

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC LẠC HỒNG



ĐỒ SĨ TRƯỜNG

**PHƯƠNG PHÁP LỰA CHỌN THUỘC TÍNH VÀ KỸ
THUẬT GOM CỤM DỮ LIỆU PHÂN LOẠI SỬ
DỤNG TẬP THÔ**

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Chuyên ngành: Khoa học máy tính

Mã số ngành: 9480101

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS.TS NGUYỄN THANH TÙNG

Đồng Nai, năm 2023

Công trình được hoàn thành tại: Trường Đại học Lạc Hồng

Người hướng dẫn khoa học:

PGS.TS. Nguyễn Thanh Tùng

Phản biện 1:

Phản biện 2:

Phản biện 3:.....

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp Trường họp tại

.....

.....

Vào hồi giờ ngày tháng năm

Có thể tìm hiểu luận án tại thư viện:

- Thư viện trường Đại học Lạc Hồng
- Thư viện Quốc Gia

MỤC LỤC

CHƯƠNG 1. MỞ ĐẦU.....	1
CHƯƠNG 2. KHÁI QUÁT VỀ LÝ THUYẾT TẬP THÔ VÀ ỨNG DỤNG TRONG KHAI PHÁ DỮ LIỆU	3
2.1 Các khái niệm cơ bản của lý thuyết tập thô	3
2.1.1 Hệ thông tin.....	3
2.1.2 Quan hệ không phân biệt được và các xấp xỉ của một tập hợp	4
2.1.3 Bảng quyết định.....	4
2.1.4 Các khái niệm lý thuyết thông tin liên quan.....	5
2.2 Khám phá tri thức từ cơ sở dữ liệu.....	7
2.2.1 Các kỹ thuật khai phá dữ liệu	7
2.3 Ứng dụng của lý thuyết tập thô trong khai phá dữ liệu	7
2.4 Kết luận chương 2.....	8
CHƯƠNG 3. LỰA CHỌN THUỘC TÍNH SỬ DỤNG LÝ THUYẾT TẬP THÔ	8
3.1 Khái quát về bài toán lựa chọn thuộc tính.....	8
3.1.1 Phương pháp tạo lập các tập con	8
3.1.2 Tiêu chuẩn đánh giá	9
3.2 Các phương pháp lựa chọn thuộc tính sử dụng lý thuyết tập thô....	10
3.2.1 Đề xuất thuật toán rút gọn thuộc tính dựa vào gom cụm ACBRC..	11
3.3 Kết luận chương 3	16

CHƯƠNG 4. GOM CỤM DỮ LIỆU SỬ DỤNG LÝ THUYẾT TẬP THÔ	16
4.1 Thuật toán MMNVI	18
4.1.1 Ý tưởng và những định nghĩa cơ bản	18
4.1.2 Thuật toán MMNVI	19
4.1.3 Độ phức tạp của thuật toán MMNVI.....	Error! Bookmark not defined.
4.1.4 Nhận xét thuật toán MMNVI	Error! Bookmark not defined.
4.1.5 Kết quả thực nghiệm thuật toán MMNVI.....	21
4.1.6 Bộ dữ liệu đánh giá	21
4.1.7 Phương pháp đánh giá hiệu suất	21
4.1.8 Kết quả gom cụm	21
4.2 Kết luận chương 4	22
CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	23
5.1 Những kết quả và đóng góp chính của luận án.....	23
5.2 Hướng phát triển của luận án	24

CHƯƠNG 1. MỞ ĐẦU

Khám phá tri thức từ CSDL là một lĩnh vực khoa học nhằm nghiên cứu để tạo ra những công cụ khai phá những thông tin, tri thức hữu ích, tiềm ẩn mang tính dự đoán trong các CSDL lớn [1, 2].

Các kết quả nghiên cứu cùng với những ứng dụng thành công thời gian qua cho thấy, khám phá tri thức từ CSDL là một lĩnh vực khoa học tiềm năng, mang lại nhiều lợi ích, đồng thời có ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống. Tuy nhiên, với tốc độ tăng trưởng của dữ liệu hiện nay cùng với việc xuất hiện các dạng dữ liệu phức tạp, việc nghiên cứu và ứng dụng các kỹ thuật khai phá dữ liệu cũng đang gặp nhiều khó khăn, thách thức, đòi hỏi các nhà nghiên cứu phải không ngừng nỗ lực nhằm tìm ra những công cụ để giải quyết các khó khăn, thách thức này.

Lý thuyết tập thô - do Zdzisaw Pawlak [3] đề được xem là công cụ hữu hiệu để giải quyết các bài toán xử lý thông tin có chứa dữ liệu mơ hồ, không chắc chắn. Do tư duy mới lạ, phương pháp độc đáo và dễ cài đặt, trong hơn ba mươi năm qua, lý thuyết tập thô đã được nghiên cứu, ứng dụng và trở thành một công cụ quan trọng trong lĩnh vực xử lý thông tin thông minh [2, 4, 5, 6, 7, 8].

Trong xu thế đó, nhiều nhóm nhà khoa học, trong đó có cả các nhà khoa học Việt nam, đã và đang quan tâm đến nghiên cứu vấn đề rút gọn thuộc tính trong bảng quyết định và gom cụm dữ liệu. Tuy nhiên, lĩnh vực nghiên cứu này vẫn còn một số vấn đề lớn cần được tiếp tục thảo luận và cải tiến. Với là lý do đó, nghiên cứu sinh chọn đề tài nghiên cứu: “Phương pháp lựa chọn thuộc tính và kỹ thuật gom cụm dữ liệu phân loại sử dụng lý thuyết tập thô”.

Mục tiêu nghiên cứu của luận án tập trung vào hai vấn đề của đề tài: (1) nghiên cứu phương pháp mới tìm tập rút gọn trong một bảng quyết định; (2) kỹ thuật gom cụm dữ liệu phân loại sử dụng tập thô.

Đối tượng nghiên cứu của luận án là các hệ thông tin, bảng quyết định có thể chứa dữ liệu mơ hồ, không chắc chắn.

Phạm vi nghiên cứu của luận án bao gồm việc nghiên cứu các phương pháp khai phá dữ liệu theo hướng tiếp cận tập thô, tập trung vào hai vấn đề chính nêu trong mục tiêu của luận án.

Phương pháp nghiên cứu các vấn đề nghiên cứu đặt ra được thực hiện bằng cách tổng hợp và đánh giá các kết quả nghiên cứu đã đạt được về lý thuyết tập thô trong khai phá dữ liệu từ các công trình đăng trên các tạp chí khoa học chuyên ngành uy tín trong và ngoài nước. Từ đó đề xuất các kỹ thuật, thuật toán mới, cài đặt, tính toán, so sánh và đánh giá kết quả thực nghiệm, chứng minh tính hiệu quả của các thuật toán.

Bố cục của luận án bao gồm chương mở đầu, ba chương nội dung chính, chương kết luận, các công trình nghiên cứu đã thực hiện và danh mục tài liệu tham khảo. Chương 2 trình bày các khái niệm cơ bản của lý thuyết tập thô cùng với một số khái niệm liên quan từ lý thuyết thông tin, khái quát về khai phá dữ liệu và tiềm năng ứng dụng lý thuyết tập thô trong khai phá dữ liệu. Chương 3 trình bày bài toán lựa chọn thuộc tính và một số thuật toán hiệu quả hiện có theo tiếp cận tập thô, khó khăn thách thức; trên cơ sở đó đề xuất thuật toán mới rút gọn thuộc tính sử dụng phương pháp gom cụm thuộc tính. Chương 4 trình bày bài toán gom cụm trong khai phá dữ liệu, một số phương pháp gom cụm hiệu quả hiện có; hạn chế của chúng và đề xuất thuật toán gom cụm dữ liệu phân loại sử dụng lý thuyết tập thô kết hợp các khái niệm entropy trong lý thuyết thông tin. Cuối cùng, chương kết luận nêu những đóng góp của luận án và các hướng phát triển.

Đóng góp chính của luận án được trình bày trong chương 3, chương 4.

Chương 3 đề xuất một thuật toán tìm toán tập rút gọn trong bảng quyết định bằng cách sử dụng phép gom cụm thuộc tính với tên gọi ACBRC (Attribute Clustering Based Reduct Computing – Tính toán tập rút gọn dựa vào gom cụm thuộc tính). Kết quả thực nghiệm cho thấy thuật toán đề xuất có khả năng tính toán tập rút gọn xấp xỉ có kích thước nhỏ và độ chính xác phân lớp cao so với các thuật toán đem so sánh, khi số cụm dùng để phân chia các thuộc tính được lựa chọn một cách thích hợp.

Chương 4 luận án đề xuất một thuật toán mới gom cụm dữ liệu phân loại với tên gọi MMNVI (Minimum Mean Normalized Variation of Information - Biến thể Thông tin Chuẩn hóa Trung bình Nhỏ nhất (MMNVI)). MMNVI thuộc loại phương pháp gom cụm phân cấp, phân phân đôi dần tập các đối tượng thành các cụm. Kết quả thử nghiệm trên các tập dữ liệu thực từ UCI cho thấy

thuật toán MMNVI có thể được sử dụng thành công trong việc phân cụm dữ liệu phân loại. Nó tạo ra kết quả phân cụm tốt hơn hoặc tương đương hơn với các thuật toán cơ bản đem so sánh.

Các đóng góp chính trên đây đã được đăng trong hai bài báo trên *Journal of Computer Science and Cybernetics*, năm 2022 [CT1] và năm 2023 [CT2]. Ngoài các đóng góp chính trình bày trong luận án, nghiên cứu sinh là đồng tác giả của có một số kết quả khác liên quan đến đề tài luận án, bao gồm một bài báo quốc tế và ba báo cáo hội thảo khoa học trong nước.

CHƯƠNG 2. KHÁI QUÁT VỀ LÝ THUYẾT TẬP THÔ VÀ ỨNG DỤNG TRONG KHAI PHÁ DỮ LIỆU

2.1 Mở đầu

Chương này trình bày khái quát về lý thuyết tập thô, quy trình khám phá tri thức từ cơ sở dữ liệu và khả năng ứng dụng của của lý thuyết về tập thô trong khai phá dữ liệu. Các vấn đề cơ bản trình bày trong chương này là cơ sở cho việc nghiên cứu đề xuất các phương pháp mới rút gọn thuộc tính, gom cụm dữ liệu phân loại trình bày các chương sau.

2.2 Các khái niệm cơ bản của lý thuyết tập thô

Lý thuyết tập thô – do Zdzisaw Pawlak [3] đề xuất vào những năm đầu thập niên tám mươi của thế kỷ hai mươi – được xem là công cụ hữu hiệu để giải quyết các bài toán chứa dữ liệu mơ hồ, không chắc chắn. Từ khi ra đời cho đến nay, lý thuyết tập thô được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau của khoa học máy tính như trí tuệ nhân tạo, hệ chuyên gia, hệ hỗ trợ quyết định, khám phá tri thức từ cơ sở dữ liệu, v.v. Mục này trình bày các khái niệm cơ bản của lý thuyết tập thô.

2.2.1 Hệ thông tin

Định nghĩa 2.1. [3] *Hệ thông tin là một bộ đôi $IS = (U, A)$, trong đó U là một tập hữu hạn, không rỗng các đối tượng, A là một tập hữu hạn, không rỗng các thuộc tính, mỗi $a \in A$ là một ánh xạ $a : U \rightarrow V_a$, trong đó V_a ký hiệu miền giá trị của a .*

2.2.2 Quan hệ không phân biệt được và các xấp xỉ của một tập hợp

Định nghĩa 2.2. [3] Cho hệ thông tin là một bộ tứ $IS = (U, A)$. Mỗi tập con các thuộc tính $B \subseteq A$ xác định một quan hệ, ký hiệu là $IND(B)$, gọi là quan hệ không phân biệt được, như sau:

$$IND(B) = \{(u, v) \in U^2: a(u) = a(v) \forall a \in B\} \quad (2.1)$$

Nếu hai đối tượng $(u, v) \in IND(B)$ thì hai đối tượng này sẽ không phân biệt được bởi các thuộc tính thuộc tập B .

Định nghĩa 2.3. [3] Cho hệ thông tin là một bộ tứ $IS = (U, A, V, f)$, $B \subseteq A$ và $X \subseteq U$, B -xấp xỉ dưới của X , ký hiệu là $\underline{B}(X)$, và B -xấp xỉ trên của X , ký hiệu là $\overline{B}(X)$, được định nghĩa tương ứng như sau:

$$\underline{B}X = \{u \in U: [u]_B \subseteq X\} \quad (2.2)$$

$$\overline{B}X = \{u \in U: [u]_B \cap X \neq \emptyset\} \quad (2.3)$$

Định nghĩa 2.4. [3] Cho hệ thông tin $IS = (U, A)$, $B \subseteq A$ và $X \subseteq U$. Độ chính xác của xấp xỉ X thông qua B được định nghĩa bởi

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} \quad (2.4)$$

Trong suốt luận án này, $|X|$ ký hiệu số phần tử của tập X .

Định nghĩa 2.5. [3] Cho hệ thông tin $IS = (U, A)$, $B \subseteq A$ và $X \subseteq U$. Độ thô (roughness) của X đối với B được định nghĩa là

$$R_B(X) = \alpha_B(X) = 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \quad (2.5)$$

2.2.3 Bảng quyết định

Định nghĩa 2.6. [3, 6] Bảng quyết định là một hệ thông tin dạng $DT = (U, C \cup \{d\})$, trong đó $d \notin C$ là một thuộc tính riêng biệt được gọi là thuộc tính quyết định. Các thuộc tính trong C được gọi là các thuộc tính điều kiện.

Định nghĩa 2.7. [3, 6] Cho $DT = (U, C \cup \{d\})$ là một bảng quyết định và tập con thuộc tính điều kiện $B \subseteq C$. Vùng dương của d đối với B , ký hiệu là $POS_B(d)$, được xác định như sau

$$POS_B(d) = \bigcup_{X \in U/d} (\underline{B}X) \quad (2.6)$$

Định nghĩa 2.8. [3, 6] Cho bảng quyết định $DT = (U, C \cup \{d\})$. Với tập con $B \subseteq C$, độ phụ thuộc $\gamma_B(d)$ của d vào B được định nghĩa như sau:

$$\gamma_B(d) = \frac{|POS_B(d)|}{|U|}. \quad (2.7)$$

Rõ ràng, $0 \leq \gamma_B(d) \leq 1$. Nếu $\gamma_B(d) = 1$, thì ta nói rằng d phụ thuộc hoàn toàn vào B , còn nếu $0 < \gamma_B(d) < 1$, thì d phụ thuộc vào B với mức độ $\gamma_B(d)$. Khi $\gamma_B(d) = 0$, ta nói rằng d không phụ thuộc vào B .

2.3 Các khái niệm lý thuyết thông tin liên quan

Cho $IS = (U, A)$ là một hệ thống thông tin, thuộc tính $a \in A$. Hệ thống thông tin IS có thể được xem như một quần thể thống kê và a là một biến ngẫu nhiên rời rạc. Giả sử $V_a = \{x_1, x_2, \dots, x_m\}$, $U/IND(a) = \{X_1, X_2, \dots, X_m\}$. Khi đó, phân phối xác suất của a có thể được xác định bởi:

$$P(a = x_i) = P(x_i) = |X_i|/|U|, \quad i = 1, \dots, m. \quad (2.8)$$

Định nghĩa 2.9. [9] Cho hệ thông tin $IS = (U, A)$ và thuộc tính $a \in A$. Shannon entropy (gọi tắt là entropy) của a là một đại lượng $H(a)$ xác định theo công thức sau:

$$H(a) = - \sum_{i=1}^m P(a = x_i) \log_2 P(a = x_i). \quad (2.9)$$

với quy ước $0 \times \log_2 0 = 0$.

Định nghĩa 2.10. [9] Cho hệ thông tin $IS = (U, A)$ và hai thuộc tính $a, b \in A$. Entropy đồng thời của a và b là một đại lượng $H(a, b)$ xác định theo công thức sau:

$$H(a, b) = - \sum_{i=1}^m \sum_{j=1}^n P(a = x_i, b = y_j) \log_2 P(a = x_i, b = y_j). \quad (2.10)$$

Entropy $H(a, b)$ biểu thị mức độ không chắc chắn của hai thuộc tính a và b .

Định nghĩa 2.11. [9] Cho hệ thông tin $IS = (U, A)$ và hai thuộc tính $a, b \in A$. Entropy có điều kiện của a khi đã biết b là đại lượng $H(a|b)$ xác định bởi:

$$H(a|b) = - \sum_{j=1}^n P(b = y_j) \sum_{i=1}^m P(a = x_i | b = y_j) \log_2 P(a = x_i | b = y_j). \quad (2.11)$$

$H(a|b)$ xác định lượng entropy (tức là độ không chắc chắn) còn lại của thuộc tính a khi đã biết giá trị của một thuộc tính b . Áp dụng các công thức (2.9), (2.10) và (2.11) ta có:

$$H(a|b) = H(a, b) - H(b) \quad (2.12)$$

Định nghĩa 2.12. [9] Cho hệ thông tin $IS = (U, A)$ và hai thuộc tính $a, b \in A$. Thông tin tương hỗ giữa hai thuộc tính a và b được định nghĩa:

$$I(a; b) = H(a) - H(a|b) = H(b) - H(b|a) \quad (2.13)$$

Định nghĩa 2.13. [9, 10] Cho hệ thông tin $IS = (U, A)$ và hai thuộc tính $a, b \in A$. Biến thể thông tin chuẩn hóa $NVI(a, b)$ giữa a và b được xác định như sau:

$$NVI(a, b) = 1 - \frac{I(a; b)}{H(a, b)} = \frac{H(a|b) + H(b|a)}{H(a, b)}. \quad (2.14)$$

Định lý 2.1. [10] $NVI(a, b)$ là một metric trên không gian của các thuộc tính, nghĩa là đối với mọi $a, b, c \in A$, ta đều có:

- (i) $NVI(a, b) \geq 0$ và đẳng thức xảy ra khi và chỉ khi $a = b$,
- (ii) $NVI(a, b) = NVI(b, a)$,
- (iii) $NVI(a, b) + NVI(b, c) \geq NVI(a, c)$.

2.4 Khám phá tri thức từ cơ sở dữ liệu

Khám phá tri thức từ CSDL (Knowledge Discovery in Databases – KDD) là một lĩnh vực khoa học nhằm nghiên cứu để tạo ra những công cụ khai phá những thông tin, tri thức hữu ích, tiềm ẩn mang tính dự đoán trong các CSDL lớn [1].

Một quá trình chuẩn khám phá tri thức từ CSDL bao gồm các công đoạn sau [1]: Lựa chọn dữ liệu; Tiền xử lý dữ liệu; Rút gọn dữ liệu; Khai phá dữ liệu; Đánh giá và diễn giải tri thức. Trong đó khai phá dữ liệu là bài toán quan trọng nhất. Các kỹ thuật khai phá dữ liệu được chia thành 2 nhóm chính :

- *Nhóm mô tả dữ liệu*: có nhiệm vụ mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong cơ sở dữ liệu hiện thời.
- *Nhóm phân lớp và dự đoán*: nhằm đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời, gồm có các kỹ thuật: phân lớp (Classification), hồi quy (Regression).

Trong đó, gom cụm dữ liệu (thuộc nhóm thứ nhất) và phân lớp (thuộc nhóm thứ hai) là hai kỹ thuật quan trọng, được sử dụng nhiều nhất trong công đoạn khai phá dữ liệu.

2.5 Ứng dụng của lý thuyết tập thô trong khai phá dữ liệu

Lý thuyết tập thô có thể được ứng dụng vào hầu hết các công đoạn của quá trình khám phá tri thức từ dữ liệu. Dưới đây là một số ứng dụng cụ thể của lý thuyết tập thô trong quá trình khám phá tri thức từ cơ sở dữ liệu [5, 4, 6, 7, 11].

(1) Tiền xử lý dữ liệu. Với giả thiết mô hình tối thiểu, lý thuyết tập thô được sử dụng để rút gọn và làm sạch dữ liệu cho các phân tích tiếp theo.

(2) Khai phá dữ liệu. Trong công đoạn khai phá dữ liệu, lý thuyết tập thô có thể được sử dụng giải quyết các vấn đề sau [5, 4, 6, 7, 11]: phân lớp dữ liệu; gom cụm dữ liệu; phát hiện luật kết hợp.

Có thể nói lý thuyết tập thô là công cụ hữu hiệu cho quá trình khám phá tri thức từ cơ sở dữ liệu. Tuy vậy, các kết quả nghiên lý thuyết và ứng dụng đến nay vẫn còn những hạn chế.

2.6 Kết luận chương 2

Nội dung chương 2 bao gồm 3 phần chính: khái quát về lý thuyết về tập thô với các khái niệm liên quan, quy trình khám phá tri thức từ cơ sở dữ liệu với các kỹ thuật khai phá dữ liệu cơ bản và ứng dụng của của lý thuyết về tập thô trong khai phá dữ liệu.

Các khái niệm cơ bản trình bày trong chương này là cơ sở để nghiên cứu đề xuất các phương pháp mới tìm tập rút gọn trong một bảng quyết định và gom cụm dữ liệu phân loại sử dụng tập thô, trình bày ở các chương sau.

CHƯƠNG 3. LỰA CHỌN THUỘC TÍNH SỬ DỤNG LÝ THUYẾT TẬP THÔ

3.1 Mở đầu

Chương này trình bày khái quát về vấn đề lựa chọn thuộc tính, các phương pháp chính tìm tập rút gọn của một bảng quyết định và đề xuất một thuật toán mới, với tên gọi ACBRC, dựa trên gom cụm các thuộc tính.

3.2 Khái quát về bài toán lựa chọn thuộc tính

Lựa chọn thuộc tính là quá trình chọn ra một tập hợp con thuộc tính từ tập hợp các thuộc tính ban đầu, với mục tiêu loại bỏ càng nhiều càng tốt các thuộc tính không liên quan và dư thừa nhằm cải thiện chất lượng dữ liệu và giảm độ phức tạp về thời gian và không gian cho việc phân tích.

Nhìn chung, một thuật toán lựa chọn thuộc tính thường bao gồm bốn bước cơ bản sau [1, 12, 13]: Tạo lập tập con để đánh giá; Đánh giá tập con; Kiểm tra điều kiện dừng; Kiểm chứng kết quả.

3.2.1 Phương pháp tạo lập các tập con

Tạo lập tập con thuộc tính là quá trình tìm kiếm liên tiếp nhằm tạo ra các tập con để tiến hành đánh giá và lựa chọn. Quy trình này bao gồm việc chọn điểm xuất phát, chọn hướng tìm kiếm và chiến lược tìm kiếm tập con.

Mỗi tập con sinh ra bởi một thủ tục sẽ được đánh giá theo một tiêu chuẩn nhất định và đem so sánh với tập con tốt nhất trước đó. Nếu tập con này tốt hơn, nó sẽ thay thế tập cũ. Quá trình tìm kiếm tập con thuộc tính tối ưu sẽ dừng khi một trong bốn điều kiện sau xảy ra [12, 13]:

- Đã thu được số thuộc tính quy định;
- Số bước lặp quy định cho quá trình lựa chọn đã hết;
- Việc thêm vào hay loại bớt một thuộc tính nào đó không cho một tập con tốt hơn;
- Đã thu được tập con tối ưu theo tiêu chuẩn đánh giá.

Tập con tốt nhất cuối cùng phải được kiểm chứng thông qua việc tiến hành các phép kiểm định, so sánh các kết quả khai phá với tập thuộc tính “tốt nhất” này và tập thuộc tính ban đầu trên các tập dữ liệu thực hoặc nhân tạo khác nhau.

Một vấn đề quan trọng khác trong lựa chọn thuộc tính là xác định cách thức đánh mức độ phù hợp của mỗi tập con.

3.2.2 Tiêu chuẩn đánh giá

Để đánh giá một tập con thuộc tính được chọn là tối ưu phải dựa trên một tiêu chuẩn đánh giá nhất định, một tập con là tối ưu theo tiêu chuẩn này chưa chắc sẽ tối ưu theo tiêu chuẩn khác. Các tiêu chuẩn đánh giá có thể phân thành hai loại: tiêu chuẩn độc lập và tiêu chuẩn phụ thuộc [12, 13].

Tiêu chuẩn độc lập (thường được dùng trong cách tiếp cận filter) đánh giá mức độ phù hợp của một hay một tập con thuộc tính một cách độc lập, không thông qua áp dụng một thuật học. Các tiêu chuẩn độc lập thường được sử dụng để đánh giá các tập con thuộc tính để lựa chọn là: số đo khoảng cách, số đo lượng thông tin thu thêm, số đo độ phụ thuộc, số đo độ nhất quán và số đo độ tương tự.

Tiêu chuẩn phụ thuộc (thường được dùng trong cách tiếp cận wrapper) đánh giá một tập con thuộc tính thông qua độ hiệu quả của một thuật học áp dụng trên chính tập thuộc tính cần đánh giá. Kết quả tập con thuộc tính được chọn dựa trên tiêu chuẩn này có khả năng dự báo cao tuy nhiên điểm hạn chế là nó sẽ mất nhiều thời gian tính toán.

3.3 Các phương pháp lựa chọn thuộc tính sử dụng lý thuyết tập thô

Trong cộng đồng tập thô, các thuật toán lựa chọn thuộc tính được thực hiện bằng việc tìm kiếm các rút gọn (reducts) của tập các thuộc tính, nghĩa là tìm cách rút gọn tối đa tập các thuộc tính ban đầu mà vẫn đảm bảo được những thông tin cần thiết đối với nhiệm vụ khai phá dữ liệu. Thật không may, việc tìm kiếm tất cả các tập rút gọn là không thể thực hiện được trong hầu hết các trường hợp vì với tập dữ liệu có n thuộc tính sẽ có $2^n - 1$ tập hợp con, khi n tăng số tập con thuộc tính sẽ tăng theo cấp số nhân. Tìm kiếm tất cả các tập rút gọn chỉ có thể được khi n tương đối nhỏ.

Tuy nhiên, trong ứng dụng thực tiễn thường không đòi hỏi tìm tất cả các tập rút gọn mà chỉ cần tìm một tập rút gọn tốt nhất theo một nghĩa nào đó là đủ. Vì vậy, trong những năm qua nhiều thuật toán heuristic tìm một tập rút gọn xấp xỉ đã được các nhà nghiên cứu đề xuất. Các thuật toán này nhằm giảm khối lượng tính toán, nhờ đó có thể áp dụng đối với các tập dữ liệu lớn. Với cách tiếp cận này, các khái niệm của lý thuyết tập thô được sử dụng để xác một tiêu chuẩn đánh giá mức độ cần thiết hay quan trọng của các thuộc tính, sau đó chuẩn đánh giá này được sử dụng như là các hàm heuristic định hướng cho quá trình lựa chọn thuộc tính trong các thuật toán.

Các thuật toán heuristic lựa chọn thuộc tính tiêu biểu bao gồm: thuật toán tìm tất cả các tập rút gọn sử dụng ma trận không phân biệt, một số thuật toán heuristic tìm tập rút gọn xấp xỉ của bảng quyết định: phương pháp dựa trên hàm đo độ phụ thuộc, phương pháp sử dụng các phép toán trong đại số quan hệ, phương pháp sử dụng entropy thông tin. Các thuật toán heuristic có độ phức tạp tính toán theo thời gian là đa thức, và do đó có thể áp dụng được trên bảng dữ liệu với kích thước lớn với độ phức tạp thời gian hợp lý.

Có thể thấy, các thuật toán trên có thể loại bỏ thành công các đặc trưng không liên quan nhưng có thể không thể loại bỏ các thuộc tính dư thừa [14, 15, 16, 17, 18]. Để khắc phục vấn đề này, một số phương pháp lựa chọn thuộc tính sử dụng gom cụm đã được đề xuất [16, 17, 18, 19, 20, 21].

Các thuật toán lựa chọn thuộc tính dựa trên gom cụm xuất phát từ một ý tưởng đơn giản. Nó chia không gian thuộc tính ban đầu thành một họ các cụm.

Nhìn chung, các hàm đo độ tương quan thường được sử dụng để đánh giá mức độ tương tự giữa hai thuộc tính trong quá trình gom cụm. Điều này làm cho các thuộc tính trong cùng một cụm là tương quan với nhau, dẫn đến việc có thể chỉ cần lựa chọn một thuộc tính để đại diện cho mỗi cụm, các thuộc tính còn lại được coi là dư thừa. Tập hợp các thuộc tính đại diện là tập các thuộc tính có liên quan, không dư thừa và được lấy làm một tập rút gọn.

Đối với cách tiếp cận bài toán lựa chọn thuộc tính dựa trên gom cụm này, có hai vấn đề cốt lõi cần được xem xét kỹ lưỡng, đó là việc lựa chọn hàm đo độ tương tự và phương pháp gom cụm để sử dụng. Hàm tương tự đo mức độ giống nhau giữa hai thuộc tính trong không gian thuộc tính; phương pháp gom cụm phân chia các thuộc tính thành các cụm bằng cách sử dụng hàm đo độ tương tự đã chọn.

Trong [16, 21] Hong và cộng sự đề xuất thuật toán lựa chọn thuộc tính dựa trên gom cụm có tên gọi là MNF (Most Neighbors First). MNF sử dụng độ phụ thuộc tương đối của một thuộc tính vào một thuộc tính khác để xây dựng thước đo độ tương tự cho một cặp thuộc tính.

Thuật toán MNF về cơ bản đã đạt được hiệu quả trong việc giảm số thuộc tính của trong một hệ thống tin. Tuy nhiên, vì hàm đo độ tương đồng (khoảng cách) giữa 2 thuộc tính $d(a_i, a_j)$ trong MNF không phải là một metric, nên sự hội tụ của thuật toán MNF có thể không được đảm bảo. Đối với nhiều tập dữ liệu, MNF có thể phải lặp lại nhiều lần các bước thực hiện mới có thể tìm ra kết quả gom cụm [14].

Mặc dù các thuật toán rút gọn thuộc tính dựa trên gom cụm thuộc tính đã nhận được nhiều sự chú ý trong thời gian gần đây, tuy nhiên số lượng các nghiên cứu vẫn còn nhiều hạn chế. Trong phần tiếp theo, luận án đề xuất một phương pháp mới rút gọn thuộc tính trong bảng quyết định dựa vào gom cụm, với tên gọi ACBRC.

3.4 Đề xuất thuật toán rút gọn thuộc tính dựa vào gom cụm ACBRC

Thuật toán ACBRC (Attribute Clustering Based Reduct Computing – Tính toán tập rút gọn dựa trên gom cụm thuộc tính) hoạt động trong ba giai

đoạn chính. Trong giai đoạn đầu, các thuộc tính không liên quan sẽ bị loại bỏ. Trong giai đoạn thứ hai, các thuộc tính có liên quan được phân chia thành một số cụm thích hợp bằng phương pháp gom cụm phân hoạch xung quanh medoids (Partitioning Around Medoids - PAM) kết hợp với một metric đặc biệt trong không gian thuộc tính là biến thể thông tin chuẩn hóa (Normalized Variation of Information). Trong giai đoạn thứ ba, một thuộc tính đại diện từ mỗi cụm được chọn là thuộc tính có độ liên quan lớn nhất với thuộc tính quyết định. Các thuộc tính được lựa chọn tạo thành một tập rút gọn xấp xỉ.

Để mô tả chính xác hơn về thuật toán, trước tiên phần này trình bày một số khái niệm liên quan:

Định nghĩa 3.1. [1, 22] Cho $DT = (U, C \cup \{d\})$ là một bảng quyết định. Gom cụm thuộc tính trong DT là phép phân chia tập C các thuộc tính có điều kiện thành một họ gồm các tập con $C_X = \{C_1, C_2, \dots, C_k\}$ của C , sao cho $C_1 \cup C_2 \cup \dots \cup C_k = C$, $C_i \neq \emptyset$, và $C_i \cap C_j = \emptyset$, với $i \neq j$.

Định nghĩa 3.2. [1, 22] Mức độ không liên quan giữa thuộc tính điều kiện $a_i \in C$ và thuộc tính quyết định d được đo bằng giá trị khoảng cách $NVI(a_i, d)$, trong đó $NVI(a_i, a_j)$ là độ đo biến thể thông tin chuẩn hóa giữa thuộc tính a_i với thuộc tính a_j đã được định nghĩa tại 2.16 Chương 2. Giá trị khoảng cách $NVI(a_i, d)$ càng lớn thì mức độ không liên quan giữa chúng càng nhỏ.

Định nghĩa 3.3 Cho G là một cụm thuộc tính. Thuộc tính $a^R \in G$ là thuộc tính đại diện của cụm nếu và chỉ nếu:

$$a^R = \operatorname{argmin}_{a \in G} NVI(a, d). \quad (3.1)$$

Công thức (3.1) nói lên a^R là thuộc tính có liên quan mạnh nhất và có thể được xem như một thuộc tính có liên quan tiêu biểu cho tất cả các thuộc tính trong cụm G .

Vì giá trị lớn nhất mà độ đo khoảng cách NVI có thể đạt được là 1, trong thuật toán đề xuất ACBRC, ta có thể quy định nếu $NVI(X_i, d)$ lớn hơn giá trị ngưỡng $\delta = 0.98$, thì có thể nói a_i là thuộc tính không liên quan với d ; trường hợp ngược lại, a_i là thuộc tính liên quan.

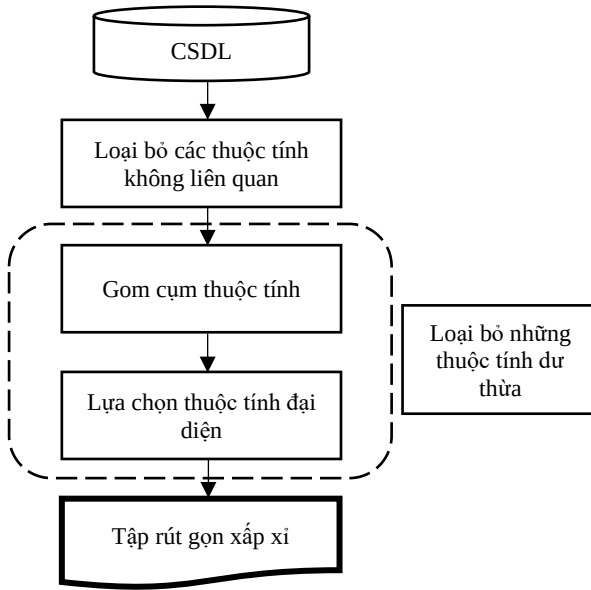
Sử dụng các định nghĩa trên, thuật toán ACBRC là quá trình bao gồm ba công đoạn liên tiếp, với khung hoạt động được thể hiện trong hình 3.1.

Các công đoạn là như sau:

(1) Đầu tiên, loại bỏ các thuộc tính không liên quan. Để thực hiện việc này, khoảng cách $NVI(a, d)$ được tính toán giữa mỗi thuộc tính điều kiện a và thuộc tính quyết định d . Thuộc tính có mức độ không liên quan lớn hơn 0,98 sẽ bị loại bỏ khỏi tập thuộc tính điều kiện ban đầu.

(2) Gom cụm các thuộc tính có liên quan bằng cách sử dụng hàm $pamk()$ trong gói R “fpc” của hệ thống R, sử dụng độ đo khoảng cách NVI . $pamk()$ là phiên bản nâng cao của $pam()$, có thể hoạt động với bất kỳ ma trận khoảng cách nào và không yêu cầu người dùng phải cung cấp số cụm k . Thay vào đó, nó thực hiện thuật toán gom cụm PAM với số cụm được ước tính bằng phương pháp chiều rộng hình bóng trung bình tối ưu, được mô tả trong [23].

(3) Cuối cùng, từ mỗi cụm thuộc tính đã được gom chọn một thuộc tính có mức độ liên quan với thuộc tính quyết định mạnh nhất. Thuộc tính này được xem như là thuộc tính đại diện cho tất cả các thuộc tính trong cụm. Khi một thuộc tính được chọn làm đại diện, các thuộc tính khác trong cùng một cụm sẽ bị loại bỏ. Tập các thuộc tính đại diện chọn được từ các cụm tạo thành tập rút gọn xấp xỉ.



Hình 3.1 Hình minh họa thuật toán ACBRC

Tựa code của thuật toán ACBRC là như sau:

Thuật toán 3.1 Thuật toán ACBRC

Đầu vào: Bảng quyết định $DT = (U, C \cup \{d\})$, ngưỡng không liên quan $\delta = 0.98$.

Đầu ra: Tập rút gọn xấp xỉ Red .

Thuật toán:

// **Loại bỏ các thuộc tính không liên quan.**

(1) Với mỗi $a \in C$

(2) Gán $irrelevance = NVI(a, d)$.

(3) Nếu $irrelevance > \delta$ thì $C^R = C \setminus \{a\}$.

(4) Hết với mỗi

// **Tính ma trận khoảng cách NVI giữa các cặp thuộc tính.**

(5) Với mọi $(a_i, a_j) \in C^R$ tính

(6) $NVI[i, j] = NVI(a_i, a_j)$

(7) Hết với mọi

//Gom cụm thuộc tính

(8) Sử dụng hàm *pamk()* trong gói R “*fpc*” để gom cụm các thuộc tính trong C^R .

(9) Với mỗi cụm G

(10) $a^R = \operatorname{argmin}_{a \in G} \operatorname{NVI}(X, d). \operatorname{Red} = \operatorname{Red} \cup a^R;$

(11) Hết với mỗi

Kết quả thực nghiệm thuật toán ACBRC

Thuật toán lựa chọn thuộc tính đề xuất ACBRC được cài đặt bằng ngôn ngữ lập trình R trên máy tính cá nhân. Tính toán thực nghiệm được thực hiện trên năm bộ dữ liệu chuẩn thu được từ kho dữ liệu chuẩn UCI.

Để đánh giá hiệu quả của thuật toán lựa chọn thuộc tính ACBRC, luận án tiến hành so sánh nó với thuật toán QuickReduct và CEBARKNC về số lượng thuộc tính được chọn, thời gian thực hiện và hiệu quả phân lớp của tập thuộc tính rút gọn được chọn. Để so sánh hiệu quả phân lớp của ACBRC, QuickReduct và CEBARKNC, luận án sử dụng C5.0 và Native Bayes, là hai thuật toán phân lớp phổ biến và được áp dụng rộng rãi trong nhiều lĩnh vực nghiên cứu khác nhau.

Để sử dụng tốt nhất dữ liệu và thu được kết quả ổn định, chiến lược xác nhận chéo 3×10 được sử dụng. Nghĩa là, đối với mỗi tập dữ liệu, mỗi thuật toán rút gọn thuộc tính và mỗi thuật toán phân lớp, xác nhận chéo 10 lần được lặp lại 3 vòng. Tại mỗi vòng xác nhận chéo 10 lần thứ tự của các đối tượng trong tập dữ liệu được ngẫu nhiên hóa. Việc sắp xếp thứ tự ngẫu nhiên của các đối tượng nhằm giảm bớt tác động của thứ tự đến kết quả phân lớp. Với mỗi thuật toán phân lớp và mỗi tập dữ liệu luận án ghi lại độ chính xác phân lớp thu được trong mỗi lần tính toán với các tập thuộc tính rút gọn cung cấp bởi ba thuật toán ACBRC, QuickReduct và CEBARKNC. Tính trung bình, ta thu được độ chính xác trung bình của từng thuật toán phân lớp theo từng thuật toán rút gọn thuộc tính trên mỗi tập dữ liệu.

Kết quả thực nghiệm thu được là như sau:

A. So sánh số lượng thuộc tính được chọn bởi ba thuật toán rút gọn thuộc tính

Kết quả thực nghiệm cho thấy cả ba thuật toán đều rút gọn đáng kể số thuộc tính ban đầu. Tuy nhiên, số thuộc tính chọn được bởi thuật toán ACBRC nó chung là nhỏ hơn.

Về thời gian thực hiện của thuật toán ACBRC lớn hơn một chút so với thuật toán QuickReduct và CEBARKNC. Tuy nhiên, thời gian thực hiện của ACBRC là chấp nhận được.

B. Đánh giá hiệu suất phân lớp của thuật toán rút gọn thuộc tính ACBRC

Đối với tập dữ liệu “Soybean”, “Lung” và “Votes”, độ chính xác phân lớp C5.0 với các thuộc tính chọn được bởi ACBRC cao hơn độ chính xác phân lớp với các thuộc tính chọn được bởi QuickReduct và CEBARKNC. Và đối với hai tập dữ liệu khác, kết quả phân lớp sau khi rút gọn thuộc tính bằng ACBRC có thể sánh với kết quả sau khi rút gọn thuộc tính bằng QuickReduct và CEBARKNC.

Đặc biệt, độ chính xác phân lớp Bayes với các thuộc tính được chọn bởi ACBRC lớn hơn độ chính xác phân lớp Bayes với các thuộc tính được chọn bởi QuickReduct và CEBARKNC.

3.5 Kết luận chương 3

Kết quả thử nghiệm cho thấy thuật toán đề xuất ACBRC là rất khả quan trong việc làm giảm số thuộc tính trong các tập dữ liệu cần khai phá các quy tắc phân lớp.

CHƯƠNG 4. GOM CỤM DỮ LIỆU SỬ DỤNG LÝ THUYẾT TẬP THÔ

4.1 Mở đầu

Gom cụm là một kỹ thuật trong khai phá dữ liệu, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tương tự nhau, đáng quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho hỗ trợ cho việc ra quyết định [1].

Cho đến nay đã có nhiều phương pháp gom cụm đã được đề xuất. Tuy nhiên, hầu hết các phương pháp gom cụm đều tập trung vào các tập dữ liệu số, trong đó mỗi thuộc tính mô tả các đối tượng đều có miền giá trị là một khoảng giá trị thực liên tục, mỗi đối tượng dữ liệu số được coi là một điểm trong không gian metric đa chiều với một metric đo khoảng cách giữa các đối tượng, chẳng hạn như metric Euclide hoặc metric Mahalanobis [1, 24, 25]. Tuy nhiên, trong các ứng dụng thực tiễn thường gặp phải những tập dữ liệu với các thuộc tính là những thuộc tính phân loại hay phạm trù (categorical), tức là những thuộc tính có miền giá trị D hữu hạn và không có thứ tự; (chẳng hạn như màu tóc, quốc tịch v.v.); trong D chỉ được phép so sánh giữa các giá trị, với bất kỳ $a, b \in D$ hoặc $a = b$ hoặc $a \neq b$. Với dữ liệu phân loại ta không thể định nghĩa hàm khoảng cách một cách tự nhiên.

Trong những năm gần đây, gom cụm dữ liệu phân loại đã thu hút nhiều sự chú ý từ cộng đồng nghiên cứu khai phá dữ liệu. Một số thuật toán gom cụm dữ liệu phân loại đã được đề xuất [1, 6, 7, 8, 26, 27]. Mặc dù những thuật toán này có những đóng góp quan trọng cho vấn đề gom cụm dữ liệu phân loại, nhưng chúng không được thiết kế để xử lý sự không chắc chắn trong quá trình gom cụm.

Một cách tiếp cận phổ biến để xử lý độ không chắc chắn là sử dụng Lý thuyết tập thô (Rough Sets Theory), đề xuất bởi Pawlak [3]. Trong những năm gần đây, một số tác giả đã đề xuất một cách tiếp cận mới bài toán gom cụm dữ liệu phân loại bằng cách sử dụng lý thuyết tập thô [28, 25, 29, 30, 31].

Gần đây, một số tác giả đã đề xuất một cách tiếp cận mới bài toán gom cụm dữ liệu phân loại bằng cách sử dụng lý thuyết tập thô và kỹ thuật phân chia dần [28, 32, 25, 29, 30]. Ý tưởng chính của cách tiếp cận này là làm thế nào để chọn ra được các thuộc tính tốt nhất từ nhiều thuộc tính ứng viên để phân chia dần các đối tượng vào các cụm tại mỗi thời điểm. Trong đó, thuật toán MGR [33] được Qin và cộng sự đề xuất, là một thuật toán gom cụm phân cấp tiêu biểu sử dụng các khái niệm lý thuyết thông tin liên quan với lý thuyết tập thô. Thuật toán MGR xác định thuộc tính gom cụm bằng cách sử dụng tỷ lệ lợi thông tin trung bình (the mean information gain ratio), điều này cho phép MGR

có khả năng xử lý sự không chắc chắn trong quá trình gom cụm dữ liệu phân loại. Tuy nhiên, Wei và các cộng sự [34] đã chỉ ra rằng MGR có thể sẽ không phù hợp để lựa chọn thuộc tính đại diện cho việc phân tách cụm.

Trong nỗ lực cải tiến chất lượng gom cụm dữ liệu, trên cơ sở đánh giá ưu nhược điểm của các thuật toán cơ sở, luận án đã đề xuất một thuật toán mới để gom cụm dữ liệu phân loại được gọi là MMNVI (Minimum Mean Normalized Variation of Information). MMNVI sử dụng Biến thể thông tin chuẩn hóa trung bình của một thuộc tính đối với các thuộc tính khác để tìm ra thuộc tính gom cụm tốt nhất và entropy của các lớp tương đương được tạo bởi thuộc tính gom cụm được chọn để phân tách nhị phân tập dữ liệu gom cụm.

4.2 Thuật toán MMNVI

4.2.1 Ý tưởng và những định nghĩa cơ bản

Như chúng ta đã thấy trong Chương 2, trong hệ thống thông tin phân loại $IS = (U, A)$, mỗi thuộc tính trong A xác định duy nhất một phân hoạch (tức là một phép gom cụm) trên tập U các đối tượng. Một thuộc tính tạo ra một phân hoạch tốt tập các đối tượng nếu phân hoạch đó chia sẽ được nhiều thông tin nhất với các phân hoạch sinh ra bởi các thuộc tính khác trong A [10, 35]. Ngoài ra, entropy của tập dữ liệu càng nhỏ thì mức độ giống nhau của các các đối tượng trong đó càng cao. Trên cơ sở hai nhận xét này, trong thuật toán MMNVI, luận án đề xuất thực hiện bước 1 và bước 2 trong khung công việc của một thuật toán gom cụm phân cấp như sau: (1) tại mỗi vòng lặp, luận án sẽ chọn thuộc tính phân cụm là thuộc tính tạo ra phân hoạch gần nhất với các phân hoạch xác định bởi các thuộc tính khác; (2) trong phân hoạch được tạo bởi thuộc tính phân cụm đã chọn, chọn lớp tương đương có độ tương tự trong lớp cao nhất làm một cụm và lấy hợp của tất cả các lớp còn lại tạo thành tập dữ liệu cần chia cụm mới. Hai bước trên sẽ được lặp lại trên tập dữ liệu cần phân cụm mới cho đến khi số cụm thu được bằng số cụm mong muốn cho trước k .

Để đo khoảng cách giữa hai thuộc tính (hai phân hoạch), luận án sử dụng số liệu NVI (Biến thể thông tin chuẩn hóa) được xác định dưới đây.

Định nghĩa 4.1. Cho hệ thống thông tin $IS = (U, A)$ và $a_i \in A$. Biến thể thông tin chuẩn hóa trung bình giữa a_i với mỗi $a_j \in A, a_j \neq a_i$ được định nghĩa như sau:

$$MNVI(a_i) = \frac{1}{|A| - 1} \sum_{j=1, j \neq i}^{|A|} NVI(a_i, a_j). \quad (4.1)$$

trong đó $NVI(a_i, a_j)$ là độ đo biến thể thông tin chuẩn hóa giữa thuộc tính a_i với thuộc tính a_j được định nghĩa tại 2.16 Chương 2.

Vì $NVI(a, b)$ là một metric trong không gian các thuộc tính, nên $MNVI(a_i)$ được xem là khoảng cách trung bình giữa a_i với mỗi $a_j \in A, a_j \neq a_i$.

Định nghĩa 4.2 [33, 36] Cho hệ thống thông tin $IS = (U, A)$ và $A = \{a_1, a_2, \dots, a_p\}$. Giả sử các thuộc tính trong A là độc lập lẫn nhau, khi đó entropy của tập dữ liệu $X \subseteq U$, ký hiệu là $Entropy(X)$, được định nghĩa bởi:

$$Entropy(X) = H_X(a_1) + H_X(a_2) + \dots + H_X(a_p). \quad (4.2)$$

trong đó $H_X(a_i)$ là entropy trên X của thuộc tính $a_i, i = 1, \dots, p$, xác định bởi công thức (2.11, Chương 2). Entropy của X càng nhỏ thì các đối tượng trong X càng giống nhau. Do đó, entropy của một cụm đã được nhiều tác giả sử dụng làm thước đo độ tương tự của các đối tượng trong cùng một cụm [33, 36].

Với ý tưởng và định nghĩa trên, thuật toán MMNVI hoạt động như sau:

4.2.2 Thuật toán MMNVI

Ở vòng lặp đầu tiên, thuật toán lấy tất cả các đối tượng trong tập U làm tập dữ liệu cần gom cụm và tại mỗi vòng lặp thực hiện ba bước chính sau đây:

Bước 1. Loại bỏ tất cả các thuộc tính có giá trị đơn lẻ.

Bước 2. Chọn ra thuộc tính có giá trị MNVI nhỏ nhất làm thuộc tính gom cụm

Bước 3. Phân hoạch tập các đối tượng cần phân chia thành các lớp tương đương; lấy lớp tương đương có Entropy nhỏ nhất trở thành một cụm, đồng thời

trả về tập dữ liệu cần phân cụm tiếp theo là tập chứa tất cả các tương đương còn lại.

Quá trình gom cụm trên được lặp lại cho đến khi số lượng nút lá thu được bằng với số lượng k cụm cho trước.

Thuật toán 4.1 Thuật MMNVI

Đầu vào: Tập dữ liệu gom cụm (Hệ thông tin) IS ; số cụm k cần gom.

Đầu ra: k cụm dữ liệu gom được.

Thuật toán:

Bước 1. $CNC = 1$ // Gán số cụm ban đầu bằng 1

$CDataSet = U$ // $CDataSet$ ký hiệu nút dữ liệu cần phân cụm.

Bước 2. $B = A$;

Với mọi $a_i \in A$

Tính phân hoạch $CDataSet/Ind\{a_i\}$;

Nếu $|CDataSet/Ind\{a_i\}| = 1$ //mọi cá thể trong
// $CNode$ có cùng một giá trị về a_i

$B = B - a_i$ // loại bỏ thuộc tính a_i

Hết nếu

Hết với mọi

Bước 3. Với mọi $a_i \in B$

Tính $MNVI\{a_i\}$ // sử dụng công thức 4.8

Hết với mọi

Bước 4. Xác định thuộc tính gom cụm a^* thỏa mãn

$$a^* = \operatorname{argmin}_{a_i \in B} MNVI(a_i)$$

Bước 5. Xác định phân hoạch $CDataSet/Ind\{a^*\} = \{X_1, X_2, \dots, X_h\}$;

$$X = \operatorname{argmin}_{X_i} (Entropy(X_i))$$

// với $X_i \in CDataSet/Ind\{a^*\}$;

Bước 6. Trả về X là một cụm

$$CNC = CNC + 1;$$

Nếu $CNC < k$

$CDataset = CDataset - X$

Quay lại Bước 2;

Ngược lại: Trả về $Cdataset$ là một cụm;

Hết nếu:

Đối với thuật toán MMNVI có hai lưu ý cho sau đây:

(1) Ở bước 4, nếu có nhiều thuộc tính cùng cho giá trị MNVI nhỏ nhất, thuật toán sẽ chọn thuộc tính đầu tiên làm thuộc tính gom cụm.

(2) Ở bước 5 và 6, sau khi chọn được thuộc tính phân tách tập dữ liệu, MMNVI sẽ chọn lớp tương đương có Entropy thấp nhất làm một cụm và lấy hợp của các lớp tương đương còn lại làm tập dữ liệu cần phân cụm tiếp.

4.2.3 Kết quả thực nghiệm thuật toán MMNVI

Để đánh giá hiệu suất gom cụm, thuật toán gom cụm MMNVI được cài đặt bằng ngôn ngữ lập trình R và tính toán thực nghiệm trên các bộ dữ liệu thực tế. Bên cạnh đó, việc đánh giá còn được so sánh với hai thuật toán cơ sở là MMR và MGR.

4.2.4 Bộ dữ liệu đánh giá

Tính toán thực nghiệm của thuật toán MMNVI đã được thực hiện trên tám bộ dữ liệu chuẩn lấy từ kho dữ liệu chuẩn UCI bao gồm: Soybean small, Zoo, Votes, Breast Cancer Wisconsin, Mushroom, Balance Scale, Car Evaluation và Chess.

4.2.5 Phương pháp đánh giá hiệu suất

Để đánh giá kết quả gom cụm một cách khách quan, ba tiêu chí thường được các nhà nghiên cứu sử dụng rộng rãi bao gồm: độ thuần khiết tổng thể (Overall Purity), chỉ số ngẫu nhiên hiệu chỉnh (Adjusted Rand Index - ARI), thông tin tương hỗ chuẩn hóa (Normalized Mutual Information - NMI).

4.2.6 Kết quả gom cụm

Trước tiên, luận án trình bày kết quả gom cụm thu được bởi thuật toán đề xuất MMNVI, sau đó trình bày các kết quả tính toán độ thuần khiết tổng thể (Overall Purity), chỉ số ngẫu nhiên hiệu chỉnh (Adjusted Rand Index - ARI) và

thông tin tương hỗ chuẩn hóa (Normalized Mutual Information – NMI thu được bởi ba thuật toán MMR, MGR và MMNVI trên 8 bộ dữ liệu.

Trong số 8 bộ dữ liệu thực nghiệm, MMNVI có độ thuần khiết tổng thể cao nhất trong năm tập dữ liệu, cụ thể là trên các tập Soybean Small, Breast Cancer Wisconsin, Car evaluation, Votes và Balance scale. MMR có độ thuần khiết tổng thể cao nhất đối với tập Mushroom. MGR có độ thuần khiết tổng thể cao nhất ở Soybean Small, Chess, và Zoo. Ngoài ra, khi so sánh độ thuần khiết tổng thể trung bình của từng thuật toán trên 8 tập dữ liệu, MMNVI cũng đạt độ tinh khiết tổng thể trung bình cao nhất.

Về chỉ số ngẫu nhiên hiệu chỉnh (ARI) của ba thuật toán cho thấy, MMNVI cũng có giá trị ARI cao nhất trên bốn tập dữ liệu Breast Cancer Wisconsin, Votes, Chess và Balance. MMR có giá trị ARI thấp nhất trên tất cả 8 tập dữ liệu. MGR có giá trị ARI cao nhất ba tập Car evaluation, Mushroom và Zoo. Thuật toán MMNVI cũng đạt giá trị ARI trung bình cao nhất.

Về chỉ số thông tin tương hỗ chuẩn hóa (NMI) của ba thuật toán, MMNVI có giá trị NMI cao nhất trên ba bộ dữ liệu Breast Cancer Wisconsin, Votes và Balance scale. MMR có NMI cao nhất trong Soybean Small, Car evaluation và Mushroom. MGR có NMI cao nhất trong Chess và Zoo.

Một chú ý quan trọng là MMNVI hoạt động tốt hơn nhiều so với 2 thuật toán còn lại trên Breast Cancer, Votes và Balance scale. Breast Cancer và Votes là các tập dữ liệu có sự phân bố lớp cân bằng. MGR hoạt động tốt hơn nhiều so với các thuật toán khác trên tập Zoo.

Kết quả thực nghiệm cho thấy, MMNVI cho thấy là thuật toán cho kết quả gom cụm tốt hơn hoặc ít ra là tương đương so với thuật toán MMR và MGR.

4.3 Kết luận chương 4

Trên cơ sở nghiên cứu các thuật toán cơ sở đã được đề xuất bởi các nhà nghiên cứu, tìm ra các thiếu sót, luận án đã đề xuất một thuật toán gom cụm phân cấp mới cho dữ liệu phân loại với tên gọi MMNVI. Kết quả thử nghiệm trên các tập dữ liệu thực tế lấy từ kho dữ liệu UCI cho thấy thuật toán MMNVI là một thuật toán ổn định, cho kết quả gom cụm tốt hơn hoặc ít ra là tương

đương so với các thuật toán cơ sở. MMNVI là thuật toán có thể được sử dụng thành công trong việc gom cụm dữ liệu phân loại.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Khám phá tri thức từ CSDL là một lĩnh vực khoa học nhằm nghiên cứu để tạo ra những công cụ khai phá những thông tin, tri thức hữu ích, tiềm ẩn mang tính dự đoán trong các CSDL lớn. Tuy nhiên, với tốc độ tăng trưởng nhanh của dữ liệu ngày nay, việc nghiên cứu và ứng dụng các kỹ thuật khai phá dữ liệu còn gặp phải nhiều khó khăn, thách thức.

Luận án tập trung vào việc ứng dụng Lý thuyết tập thô nhằm giải quyết hai vấn đề của khám phá tri thức: (1) nghiên cứu thuật toán hiệu quả tìm tập rút gọn thuộc tính trong một bảng quyết định; (2) kỹ thuật gom cụm dữ liệu phân loại cho phép xử lý sự không chắc chắn trong quá trình gom cụm.

5.1 Những kết quả và đóng góp chính của luận án

(1) Bảng việc nghiên cứu các thuật toán đã được đề xuất bởi các nhà nghiên cứu, tìm ra các thiếu sót, luận án đã đề xuất một thuật toán mới lựa chọn thuộc tính trong một bảng quyết định dựa trên gom cụm.

Kết quả thử nghiệm trên các tập dữ liệu thực tế lấy từ kho dữ liệu UCI cho thấy thuật toán đề xuất ACBRC là rất khả quan trong việc làm giảm số thuộc tính trong các bảng quyết định, đồng thời nâng cao được độ chính xác phân lớp.

(2) Bảng việc nghiên cứu các thuật toán cơ sở đã được đề xuất bởi các nhà nghiên cứu, phân tích các thiếu sót, luận án đã đề xuất thuật toán gom cụm dữ liệu phân loại MMNVI theo phương pháp phân cấp.

Kết quả thử nghiệm trên các tập dữ liệu thực tế lấy từ kho dữ liệu UCI cho thấy thuật toán MMNVI là một thuật toán ổn định, cho kết quả gom cụm tốt hơn hoặc ít ra là tương đương so với các thuật toán cơ sở. MMNVI là thuật toán có thể được sử dụng thành công trong việc gom cụm dữ liệu phân loại.

Các đóng góp chính trên đây đã được đăng trong hai bài báo trên Journal of Computer Science and Cybernetics, năm 2022 [CT2] và năm 2023 [CT3]. Ngoài các đóng góp chính trình bày trong luận án, nghiên cứu sinh là

đồng tác giả của có một số kết quả khác liên quan đến đề tài luận án, bao gồm một bài báo quốc tế và ba báo cáo hội thảo khoa học trong nước.

5.2 Hướng phát triển của luận án

Đối với bài toán lựa chọn thuộc tính, tiếp tục nghiên cứu:

- Giải pháp tìm tập thuộc tính rút gọn trong bảng quyết định có thông tin bị thiếu;
- Ứng dụng thuật toán lựa chọn thuộc tính ACBRC vào phân loại văn bản.

Đối với bài toán gom cụm dữ liệu, tiếp tục nghiên cứu:

- Cải tiến thuật toán gom cụm MMNVI để nó có khả năng tự động xác định số lượng cụm thích hợp cho tập dữ liệu, thay vì phải chỉ định trước số lượng cụm cần gom. Chẳng hạn, chúng ta có thể để thuật toán MMNVI dừng quá trình lặp khi entropy của tất cả các nút lá thấp hơn một giá trị ngưỡng thích hợp;
- Phát triển MMNVI để có thể xử lý cả dữ liệu số và dữ liệu phân loại;
- Nghiên cứu cải thiện độ phức tạp tính toán.

DANH MỤC CÔNG TRÌNH ĐÃ CÔNG BỐ

Tạp chí quốc tế

[CT1] Do Si Truong, Nguyen Thanh Tung, Lam Thanh Hien, Improved minimum-minimum roughness algorithm for clustering categorical data, *International Journal of Advanced and Applied Sciences*, 8(10), Pages 43-50, 2021.

<https://doi.org/10.21833/ijaas.2021.10.006>

Tạp chí trong nước

[CT2] Do Si Truong, Lam Thanh Hien, Nguyen Thanh Tung, An effective algorithm for computing reducts decision table. *Journal of Computer Science and Cybernetics*, V.38, N.3 (2022), Pages 277–292, 2022.

DOI: <https://doi.org/10.15625/1813-9663/38/3/17450>.

[CT3] Do Si Truong, Lam Thanh Hien, Nguyen Thanh Tung, A new information theory based algorithm for clustering categorical data. *Journal of Computer Science and Cybernetics*, V.39, N.3 (2023), Pages 259-278, 2023.

DOI: <https://doi.org/10.15625/1813-9663/18568>

Hội thảo khoa học trong nước

[CT4] Pham Cong Xuyen, Do Si Truong, Nguyen Thanh Tung, An information-Theoretic metric based method for selecting clustering attribute. *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, trang 31-40, 2016. DOI: 10.15625/vap.2016.0005.

[CT5] Nguyễn Minh Huy, Đỗ Sĩ Trường, Nguyễn Thanh Tùng, Về hàm đo độ phụ thuộc thuộc tính suy rộng. *Kỷ yếu Hội thảo quốc gia lần thứ XX: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Quy Nhơn, 23-24/11/2017, trang 396-403, 2017.

[CT6] Đỗ Sĩ Trường, Trần Thanh Phương, Phạm Công Xuyên, Nguyễn Thanh Tùng, Thuật toán MMR cải tiến gom cụm dữ liệu phân loại, *Kỷ yếu Hội thảo “Fundamental and Applied IT Research 12/2021 (FAIR XIV)”*, Trường Đại học Công nghiệp Thực phẩm Tp.HCM, 2021.

TÀI LIỆU THAM KHẢO

- [1] J. Han, J. Pei, H. Tong, "Data Mining: Concepts and Techniques," *Morgan Kanufmann*, vol. 3th Edition, 2012.
- [2] A. Skowron, S. Dutta, "Rough sets: past, present, and future," *Natural computing*, vol. 17(4), pp. 855-876, 2018.
- [3] Z. Pawlak, "Rough Sets - Theoretical Aspects of Reasoning about Data," *Kluwer Academic Publishers, Dordrecht*, 1991.
- [4] S. Pal, "Rough set and deep learning: Some concepts," *Academia Letters*, 1849, pp. 1-6, 2021.
- [5] P. Pieta, T. Szmuc, "Applications of rough sets in big data analysis: An overview," *International Journal of Applied Mathematics and Computer Science*, vol. 31, no. 4, p. 659–683 , 2021.
- [6] Q. Zhang, Q. Xie, G. Wang, "A survey on rough set theory and its applications," *CAAI Transactions on Intelligence Technology*, 1, pp. 323-333, 2016.
- [7] R. Bello, R. Falcon, "Rough sets in machine learning: a review," *Thriving Rough Sets*, pp. 87-118, 2017.
- [8] R. Słowiński, S. Greco, B. Matarazzo, "Rough-set-based decision support In Search Methodologies," *Springer, Boston, MA.*, pp. 557-609, 2014.
- [9] Y. Y. Yao, "Information-Theoretic Measures for Knowledge Discovery and Data Mining," *Part of the Studies in Fuzziness and Soft Computing book series (STUDFUZZ)*, vol. 119.
- [10] N. X. Vinh, J. Epps, J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties Normalization and Correction for Chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837-2854, 2012.
- [11] A. Skowron, C. Rauszer, "The Discernibility Matrices and Functions in Information Systems. Intelligent Decision Support," *Handbook of Applications and Advances of the Rough Sets Theory*, *Kluwer, Dordrecht*, p. 331–362, 1992.

- [12] U. Stańczyk, L. C. Jain, "Feature selection for data and pattern recognition: an introduction," *In Feature Selection for Data and Pattern Recognition*, vol. 584, pp. 1-7, 2015.
- [13] P. Dhal, C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence*, pp. 1-39, 2021.
- [14] M. Alimoussa, A. Porebski, N. Vandenbroucke, R. O. H. Thami, S. El Fkihi, "Clustering-based Sequential Feature Selection Approach for High Dimensional Data Classification," *VISIGRAPP (4: VISAPP)*, pp. 122-132, 2021.
- [15] S. Chormunge , S. Jena, "Correlation based feature selection with clustering for high dimensional data.," *Journal of Electrical Systems and Information Technology*, 5, p. 542–549, 2018.
- [16] T. P. Hong, Y. L. Liou, S. L. Wang, Bay Vo, "Feature selection and replacement by clustering attributes," *Vietnam J Comput Sci*, vol. 1, p. 47–55, 2014.
- [17] A. Janusz, D. Slezak, "Utilization of Attribute Clustering Methods for Scalable Computation of Reducts from High-Dimensional Data.," *Proceedings of the Federated Conference on Computer Science and Information Systems*, p. 295–302, 2012.
- [18] A. Janusz, D. Slezak, "Rough set methods for attribute clustering and selection," *Applied Artificial Intelligence*, vol. 28, p. 220–242, 2014.
- [19] T. P. Hong, C. H. Chen, F. S. Lin, "Using group genetic algorithm to improve performance of attribute clustering, Appl.," *Soft Comput. J.*, p. <http://dx.doi.org/10.1016/j.asoc.2015.01.001>, 2015.
- [20] F. Pacheco, M. Cerrada, R. V. Sánchez, D. Cabrera, C. Li, J. V. de Oliveira, "Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery.," *Expert Systems with Application*, pp. 69-86, 2017.
- [21] T. P. Hong, P. C. Wang, C.K. Ting, "An evolutionary attribute clustering and selection method based on feature similarity," *The IEEE Congress on Evolutionary Computation*, 2010.
- [22] G. Chandrashekar, F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40 (1), no. 40th-year commemorative issue, p. 16 – 28, 2014.

- [23] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.
- [24] S. Naouali, S. B. Salem, Z. Chtourou, "Clustering Categorical Data: A Survey," *International Journal of Information Technology & Decision Making*, Vols. Vol. 19, No. 01, pp. 49-96, 2020.
- [25] J. Uddin, R. Ghazali, J. H. Abawajy, H. Shah, N. A. Husaini, A. Zeb, "Rough set based information theoretic approach for clustering uncertain categorical data," *PLOS ONE*, 2022.
- [26] G. Khandelwal, R. Sharma, "A simple yet fast clustering approach for categorical data.," *International Journal of Computer Applications*, 120(17)., pp. 25-30, 2015.
- [27] W. A. Hassanein, "Clustering algorithms for categorical data using concepts of significance and dependence of attributes," *European Scientific Journal*, vol. 10, pp. 381-400, 2014.
- [28] G. K. Singh, S. Mandal, "Cluster Analysis using Rough Set Theory," *Journal of Informatics and Mathematical Sciences*, Vols. 9, No. 3, pp. 509–520, , 2017.
- [29] S. Raheem, S. Al Shehabi, and A. M. Nassief, "MIGR: A Categorical Data Clustering Algorithm Based on Information Gain in Rough Set Theory," *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, Vols. 30, No. 05, pp. 757-771, 2022.
- [30] T. Herawan, M. M. Deris, J. H. Abawajy, "A rough set approach for selecting clustering attribute," *Knowledge-Based Systems*, vol. 23, p. 220–231, 2010.
- [31] J. Uddin, R. Ghazali, M. M. Deris, "An empirical analysis of rough set categorical clustering techniques," *PloS one*, 12(1), 2017.
- [32] J. Uddin, Rozaida Ghazali, Mustafa Mat Deris, "An empirical analysis of rough set categorical clustering techniques," *PLOS ONE*, Vols. 12, No. 1, 2017.
- [33] H. Qin, Xiuqin Ma, T. Herawan, J. M. Zain, "MGR: An information theory based hierarchical divisive clustering algorithm for categorical data," *Knowledge-Based Systems*, vol. 67, p. 401–411, 2014.

- [34] W. Wei, J. Liang, X. Guo, P. Song, Y. Sun, "Hierarchical division clustering framework for categorical data," *Neurocomputing*, vol. 341, p. 118–134, 2019.
- [35] Y. Y. Yao, Y. Zhao, J. Wang, "On reduct construction algorithms, Rough Sets and Knowledge Technology," *First International Conference, RSKT 2006, Proceedings, LNAI 4062*, pp. 297-304, 2006.
- [36] J. McCaffrey, "Data Clustering Using Entropy Minimization.," <http://visualstudiomagazine.com/Articles/2013/02/01/Data-Clustering-Using-Entropy-Minimization.aspx?Page=2&p=1>, 2018.
- [37] Mazlack, L. J., He, A., Zhu, Y., Coppock, S., "A rough set approach in choosing clustering attributes," *Proceedings of the ISCA 13th International Conference (CAINE 2000)*, p. 1–6, 2000.
- [38] Hoàng Thị Lan Giao, Khía cạnh đại số và logic phát hiện luật theo tiếp cận tập thô, Luận án Tiến sĩ Toán học, Viện Công Nghệ Thông Tin, 2007.
- [39] Nguyễn Đức Thuận, Phủ tập thô và độ đo đánh giá hiệu năng tập luật quyết định, Luận án Tiến sĩ Toán học, Viện Khoa học và Công nghệ Việt Nam, 2010.
- [40] Nguyễn Long Giang, Nghiên cứu một số phương pháp khai phá dữ liệu theo hướng tiếp cận tập thô, Luận án Tiến sĩ Toán học, Viện Công Nghệ Thông Tin, 2012.
- [41] J. Liang, K. S. Chin, C. Dang, R. C. M. Yam, "A new method for measuring uncertainty and fuzziness in rough set theory," *International Journal of General Systems*, vol. 31(4), pp. 331-342, 2002.
- [42] Y. Zhao, "R and Data Mining: Examples and Case Studies. <https://www.webpages.uidaho.edu/~stevel/517/RDataMining-book.pdf>," *Published by Elsevier*, December 2012.
- [43] M. J. Wierman, "Measuring uncertainty in rough set theory," *International Journal of General System*, 28(4-5), pp. 283-297, 1999.
- [44] J. Han , R. Sanchez , X. Hu, "Feature Selection Based on Relative Attribute Dependency: An Experimental Study," *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pp. 214-223, 2005.
- [45] J. Zhou, D. Miao, W. Pedrycz, H. Zhang, "Analysis of alternative objective functions for attribute reduction in complete decision tables," *Soft Comput* 15, p. 1601–1616, 2011.

- [46] Q. Song, J. Ni, G. Wang, "A fast clustering based feature subset selection algorithm for highdimensional data.," *IEEE Transactions on Knowledge and Data Engineering*, 25(1), p. 2013, 1–14.
- [47] K. Zhu, J. Yang, "A cluster-based sequential feature selection algorithm.," *In 2013 Ninth International Conference on Natural Computation (ICNC)*, p. 848–852, 2013.
- [48] C. Shannon , "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.
- [49] X. Hu, J. Han, T. Y. Lin, "A New Rough Sets Model Based on Database Systems," *Fundamenta Informaticae* 59 no. 2-3, pp. 135-152, 2004.
- [50] N. Xie, M. Liu, Z. Li, G. Zhang, "New measures of uncertainty for an interval-valued information system," *Information Sciences*, vol. 470, pp. 156-174, 2019.
- [51] M. Zhang, J. T. Yao, "A rough sets based approach to feature selection," *IEEE Annual Meeting of the Fuzzy Information*, 2004.
- [52] M. S. Raza, U. Qamar, "Feature selection using rough set-based direct dependency calculation by avoiding the positive region," *International Journal of Approximate Reasoning* 92, p. 175–197, 2018.
- [53] S. R. A. Ahmed, I. A. Barazanchi, Z. A. Jaaz, H. R. Abdulshaheed, "Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set," *Periodicals of Engineering and Natural Sciences (PEN)*, 7(2), pp. 448-457, 2019.
- [54] K. Thangavel, A. Pethalakshmi, "Dimensionality reduction based on rough set theory: A review," *Applied Soft Computing* 9, p. 1–12, 2009.
- [55] Z. Abbas, A. Burney, "A survey of software packages used for rough set analysis," *Journal of Computer and Communications*, 4 (9), pp. 10-18, 2016.
- [56] S. L. M. Belaidan, L. Y. Yee, N. A. Abd Rahman, K. S. Harun, "Implementing k-means clustering algorithm in collaborative trip advisory and planning system," *Periodicals of Engineering and Natural Sciences (PEN)*, 7(2), pp. 723-740, 2019.
- [57] B. K. Tripathy, A. Goyal, R. Chowdhury, A. S. Patra, "MMeMeR: An algorithm for clustering heterogeneous data using rough set theory,"

International Journal of Intelligent Systems and Applications, 9(8), 25., pp. 25-33, 2017.

- [58] R. Kohavi, G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, 97(1-2), pp. 273-324, 1997.
- [59] A. Jakulin, "Machine Learning Based on Attribute Interactions," *PhD Dissertation*, 2005.
- [60] A. Sandro, A. Pessoa, S. Stephany, "An Innovative Approach for Attribute Reduction in Rough Set Theory," *Intelligent Information Management*, pp. 223-239, 2014.
- [61] G. Y. Wang, H. Yu, D. C. Yang, "Decision table reduction based on conditional information entropy," *Chinese Journal Of Computers-Chinese* , vol. 25 (7), pp. 759-766, 2002.
- [62] D. Harris, A. V. Niekerk, "Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data.," *International Journal of Remote Sensing*, 39(23), p. 8934–8949, 2018.
- [63] T. P. Hong, Y. L. Liou, "Attribute clustering in high dimensional feature spaces," *International Conference on Machine Learning and Cybernetics*, p. 19–22, 2007.
- [64] T. P. Hong, P. C. Wang, Y. C. Lee, "Cybernetics and Systems: An International Journal," *An effective attribute clustering approach for feature selection and replacement.*, p. 657–669, 2009.
- [65] D. E. Goldberg, "Genetic algorithms in search, optimization & machine learning," *Boston, MA, USA: Addison Wesley*, 1989.
- [66] L. Kaufman, P. J. Rousseeuw, "Computing groups in data: an introduction to cluster analysis," *John Wiley & Sons, Toronto*, 1990.
- [67] D. Parmar, T. Wu, J. Blackhurst, "MMR: an algorithm for clustering categorical data using rough set theory," *Data and Knowledge Engineering*, vol. 63, p. 879–893, 2007.
- [68] I. K. Park, G. S. Choi, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems*, vol. 48, pp. 289-295, 2015.
- [69] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, pp. 107-145, 2001.

- [70] F. M. Reza, "An introduction to information theory," *Dover Publications, New York*, 1994.
- [71] H. Qin, Xiuqin Ma, J. M. Zain, T. Herawan, "A novel soft set approach in selecting clustering attribute," *Knowledge-Based Systems*, vol. 36, p. 139–149, 2012.
- [72] n.t. truong, "aaa," *abc*, 2023.
- [73] Pal, S., "Rough set and deep learning: Some concepts," *Academia Letters, 1849*, pp. 1-6, 2021.
- [74] Han, J., Pei, J., Tong, H, "Data Mining: Concepts and Techniques," *Morgan Kanufmann*, 2022.
- [75] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, Q. Song, "A new unsupervised feature selection algorithm using similarity-based feature clustering," *Computational Intelligence*, vol. 35 (1), p. 2–22, 2019.
- [76] S. Mesakar, M. S. Chaudhari, "Review Paper On Data Clustering Of Categorical Data," *International Journal of Engineering Research & Technology*, vol. 1, no. 10, December, 2012.
- [77] M. M. Baroud, S. Z. M. Hashim, J. U. Ahsan, A. Zainal, "Positive region: An enhancement of partitioning attribute based rough set for categorical data," *Periodicals of Engineering and Natural Sciences*, Vols. Vol. 8, No. 4, December 2020.
- [78] S. Raheem, S. Al Shehabi, and A. M. Nassief, "MIGR: A Categorical Data Clustering Algorithm Based on Information Gain in Rough Set Theory," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vols. Vol. 30, No. 05, pp. 757-771, 2022.
- [79] R. Jensen, Q. Shen, "A Rough Set-Aided System for Sorting WWW Bookmarks," *Lecture Notes in Computer Science*, vol. 2198, 2001.
- [80] R. Jensen, Q. Shen, "Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches," *Wiley-IEEE Press eBook*, 2008.
- [81] R. Jensen, Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Trans. Fuzzy Syst.* 17 (4), p. 824–838, 2009.
- [82] T. Herawan, "Rough Set Approach for Categorical Data Clustering," *A thesis submitted in fulfillment of requirements for the award of the Doctor of Philosophy*, 2010.

- [83] M. M. Baroud, S. Z. m. Hashim, A. Zainul, J. Ahnad, "An New Algorithm-based Rough Set for Selecting Clustering Attribute in Categorical Data," *6th International Conference on Advanced Computing & Commucation Systems (ICACCS)*, pp. 1358-1364, 2020.
- [84] B. Tripathy, A. Ghosh, "SDR: An algorithm for clustering categorical data using rough set theory," *Recent Advances in Intelligent Computational Systems, IEEE*, pp. 867-872, 2011.
- [85] T. Herawan, "Rough clustering for cancer data sets," *International Journal of Modern Physics: Conference Series*, vol. 09, pp. 240-258, 2012.
- [86] N. T. T. Hien, H. V. Nam, "A k-Means-Like Algorithm for Clustering Categorical Data Using an Information Theoretic-Based Dissimilarity Measure," *Foundations of Information and Knowledge Systems: 9th International Symposium, FoIKS, Linz, Austr*, 2016.