> (d,3) }

(iii) Khuynh hướng tư vấn ở mức thấp hơn năng lực học tập của sinh viên, có tập 4 luật sau: { (4,5) -> (d,1), (4,6) -> (d,1), (3,6) -> (d,1), (3,5) -> (d,1) }

Bảng 9. Kết quả sắp xếp các luật quyết định đa trị dựa trên giá trị chỉ số hàm ý giảm dần.

Stt	Luat quyet dinh	n	na	nb	nab_	I_indice	Stt	Luat quyet dinh	n	na	nb		I_indice
1	(1,{5,5}) -> (d,2)	297	27	170	26	4.2540	16	(5,6) -> (d, 2)	297	18	170	6	-0.6117
2	(4,5) -> (d, 1)	297	78	110	72	3.2661	17	(3,4) -> (d, 1)	297	23	110	11	-0.9149
3	(4,6) -> (d, 1)	297	81	110	74	3.2206	18	(2,7) -> (d, 2)	297	9	170	2	-0.9423
4	(2,5) -> (d, 2)	297	118	170	71	2.8919	19	(5,6) -> (d, 3)	297	18	17	12	-1.2064
5	(1,6) -> (d, 2)	297	17	170	14	2.4964	20	(2,6) -> (d, 1)	297	58	110	29	-1.2442
6	(2,4) -> (d, 2)	297	25	170	16	1.6240	21	(3,6) -> (d, 2)	297	72	170	23	-1.4036
7	(2,6) -> (d, 2)	297	58	170	29	0.8431	22	(2,4) -> (d, 1)	297	25	110	9	-1.6990
8	(3,4) -> (d, 2)	297	23	170	12	0.6903	23	(4,{7,7}) ->(d,2)	297	15	170	1	-2.1378
9	(2,7) -> (d, 1)	297	9	110	7	0.5601	24	(1,6) -> (d, 1)	297	17	110	3	-2.3547
10	(3,6) -> (d, 1)	297	72	110	49	0.5446	25	(6,7) -> (d, 3)	297	9	17	0	-2.9129
11	(4,5) -> (d, 3)	297	78	17	77	0.4040	26	(2,5) -> (d, 1)	297	118	110	47	-3.1668
12	(4,6) -> (d, 3)	297	81	17	78	0.1873	27	(1,{5,5})->(d,1)	297	27	110	1	-3.8806
13	(4,{7,7}) -> (d,3)	297	15	17	14	-0.0376	28	(4,6) -> (d, 2)	297	81	170	10	-4.1861
14	(3,5) -> (d, 2)	297	131	170	54	-0.2695	29	(4,5) -> (d, 2)	297	78	170	7	-4.5632
15	(3,5) -> (d, 1)	297	131	110	77	-0.6036							

VI. KẾT LUẬN

Đánh giá chất lượng tập luật quyết định đa trị dựa trên tiếp cận hàm ý thống kê là một hướng tiếp cận mới, được thực hiện nhằm chỉ ra vai trò của các luật quyết định đa trị dựa trên các giá trị độ đo hàm ý thống kê. Thông qua giá trị độ đo hàm ý thống kê, các luật quyết định đa trị được đánh giá và sắp xếp thứ tự. Để có thể chọn được các luật tốt, chúng tôi đề xuất việc chọn lựa theo hai hướng, đó là dựa vào một giá trị ngưỡng β cho trước hay chọn m luật quyết định đa trị đầu tiên có giá trị độ đo hàm ý thống kê cao nhất. Ngoài ra, chúng tôi cũng đề xuất có thể xem xét thêm các luật quyết định đa trị "xấu nhất" dựa vào một ngưỡng θ cho trước hoặc lấy k luật quyết định đa trị có giá trị độ đo hàm ý thống kê thấp nhất. Phân tích hàm ý thống kê cho phép đánh giá sự hấp dẫn của các luật và cấu trúc chúng để phát hiện những mối quan hệ ở các mức độ chi tiết khác nhau và làm nổi bật các thuộc tính nổi trội của các luật quyết định đa trị. Bài viết cũng đề xuất giải thuật tổng quát đánh giá chất lượng các luật quyết định đa trị dựa trên tiếp cận hàm ý thống kê.

VII. LỜI CẢM ƠN

Nhóm tác giả xin chân thành cảm ơn sự hỗ trợ của Nguyễn Minh Kỳ, Trường Đại học Kỹ thuật và Công nghệ Cần Thơ, trong việc lập trình R.

VIII. TÀI LIỆU THAM KHẢO

[1] Abedin, J., Das, K., K., "Data Manipulation with R" (ISBN 978-1-78528-881-4), Packt, Second Edition, 2015.

[2] B. Walczak, B., Massart, D.,L., Rough sets theory, International Journal of Information Sciences 47, Springer-Verlag, pp.1-16,1999.

[3] Cios, K. J., Pedrycz, W., Swiniarski, R.W.: "Rough Sets: Data Mining: Methods for Knowledge Discovery", pp.27-45. Kluwer Academic Publishers, Boston/Dordrecht/London, 1998.

[4] Gras, R., Kuntz, P.: "An overview of the Statistical Implication Analysis (SIA) development", Statistical Implicative Analysis - Studies in Computational Intelligence (Volume 127), Springer-Verlag, pp.11-40, 2008.

[5] Gras, R., Suzuki, E., Guillet, F., Spagnolo, F., "Statistical Implicative Analysis, Methodology and concepts for SIA", Springer-Verlag Berlin Heidelberg (ISBN 978-3-540-78982-6), pp.8-70, 2008.

[6] Greco, S., *et al.:* "Rough sets theory for multicriteria decision analysis", European Journal of Operational Research 129, pp.1-47, 2001.

[7] Guan, Y.-Y., Wang, H.-K.: "Set-valued Information Systems", International Journal of Information Sciences 176, Elsevier, pp.2507-2525, 2006.

[8] Hà, T., T., *el al.:* Hệ thống quản lý (https://htql.ctu.edu.vn/htql/login.php), Trường Đại học Cần Thơ, 2015.

[9] Hà, T., T., Qui định về công tác học vụ, Trường Đại học Cần Thơ (Số: 2736 /QĐ-ĐHCT), 2014.

[10] Huỳnh, M.T., Phan, T.T., Huỳnh, X.H., "Cảnh báo cháy rừng với luật quyết định đa trị", Hội thảo quốc gia lần thứ XVII về một số vấn đề chọn lọc của công nghệ thông tin & truyền thông, Nhà xuất bản khoa học và kỹ thuật, pp.402-408, 2014.

[11] Kryszkiewicz, M.: "Rules in Incomplete Information Systems", International Journal of Information Sciences 113, pp.271-292, 1999.

[12] Leung, Y., Li, D.: "Maximal Consistent Block Technique for Rule Acquisition in Incomplete Information systems", International Journal of Information Sciences 153, pp.85-106, 2003.

[13] Phan, T.T. *et al.*: "Tư vấn học tập nhóm bậc đại học trên cơ sở tiếp cận lớp tương đồng lớn nhất trong hệ thống thông tin đa trị", Tạp chí khoa học Đại học Cần Thơ (ISSN 1859-2333), pp.123-133, 2013.

[14] Phan, T.T., Huynh, X.H.: "Mô hình trao đổi thông tin nhiều chiều dựa trên tiếp cận hệ thống thông tin quyết định đa trị", Hội thảo khoa học hệ thống thông tin, Đại học Đà Năng, pp.23-31, 2014.

[15] Song, X.,-X., Zhang,W.,-X., "Rough Sets and Knowledge Technology", Formal Concept Analysis and Set-Valued Information Systems (Volume 4481), Springers (ISBN: 9783540724582 – 9783540724575), pp.395-402, 2007.

[16] Spector, P., "Data Manipulation with R", Springers (ISBN: 978-0-387-74730-9), 2008.

[17] Wang, G.,Yang, Q., Zhang, Q., "Rough Sets, Fuzzy Sets, Data Mining and Granular Computing", Disjunctive Set-Valued Ordered Information Systems Based on Variable Precision Dominance Relation (Volume 6743), Springer, pp.207-210, 2011.

[18] Zhu, D., Feng, B., Guan, T., "MICAI 2005: Advances in Artificial Intelligence", Rough Sets and Decision Rules in Fuzzy Set-Valued Information Systems (Volume 3789), Springer, pp.204-213, 2005.

EVALUATING THE QUALITY OF SET-VALUED DECISION RULES BASED ON STATISTICAL IMPLICATIVE APPROACH

Phan Tan Tai, Le Duc Thang, Huynh Xuan Hiep

ABSTRACT - *In this paper, we introduce a new approach for evaluating the quality of set-valued decision rules based on statistical implicative analysis. The research is begun to consider the set-values decision information system and appropriate generated set-values decision rules. From here, the obtained set-values decision rules will consider and assess the quality of the rules based on statistical implicative analysis. Through this result, the set-valued decision rules will be arranged according to the different priority levels based on two measures: implication indice, implication intensity. The experimental scenarios are deployed on good set-valued decision rules, bad set-valued decision rules and the role of the set-valued decision rules. This is a problem that the set-valued decision rules model not previously researched.*