Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00019

GIẢI PHÁP HỖ TRỢ SINH VIÊN LẬP KẾ HOẠCH HỌC TẬP DỰA TRÊN TIẾP CẬN TẬP THÔ

Lê Đức Thắng[1], Trương Thị Hải[2], Nguyễn Thái Nghe[1], Huỳnh Xuân Hiệp[1]

[1]Khoa CNTT&TT, Trường Đại học Cần Thơ
[2]Khoa Công Nghệ, Trường Đại học Phú Xuân
{ldthang,ntnghe,hxhiep}@cit.ctu.edu.vn, trnhai208@gmail.com

TÓM TẮT—*Lý thuyết tập thô (rough set theory) đã được xây dựng trên một nền tảng toán học vững chắc vì thế được ứng dụng trong khá nhiều lĩnh vực, đặc biệt là ứng dụng trong khoa học máy tính như giải quyết bài toán phân lớp (đa nhãn hoặc đơn nhãn), phân cụm và luật kết hợp. Dựa trên tiếp cận tập thô, bài viết này đề xuất một phương pháp mới trong dự đoán kết quả học tập của sinh viên nhằm hỗ trợ sinh viên lập kế hoạch học tập phù hợp. Thực nghiệm trên dữ liệu thực tế để xác định các sinh viên có thuộc diện "cảnh báo" hay "không cảnh báo" đã cho thấy phương pháp này có khả năng dự đoán tốt hơn so với các phương pháp khai phá dữ liệu tiêu biểu như Cây quyết định, láng giềng lân cận và một số phương pháp sinh luật khác. Tiếp cận tập thô cũng đã cho thấy nó rất hiệu quả trong trường hợp dữ liệu mất cân bằng.*

Từ khóa—*Lý thuyết tập thô; bảng quyết định; luật quyết định; dữ liệu mất cân bằng; lựa chọn môn học.*

I. GIỚI THIỆU

Lý thuyết tập thô (rough set theory) - do Zdzislaw Pawlak đề xuất năm 1982 [14] - được xây dựng trên một nền tảng toán học vững chắc vì thế được ứng dụng trong khá nhiều lĩnh vực, đặc biệt là ứng dụng trong khoa học máy tính như giải quyết bài toán phân lớp (đa nhãn hoặc đơn nhãn), phân cụm và luật kết hợp. Bài viết này sẽ đề xuất sử dụng lý thuyết tập thô trong xây dựng hệ thống hỗ trợ giáo dục đào tạo, đặc biệt là việc dự đoán kết quả học tập của sinh viên.

Thật vậy, dự đoán kết quả học tập của sinh viên một các chính xác là rất hữu ích trong nhiều ngữ cảnh khác nhau ở các trường đào tạo đại học và sau đại học. Chẳng hạn, xác định các ứng viên xuất sắc để tham gia các đội tuyển tin học, hoặc cấp học bổng nhằm khuyến khích họ nỗ lực hơn nữa trong học tập, hay việc xác định các sinh viên có năng lực yếu kém để có những biện pháp thích hợp nhằm hỗ trợ họ học tập tốt hơn.

Thời gian gần đây, số lượng sinh viên bị buộc thôi học có chiều hướng tăng ở nhiều trường đại học (chẳng hạn như tại Đại học Cần Thơ, hàng năm có trên dưới 150 sinh viên thuộc diện buộc thôi học do kết quả học tập yếu kém [4]) và thường tập trung vào những sinh viên học năm thứ ba và năm thứ tư. Một phần nguyên nhân là do sinh viên không có kế hoạch học tập phù hợp. Hiện tượng này đã gây tổn thất lớn cho bản thân sinh viên, gia đình và toàn xã hội. Chính vì thế việc phát hiện sớm các học viên yếu kém để giúp họ lập kế hoạch học tập sao cho phù hợp là một nhu cầu rất cần thiết.

Dự đoán kết quả học tập của sinh viên cũng đã được nhiều nhà nghiên cứu quan tâm, như trong [1][2][3][5][7][8][10][11]. Hàng năm đều có hội thảo quốc tế chuyên về lĩnh vực này (xem chi tiết tại www.educationaldatamining.org), ở đó ta có thể tìm thấy những nghiên cứu gần nhất. Tuy nhiên phần lớn các nghiên cứu trên tập trung vào việc ứng dụng các kỹ thuật trong khai phá dữ liệu (data mining) trong dự đoán kết quả học tập của sinh viên [12][13].

Nghiên cứu này sẽ **đề xuất một phương pháp mới trong dự đoán kết quả học tập của sinh viên dựa trên tiếp cận tập thô**. Trước hết, bài viết sẽ giới thiệu một trong những ưu điểm của lý thuyết tập thô, là dùng để đơn giản hóa thông tin trước khi sinh ra luật quyết định, do đó tập luật thu được khá tối tiểu nhưng vẫn bao hàm được thông tin mong muốn. Sau đó đề xuất ứng dụng cho bài toán dự đoán/phân lớp (prediction/classification) sinh viên thuộc diện cảnh báo học vụ hay không, từ đó hỗ trợ họ lập kế hoạch học tập phù hợp. Thực nghiệm trên dữ liệu thực tế đã cho thấy phương pháp này có khả năng dự đoán tốt hơn so với các phương pháp khai phá dữ liệu truyền thống khác như Cây quyết định, láng giềng lân cận, SVM,… đặc biệt là trong trường hợp dữ liệu mất cân bằng (imbalanced data).

Trong các phần tiếp theo, bài viết sẽ giới thiệu ngắn gọn về lý thuyết tập thô cũng như minh họa cho bài toán phân lớp dựa trên các luật sinh ra, sau cùng là mô hình đề xuất và kết quả thực nghiệm.

II. LÝ THUYẾT TẬP THÔ

A. Hệ thống thông tin quyết định

Thông thường một hệ thống biểu diễn tri thức được mô tả bằng hệ thống thông tin quyết định, đó là một bộ bốn DIS = (U, A∪{d}, V, f). Ở đó, U là một tập hợp hữu hạn các đối tượng; A là một tập hợp hữu hạn các thuộc tính điều kiện, d∉A là thuộc tính quyết định. Mỗi thuộc tính a ∈ A∪{d} định nghĩa một hàm thông tin $f_a : U \to V_a$, với V_a là miền giá trị của a, và V là tập hợp các V_a.

Một hệ thống thông tin quyết định được mô tả bằng một bảng quyết định như ví dụ sau.

Bảng 1. Bảng quyết định

U	A			{d}
	a_1	a_2	a_3	
x_1	2	1	3	1
x_2	3	2	1	2
x_3	2	1	3	1
x_4	2	2	3	2
x_5	1	1	4	3
x_6	1	1	2	3
x_7	3	2	1	2
x_8	1	1	4	3
x_9	2	1	3	1
x_{10}	3	2	1	2

B. Quan hệ đẳng trị

Với tập con các thuộc tính điều kiện $B \subseteq A$, quan hệ theo B, ký hiệu $Ind(B)$, được định nghĩa như sau [16]: $Ind(B) = \{(x, y) \in U \times U / f_b(x) = f_b(y) \ \forall b \in B\}$ Ind(B) là quan hệ tương đương. Ký hiệu U/Ind(B) là tập hợp tất cả các lớp tương đương, mỗi lớp tương đương được gọi là tập hợp sơ cấp. Với mỗi thuộc tính $b \in B$ và với mỗi tập hợp sơ cấp $S^B \in U / Ind(B)$ thì mọi đối tượng trong S^B có cùng một giá trị đối với b, ký hiệu là $f_b(S^B)$. Với thuộc tính điều kiện d, tập hợp các lớp tương đương U/Ind(d) được gọi là một phân hoạch của U. Theo Bảng 1 ta có:

$$U / Ind(A) = \{\{x_1, x_3, x_9\}, \{x_2, x_7, x_{10}\}, \{x_4\}, \{x_5, x_8\}, \{x_6\}\} \; U / Ind\{d\} = \{\{x_1, x_3, x_9\}, \{x_2, x_4, x_7, x_{10}\}, \{x_5, x_6, x_8\}\}$$

C. Xấp xỉ dưới và xấp xỉ trên - Tập thô

Với mỗi tập con $X \subset U$, xấp xỉ dưới và xấp xỉ trên của X theo $B \subseteq A$ lần lượt được ký hiệu và định nghĩa như sau: $\underline{B}(X)$ tập các đối tượng thuộc về các tập hợp sơ cấp theo B nằm trong X, $\overline{B}(X)$ là tập các đối tượng thuộc về tập hợp sơ cấp theo B có phần tử chung với X. Với $X = \{x_1, x_3, x_4, x_5, x_9\}$ và $B \subseteq A$ ta có : $\underline{A}(X) = \{x_1, x_3, x_4, x_9\}$ và $\overline{A}(X) = \{x_1, x_3, x_4, x_5, x_8, x_9\}$.

Để diễn đạt một cách chính xác tập con $X \subset U$ chúng ta dùng bộ $(\underline{B}(X), \overline{B}(X))$, gọi là tập thô của X.

Giả sử phân hoạch $U / Ind\{d\}$ có r lớp: $D_1, D_2, ..., D_r$, xấp xỉ dưới và xấp xỉ trên theo $B \subseteq A$ của phân hoạch được ký hiệu và định nghĩa như sau:

$$\underline{B}(U / Ind\{d\}) = \{\underline{B}(D_1), \underline{B}(D_2), ..., \underline{B}(D_r)\} \; \overline{B}(U / Ind\{d\}) = \{\overline{B}(D_1), \overline{B}(D_2), ..., \overline{B}(D_r)\}$$

D. Thuộc tính d-thừa; d-nhân và d-rút gọn của các thuộc tính

Thuộc tính điều kiện $a_i \in B \subseteq A$ được là thuộc tính d-thừa nếu việc sử dụng nó không ảnh hưởng đến xấp xỉ dưới của phân hoạch $U / Ind\{d\}$ theo B, tức là: $\underline{B}(U / Ind\{d\}) = \underline{(B - \{a_i\})}(U / Ind\{d\})$, ngược lại, a_i là thuộc tính d-không thể thiếu được. Tập hợp tất cả các thuộc tính điều kiện d-không thể thiếu được được gọi là d-nhân của A. Tập con tối tiểu các thuộc tính điều kiện phân biệt tất cả các lớp tương đương trong $U / Ind\{d\}$ được gọi là d-rút gọn của A.

Để tìm ra các d-rút gọn và d-nhân của A, trước tiên ma trận d-phân biệt α được xây dựng trên tập hợp các đối tượng được sử dụng. Đây là ma trận vuông, đối xứng, có số dòng và số cột là số đối tượng. Phần tử $\alpha(x_i, x_j)$ của ma trận này là tập hợp tất cả các thuộc tính điều kiện phân biệt được đối tượng x_i và x_j: $\alpha(x_i, x_j) = \{a \in A / f_a(x_i) \neq f_a(x_j)\}$. d-nhân của A chính là tập hợp các phần tử đơn trong ma trận d-phân biệt.

Để tính các d-rút gọn của A, người ta dùng hàm d-phân biệt $f^{\{d\}}(A)$, là hàm bool có công thức là dạng tuyển chuẩn tắc được định nghĩa như sau: $f^{\{d\}}(A) = \bigwedge_{x_i \in U} \bigwedge_{x_j \in U, j>i} \vee \alpha(x_i, x_j)$. Mỗi nguyên nhân nguyên tố trong công thức của $f^{\{d\}}(A)$ tương ứng với một d-rút gọn của A.

E. d- rút gọn và d- nhân của các giá trị thuộc tính

Sau khi đơn giản hóa hệ thống thông tin bằng rút gọn của tập hợp các thuộc tính A, chúng ta tiếp tục đơn giản hóa nữa bằng cái rút gọn và nhân của các giá trị thuộc tính. Cách tìm cái d-rút gọn và d-nhân của giá trị thuộc tính giống như

tìm cái d-rút gọn và d-nhân của tập hợp thuộc tính, đều dựa trên ma trận phân biệt α. Nhưng thay vì chỉ tính một hàm d-phân biệt thì phải tính nhiều hàm d-phân biệt cho từng đối tượng theo công thức: $f_i^{\{d\}}(A) = \underset{x_j \in U, j \neq i}{\wedge} \vee \alpha(x_i, x_j)$, mỗi nguyên nhân nguyên tố trong đó là một d-rút gọn của giá trị thuộc tính. Khi đó chúng ta chỉ quan tâm đến giá trị của các thuộc tính trong d-rút gọn của giá trị thuộc tính.

III. SINH LUẬT TỪ BẢNG QUYẾT ĐỊNH

Từ các kết quả trên chúng tôi đề xuất thủ tục phân tích bảng quyết định (U, A∪{d}, V, f) để nhận được các luật quyết định tối ưu phục vụ việc phân lớp/dự đoán như sau:

1. Xây dựng ma trận d-phân biệt α có các phần tử được xác định bởi:
   $\alpha(x_i, x_j) = \{a \in A / f_a(x_i) \neq f_a(x_j)\}$
2. Xây dựng hàm d- phân biệt:
   $f^{\{d\}}(A) = \underset{x_i \in U}{\wedge} \underset{x_j \in U, j>i}{\wedge} \vee \alpha(x_i, x_j)$
3. Chọn một nguyên nhân nguyên tố của $f^{\{d\}}(A)$ làm d- rút gọn, gọi là B.
4. Rút gọn ma trận d-phân biệt α theo B
5. Xây dựng hàm d- phân biệt cho mỗi đối tượng $x_i \in U$:
   $f_i^{\{d\}}(A) = \underset{x_j \in U, j \neq i}{\wedge} \vee \alpha(x_i, x_j)$
   $\alpha(x_i, x_j) = \{b \in B / f_b(x_i) \neq f_b(x_j)\}$

   Mỗi nguyên nhân nguyên tố trong $f_i^{\{d\}}(A)$ cho biết các thuộc tính có giá trị cần quan tâm đối với đối tượng x_i (các giá trị không cần quan tâm được thay thế bằng *).
6. Xây dựng bảng quyết định rút gọn (U, B∪{d}, V, f) với các giá trị thuộc tính cần quan tâm.
7. Xây dựng quan hệ trên bảng quyết định rút gọn:

$$Ind(B) = \left\{ \begin{array}{l} (x, y) \in U \times U / \forall b \in B : f_b(x) = f_b(y) \\ | f_b(x) = * \\ | f_b(y) = * \end{array} \right\}$$

Mỗi lớp $S \in U / Ind(B)$ sinh ra một luật theo cách như sau:

$$\underset{b \in S}{\wedge} (b = f_b(S) \neq *) \rightarrow (d = f_d(S))$$

Ví dụ minh họa

Chúng ta bắt đầu từ Bảng 1, với thuộc tính điều kiện $A = \{a_1, a_2, a_3\}$ và thuộc tính quyết định $\{d\}$ ta có:

1. Trước tiên ma trận d-phân biệt α được xây dựng như sau (ký hiệu $a_1a_2a_3$ nghĩa là $\{a_1, a_2, a_3\}$)

Bảng 2. Ma trận d-phân biệt

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	-									
x_2	$a_1a_2a_3$	-								
x_3	-	$a_1a_2a_3$	-							
x_4	a_2	-	a_2	-						
x_5	a_1a_3	$a_1a_2a_3$	a_1a_3	$a_1a_2a_3$	-					
x_6	a_1a_3	$a_1a_2a_3$	a_1a_3	$a_1a_2a_3$	-	-				
x_7	$a_1a_2a_3$	-	$a_1a_2a_3$	-	$a_1a_2a_3$	$a_1a_2a_3$	-			
x_8	a_1a_3	$a_1a_2a_3$	a_1a_3	$a_1a_2a_3$	-	-	$a_1a_2a_3$	-		
x_9	-	$a_1a_2a_3$	-	a_2	a_1a_3	a_1a_3	$a_1a_2a_3$	a_1a_3	-	
x_{10}	$a_1a_2a_3$	-	$a_1a_2a_3$	-	$a_1a_2a_3$	$a_1a_2a_3$	-	$a_1a_2a_3$	$a_1a_2a_3$	-

2. Hàm d-phân biệt tính được là:

$$f^{\{d\}}(A) = a_1a_2 \vee a_2a_3 .$$

3. Theo kết quả này thì có hai d-rút gọn là $\{a_1, a_2\}$, $\{a_2, a_3\}$ và một d-nhân là $\{a_2\}$. Hai d-rút gọn này có thể được chọn lần lượt để đơn giản hóa Bảng 1. Giả sử chúng ta chọn $\{a_1, a_2\}$

4. Rút gọn ma trận d-phân biệt α theo $\{a_1, a_2\}$, kết quả như trong Bảng 3

Bảng 3. Ma trận d-phân biệt rút gọn

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	-	a_1a_2	-	a_2	a_1	a_1	a_1a_2	a_1	-	a_1a_2
x_2	a_1a_2	-	a_1a_2	-	a_1a_2	a_1a_2	-	a_1a_2	a_1a_2	-
x_3	-	a_1a_2	-	a_2	a_1	a_1	a_1a_2	a_1	-	a_1a_2
x_4	a_2	-	a_2	-	a_1a_2	a_1a_2	-	a_1a_2	a_2	-
x_5	a_1	a_1a_2	a_1	a_1a_2	-	-	a_1a_2	-	a_1	a_1a_2
x_6	a_1	a_1a_2	a_1	a_1a_2	-	-	a_1a_2	-	a_1	a_1a_2
x_7	a_1a_2	-	a_1a_2	-	a_1a_2	a_1a_2	-	a_1a_2	a_1a_2	-
x_8	a_1	a_1a_2	a_1	a_1a_2	-	-	a_1a_2	-	a_1	a_1a_2
x_9	-	a_1a_2	-	a_2	a_1	a_1	a_1a_2	a_1	-	a_1a_2
x_{10}	a_1a_2	-	a_1a_2	-	a_1a_2	a_1a_2	-	a_1a_2	a_1a_2	-

5. Xây dựng các hàm {d}-phân biệt cho các đối tượng:

$f_1^{\{d\}}(A) = a_1a_2$, $f_2^{\{d\}}(A) = a_1 \vee a_2$, $f_3^{\{d\}}(A) = a_1a_2$, $f_4^{\{d\}}(A) = a_2$, $f_5^{\{d\}}(A) = a_1$, $f_6^{\{d\}}(A) = a_1$, $f_7^{\{d\}}(A) = a_1 \vee a_2$, $f_8^{\{d\}}(A) = a_1$, $f_9^{\{d\}}(D) = a_1a_2$, $f_{10}^{\{d\}}(A) = a_1 \vee a_2$.

6. Xây dựng bảng quyết định rút gọn sau cùng:

Bảng 4. Rút gọn Bảng 1

U	a_1	a_2	{d}
x_1	2	1	1
x_2	*	2	2
x_3	2	1	1
x_4	*	2	2
x_5	1	*	3
x_6	1	*	3
x_7	*	2	2
x_8	1	*	3
x_9	2	1	1
x_{10}	*	2	2

7. Quan hệ $Ind\{a_1, a_2\}$ cho kết quả như sau:

$U / Ind\{a_1, a_2\} = \{\{x_1, x_3, x_9\}, \{x_2, x_4, x_7, x_{10}\}, \{x_5, x_6, x_8\}\}$. Khi đó chúng ta có các luật quyết định:

$a_1 = 2 \wedge a_2 = 1 \rightarrow d = 1$

$a_2 = 2 \rightarrow d = 2$

$a_1 = 1 \rightarrow d = 3$

Trong bài toán dự đoán kết quả học tập của sinh viên, thì a_1, a_2,…a_n sẽ là các thuộc tính đầu vào (predictors) và d sẽ là thuộc tính kết quả cần dự đoán (target attribute).

Ví dụ, một luật sinh ra có dạng:

"Giới tính" = "Nam" ∧ "trình độ anh văn" = "chưa có chứng chỉ" ∧ ... ∧ "điểm TB học kỳ trước" = "Cảnh báo" → "Điểm TB học kỳ này" = "Cảnh báo".

Dựa trên những luật như thế, ta có thể phân lớp (dự đoán) kết quả của các sinh viên mới (tương tự như luật sinh ra từ cây quyết định). Chi tiết về mô hình và thuộc tính, dữ liệu sẽ được mô tả trong phần tiếp theo.

IV. DỰ ĐOÁN KẾT QUẢ HỌC TẬP DỰA TRÊN TIẾP CẬN TẬP THÔ

Tương tự như những bài toán khác trong khai phá dữ liệu, việc xây dựng hệ thống dự đoán kết quả học tập cũng được thực hiện dựa trên quy trình chuẩn CRISP-DM (CRoss Industry Standard Process for Data Mining). Quy trình này bao gồm sáu giai đoạn, bao gồm: Tìm hiểu vấn đề, tìm hiểu dữ liệu, tiền xử lý dữ liệu, mô hình hóa, đánh giá mô hình và triển khai ứng dụng. Trong bài viết này, chúng tôi tập trung trên việc đề xuất và đánh giá mô hình, những chi tiết khác vui lòng xem thêm trong [4][8].

A. Phát biểu vấn đề

Vấn đề cần dự đoán ở đây là dựa trên các thông tin nhân khẩu học của sinh viên (như giới tính, độ tuổi, trình độ Anh văn, điểm tuyển sinh đầu vào,…) và điểm trung bình của học kỳ trước để dự đoán điểm trung bình học kỳ tiếp theo. Việc dự đoán này sẽ giúp bản thân sinh viên và giáo viên cố vấn học tập hỗ trợ sinh viên lập kế hoạch học tập sao cho phù hợp nhằm giảm bớt tình trạng cảnh báo học vụ và buộc thôi học, do trong quy chế đào tạo nếu mỗi sinh viên trong hai học kỳ liên tiếp có điểm trung bình dưới 0.8 (thang điểm 4) thì sẽ bị buộc thôi học. Cụ thể, nghiên cứu này sẽ dự đoán điểm trung bình của học kỳ 6 (cuối năm thứ 3) dựa trên thông tin nhân khẩu học và điểm trung bình của học kỳ 5. Tuy nhiên, việc dự đoán cho các học kỳ khác là hoàn toàn tương tự [3] [8].

B. Dữ liệu

Để có được tập dữ liệu cho mô hình dự đoán, nhóm tác giả đã tìm hiểu và thu thập dữ liệu từ hệ thống thực tế của trường Đại học Cần Thơ [3][8], từ đó tiến hành tiền xử lý dữ liệu bằng cách loại bỏ các giá trị dư thừa và thiếu (missing), số lượng mẫu tin (mỗi mẫu tin tương ứng với một sinh viên) còn lại là **19530** mẫu tin. Sau khi loại bỏ các thuộc tính thừa, 14 thuộc tính quan trọng còn lại được dùng cho việc dự đoán, mô tả trong bảng dưới đây (dữ liệu này kế thừa từ nghiên cứu [8]).

Bảng 5. Các thuộc tính dùng cho dự đoán

STT	Thuộc tính	Infomation Gain	Diễn giải
1	GPA_Sem5	0.429	Điểm trung bình học kỳ 5
2	FOS	0.177	Ngành học
3	Faculty	0.131	Khoa học
4	Gender	0.089	Giới tính
5	EntryMark	0.039	Điểm tuyển sinh đại học
6	Age Range	0.032	Độ tuổi
7	English Skill	0.023	Trình độ anh văn
8	Policy Priority	0.016	Gia đình diện chính sách
9	Family Job	0.014	Nghề nghiệp gia đình
10	School Rank	0.012	Trong dữ liệu thu thập được, có 285 trường phổ thông trung học mà sinh viên học trước khi vào trường đại học Cần Thơ. Vì thế các trường học đó được sắp xếp theo các giá trị liên tục dựa trên sự chênh lệch giữa tỷ lệ tốt nghiệp. Tỷ lệ đó được tính như sau: Rank = AVG(Tỷ lệ tốt nghiệp đại học) – AVG(Tỷ lệ đầu vào đại học) [8] Giá trị rank từ 1 đến 10
11	Province	0.010	Quê quán
12	Area Priority	0.004	Khu vực ưu tiên
13	Ethnic	0.001	Dân tộc
14	Religious	0.001	Tôn giáo

C. Phương pháp

- Đầu vào: 14 thuộc tính đã mô tả ở Bảng 5 (gồm kết quả của học kỳ trước).
- Kỹ thuật: Sinh luật quyết định theo tiếp cận tập thô, từ đó tiến hành phân lớp kết quả học tập.
- Đầu ra: Kết quả học tập học kỳ kế tiếp. Trong nghiên cứu này, chúng tôi quan tâm đến việc phân lớp nhị phân (binary classification) với hai giá trị là "Cảnh báo" hoặc "Không cảnh báo".

V. KẾT QUẢ THỰC NGHIỆM

A. Dữ liệu thực nghiệm

Như đã mô tả ở phần trước, tập dữ liệu gồm có **19530** dòng và 14 thuộc tính. Thuộc tính cần dự đoán có phân phối 1565/17965 tương ứng với hai lớp 'cảnh báo'/ 'không cảnh báo'. Tập dữ liệu này thuộc dạng mất cân bằng (imbalanced data) do chỉ có 8.01% thuộc lớp số ít (minority class) [7][9].

B. Các kỹ thuật khác dùng để so sánh

Chúng tôi sẽ so sánh phương pháp đề xuất dùng lý thuyết tập thô (đặt tên là RSRule) với các phương pháp phổ biến khác trong data mining như: láng giềng lân cận (kNN), máy học véctơ hỗ trợ (SVM) và các phương pháp sinh luật khác như Decision Tree, Conjunctive Rule, Decision Table và PART [15]. Các phương pháp này đã được cài đặt sẵn trong công cụ Weka (www.cs.waikato.ac.nz/ml/weka).

C. Kết quả

Pháp kiểm tra chéo 5 đường (5-folds cross validation) được sử dụng để so sánh kết quả. Ở đây, do tập dữ liệu khá mất cân bằng nên độ đo chính xác (accuracy) tỏ ra không phù hợp. Thay vào đó, chúng tôi trình bày chi tiết kết quả của ma trận nhầm lẫn (confusion matrix) như trong Hình 1, tỷ lệ True Positive (true positive trong trường hợp này chính là số sinh viên thuộc diện "Cảnh báo" được dự đoán đúng – do ta sẽ quan tâm đến các đối tượng này nhiều hơn) và độ đo G-Mean [3][9].

Rõ ràng rằng nếu sử dụng độ chính xác thì từ Hình 1 ta dễ dàng xác định được kỹ thuật ConjunctiveRule có độ chính xác là 17965/19530 = 91.98% cao hơn RSRule, do RSRule chỉ đạt (523+17347)/19530 = 91.5%. Tuy nhiên kết quả này không có ý nghĩa do tất cả các sinh viên thuộc diện "Cảnh báo" đã bị dự đoán sai (phương pháp Conjunctive Rule), mặc dù đây mới chính là đối tượng mà ta cần dự đoán. Do vậy, trong Bảng 6 chúng tôi trình bày số lượng và tỷ lệ sinh viên thuộc diện "Cảnh báo" học vụ được các mô hình dự đoán đúng, ở đây ta thấy RSRule tỏ ra hiệu quả hơn các phương pháp khác.

kNN	Dự đoán	
Thực tế	Cảnh báo	Không
Cảnh báo	277	1288
Không	259	17706

ConjunctiveRule	Dự đoán	
Thực tế	Cảnh báo	Không
Cảnh báo	0	1565
Không	0	17965

DecisionTree	Dự đoán	
Thực tế	Cảnh báo	Không
Cảnh báo	468	1097
Không	304	17661

DecisionTable	Dự đoán	
Thực tế	Cảnh báo	Không
Cảnh báo	507	1058
Không	331	17634

SVM	Dự đoán	
Thực tế	Cảnh báo	Không
Cảnh báo	333	1232
Không	242	17723

PART	Dự đoán	
Thực tế	Cảnh báo	Không
Cảnh báo	488	1077
Không	466	17499

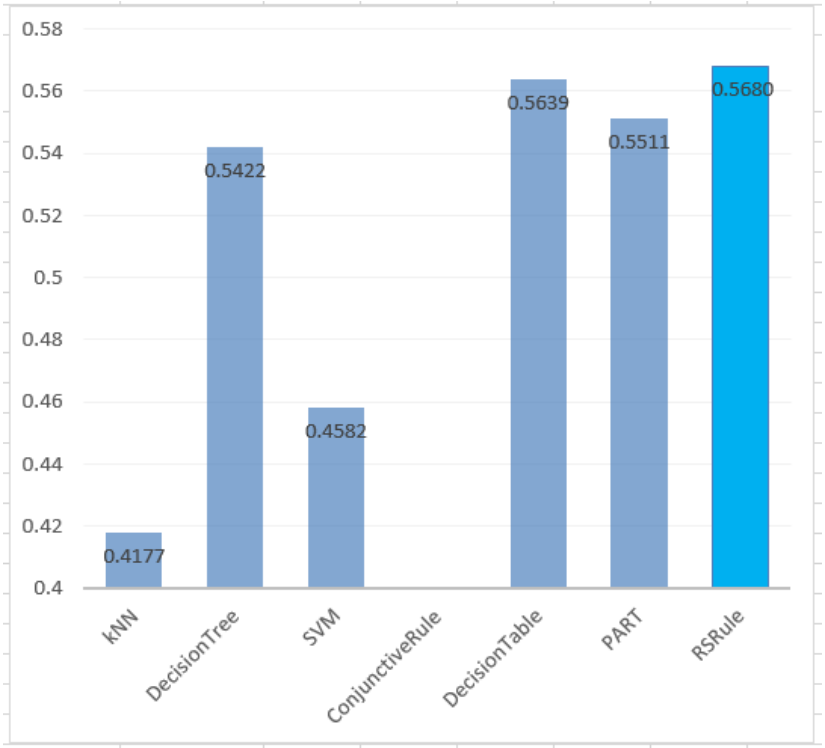
RSRule	Dự đoán	
Thực tế	Cảnh báo	Không
Cảnh báo	523	1042
Không	618	17347

Hình 1. Kết quả dự đoán thể hiện bằng Confusion matrix

Bên cạnh đó chúng tôi cũng trình bày kết quả so sánh của các phương pháp bằng độ đo G-Mean, đây là độ đo hay được sử dụng trong trường hợp dữ liệu mất cân bằng [3][9], kết quả như trong Hình 2. Thực nghiệm cũng cho thấy rằng tiếp cận tập thô RSRule cho kết quả dự đoán tốt hơn.

Bảng 6. Tỷ lệ SV bị "Cảnh báo" được các mô hình dự đoán đúng

Kỹ thuật	Số lượng SV bị "Cảnh báo" được dự đoán đúng	Tỷ lệ % (trong tổng số 1565 SV bị "Cảnh báo")
Conjunctive Rule	0	0.00
kNN	277	17.70
SVM	333	21.28
Decision Tree	468	29.90
PART	488	31.18
Decision Table	507	32.40
RSRule	**523**	**33.42**



Hình 2. So sánh bằng độ đo G-Mean

Từ những kết quả trên đã cho thấy tiếp cận tập thô đã được đề xuất cho bài toán dự đoán kết quả học tập là phù hợp. Đặc biệt là khi phân lớp trong môi trường dữ liệu mất cân bằng. Tuy nhiên, việc so sánh với các phương pháp dành riêng cho loại dữ liệu này sẽ được thực hiện trong tương lai.

VI. KẾT LUẬN

Bài viết này đã đề xuất phương pháp sử dụng lý thuyết tập thô trong dự đoán kết quả học tập của sinh viên nhằm hỗ trợ sinh viên lập kế hoạch học tập phù hợp. Thực nghiệm trên dữ liệu thực tế để xác định các sinh viên có thuộc diện "cảnh báo" hay "không cảnh báo" đã cho thấy phương pháp này có khả năng dự đoán tốt hơn so với các phương pháp khai phá dữ liệu truyền thống khác như Cây quyết định, láng giềng lân cận,.. đặc biệt là trong trường hợp dữ liệu mất cân bằng. Chúng tôi sẽ tiếp tục mở rộng sang các lĩnh vực khác trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Bekele, R. and Menzel, W. 2005. A Bayesian approach to predict performance of a student (BAPPS): A case with Ethiopian students. Proceedings of the International Conference on Artifi. Intell. and Applications (AIA-2005).

[2] Delavari N. & Beikzadeh M. R & Shirazi M. R. A. 2004. A New Model for Using Data Mining in Higher Educational System. Proceedings of 5th Inter. Conf. on Information Technology Based Higher Education and Training.

[3] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, September 2009.

[4] Huỳnh Lý Thanh Nhàn, Nguyễn Thái Nghe. 2013. Hệ thống dự đoán kết quả học tập và gợi ý lựa chọn môn học. Kỷ yếu Hội thảo quốc gia lần thứ XVI: Một số vấn đề chọc lọc của CNTT&TT, 110-118. NXB Khoa học và Kỹ thuật.

[5] Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., and Punch, W. F. 2003. Predicting student performance: an application of data mining methods with an educational web-based system. Proceedings of 33rd Annual Conference on Frontiers in Education (FIE 2003).

[6] Nguyễn Thái Nghe, Huỳnh Xuân Hiệp. 2012. Ứng dụng kỹ thuật phân rã ma trận đa quan hệ trong xây dựng hệ trợ giảng thông minh. Kỷ yếu Hội thảo quốc gia lần thứ XV: Một số vấn đề chọn lọc của CNTT&TT, 470-477. NXB Khoa học và Kỹ thuật. ISBN: 893-5048-931578

[7] Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme. 2009. Improving Academic Performance Prediction by Dealing with Class Imbalance, in Proceedings of the 9th IEEE Inter. Conf. on Intell. Syst. Design and Applications (ISDA 2009), 878-883. IEEE CS.

[8] Nguyen Thai-Nghe, Paul Janecek, and Peter Haddawy. 2007. A comparative analysis of techniques for predicting academic performance, Proceedings of the 37th IEEE Frontiers in Education, 7-12. IEEE Xplore.

[9] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. 2010. Cost-Sensitive Learning Methods for Imbalanced Data, Proceedings of IEEE Inter. Joint Conf. on Neural Networks, ISBN 978-1-4244-6916-1. IEEE Xplore.

[10] Nguyễn Thị Thanh Thủy, Nguyễn Trần Quốc Vinh. Ứng dụng khai phá dữ liệu xây dựng công cụ dự đoán kết quả học tập của sinh viên. Kỷ yếu Hội nghị SV NCKH lần thứ 8, Đại học Đà Nẵng, 2012.

[11] Romero, C., Ventura, S., Espejo, P.G., Hervas, C. 2008. Data Mining Algorithms to Classify Students. Proceedings of the First Inter. Conf. on Educational Data Mining, 8-17.

[12] Romero, Cristobal, and Sebastian Ventura. 2013. Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3.1 (2013): 12-27.

[13] V Ramesh, P Parkavi and K Ramar. Article: Predicting Student Performance: A Statistical and Data Mining Approach. International Journal of Computer Applications 63(8):35-39, February 2013.

[14] Walczak, B., and D. L. Massart. "Rough sets theory." Chemometrics and intelligent laboratory systems, 47.1 (1999): 1-16

[15] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[16] Walczak, B., and D. L. Massart. "Rough sets theory." Chemometrics and intelligent laboratory systems 47.1 (1999): 1-16.

DECISION SUPPORT FOR ACADEMIC STUDY PLANNING BASED ON ROUGH SET APPROACH

Le Duc Thang, Truong Thi Hai, Nguyen Thai Nghe, Huynh Xuan Hiep

ABSTRACT*—Rough set theory was conducted on stability mathematic background, thus, it has been applied in many areas, especially in computer science for machine learning problems (e.g., classification, clustering, and association rules). Based on rough set theory, this work proposes a new approach in predicting student study results to support their academic study planning. Experimental results show that the proposed approach work well on binary classification problem, especially when the data set is imbalanced.*

Keywords*— Rough set theory; decision table; decision rule; imbalanced data; academic study planning.*