

Về một phương pháp rút gọn thuộc tính cho bảng quyết định theo tiếp cận topo mờ trực cảm

Trần Thanh Đại¹, Nguyễn Long Giang², Trần Thị Ngân³, Hoàng Thị Minh Châu⁴, Vũ Thu Uyên⁵, Vương Trung Hiếu⁶

^{1,4,5} Trường Đại học Kinh tế Kỹ thuật Công nghiệp, Hà Nội

² Viện CNTT - Viện Hàn lâm Khoa học Công nghệ Việt Nam, Hà Nội

³ Trường Đại học Thủy Lợi, Hà Nội

⁶ Trường Đại học Công nghiệp, Hà Nội

Tác giả liên hệ: Trần Thanh Đại, tt Daiuneti@gmail.com

Ngày nhận bài: 14/08/2022, ngày sửa chữa: 15/11/2022, ngày duyệt đăng: 25/11/2022

Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n2.1132

Tóm tắt: Hầu hết các phương pháp rút gọn thuộc tính theo tiếp cận tính toán hạt của tập thô và tập thô mở rộng hiện nay đều sử dụng các độ đo để đánh giá độ quan trọng của thuộc tính cũng như định nghĩa tập rút gọn. Các độ đo này chủ yếu lấy xấp xỉ độ tương tự giữa các hạt thông tin mờ trực cảm mà không thể hiện đầy đủ mức độ tương tự về mặt cấu trúc, do đó tập rút gọn thu được còn chưa hiệu quả về kích thước. Do đó, trong bài báo này chúng tôi đề xuất mô hình rút gọn thuộc tính theo tiếp cận topo mờ trực cảm (Intuitionistic Fuzzy Topology - IFT). Trong đó độ đo độ khác biệt giữa các cơ sở con (subbase) của IFT được định nghĩa để làm công cụ phân loại thuộc tính và cấu trúc cơ sở (base) của IFT đơn vị được sử dụng để định nghĩa tập rút gọn. Các kết quả phân tích về phương diện lý thuyết và thực nghiệm cho thấy phương pháp rút gọn thuộc tính theo tiếp cận IFT cho tập rút gọn có kích thước nhỏ hơn đáng kể so với tiếp cận độ đo truyền thống, trong khi độ chính xác phân lớp của tập rút gọn thu được có thể chấp nhận được trong một số bài toán thực tế.

Từ khóa: rút gọn thuộc tính, tập thô, tập mờ trực cảm, topo, topo mờ trực cảm.

Title: Intuitionistic Fuzzy Topology-Based Attribute Reduction on the Decision Table

Abstract: Most attribute reduction methods following the granular computing approach of rough sets and extended rough sets today use metrics to evaluate the attribute's importance and define the reduced set. These measures mainly approximate the similarity between intuitionistic fuzzy granular but do not fully represent the structural similarity so that the resulting reduct could be more efficient in size. This paper proposes an attribute reduction model according to the intuitionistic fuzzy topological approach. The measure of the difference between the subbases is defined to evaluate the attribute's importance, and the unit base of the unit topology is used to define the reduced set. The results of the theoretical and experimental analysis show that the attribute proposed method has a significantly smaller size than the traditional measure approach, while the classification accuracy of the reduct is acceptable in some real problems.

Keywords: attribute reduction, rough set, intuitionistic fuzzy set, topology, intuitionistic fuzzy topology.

I. GIỚI THIỆU

Rút gọn thuộc tính là một quá trình tiền xử lý dữ liệu quan trọng trong các lĩnh vực nhận dạng mẫu, khai thác dữ liệu và học máy. Trong thực tế, nhiều thuộc tính (attribute) hay các đặc trưng (features) có thể loại bỏ khỏi tập dữ liệu mà không ảnh hưởng đến chất lượng của mô hình được xây dựng so với tập dữ liệu gốc [2].

Đối với các bảng quyết định có miền giá trị số (liên tục). Các thuật toán hiện nay đều rút gọn trực tiếp trên các

bảng quyết định gốc mà không phải trải qua quá trình rời rạc hóa dữ liệu. Khi đó khái niệm các hạt thông tin mờ dựa trên nền tập mờ (Fuzzy Set - FS) [7] và hạt thông tin mờ trực cảm dựa trên nền tập mờ trực cảm (Intuitionistic Fuzzy Set - IFS) [9, 10, 12] được sử dụng. Đây là các khái niệm mở rộng từ hạt thông tin truyền thống để xây dựng các độ đo mở rộng như: miền dương mờ trực cảm (Intuitionistic Fuzzy POS - IFPOS) [10], entropy mờ trực cảm (Intuitionistic Fuzzy Entropy - IFE) [12], khoảng cách mờ trực cảm (Intuitionistic Fuzzy Distance - IFD) [9]. Trên

cơ sở đó, các thuật toán rút gọn thuộc tính trên tập nền FS và IFS được phát triển [9, 10, 12]. Tuy nhiên các thuật toán này vẫn sử dụng các độ đo để đánh giá độ quan trọng của thuộc tính cũng như định nghĩa tập rút gọn. Các độ đo này chủ yếu lấy xấp xỉ độ tương tự giữa các hạt thông tin mờ trực cảm mà không thể hiện đầy đủ mức độ tương tự về mặt cấu trúc, do đó tập rút gọn thu được còn chưa hiệu quả về kích thước.

Gần đây tiếp cận rút gọn thuộc tính trên nền không gian topo được nhiều nhà nghiên cứu quan tâm do những mối liên hệ về cấu trúc giữa tập thô và topo nói chung, tập thô mờ trực cảm (Intuitionistic Fuzzy Rough Set - IFRS) và IFT [2–4] nói riêng. Trong đó, một số phương pháp xây dựng topo đã được đề xuất thông qua các phương pháp tiếp cận như: phương pháp dựa trên quan hệ ưu tiên [11], phương pháp dựa trên tập thô (RS) [2–4]. Tuy nhiên các nghiên cứu này chỉ dừng lại ở khía cạnh lý thuyết, chưa có ứng dụng rút gọn thuộc tính hoàn chỉnh cho bảng quyết định nói chung và đặc biệt đối với bảng quyết định có miền giá trị số/liên tục nói riêng.

Trong bài báo này chúng tôi sử dụng cấu trúc topo mờ trực cảm để xây dựng mô hình rút gọn thuộc tính theo tiếp cận Filter. Trong đó tập nền IFS được sử dụng do IFS được cộng đồng nghiên cứu đánh giá cao về khả năng xử lý tốt các trường hợp dữ liệu nhiễu [8, 9, 10]. Sử dụng độ đo độ tương tự giữa các cấu trúc cơ sở con mờ trực cảm (IF-subbase) của IFT để đánh giá thuộc tính và sử dụng cấu trúc cơ sở mờ trực cảm (IF-base) đơn vị của IFT làm điều kiện dừng của thuật toán. Phần còn lại của bài báo được tổ chức như sau:

Phần II nhắc lại một số kiến thức cơ bản về tập nền IFS và cấu trúc IFT. Phần III đề xuất phương pháp xây dựng cấu trúc IFT và các phép toán tương ứng trên các cấu trúc IF-base và IF-subbase. Phần IV đề xuất mô hình rút gọn thuộc tính theo tiếp cận Filter. Phần V trình bày một số kết quả thực nghiệm và kết luận được trình bày trong Phần VI của bài báo.

II. CÁC KHÁI NIỆM CƠ BẢN

Trước khi đi vào cấu trúc IFT được đề xuất trong Phần III của bài báo, phần này nhắc lại một số khái niệm về bảng quyết định miền giá trị số, các định nghĩa và các tính chất về tập nền IFS và không gian IFT.

Bảng quyết định miền giá trị số được kí hiệu bởi trong đó O là tập không rỗng các đối tượng, X là tập không rỗng các thuộc tính điều kiện và Y là thuộc tính quyết định. Đối với miền giá trị của mỗi thuộc tính điều kiện trong X là các giá trị số liên tục trong khi miền giá trị của thuộc tính Y là các giá trị rời rạc.

Định nghĩa 2.1: [3] Cho O là một tập khác rỗng các đối tượng. Tập mờ trực cảm (IFS) $A(A_{IF})$ có dạng:

Bảng I
BẢNG QUYẾT ĐỊNH MIỀN GIÁ TRỊ SỐ

U	x_1	x_2	x_3	x_4	x_5	x_6	Y
o_1	1.0	0.4	0.8	0.2	1.0	0.0	0
o_2	1.0	0.4	0.2	0.4	0.2	0.8	1
o_3	0.8	0.6	1.0	0.0	0.6	0.4	0
o_4	0.2	0.6	0.8	0.2	0.0	1.0	1
o_5	0.2	0.8	0.8	0.2	0.0	1.0	1
o_6	0.2	0.8	0.2	0.8	0.0	1.0	0

$$A_{IF} = \{ \langle o, \mu_{A_{IF}}(o), \gamma_{A_{IF}}(o) \rangle : o \in O \}$$

Với mỗi $o \in O$, độ thành viên của o in A_{IF} được kí hiệu bởi $\mu_{A_{IF}} : O \rightarrow [0, 1]$, và độ không thành viên của o in A_{IF} được kí hiệu bởi $\gamma_{A_{IF}} : O \rightarrow [0, 1]$, trong đó $0 \leq \mu_{A_{IF}}(o) + \gamma_{A_{IF}}(o) \leq 1$.

Mệnh đề 2.1: [3] Cho A_{IF} và B_{IF} là hai tập mờ trực cảm xác định trên O . Ta có:

(1) $A_{IF} \subseteq B_{IF}$ iff $\mu_{A_{IF}}(o) \leq \mu_{B_{IF}}(o)$ and $\gamma_{A_{IF}}(o) \geq \gamma_{B_{IF}}(o)$ for all $o \in O$.

(2) $A_{IF} = B_{IF}$ iff $A_{IF} \subseteq B_{IF}$ and $B_{IF} \subseteq A_{IF}$.

(3) $\bar{A}_{IF} = \{ \langle o, \gamma_{A_{IF}}(o), \mu_{A_{IF}}(o) \rangle : o \in O \}$

(4) $A_{IF} \cap B_{IF} = \left\{ \left\langle o, \mu_{A_{IF}}(o) \wedge \mu_{B_{IF}}(o), \gamma_{A_{IF}}(o) \vee \gamma_{B_{IF}}(o) \right\rangle : o \in O \right\}$

(5) $A_{IF} \cup B_{IF} = \left\{ \left\langle o, \mu_{A_{IF}}(o) \vee \mu_{B_{IF}}(o), \gamma_{A_{IF}}(o) \wedge \gamma_{B_{IF}}(o) \right\rangle : o \in O \right\}$

Định nghĩa 2.2: [2] Topo mờ trực cảm (IFT) xác định trên O là một họ $\mathcal{T}$ các tập IFS trên O thỏa mãn các tiên đề sau:

(T₁) $0_{IF}, 1_{IF} \in \mathcal{T}$

(T₂) $G_1 \cap G_2 \in \mathcal{T} : G_1, G_2 \in \mathcal{T}$

(T₃) $\cup G_i \in \mathcal{T} : \{G_i : G_i \in \mathcal{T}, i \in I\}$

Khi đó, cặp $(O, \mathcal{T})$ được gọi là không gian IFT.

Định nghĩa 2.3: [2] Cho O là tập không rỗng các đối tượng, khi đó cơ sở (base) của topo $\mathcal{T}$ trên O là một tập hợp $\mathcal{B}$ của O sao cho:

(1) Với mọi $o \in O$, tồn tại $G \in \mathcal{B}$ sao cho $o \in G$.

(2) Nếu $o \in G_1 \cap G_2$, thì tồn $G_3 \in \mathcal{B}$ sao cho $o \in G_3$. Trong đó $G_3 = G_1 \cap G_2$, và $G_1, G_2 \in \mathcal{B}$.

Định nghĩa 2.4: [2] Cho không gian topo mờ trực cảm $(O, \mathcal{T})$. Khi đó $\mathcal{S} \subseteq \mathcal{T}$ được gọi là một cơ sở con (subbase) của $\mathcal{T}$ nếu giao hữu hạn các phần tử của $\mathcal{S}$ tạo thành cơ sở (base) $\mathcal{B}$ của $\mathcal{T}$.

III. CẤU TRÚC TOPO MỜ TRỰC CẢM SỬ DỤNG QUAN HỆ ƯU TIÊN MỜ TRỰC CẢM

Trước khi đi vào đề xuất phương pháp rút gọn thuộc tính cho bảng quyết định theo tiếp cận IFT trong Phần IV của bài báo, phần này đề xuất phương pháp xây dựng các tập cơ sở mờ trực cảm (IF-base) và cơ sở con mờ trực cảm

(IF-subbase). Sau đó là một số phép toán và tính chất hoạt động trên các IF-base và IF-subbase được khảo sát.

Định nghĩa 3.1: Cho bảng quyết định $DT = (O, X, Y)$. Với mọi $(m, n) \in O$ và $\delta \in [0.5, 1]$, quan hệ ưu tiên $IFR_a^\geq$ tương ứng với thuộc tính $a \subseteq X$ được xác định như sau:

$$IFR_a^\geq(m, n) = \langle n, \mu_n, \gamma_n \rangle$$

Trong đó:

$$\mu_n = \begin{cases} 1 - |a(m) - a(n)| & \text{if } p_a(m, n) \geq \delta \\ 0 & \text{if } other \end{cases} \quad (1)$$

$$\gamma_n = 1 - \mu_n$$

Với: $p_a(m, n) = \frac{a(m) - a(n) + 1}{2}$; $a(m)$ và $a(n)$ là các giá trị của m và n tương ứng với thuộc tính a .

Định nghĩa 3.2: Cho bảng quyết định $DT = (O, X, Y)$ và quan hệ ưu tiên mờ trực cảm $IFR_a^\geq$ tương ứng với thuộc tính $a \subseteq X$. Khi đó với mọi $(x, y) \in O$, quan hệ $IFR_a^\geq$ có thể được biểu diễn dưới ma trận quan hệ mờ trực cảm như sau: $M_a = [IFR_a^\geq(x, y)]_{O \times O}$.

Ví dụ 1: Cho bảng quyết định $DT = (O, X, Y)$ như trong Bảng I, với $\delta = 0.5$ ta có:

$$M_{x_1}^\geq = \begin{bmatrix} (1,0) & (1,0) & (0.8,0.2) & (0.2,0.8) & (0.2,0.8) & (0.2,0.8) \\ (1,0) & (1,0) & (0.8,0.2) & (0.2,0.8) & (0.2,0.8) & (0.2,0.8) \\ (0,1) & (0,1) & (1,0) & (0.4,0.6) & (0.4,0.6) & (0.4,0.6) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \end{bmatrix}$$

Định nghĩa 3.3: Cho bảng quyết định $DT = (O, X, Y)$ và ma trận quan hệ $M_a^\geq$, khi đó IF-subbase S_a được xác định như sau: $S_a = \{S_a^L, S_a^R\}$.

Trong đó $S_a^L = M_a^\geq$ và $S_a^R = (M_a^\geq)^T$, với $(M_a^\geq)^T$ là ma trận chuyển vị của ma trận $M_a^\geq$.

Ví dụ 2: Tiếp theo Ví dụ 1, ta có:

$$(M_{x_1}^\geq)^T = \begin{bmatrix} (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (0.8,0.2) & (0.8,0.2) & (1,0) & (0,1) & (0,1) & (0,1) \\ (0.2,0.8) & (0.2,0.8) & (0.4,0.6) & (1,0) & (1,0) & (1,0) \\ (0.2,0.8) & (0.2,0.8) & (0.4,0.6) & (1,0) & (1,0) & (1,0) \\ (0.2,0.8) & (0.2,0.8) & (0.4,0.6) & (1,0) & (1,0) & (1,0) \end{bmatrix}$$

Định nghĩa 3.4: Cho bảng quyết định $DT = (O, X, Y)$ và hai IF-subbase S_p, S_q tương ứng của $p, q \in X$. Khi đó giao của S_p và S_q được xác định như sau:

$$S_p \cap S_q = \{S_p^L \cap S_q^L, S_p^R \cap S_q^R\} \quad (2)$$

Định nghĩa 3.5: Cho bảng quyết định $DT = (O, X, Y)$ và hai IF-subbase S_p, S_q tương ứng của $p, q \in X$. Khi đó hợp của S_p và S_q được xác định như sau:

$$S_p \cup S_q = \{S_p^L \cup S_q^L, S_p^R \cup S_q^R\} \quad (3)$$

Định nghĩa 3.6: Cho bảng quyết định $DT = (O, X, Y)$ và $S_a = \{S_a^L, S_a^R\}$, trong đó S_a^L là IF-subbase trái, và S_a^R là IF-subbase phải tương ứng của $a \in X$. Khi đó IF-base B_a được xác định như sau:

$$B_a = S_a^L \cap S_a^R \quad (4)$$

Mệnh đề 3.1: Cho bảng quyết định $DT = (O, X, Y)$ và IF-subbase B_a tương ứng với thuộc tính $a \in X$ được định nghĩa bởi công thức (4). Khi đó B_a là một cơ sở của IFT $\mathcal{T}_a$.

Ví dụ 3: Tiếp theo ví dụ 2, ta có:

$$B_{x_1} = \begin{bmatrix} (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (0,1) & (0,1) & (1,0) & (0,1) & (0,1) & (0,1) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \end{bmatrix}$$

IV. PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH THEO TIẾP CẬN TOPO MỜ TRỰC CẢM

Trong phần này chúng tôi đề xuất thuật toán rút gọn thuộc tính theo tiếp cận Filter, sử dụng cấu trúc IFT được định nghĩa trong Phần III của bài báo. Trong đó độ đo độ khác biệt giữa các IF-subbase được xây dựng để đánh giá thuộc tính và sử dụng IF-base của IFT đơn vị để làm điều kiện dừng của thuật toán. Cuối cùng là phần đánh giá độ phức tạp của thuật toán.

1. Độ đo độ khác biệt giữa các cơ sở con của topo mờ trực cảm

Định nghĩa 4.1: Cho bảng quyết định $DT = (O, X, Y)$ và hai IF-subbase S_p, S_q tương ứng của thuộc tính $p, q \in X$. Khi độ khác biệt của S_p và S_q được xác định như sau:

$$\begin{aligned} \zeta(S_p, S_q) &= \frac{1}{|O|} \sum_{i=1}^{|O|} (|S_p^L[i] \cup S_q^L[i]| - |S_p^L[i] \cap S_q^L[i]|) \\ &+ \frac{1}{|O|} \sum_{i=1}^{|O|} (|S_p^R[i] \cup S_q^R[i]| - |S_p^R[i] \cap S_q^R[i]|) \end{aligned} \quad (5)$$

Mệnh đề 4.1: Cho bảng quyết định $DT = (O, X, Y)$ và hai IF-subbase S_p, S_q tương ứng của thuộc tính $p, q \in X$. Khi độ công thức (5) là tương đương với công thức (6) sau đây:

$$\begin{aligned} \zeta(S_p, S_q) &= \frac{2}{|O|} \sum_{i=1}^{|O|} (|S_p^L[i] \cup S_q^L[i]| - |S_p^L[i] \cap S_q^L[i]|) \end{aligned} \quad (6)$$

Mệnh đề 4.2: Cho bảng quyết định $DT = (O, X, Y)$ và hai IF-subbase $S_X, S_{X \cup Y}$. Khi đó:

$$\zeta(S_X, S_{X \cup Y}) = \frac{2}{|O|} \sum_{i=1}^{|O|} (|S_Y^L[i] - S_Y^L[i] \cap S_X^L[i]|) \quad (7)$$

là độ đo khác biệt giữa S_X và $S_{X \cup Y}$.

Mệnh đề 4.3: Cho bảng quyết định $DT = (O, X, Y)$ và hai IF-subbase S_X, S_R , với $R \subseteq X$. Khi đó:

$$\zeta(S_Y, S_{Y \cup X}) \leq \zeta(S_Y, S_{Y \cup R})$$

2. Rút gọn thuộc tính cho bảng quyết định sử dụng cấu trúc topo mờ trực cảm.

Định nghĩa 4.2: Cho bảng quyết định $DT = (O, X, Y)$ và $R \subseteq X$. Khi đó độ quan trọng của thuộc tính $a \in X$ với R được xác định như sau:

$$Sig_R(a) = \zeta(S_Y, S_{Y \cup R \cup a}) - \zeta(S_Y, S_{Y \cup R}) \quad (8)$$

Mệnh đề 4.4: Cho bảng quyết định $DT = (O, X, Y)$ và hai IF-subbase S_X, S_R , với $R \subseteq X$. Khi đó: nếu $\mathcal{B}_R = \mathcal{B}_I$ thì $\mathcal{B}_X = \mathcal{B}_I$.

Trong đó $\mathcal{B}_I$ là một IF-base của IFT đơn vị, với $\mathcal{B}_I$ được xác định như sau:

$$B_I = \left\{ \begin{array}{l} (1, 0) : i = j \\ (0, 1) : i \neq j \end{array} \right\}_{0 \leq i, j \leq |O|}$$

Định nghĩa 4.3: Cho bảng quyết định $DT = (O, X, Y)$ và $R \subseteq X$. Khi đó R được gọi là tập rút gọn của X khi và chỉ khi:

- (1) $\mathcal{B}_R = \mathcal{B}_I$;
- (2) $\mathcal{B}_{R-c} \neq \mathcal{B}_I$ với mọi $c \in R$.

Thuật toán 1: Thuật toán Filter thuộc tính theo tiếp cận topo mờ trực cảm (F_IFT).	
1	Dữ liệu vào: Decision Table $DT = (O, X, Y)$ với $\delta = 0.5$.
2	Dữ liệu ra: Tập rút gọn R .
3	begin
4	$R \leftarrow \emptyset$;
5	$\mathcal{B}_R \leftarrow [1 \sim]_{O \times O}$;
6	Init $\mathcal{B}_I$;
7	for $c \in X \cup Y$ do
8	calculate S_c dựa trên công thức (1);
9	end
10	while $\mathcal{B}_R \neq \mathcal{B}_I$ do
11	for $c \in X - R$ do
12	calculate $Sig_R(c)$ dựa trên công thức (8);
13	end
14	select $c_m \in X - R : c_m = \underset{c \in X-R}{Max} \{Sig_R(c)\}$;
15	$R \leftarrow R \cup \{c_m\}$;
16	update $\mathcal{B}_R$ dựa trên công thức (4);
17	end
18	Return R ;
19	end

Tiếp theo chúng tôi đánh giá độ phức tạp tính toán của giải thuật rút gọn đề xuất F_IFT. Trước tiên chúng tôi kí hiệu $|X|, |O|$ tương ứng là kích thước của tập thuộc tính X và kích thước của tập đối tượng O :

- (1). Độ phức tạp tính toán tại các dòng lệnh 7-9 là $O(|X||O|^2)$;
- (2). Độ phức tạp tính toán tại các dòng lệnh 11-13 là $O(|X - R||O|^2)$, độ phức tạp tại dòng lệnh 14-15 là $O(|X - R|)$, độ phức tạp tại dòng lệnh 16 là $O(|O|^2)$.

Do đó độ phức tạp tính toán của vòng lặp **while** là: $O(|X - R||O|^2)$;

Từ (1) and (2), ta có độ phức tạp tính toán của thuật toán F_IFT là $O(|X||O|^2)$.

V. MỘT SỐ KẾT QUẢ THỰC NGHIỆM

Để củng cố cơ sở lý thuyết về phương pháp rút gọn thuộc tính cho bảng quyết định theo tiếp cận IFT và đánh giá hiệu năng của thuật toán đề xuất F_IFT được trình bày trong Phần IV của bài báo, phần này trình bày kế hoạch thực nghiệm và các kết quả thực nghiệm của thuật toán F_IFT trên các bộ dữ liệu thực.

Bảng II
BẢNG MÔ TẢ DỮ LIỆU THỰC NGHIỆM

ID	Data	Describe	O	X	Y
1	Wine	Wine	178	13	3
2	Heart	Statlog (Heart)	270	13	2
3	Wdbc	Breast Cancer Wisconsin (Diagnostic)	569	30	2
4	Wpbc	Breast Cancer Wisconsin (Prognostic)	198	33	2
5	Iono	Ionosphere	351	34	2
6	UFDC	Ultrasonic flowmeter diagnostics (C)	181	43	4
7	Sona	Connectionist Bench	208	60	2
8	Libras	Libras Movement	360	90	15
9	Musk	Musk	476	166	2
10	LVB	Voice Rehabilitation(Binary)	126	310	2
11	LVG	Voice Rehabilitation(Gender)	126	310	2
12	PD	Parkinson's Disease Classification	756	754	2

1. Kế hoạch và môi trường thực nghiệm

Việc thực nghiệm nhằm đánh giá tính hiệu quả của thuật toán đề xuất bằng cách so sánh thuật toán đề xuất F_IFT với các thuật toán F_IFPOS [10], F_IFE [12], và F_IFD [9]. Tất cả các thuật toán trên đều được xây dựng theo tiếp cận Filter thuộc tính và sử dụng tập nền IFS để xây dựng các độ đo. Môi trường thực nghiệm cho các thuật toán được thực hiện trên nền hệ điều hành Window 10, CPU core i5 Processor – RAM 8Gb. Tất cả các thuật toán đều được cài đặt bằng ngôn ngữ lập trình Python và thực nghiệm với 12 bộ dữ liệu thực từ UCI [8]. Chi tiết về các tập dữ liệu được mô tả trong Bảng II.

Để đánh giá các thuật toán, chúng tôi sử dụng ba tiêu chí để đánh giá đó là:

- 1) Kích thước của tập rút gọn;
- 2) Độ chính xác phân lớp thuộc đoạn $[0,1]$;
- 3) Thời gian thực hiện của thuật toán (s).

Trong đó độ chính xác phân lớp được thống kê trên hai mô hình phân lớp SVM và K-láng giềng gần (KNN)

với số lượng láng giềng là $K=10$ và $p = 2$ là khoảng cách Euclidean được sử dụng. Độ đo *accuracy* kết hợp với phương pháp đánh giá chéo 10-fold được sử dụng trên từng mô hình phân lớp. Các tập dữ liệu trước khi rút gọn được chuẩn hóa về miền giá trị $[0,1]$ và công thức (1) được sử dụng với $\delta = 0.5$. Kết quả so sánh trong các Bảng là giá trị trung bình của 10 lần chạy với 10 lần lấy mẫu ngẫu nhiên trên từng tập dữ liệu.

2. Đánh giá thuật toán F_IFT

Bảng III mô tả các thuộc tính của tập rút gọn thu được từ thuật toán đề xuất F_IFT. Bảng IV so sánh kích thước của tập rút gọn giữa thuật toán đề xuất F_IFT so với nhóm các thuật toán Filter được mô tả ở bên trên. Trong khi đó Bảng V và Bảng VI được dùng để so sánh độ chính xác phân lớp trên hai mô hình SVM và KNN giữa các tập rút gọn thu được. Bảng VII được dùng để so sánh về thời gian tính toán tập rút gọn giữa các thuật toán.

Như chúng ta có thể quan sát thấy tại Bảng IV, thuật toán đề xuất F_IFT cho tập rút gọn có kích thước tốt hơn đáng kể so với các thuật toán khác. Thông qua kích thước trung bình (5.33) trên 12 tập dữ liệu, chúng ta có thể khẳng định thuật toán đề xuất F_IFT cho tập rút gọn có kích thước tốt hơn so với các thuật toán trên toàn bộ 12 tập dữ liệu thực nghiệm. Hơn nữa thời gian thực hiện của thuật toán đề xuất F_IFT cũng rất tốt, cụ thể với 12 tập dữ liệu thực nghiệm, thuật toán đề xuất F_IFT có thời gian thực hiện trung bình nhỏ nhất (1.25s). Điều này có thể lý giải được là do tốc độ hội tụ của thuật toán tỉ lệ thuận với kích thước của tập rút gọn thu được.

Tuy nhiên, độ chính xác phân lớp của tập rút gọn thu được từ thuật toán đề xuất F_IFT còn thấp hơn so với các thuật toán khác trên cả hai mô hình SVM và KNN được xử dụng, điều này có thể lý giải do số lượng thuộc tính của tập rút gọn khá nhỏ, ảnh hưởng tới chất lượng phân lớp của mô hình. Thông qua độ chính xác trung bình (0.7) với SVM và (0.71) với KNN, chúng ta có thể khẳng định việc đánh đổi giữa kích thước của tập rút gọn và độ chính xác phân lớp hoàn toàn có thể chấp nhận được vì độ chính xác phân lớp của mô hình vẫn có thể chấp nhận được trong một số bài toán trong thực tế.

VI. KẾT LUẬN

Trong bài báo này, chúng tôi đã sử dụng cấu trúc IFT để rút gọn thuộc tính cho bảng quyết định miền giá trị số. Trong đó, cấu trúc IF-subbase được sử dụng để xây dựng độ đo đánh giá thuộc tính và cấu trúc IF-base của IFT đơn vị được sử dụng để làm điều kiện dừng cho thuật toán. Kết quả thực nghiệm cho thấy phương pháp đề xuất là hiệu quả về thời gian tính toán và kích thước của tập rút gọn. Điều

Bảng III
TẬP RÚT GỌN THU ĐƯỢC TỪ THUẬT TOÁN F_IFT

ID	Data	Reduct sets
1	Wine	[7, 11, 9]
2	Heart	[0, 7, 9]
3	Wdbc	[21, 4]
4	Wpbc	[0, 1]
5	Iono	[18, 9, 3, 25, 30, 6, 7, 5, 2, 19, 23, 32, 12, 4, 27, 8]
6	UFDC	[3, 26, 10, 42, 6, 29]
7	Sona	[19, 30]
8	Libras	[66, 1, 71, 0, 88, 39, 89, 51, 58]
9	Musk	[31, 62, 162, 164, 89, 25, 115, 106]
10	LVB	[58, 51]
11	LVG	[79, 82, 56, 84, 47, 49]
12	PD	[0, 420, 421, 422, 423, 424, 426, 427, 430, 437, 502, 613]

Bảng IV
BẢNG SO SÁNH KÍCH THƯỚC GIỮA CÁC TẬP RÚT GỌN

ID	Data	F_IFT	[9]	[10]	[12]	X
1	Wine	3	10	13	11	13
2	Heart	3	13	13	13	13
3	Wdbc	2	16	21	7	30
4	Wpbc	2	14	22	18	33
5	Iono	16	16	29	17	34
6	UFDC	6	6	16	5	43
7	Sona	2	10	44	12	60
8	Libras	9	10	22	30	90
9	Musk	8	8	61	10	166
10	LVB	2	5	60	5	310
11	LVG	6	6	60	6	310
12	PD	5	23	90	10	754
Average		5.33	11.42	37.58	12.00	100.18

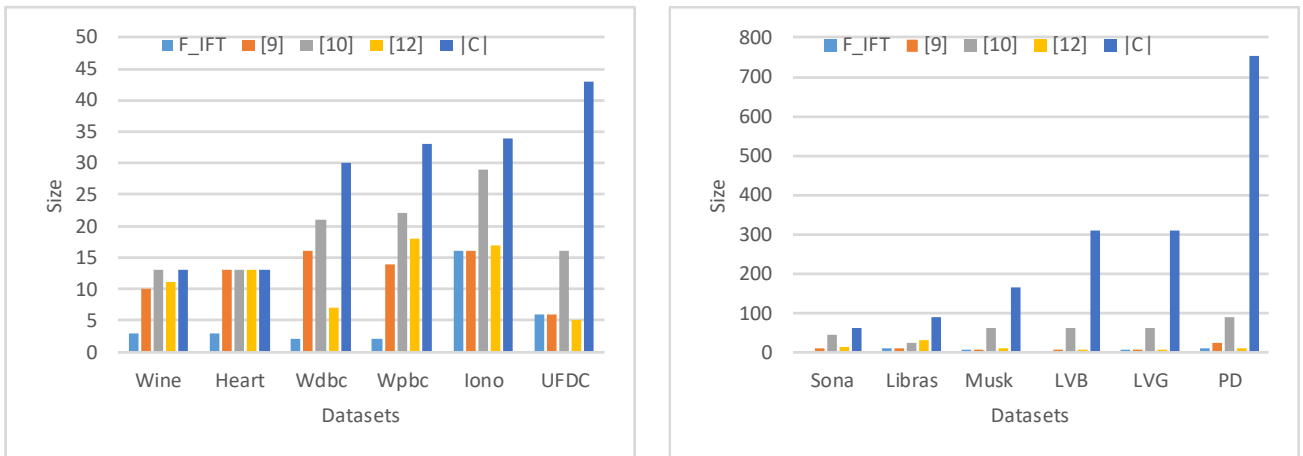
Bảng V
BẢNG SO SÁNH ĐỘ CHÍNH XÁC PHÂN LỚP TRÊN MÔ HÌNH SVM CỦA CÁC THUẬT TOÁN.

ID	Data	F_IFT	[9]	[10]	[12]	X
1	Wine	0.89	0.97	0.98	0.93	0.98
2	Heart	0.75	0.84	0.84	0.84	0.84
3	Wdbc	0.8	0.96	0.97	0.92	0.98
4	Wpbc	0.77	0.76	0.77	0.76	0.78
5	Iono	0.87	0.84	0.88	0.86	0.88
6	UFDC	0.36	0.35	0.55	0.43	0.43
7	Sona	0.64	0.64	0.67	0.72	0.65
8	Libras	0.57	0.59	0.62	0.65	0.71
9	Musk	0.66	0.63	0.72	0.72	0.75
10	LVB	0.67	0.67	0.86	0.68	0.83
11	LVG	0.67	0.67	0.85	0.82	0.89
12	PD	0.84	0.84	0.88	0.86	0.81
Average		0.70	0.72	0.79	0.76	0.79

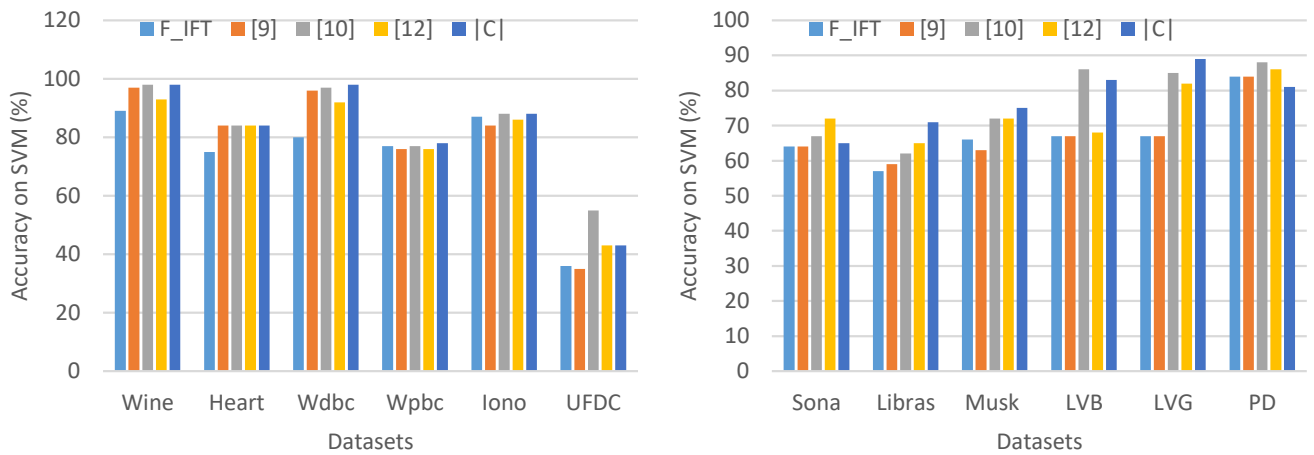
đó càng khẳng định phương pháp chọn điều kiện dừng cho thuật toán đề xuất F_IFT theo tiếp cận IFT là hoàn toàn phù hợp với cách định nghĩa tập rút gọn được phân tích trong Phần IV của bài báo.

Tuy nhiên độ chính xác phân lớp còn hạn chế so với các thuật toán khác, do đó chúng ta có thể cân nhắc việc đánh đổi độ chính xác phân lớp để có tập rút gọn có kích thước tốt và thời gian tính toán hiệu quả.

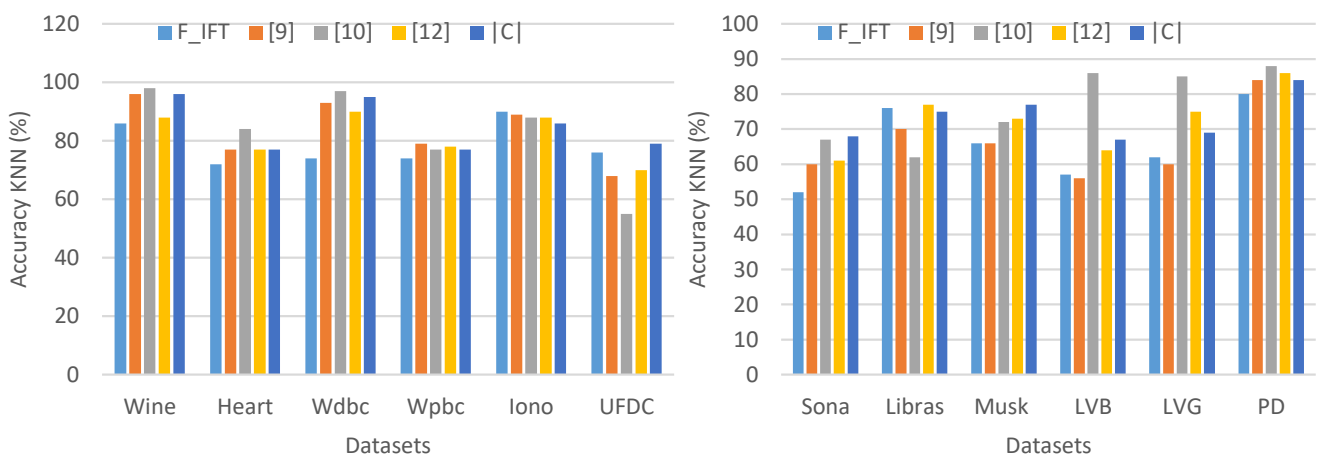
Hơn nữa, cách tiếp cận chọn lọc thuộc tính cho tập rút



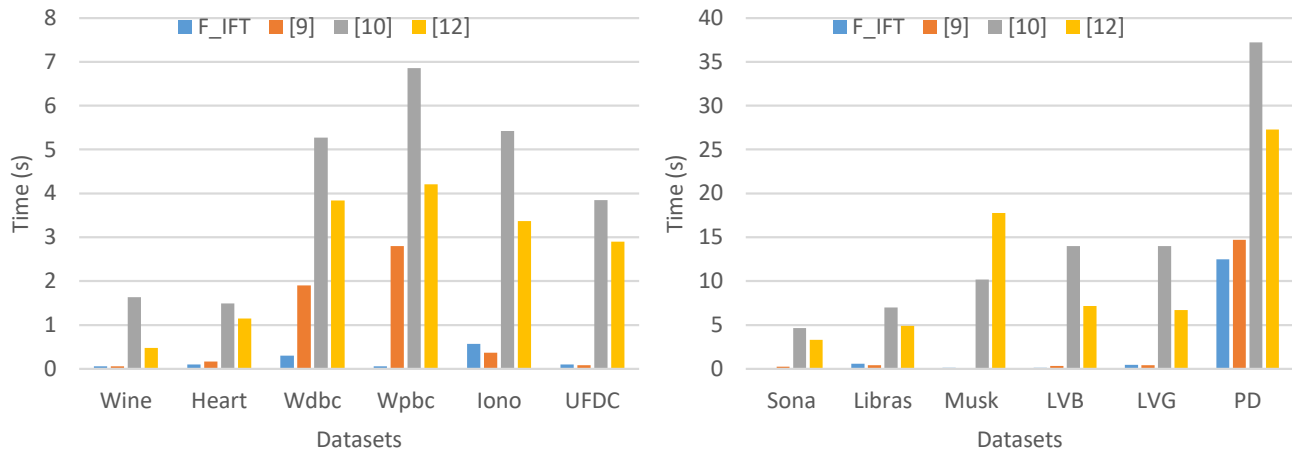
Hình 1. So sánh kích thước của các tập rút gọn thu được



Hình 2. So sánh độ chính xác phân lớp của các tập rút gọn trên mô hình SVM



Hình 3. So sánh độ chính xác phân lớp của các tập rút gọn trên mô hình KNN



Hình 4. So sánh thời gian tính toán tập rút gọn của các thuật toán

Bảng VI
BẢNG SO SÁNH ĐỘ CHÍNH XÁC PHÂN LỚP
TRÊN MÔ HÌNH KNN CỦA CÁC THUẬT TOÁN.

ID	Data	F_IFT	[9]	[10]	[12]	[X]
1	Wine	0.86	0.96	0.98	0.88	0.96
2	Heart	0.72	0.77	0.84	0.77	0.77
3	Wdbc	0.74	0.93	0.97	0.9	0.95
4	Wpbc	0.74	0.79	0.77	0.78	0.77
5	Iono	0.9	0.89	0.88	0.88	0.86
6	UFDC	0.76	0.68	0.55	0.7	0.79
7	Sona	0.52	0.6	0.67	0.61	0.68
8	Libras	0.76	0.7	0.62	0.77	0.75
9	Musk	0.66	0.66	0.72	0.73	0.77
10	LVB	0.57	0.56	0.86	0.64	0.67
11	LVG	0.62	0.6	0.85	0.75	0.69
12	PD	0.82	0.84	0.88	0.86	0.84
Average		0.71	0.74	0.79	0.76	0.79

Bảng VII
BẢNG SO SÁNH THỜI GIAN TÍNH TOÁN
TẬP RÚT GỌN GIỮA CÁC THUẬT TOÁN.

ID	Data	F_IFT	[9]	[10]	[12]
1	Wine	0.06	0.06	1.63	0.48
2	Heart	0.1	0.17	1.49	1.15
3	Wdbc	0.3	1.9	5.27	3.84
4	Wpbc	0.06	2.8	6.86	4.21
5	Iono	0.57	0.37	5.42	3.37
6	UFDC	0.1	0.08	3.85	2.9
7	Sona	0.07	0.24	4.63	3.31
8	Libras	0.57	0.43	7	4.9
9	Musk	0.12	0.07	10.2	17.77
10	LVB	0.12	0.34	14	7.15
11	LVG	0.45	0.4	14	6.68
12	PD	12.5	14.7	37.2	27.3
Average		1.25	1.80	9.30	6.92

gọn trong bài vẫn dừng lại ở tiếp cận độ đo, chỉ có điều kiện dừng của thuật toán là sử dụng tiếp cận IFT. Do đó, trong tương lai chúng tôi sẽ xây dựng phương pháp chọn lọc thuộc tính theo tiếp cận IFT và phát triển mô hình lai ghép filter-wrapper để tập rút gọn thu được tốt cả về kích thước và độ chính xác phân lớp.

TÀI LIỆU THAM KHẢO

- [1] Atanassov, K. T. (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*. An International Journal in Information Science and Engineering, 20(1), 87–96.
- [2] Yun, S. M., Eom, Y. S., & Lee, S. J. (2021). Topology of the redefined intuitionistic fuzzy rough sets. *International Journal of Fuzzy Logic and Intelligent Systems*, 21(4), 369–377.
- [3] Yun, S. M., Eom, Y. S., & Lee, S. J. (2021). Topology of the redefined intuitionistic fuzzy rough sets. *International Journal of Fuzzy Logic and Intelligent Systems*, 21(4), 369–377.
- [4] Bashir, Z., Abbas Malik, M. G., Asif, S., & Rashid, T. (2020). The topological properties of intuitionistic fuzzy rough sets. *Journal of Intelligent & Fuzzy Systems*, 38(1), 795–807.
- [5] Abo-Elhamayel, M., & Yang, Y. (2021). Generalizations of rough sets via topology. *Afrika Matematika*, 32(1–2), 41–50.
- [6] Rebecca Paul, N. (2019). Rough sets applicable in a topology. *Malaya Journal of Matematik*, S(1), 61–65.
- [7] Giang, N. L., Son, L. H., Ngan, T. T., Tuan, T. M., Phuong, H. T., Abdel-Basset, M., de Macedo, A. R. L., & de Albuquerque, V. H. C. (2020). Novel incremental algorithms for attribute reduction from dynamic decision tables using hybrid filter-wrapper with fuzzy partition distance. *IEEE Transactions on Fuzzy Systems: A Publication of the IEEE Neural Networks Council*, 28(5), 858–873.
- [8] UCI. machine learning repository (2021). Data. <https://archive.ics.uci.edu/ml/index.php>
- [9] Nguyen, T. T., Giang, N. L., Tran, D. T., Nguyen, T. T., Nguyen, H. Q., Pham, A. V., & Vu, T. D. (2021). A novel filter-wrapper algorithm on intuitionistic fuzzy set for attribute reduction from decision tables. *International journal of data warehousing and mining*, 17(4), 67–100.
- [10] Tan, A., Wu, W.-Z., Qian, Y., Liang, J., Chen, J., & Li, J. (2019). Intuitionistic fuzzy rough set-based granular structures and attribute subset selection. *IEEE Transactions on Fuzzy Systems: A Publication of the IEEE Neural Networks Council*, 27(3), 527–539.
- [11] Mishra, S., & Srivastava, R. (2018). Fuzzy topologies are generated by fuzzy relations. *Soft Computing*, 22(2), 373–385.
- [12] Tan, A., Shi, S., Wu, W.-Z., Li, J., & Pedrycz, W. (2022). Granularity and entropy of intuitionistic fuzzy information and their applications. *IEEE Transactions on Cybernetics*, 52(1), 192–204.

SƠ LƯỢC VỀ TÁC GIẢ



Trần Thanh Đại nhận bằng Thạc sỹ khoa học máy tính năm 2011 tại Học viện Kỹ thuật Quân Sự. Hiện là nghiên cứu sinh tại Đại học Khoa học Công nghệ - Viện Hàn Lâm Khoa học Công nghệ Việt Nam. Lĩnh vực nghiên cứu: Tính toán mềm, tính toán hạt, tính toán thô, khai phá dữ liệu và học máy.

Email: ttaiuneti@gmail.com



Hoàng Thị Minh Châu nhận bằng Thạc sỹ Bảo đảm toán học cho máy tính và hệ thống tính toán năm 2010 tại Đại học Khoa Học Tự Nhiên – ĐHQGHN. Hiện đang là giảng viên khoa Công nghệ thông tin - Đại học Kinh tế Kỹ thuật Công nghiệp. Lĩnh vực nghiên cứu: IOT, AI, Khai phá dữ liệu và học máy.

Email: htmchau@uneti.edu.vn



Nguyễn Long Giang nhận bằng Tiến Sỹ Toán học năm 2012, Phó Giáo sư tại Viện Công nghệ thông tin - Viện Hàn lâm khoa học Việt Nam năm 2017. Tổng biên tập tạp chí Tin học và điều khiển (JCC). Tham gia điều hành tại các hội thảo quốc tế như: IJCRS 2019, AICI 2019, MARR 2019, IEEE-RIVF 2019, SoICT 2019.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo (AI), tính toán mềm, tính toán mờ, khai phá dữ liệu:

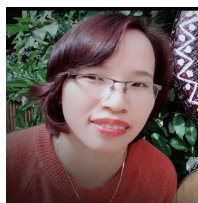
Email: nlgiang@ioit.ac.vn



Vũ Thu Uyên nhận bằng Thạc sỹ ngành công nghệ phần mềm năm 2009 tại Trường Đại học công nghệ - Đại học Quốc Gia. Hiện là giảng viên Khoa Công nghệ thông tin - Trường Đại học Kinh tế Kỹ thuật Công nghiệp.

Lĩnh vực nghiên cứu chính: Khai phá dữ liệu và học máy

Email: vtuyen@uneti.edu.vn



Trần Thị Ngân nhận bằng Đại học toán ứng dụng tại Trường Đại học Khoa học tự nhiên - Đại học Quốc Gia năm 2003, bằng Thạc sỹ khoa học máy tính năm 2007 tại Đại học Thái Nguyên và bằng Tiến sỹ năm 2014. Được phong Phó giáo sư năm 2021. Hiện đang công tác tại Khoa công nghệ thông tin - Trường Đại học Thủy Lợi.

Lĩnh vực nghiên cứu: Tính toán mờ, tối ưu, khai phá dữ liệu và học máy.

Email: ngantt@tlu.edu.vn



Vương Trung Hiếu nhận bằng Cử nhân ngành Kỹ thuật máy tính tại Đại học Osmania, Ấn Độ. Nghiên cứu sinh tại Học viện Khoa học Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Hiện đang là viên chức Phòng Tổ chức - Hành chính, Trường Đại học Công nghiệp Hà Nội.

Lĩnh vực nghiên cứu: Tính toán mềm, khai phá dữ liệu và học máy.

Email: vthieu@haui.edu.vn