LIÊN HIỆP CÁC HỘI KHOA HỌC VÀ KỸ THUẬT VIỆT NAM

FAIR

KỶ YẾU HỘI NGHỊ KHOA HỌC CÔNG NGHỆ
QUỐC GIA LẦN THỨ XII

NGHIÊN CỨU CƠ BẢN VÀ ỨNG DỤNG CÔNG NGHỆ THÔNG TIN

HUẾ, 07-08/6/2019

Proceedings of the 12th National Conference on Fundamental and Applied Information Technology Research

(FAIR'2019)

MỤC LỤC

Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.00038

MỘT PHƯƠNG PHÁP LỰA CHỌN THUỘC TÍNH GOM CỤM SỬ DỤNG LÝ THUYẾT THÔNG TIN

Phạm Công Xuyên, Nguyễn Thanh Tùng

Lac Hong University

pcxuyen@lhu.edu.vn, nttung@lhu.edu.vn

TÓM TẮT: *Bài toán gom cụm dữ liệu xuất hiện trong nhiều lĩnh vực khác nhau. Mục tiêu cơ bản của gom cụm là nhóm đối tượng thành các cụm sao cho các đối tượng trong cùng một cụm thì tương tự như nhau hơn là các đối tượng từ các cụm khác nhau. Gần đây, nhiều nhà nghiên cứu quan tâm đến vấn đề gom cụm dữ liệu phạm trù (categorical), trong đó các đối tượng dữ liệu được mô tả bởi các thuộc tính không phải thuộc tính số. Đặc biệt, phương pháp gom cụm phân cấp dữ liệu phạm trù sử dụng lý thuyết tập thô đã thu hút nhiều sự chú ý. Chìa khóa của các phương pháp này là làm thế nào để chọn được một thuộc gom cụm tốt nhất tại mỗi thời điểm trong số nhiều thuộc tính ứng viên.*

Trong bài báo này, chúng tôi xem xét ba kỹ thuật dựa trên lý thuyết tập thô: TR (Total Roughness), MMR (Min-Min Roughness) và MDA (Maximum Dependency Attribute), và đề xuất một thuật toán mới MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), cho việc lựa chọn thuộc tính phân cụm theo tiếp cận phân cấp. MAX-MEAN-SU sử dụng độ đo SU (Symmetric Uncertainty), một độ đo lý thuyết thông tin cho phép lượng hóa mức độ tương quan lẫn nhau giữa hai thuộc tính; và tìm cách xác định thuộc tính gom cụm sao cho độ tương quan trung bình của nó với các thuộc tính khác đạt giá trị lớn nhất. Để đánh giá và so sánh MAX-MEAN-SU với ba kỹ thuật dựa trên lý thuyết tập thô, chúng tôi sử dụng khái niệm "Độ tương tự trung bình bên trong các cụm" của một phép gom cụm để đo lường chất lượng gom cụm của mỗi thuộc tính được chọn bởi mỗi phương pháp. Kết quả thực nghiệm cho thấy chất lượng gom cụm của thuộc tính chọn được bằng phương pháp MAX-MEAN-SU là cao hơn so với các thuộc tính chọn bởi các phương pháp TR, MMR và MDA. Do đó, MAX-MEAN-SU có thể được sử dụng như là một kỹ thuật hiệu quả lựa chọn thuộc tính trong phân cụm phân cấp dữ liệu phạm trù.

Từ khóa: *Gom cụm, Dữ liệu phân loại, Gom cụm phân cấp, Lý thuyết tập thô, Lựa chọn thuộc tính gom cụm, Độ không chắc chắn đối xứng.*

I. GIỚI THIỆU

Gom cụm là một trong những nhiệm vụ chính của khai thác dữ liệu. Nó có thể được định nghĩa như sau. Cho tập gồm n đối tượng $D = \{x_1, x_2, \dots, x_N\}$, trong đó mỗi đối tượng x_i được mô tả bằng một véc tơ M chiều trong không gian đặc trưng đã cho. Gom cụm là việc tìm ra các cụm/nhóm các đối tượng sao cho các đối tượng trong cùng một cụm thì có độ tương tự cao, còn các đối tượng trong các cụm khác nhau thì có độ tương tự thấp [6].

Bài toán gom cụm xuất hiện trong nhiều lĩnh vực khác nhau như Nhận dạng mẫu, Thị giác máy tính, Sinh học, Y học, Truy tìm thông tin,.... Hiện nay, có nhiều thuật toán gom cụm đã được xây dựng và giới thiệu trong các tài liệu. Các thuật toán này có thể được phân đại thể thành hai loại: phân hoạch và phân cấp. Các phương pháp phân hoạch tạo ra một phân hoạch trên tập dữ liệu, tối ưu hóa một hàm tiêu chuẩn. Các phương pháp phân cấp sinh ra một dãy các phân hoạch lồng nhau của tập dữ liệu.

Hầu hết các công trình về gom cụm đều tập trung vào dữ liệu số, nơi mà các tính chất hình học sẵn có có thể được khai thác để định nghĩa một cách tự nhiên khoảng cách giữa các điểm dữ liệu. Tuy nhiên, các ứng dụng khai thác dữ liệu thực tiễn thường gặp phải những tập dữ liệu, trong đó các thuộc tính là những thuộc tính phạm trù (categorical), mà với chúng ta không thể định nghĩa hàm khoảng cách một cách tự nhiên. Gần đây, gom cụm dữ liệu phạm trù đã thu hút sự quan tâm lớn của cộng đồng nghiên cứu khai thác dữ liệu [1, 4, 8, 10, 11, 12, 13]. Một trong các kỹ thuật *gom cụm phân cấp dữ liệu phạm trù* là sử dụng một dãy các thuộc tính gom cụm. Đầu tiên, coi tất cả các đối tượng là một cụm. Tại mỗi bước tiếp theo một thuộc tính được lựa chọn để phân chia một nút "lá". Quá trình lặp lại cho đến khi đạt được số cụm yêu cầu. Để áp dụng được kỹ thuật này, một vấn đề đặt ra là tại mỗi bước ta phải lựa chọn được, trong số nhiều thuộc tính ứng viên, một thuộc tính tốt nhất theo một tiêu chuẩn xác định để phân cụm các đối tượng.

Gần đây, xuất hiện một số công trình áp dụng lý thuyết tập thô, một công cụ xử lý sự không chắc chắn, trong quá trình lựa chọn thuộc tính gom cụm [8, 10, 11, 12, 13]. Mazlack và cộng sự [12] đề xuất kỹ thuật sử dụng Độ thô toàn phần (Total Roughness - TR) trong lý thuyết tập thô. Theo đó, độ thô toàn phần càng lớn thì chất lượng gom cụm của thuộc tính được chọn sẽ càng cao. Parmar và cộng sự [13] đề xuất MMR (Min-Min-Roughness), một thuật toán thuần túy dựa vào lý thuyết tập thô. Thuật toán MMR xác định thuộc tính gom cụm dựa vào tiêu chuẩn Độ thô cực tiểu (Min-Roughness, MR). Thế nhưng, Herawan và cộng sự [8] đã chỉ ra rằng, MMR chỉ là sự bổ sung của TR, hơn nữa kỹ thuật này có độ phức tạp tính toán cao. Để khắc phục vấn đề này, Herawan và cộng sự [8] đã đề xuất một kỹ thuật khác gọi là MDA (Maximum Dependency Attributes). MDA sử dụng độ phụ thuộc giữa các thuộc tính trong lý thuyết tập thô. Theo Herawan và cộng sự [8], kỹ thuật MDA có hiệu năng tốt hơn TR và MMR. Tuy nhiên, giữa TR, MMR và MDA có sự giống nhau về bản chất. Sự giống nhau này nằm ở chỗ tiêu chuẩn lựa chọn thuộc tính gom cụm của cả ba kỹ thuật đều được xác định chủ yếu thông qua số phần tử có trong xấp xỉ dưới của một thuộc tính đối với các thuộc tính khác.

Trong bài báo này, chúng tôi xem xét ba kỹ thuật dựa trên lý thuyết tập thô: TR (Total Roughness), MMR (Min-Min Roughness) và MDA (Maximum Dependency Attribute), và đề xuất MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), một thuật toán mới cho việc lựa chọn thuộc tính gom cụm theo tiếp cận phân cấp. MAX-MEAN-SU sử dụng Độ không chắc chắn đối xứng (Symmetric Uncertainty - *SU*), một độ đo của Lý thuyết thông tin cho phép lượng hóa mức độ tương quan lẫn nhau giữa hai thuộc tính. Thuộc tính gom cụm được chọn là thuộc tính có độ tương quan trung bình với các thuộc tính khác lớn nhất. Ưu điểm của *SU* so với Độ lợi thông tin (Information Gain - *IG*), là *SU không thiên vị các thuộc tính có nhiều giá trị*. Để đánh giá và so sánh MAX-MEAN-SU với ba kỹ thuật dựa trên lý thuyết tập thô, chúng tôi sử dụng khái niệm "Độ tương tự trung bình bên trong các cụm" của một phép gom cụm để đo lường chất lượng gom cụm của mỗi thuộc tính được chọn bởi mỗi phương pháp. Kết quả thực nghiệm cho thấy chất lượng gom cụm của thuộc tính chọn được bằng phương pháp MAX-MEAN-SU là cao hơn so với các thuộc tính chọn bởi các phương pháp TR, MMR và MDA. Do đó, MAX-MEAN-SU có thể được sử dụng như là một kỹ thuật hiệu quả lựa chọn thuộc tính trong phân cụm phân cấp.

II. CÁC KHÁI NIỆM CƠ BẢN

Mục này sẽ trình bày một cách vắn tắt một số khái niệm liên quan đến Lý thuyết tập thô [14] và Lý thuyết thông tin [16].

Một tập dữ liệu gom cụm có thể được biểu diễn dưới dạng một bảng, trong đó mỗi hàng biểu diễn một đối tượng, một trường hợp hay một sự kiện, mỗi cột biểu diễn một thuộc tính, một tính chất hay một số đo có thể đo được trên mỗi đối tượng. Trong lý thuyết tập thô, một bảng dữ liệu như vậy được gọi là một hệ thông tin. Một cách hình thức, người ta định nghĩa hệ thông tin như sau.

Định nghĩa 1. Hệ thông tin là một bộ tứ $S = (U, A, V, f)$, trong đó U là một tập hữu hạn, không rỗng các đối tượng, A là một tập hữu hạn, không rỗng các thuộc tính, $V = \bigcup_{a \in A} V_a$ với V_a là tập tất cả các giá trị của thuộc tính a, và $f: U \times A \to V$ là hàm thông tin, gán giá trị $f(u, a) \in V_a$ cho mỗi cặp $(u, a) \in U \times A$.

Định nghĩa 2. Cho $S = (U, A, V, f)$ là một hệ thông tin, $B \subseteq A$. Hai phần tử $x, y \in U$ được gọi là B-không phân biệt được trong S nếu và chỉ nếu $f(x, a) = f(y, a)$ với mọi $a \in B$.

Ta ký hiệu quan hệ không phân biệt sinh bởi tập thuộc tính B bởi $IND(B)$. Dễ thấy, $IND(B)$ là một quan hệ tương đương và nó sinh ra một phân hoạch (một phép gom cụm) trên U. Phân hoạch trên U sinh bởi $IND(B)$ trong $S = (U, A, V, f)$ được ký hiệu là π_B và lớp tương đương trong π_B chứa $x \in U$, được ký hiệu là $[x]_B$.

Định nghĩa 3. Cho $S = (U, A, V, f)$ là một hệ thông tin, $B \subseteq A$ 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:

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \emptyset\} \quad (1)$$

Định nghĩa trên nói rằng nếu đối tượng $x \in \underline{B}X$ thì nó *chắc chắn* thuộc vào tập X, còn khi $x \in \overline{B}X$ thì nó *có thể* thuộc vào tập X. Hiển nhiên, ta có $\underline{B}X \subseteq X \subseteq \overline{B}X$. X được gọi là định nghĩa được nếu $\underline{B}X = \overline{B}X$. trường hợp ngược lại, X được gọi là tập thô với B-biên $BN_B(X) = \overline{B}X - \underline{B}X$.

Định nghĩa 4. Cho $S = (U, A, V, f)$ là một hệ thông tin, $B \subseteq A$ và $X \subseteq U$. Độ chính xác của xấp xỉ X thông qua B được định nghĩa bởi

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} \quad (2)$$

Trong suốt bài báo này, $|X|$ ký hiệu số phần tử của tập X.

Hiển nhiên, $0 \leq \alpha_B(X) \leq 1$. Nếu $\alpha_B(X) = 1$, thì $\underline{B}X = \overline{B}X$, B-biên của X là tập rỗng, và X là tập rõ đối với B. Nếu $\alpha_B(X) < 1$, thì $\underline{B}X \subset \overline{B}X$, B-biên của X là khác rỗng, và X là tập thô đối với B.

Định nghĩa 5. Cho $S = (U, A, V, f)$ là một hệ thông tin, $B \subseteq A$ và $X \subseteq U$. Độ thô (roughness) của X đối với B được định nghĩa là

$$\rho_B(X) = 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \quad (3)$$

Định nghĩa 6. Cho $S = (U, A, V, f)$ là một hệ thông tin. Với $P, Q \subseteq A$, ta nói Q phụ thuộc vào P với mức k ($0 \leq k \leq 1$), ký hiệu $P \Rightarrow_k Q$, nếu

$$k = \frac{\sum_{X \in \pi_Q} |\underline{P}(X)|}{|U|} \quad (4)$$

Định nghĩa 7. Cho $S = (U, A, V, f)$ là một hệ thông tin, $P \subseteq A$. Giả sử P sinh ra phân hoạch $\pi_P = \{P_1, P_2, \dots, P_m\}$ trên U. Khi đó, entropy của P được xác định bởi

$$E(P) = -\sum_{i=1}^{m} Pr(P_i)\log_2 Pr(P_i) \quad (5)$$

trong đó $Pr(P_i) = |P_i|/|U|$, $i = 1,2,\dots,m$ và quy định $0\log_2 0 = 0$.

Định nghĩa 8. Cho $S = (U, A, V, f)$ là một hệ thông tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Entropy có điều kiện của Q đối với P được định nghĩa bởi

$$E(Q|P) = -\sum_{j=1}^{n} Pr(P_i) \sum_{i=1}^{m} Pr(Q_j|P_i)\log_2 Pr(Q_j|P_i) \quad (6)$$

trong đó $Pr(Q_j|P_i) = |X_i \cap Y_j|/|Y_j|$, $i = 1,2,\dots,m$ và $j = 1,2,\dots,n$.

Định nghĩa 9. Cho $S = (U, A, V, f)$ là một hệ thông tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Entropy đồng thời của P và Q được định nghĩa bởi

$$E(P, Q) = -\sum_{i=1}^{m}\sum_{j=1}^{n} Pr(P_i, Q_j)\log_2 Pr(P_i, Q_j) \quad (7)$$

trong đó $Pr(P_i, Q_j) = |X_i \cap Y_j|/|U|$, $i = 1,2,\dots,m$ và $j = 1,2,\dots,n$.

Từ các công thức (5), (6) và (7) ta có

$$E(Q|P) = E(P, Q) - E(P) \quad (8)$$

Định nghĩa 10. Cho $S = (U, A, V, f)$ là một hệ thông tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Thông tin tương hỗ (Mutual information) giữa P và Q được định nghĩa bởi

$$I(P; Q) = E(Q) - E(Q|P) = E(P) - E(P|Q) = E(P) + E(Q) - E(P, Q) \quad (9)$$

Định nghĩa 11. [16] Cho $S = (U, A, V, f)$ là một hệ thông tin; $P, Q \subseteq A$, $\pi_P = \{P_1, P_2, \dots, P_m\}$, và $\pi_Q = \{Q_1, Q_2, \dots, Q_n\}$. Độ không chắc chắn đối xứng (Symmetrical uncertainty - SU) giữa P và Q được định nghĩa bởi

$$SU(P, Q) = 2\frac{I(P; Q)}{E(P) + E(Q)}. \quad (10)$$

SU là một độ đo chuẩn hóa, cho phép lượng hóa sự phụ thuộc lẫn nhau của hai thuộc tính P và Q. SU cho phép phản ánh mức độ tương quan ngay cả khi sự tương quan giữa P và Q không phải là tuyến tính [16]. Giá trị của độ đo này nằm trên đoạn [0,1]. Nếu SU bằng 1, thì giá trị của thuộc tính này sẽ hoàn toàn dự đoán được nếu biết giá trị của thuộc tính kia; trường hợp SU bằng 0, hai thuộc tính P và Q là độc lập nhau.

III. BA KỸ THUẬT DỰA TRÊN LÝ THUYẾT TẬP THÔ

Cho $S = (U, A, V, f)$ là một hệ thông tin, $a_i \in A$, $V(a_i)$ là tập các giá trị của thuộc tính a_i; $X(a_i = \alpha)$ là tập các đối tượng có cùng giá trị thuộc tính a_i là α, nghĩa là $X(a_i = \alpha)$ là một lớp tương đương trong U sinh bởi quan hệ không phân biệt $IND(a_i)$; $\underline{X_{a_j}(a_i = \alpha)}$ là xấp xỉ dưới, và $\overline{X_{a_j}(a_i = \alpha)}$ là xấp xỉ trên của $X(a_i = \alpha)$ đối với a_j.

A. Kỹ thuật TR (Total Roughness) [12]

Input: Tập dữ liệu gom cụm (Hệ thông tin) S.

Output: Thuộc tính gom cụm.

Begin

Bước 1: Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2: Với mỗi thuộc tính a_i xác định độ thô trung bình $Rough_{a_j}(a_i)$ của nó đối với mỗi thuộc tính a_j với $j \neq i$, theo công thức

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X \mid a_i = a_k)}{|V(a_i)|} \quad (11)$$

trong đó

$$R_{a_j}(X \mid a_i = \alpha) = \frac{\left|X_{a_j}(a_i = \alpha)\right|}{\left|\overline{X_{a_j}(a_i = \alpha)}\right|} \quad (12)$$

Bước 3: Với mỗi $a_i \in A$ tính **độ thô toàn phần** của nó đối với mọi a_j, $i \neq j$, theo công thức

$$TR(a_i) = \frac{\sum_{(j=1)\wedge(i\neq j)}^{|A|} Rough_{a_j}(a_i)}{|A| - 1} \quad (13)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị TR lớn nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \text{argmax}_{a_j \in A}\{TR(a_j)\} \quad (14)$$

End

B. Kỹ thuật MMR (Min-Min-Roughness) [13]

Input: Tập dữ liệu gom cụm (Hệ thông tin) S.

Output: Thuộc tính gom cụm.

Begin

Bước 1: Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2: Với mỗi thuộc tính a_i xác định độ thô trung bình $Rough_{a_j}(a_i)$ của nó đối với mỗi thuộc tính a_j với $j \neq i$, theo công thức

$$Rough_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X \mid a_i = a_k)}{|V(a_i)|} \quad (15)$$

trong đó

$$R_{a_j}(X \mid a_i = \alpha) = 1 - \frac{\left|X_{a_j}(a_i = \alpha)\right|}{\left|\overline{X_{a_j}(a_i = \alpha)}\right|} \quad (16)$$

Bước 3: Với mỗi $a_i \in A$ tính **độ thô nhỏ nhất** của nó $MR(a_i)$ theo công thức

$$MR(a_i) = \min_{(a_j \in A)\wedge(j\neq i)} \left(Rough_{a_j}(a_i)\right) \quad (17)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị MR nhỏ nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \text{argmin}_{a_j \in A}\{MR(a_j)\} \quad (18)$$

End

C. Kỹ thuật MDA (Maximum degree of Dependency of Attributes) [8]

Input: Tập dữ liệu gom cụm (Hệ thông tin) S.

Output: Thuộc tính gom cụm.

Begin

Bước 1: Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2: Với mỗi thuộc tính a_i xác định độ phụ thuộc của a_i vào mỗi a_j với $j \neq i$, theo công thức

$$\gamma_{a_j}(a_i) = \frac{\sum_{X \in U/a_i} \left|\underline{a_j X}\right|}{|U|} \quad (19)$$

Bước 3. Chọn độ phụ thuộc lớn nhất $MD(a_i)$ của mỗi thuộc tính a_i ($a_i \in A$) như sau

$$MD(a_i) = \max_{(a_j \in A)\wedge(i\neq j)} \left(\gamma_{a_j}(a_i)\right) \quad (20)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị MD lớn nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \text{argmax}_{a_j \in A}\{MD(a_j)\} \quad (21)$$

IV. KỸ THUẬT MAX-MEAN-SU

Mục này đề xuất MAX-MEAN-SU, một kỹ thuật dựa vào độ không chắc chắn đối xứng trung bình lớn nhất (Maximum Mean of Symmetric Uncertainties) để lựa chọn thuộc tính gom cụm.

Định nghĩa 12. Cho $S = (U, A, V, f)$ là một hệ thông tin, và $a_i \in A$ là một thuộc tính. Độ không chắc chắn đối xứng trung bình giữa a_i với mỗi $a_j \in A, j \neq i$, được xác định bởi

$$MSU(a_i) = \frac{\sum_{j=1, j\neq i}^{|A|} SU(a_i, a_j)}{|A| - 1}. \quad (22)$$

Rõ ràng, nếu giá trị MSU của một thuộc tính nào đó càng cao thì tính đại diện của nó cho các thuộc tính còn lại càng lớn. Do đó, nếu sử dụng nó để gom cụm thì chất lượng gom cụm thu được sẽ càng cao. Dựa vào định nghĩa trên, chúng tôi đề xuất thuật toán gom cụm MAX-MEAN-SU như sau.

Input: Tập dữ liệu gom cụm (Hệ thông tin) S.

Output: Thuộc tính gom cụm.

Begin

Bước 1. Tính các lớp tương đương sinh bởi quan hệ không phân biệt trên mỗi thuộc tính.

Bước 2. Với mỗi thuộc tính a_i xác định độ không chắc chắn đối xứng $SU(a_i, a_j)$ giữa a_i và mỗi $a_j, j \neq i$ theo công thức

$$SU(a_i, a_j) = 2 \times \frac{I(a_i; a_j)}{E(a_i) + E(a_j)} \quad (23)$$

Bước 3. Tính độ không chắc chắn đối xứng trung bình $MSU(a_i)$ của mỗi thuộc tính $a_i \in A$ như sau

$$MSU(a_i) = \frac{\sum_{j=1, j\neq i}^{|A|} SU(a_i, a_j)}{|A| - 1} \quad (24)$$

Bước 4. Chọn thuộc tính a_i^* cho giá trị MSU lớn nhất làm thuộc tính gom cụm, nghĩa là

$$a_i^* = \text{argmax}_{a_j \in A}\{MSU(a_j)\}$$

End

Ta hãy minh họa thuật toán Max-Mean-SU bằng một ví dụ.

Ví dụ. Bảng 1 mô tả tập dữ liệu Animal world lấy từ **[8]**. Tập dữ liệu này gồm 9 đối tượng với 10 thuộc tính: Animal, Hair, Teeth, Eye, Feather, Feet, Eat, Milk, Fly và swim.

Bảng 1. Tập dữ liệu "Animal world"

Animal	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim
Tiger	Y	pointed	forward	N	claw	meat	Y	N	Y
Cheetah	Y	pointed	forward	N	claw	meat	Y	N	Y
Giraffe	Y	blunt	side	N	hoof	grass	Y	N	N
Zebra	Y	blunt	side	N	hoof	grass	Y	N	N
Ostrich	N	N	side	Y	claw	grain	N	N	N
Penguin	N	N	side	Y	web	fish	N	N	Y
Albatross	N	N	side	Y	craw	grain	N	Y	Y
Eagle	N	N	forward	Y	craw	meat	N	Y	N
Viper	N	pointed	forward	N	N	meat	N	N	N

Trước tiên, ta làm việc với thuộc tính Hair. Phân hoạch của U sinh bởi Hair là

π_{Hair} ={{Tiger, Cheetah, Giraffe, Zebra}, {Ostrich, Penguin, Albatross, Eagle, Viper}}.

Ta có

$$E(\text{Hair}) = -\frac{4}{9} log_2 \frac{4}{9} - \frac{5}{9} log_2 \frac{5}{9} = 0{,}991 .$$

Phân hoạch của U sinh bởi Teeth là

π_{Teeth} ={{Tiger,Cheetah,Viper},{Giraffe, Zebra},{Ostrich, Penguin, Albatross, Eagle}}.

$$E(\text{Teeth}) = -\frac{3}{9}log_2\frac{3}{9} - \frac{2}{9}log_2\frac{2}{9} - \frac{4}{9}log_2\frac{4}{9} = 1{,}53.$$

Phân hoạch của U sinh bởi {Hair, Teeth} là

$\pi_{\{\text{Hair,Teeth}\}}$ = {{Tiger, Cheetah}, {Giraffe, Zebra}, {Ostrich, Penguin, Albatross, Eagle}, {Viper}}

$$E(\text{Hair, Teeth}) = -\frac{2}{9}log_2\frac{2}{9} - \frac{2}{9}log_2\frac{2}{9} - \frac{4}{9}log_2\frac{4}{9} - \frac{1}{9}log_2\frac{1}{9} = 1{,}837.$$

Áp dụng công thức (9), ta có thông tin tương hỗ giữa Hair và Teeth như sau

$$I(\text{Hair; Teeth}) = E(\text{Hair}) + E(\text{Teeth}) - E(\text{Hair, Teeth}) = 0{,}991 + 1{,}53 - 1{,}837 = 0{,}684.$$

Áp dụng công thức (10), ta có độ không chắc chắn đối xứng giữa Hair và Teeth như sau

$$SU(\text{Hair, Teeth}) = 2\frac{I(\text{Hair; Teeth})}{E(\text{Hair}) + E(\text{Teeth})} = 2\frac{0{,}684}{0{,}991 + 1{,}53} = 0{,}543.$$

Bằng cách tương tự, ta có thể thu được độ không chắc chắn đối xứng giữa Hair và các thuộc tính khác:

$SU(Hair, Eye) = 0{,}007, SU(Hair, Feather) = 0{,}595, SU(Hair, Feet) = 0{,}341,$

$SU(Hair, Eat) = 0{,}387, SU(Hair, Milk) = 1, SU(Hair, Fly) = 0{,}256, SU(Hair, Swim) = 0{,}007.$

Độ không chắc chắn đối xứng trung bình MSU của thuộc tính Hair được tính theo công thức (22) như sau

$$MSU(Hair) = \frac{0{,}543 + 0{,}007 + 0{,}595 + 0{,}341 + 0{,}387 + 1 + 0{,}256 + 0{,}007}{8} = 0{,}392.$$

Bằng quá trình tương tự như đối với Hair, ta có thể tính toán với các thuộc tính khác. Giá trị MSU của tất cả các thuộc tính được trình bày trong Bảng 2.

Bảng 2. Độ không chắc chắn đối xứng và độ không chắc chắn trung bình giữa các thuộc trong Bảng 1

Thuộc tính	SU									MSU
	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim	
Hair	1	0,543	0,007	0,595	0,341	0,387	1	0,257	0,007	0,392
Teeth	0,543	1	0,5	0,786	0,622	0,695	0,543	0,279	0,191	**0,520**
Eye	0,007	0,5	1	0,092	0,341	0,701	0,007	0,003	0,007	0,207
Feather	0,595	0,786	0,092	1	0,341	0,446	0,595	0,364	0,007	0,403
Feet	0,341	0,622	0,341	0,341	1	0,743	0,341	0,186	0,341	0,407
Eat	0,387	0,695	0,701	0,446	0,743	1	0,387	0,14	0,23	0,466
Milk	1	0,543	0,007	0,595	0,341	0,387	1	0,257	0,007	0,392
Fly	0,257	0,279	0,003	0,364	0,186	0,14	0,257	1	0,003	0,186
Swim	0,007	0,191	0,007	0,007	0,341	0,23	0,007	0,003	1	0,099

Từ Bảng 2, có thể thấy thuộc tính Teeth có giá trị MSU lớn nhất, do đó Teeth được chọn làm thuộc tính gom cụm bởi thuật toán MAX-MEAN-SU.

V. ĐÁNH GIÁ THỰC NGHIỆM

A. Tiêu chuẩn đánh giá

Các kỹ thuật TR, MMR, MDA và MAX-MEAN-SU sử dụng các phương pháp khác nhau trong việc lựa chọn thuộc tính gom cụm. Để đo đạc được một cách chính xác chất lượng gom cụm của các thuộc tính được chọn bởi các kỹ thuật khác nhau là công việc không đơn giản. Vì mục tiêu của phân tích gom cụm là nhóm các đối tượng có cùng các đặc trưng, chúng tôi sử dụng *độ tương tự trung bình trong mỗi cụm* (average intra-class similarity) để đánh giá chất lượng gom cụm của các thuộc tính được chọn.

Định nghĩa 13. Cho hệ thông tin $S = (U, A, V, f)$ và giả sử tất cả các thuộc tính trong A đều là những thuộc tính phạm trù. Khi đó độ tương tự của hai đối tượng x_i và x_j trong U được định nghĩa như sau:

$$s(x_i, x_j) = \frac{\left|\left\{\{a_k \in A\} \middle| f(x_i, a_k) = f(x_j, a_k)\right\}\right|}{|A|} \quad (25)$$

Định nghĩa 14. Cho hệ thông tin $S = (U, A, V, f)$. Giả sử $a_j \in A$ được chọn làm thuộc tính gom cụm và phép gom cụm (phân hoạch) sinh ra bởi a_j là $\pi_{a_j} = \{X_1, X_2, \dots, X_m\}$ trong đó $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. Nếu $|X_i| > 1$ thì độ tương tự trung bình (AS) của đối tượng x_{ij} với các đối tượng khác trong X_i được định nghĩa như sau:

$$AS(x_{ij}) = \frac{\sum_{k=1, k\neq j}^{|X_i|} s(x_{ij}, x_{ik})}{|X_i| - 1} \quad (26)$$

Định nghĩa 15. Cho hệ thông tin $S = (U, A, V, f)$. Giả sử $a_j \in A$ được chọn làm thuộc tính gom cụm và phép gom cụm (phân hoạch) sinh ra bởi a_j là $\pi_{a_j} = \{X_1, X_2, \dots, X_m\}$ trong đó $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. Khi đó, độ tương tự bên trong (CS) của cụm X_i được định nghĩa bởi:

$$CS(X_i) = \begin{cases} 1 \text{ nếu } |X_i| = 1, \\ \dfrac{\sum_{j=1}^{|X_i|} AS(x_{ij})}{|X_i|} \text{ trong các trường hợp khác} \end{cases} \quad (27)$$

Định nghĩa 16. Cho hệ thông tin $S = (U, A, V, f)$. Giả sử $a_j \in A$ được chọn làm thuộc tính gom cụm và phép gom cụm (phân hoạch) sinh ra bởi a_j là $\pi_{a_j} = \{X_1, X_2, \dots, X_m\}$ trong đó $X_i = \{x_{i1}, x_{i2}, \dots, x_{i|X_i|}\}$, $i = 1, 2, \dots, m$. Độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm tạo ra bởi a_j được định nghĩa bởi:

$$ACS(a_j) = \frac{\sum_{i=1}^{m} CS(X_i)}{m}. \quad (28)$$

Độ tương tự trung bình bên trong cụm (ACS) càng cao thì chất lượng gom cụm của thuộc tính gom cụm được chọn càng lớn.

B. Dữ liệu và kết quả tính toán đánh giá

Để đánh giá và so sánh MAX-MEAN-SU với các kỹ thuật TR, MMR, MDA, chúng tôi đã sử dụng bốn tập dữ liệu:

Animal world lấy từ tài liệu [8], được trình bày ở Bảng 1.

Zoo lấy từ UCI database, có 101 đối tượng, trình bày thông tin về các loài thú thông qua 18 thuộc tính.

Soybean lấy từ UCI database, có 47 đối tượng, 35 thuộc tính và một thuộc tính phân lớp.

Statlog (Heart) lấy từ UCI database, có 270 đối tượng, 13 thuộc tính và một thuộc tính phân lớp.

Chúng tôi đã lập trình cả bốn kỹ thuật TR, MMR, MDA và MAX-MEAN-SU bằng ngôn ngữ R với sự hỗ trợ của gói chương trình "RoughSets".

Bảng 3 thống kê các thuộc tính gom cụm được lựa chọn bởi các kỹ thuật trong mỗi tập dữ liệu.

Bảng 3. Các thuộc tính gom cụm được chọn bởi các kỹ thuật trong mỗi tập dữ liệu

	Datasets			
	Animal world	Zoo	Soybean	Statlog (Heart)
TR	Hair	Feathers	Plant-growth	Serum cholestoral
MMR	Hair	Feathers	Plant-growth	Serum cholestoral
MDA	Hair	Feathers	Plant-growth	Serum cholestoral
MAX-MEAN-SU	Teeth	Type	Class	Class

Từ Bảng 3, có thể thấy trong cả bốn tập dữ liệu xem xét, ba thuật kỹ thuật TR, MMR and MDA đều chọn cùng một thuộc tính. Kỹ thuật MAX-MEAN-SU của chúng tôi chọn thuộc tính khác trong cả bốn tập dữ liệu.

Bây giờ, chúng ta hãy đánh giá chất lượng gom cụm của các thuộc tính khác nhau được chọn bởi TR, MMR, MDA và MAX-MEAN-SU trong bốn tập dữ liệu, thông qua tính toán độ tương tự trung bình bên trong cụm (ACS).

Để minh họa công việc tính toán độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm tạo ra bởi một thuộc tính, chúng ta lấy thuộc tính Hair trong tập dữ liệu "Animal" làm ví dụ.

Phân hoạch của tập động vật U sinh bởi thuộc tính Hair gồm hai lớp tương đương:

$X_1 = X(\text{Hair} = \text{Y}) = \{\text{Tiger, Cheetah, Giraffe, Zebra}\}$,

$X_2 = X(\text{Hair} = \text{N}) = \{\text{Ostrich, Penguin, Albatross, Eagle, Viper}\}$.

Áp dụng công thức (25), ta có

$s(\text{Tiger, Cheetah}) = 1$, $s(\text{Tiger, Giraffe}) = 0.444$, $s(\text{Tiger, Zebra}) = 0.444$.

Áp dụng công thức (26), ta có *độ tương tự trung bình* của Tiger với các động vật khác trong X_1 là:

$$AS(\text{Tiger}) = \frac{1 + 0.444 + 0.444}{3} = 0.630$$

Bằng thực hiện quá trình tính toán tương tự, ta thu được độ tương tự trung bình của các động vật khác trong X_1. Các kết quả tính toán cho trong Bảng 4.

Bảng 4. Độ tương tự, *AS* và *CS* của các động vật trong X_1 sinh ra bởi Hair

Animal	Tiger	Cheetah	Giraffe	Zebra	*AS*	*CS*
Tiger	-	1.000	0.444	0.444	0.630	0.630
Cheetah	1.000	-	0.444	0.444	0.630	
Giraffe	0.444	0.444	-	1.000	0.630	
Zebra	0.444	0.444	1.000	-	0.630	

Áp dụng công thức (27), độ tương tự bên trong của X_1 được xác định như sau.

$$CS(X_1) = \frac{0.630 + 0.630 + 0.630 + 0.630}{4} = 0.630$$

Bằng cách tương tự, ta thu được $CS(X_2) = 0.544$.

Cuối cùng, sử dụng công thức (28), ta thu được độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm sinh bởi Hair là:

$$ACS(\text{Hair}) = \frac{0.630 + 0.544}{2} = 0.587.$$

Bằng cách tương tự, chúng tôi đã tính toán được độ tương tự trung bình bên trong cụm (ACS) của phép gom cụm sinh bởi Teeth trong "Animal world", bởi Feathers và bởi Type trong Zoo, bởi Plant-growth và bởi Class trong Soybean, bởi Serum cholestoral và bởi Class trong Statlog (Heart). Kết quả tính toán thu được cho trong Bảng 5.

Bảng 5. Độ tương tự trung bình bên trong cụm (ACS) của các thuộc tính

	Thuộc tính được chọn và giá trị *ACS* của nó			
	trên Animal world	trên Zoo	trên Soybean	trên Statlog (Heart)
TR	Hair **0.587**	Feathers **0.740**	Plant-growth **0.681**	Serum cholestoral **0.553**
MMR	Hair **0.587**	Feathers **0.740**	Plant-growth **0.681**	Serum cholestoral **0.553**
MDA	Hair **0.587**	Feathers **0.740**	Plant-growth **0.681**	Serum cholestoral **0.553**
MAX-MEAN-SU	Teeth **0.784**	Type **0.866**	Class **0.853**	Class **0.606**

Các kết quả tính toán trong Bảng 5 cho thấy chất lượng gom cụm của các thuộc tính được chọn bởi MAX-MEAN-SU là tốt hơn chất lượng gom cụm của các thuộc tính được chọn bởi TR, MMR and MDA techniques.

VI. KẾT LUẬN

Gần đây, một số công trình áp dụng lý thuyết tập thô vào việc lựa chọn thuộc tính gom cụm theo tiếp cận phân cấp đã được đề xuất bởi một số tác giả. Trong bài báo này, chúng tôi xem xét ba kỹ thuật dựa trên lý thuyết tập thô: TR (Total Roughness), MMR (Min-Min Roughness) và MDA (Maximum Dependency Attribute), và đề xuất MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), một thuật toán mới cho việc lựa chọn thuộc tính gom cụm theo tiếp cận phân cấp. MAX-MEAN-SU sử dụng Độ không chắc chắn đối xứng (Symmetric Uncertainty - *SU*), một độ đo của Lý thuyết thông tin cho phép lượng hóa mức độ tương quan lẫn nhau giữa hai thuộc tính. Thuộc tính gom cụm được chọn là thuộc tính có độ tương quan trung bình với các thuộc tính khác lớn nhất. Ưu điểm của *SU* so với Độ lợi thông tin (Information Gain - *IG*), là *SU không thiên vị các thuộc tính có nhiều giá trị*. Để đánh giá và so sánh MAX-MEAN-SU với ba kỹ thuật dựa trên lý thuyết tập thô, chúng tôi sử dụng khái niệm "Độ tương tự trung bình bên trong các cụm" của một phép gom cụm để đo lường chất lượng gom cụm của mỗi thuộc tính được chọn bởi mỗi phương pháp. Kết quả thực nghiệm cho thấy chất lượng gom cụm của thuộc tính chọn được bằng phương pháp MAX-MEAN-SU là cao hơn so với các thuộc tính chọn bởi các phương pháp TR, MMR và MDA. Do đó, MAX-MEAN-SU có thể được sử dụng như là một kỹ thuật hiệu quả lựa chọn thuộc tính trong phân cụm phân cấp.

TÀI LIỆU THAM KHẢO

[1] Barbara, D., Li, Y., Couto, J.: COOLCAT: an entropy-based algorithm for categorical clustering. In: *Proc. of CIKM 2002*, 582-589, 2002.

[2] Cao F. Y., Liang J. Y. , Li D. Y., Bai L., A new initialization method for categorical data clustering. *Expert Syst. Appl.* 36, 10223-10228, 2009.

[3] Ganti, V., Gehrke, J., Ramakrishnan, R.: CACTUS - clustering categorical data using summaries. In: *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 73-83, 1999.

[4] Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes. *Information Systems*, 25(5), 345-366, 2000.

[5] Huang, Z.: Extensions to the k-averages algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3), 283-304, 1998.

[6] Han J., and Kamber M., *Data Mining: Concepts and Techniques*, 2nd Morgan Kanufmann Publishers, 2006.

[7] Hassanein W. A., clustering algorithms for categorical data using concepts of significance and dependence of attributes. European Scientific Journal, Vol. 10, No 3, 381-400, 2014.

[8] Herawan, T., Deris, M. M., Abawajy, J. H.: A rough set approach for selecting clustering attribute. *Knowledge-Based Systems* 23, 220-231, 2010.

[9] Jain A. K., Data clustering: 50 years beyond k-averages, Pattern Recogn. Lett., 31(8), 651-666, 2010.

[10] Jyoti Dr., Clustering categorical data using rough sets: a review. *International Journal of Advanced Research in IT and Engineering*, Vol. 2, No. 12, December, 30-37, 2013.

[11] Khandelwal G., and Sharma R., "A Simple Yet Fast Clustering Approach for Categorical Data", *International Journal of Computer Applications* (0975 - 8887), Volume 120 - No.17, 25-30, June 2015.

[12] Mazlack, L. J., He, A., Zhu, Y., Coppock, S.: A rough set approach in choosing clustering attributes. In: *Proceedings of the ISCA 13th International Conference* (CAINE 2000), 1-6, 2000.

[13] Parmar, D., Wu, T., Blackhurst, J., MMR: an algorithm for clustering categorical data using rough set theory. *Data and Knowledge Engineering*, 63, 879-893, 2007.

[14] Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science*, 11, 341-356, 1982.

[15] Suchita S. Mesakar, M. S. Chaudhari, Review Paper On Data Clustering Of Categorical Data. *International Journal of Engineering Research & Technology*, Vol. 1 Issue 10, December, 2012.

[16] Yu L., Liu H.: Feature Selection for High-Dimensional Data: A Fast Correlation Based Filter Solution. ICML 856-863, 2003.

AN APPROACH FOR SELECTING CLUSTERING ATTRIBUTE USING INFORMATION THEORY

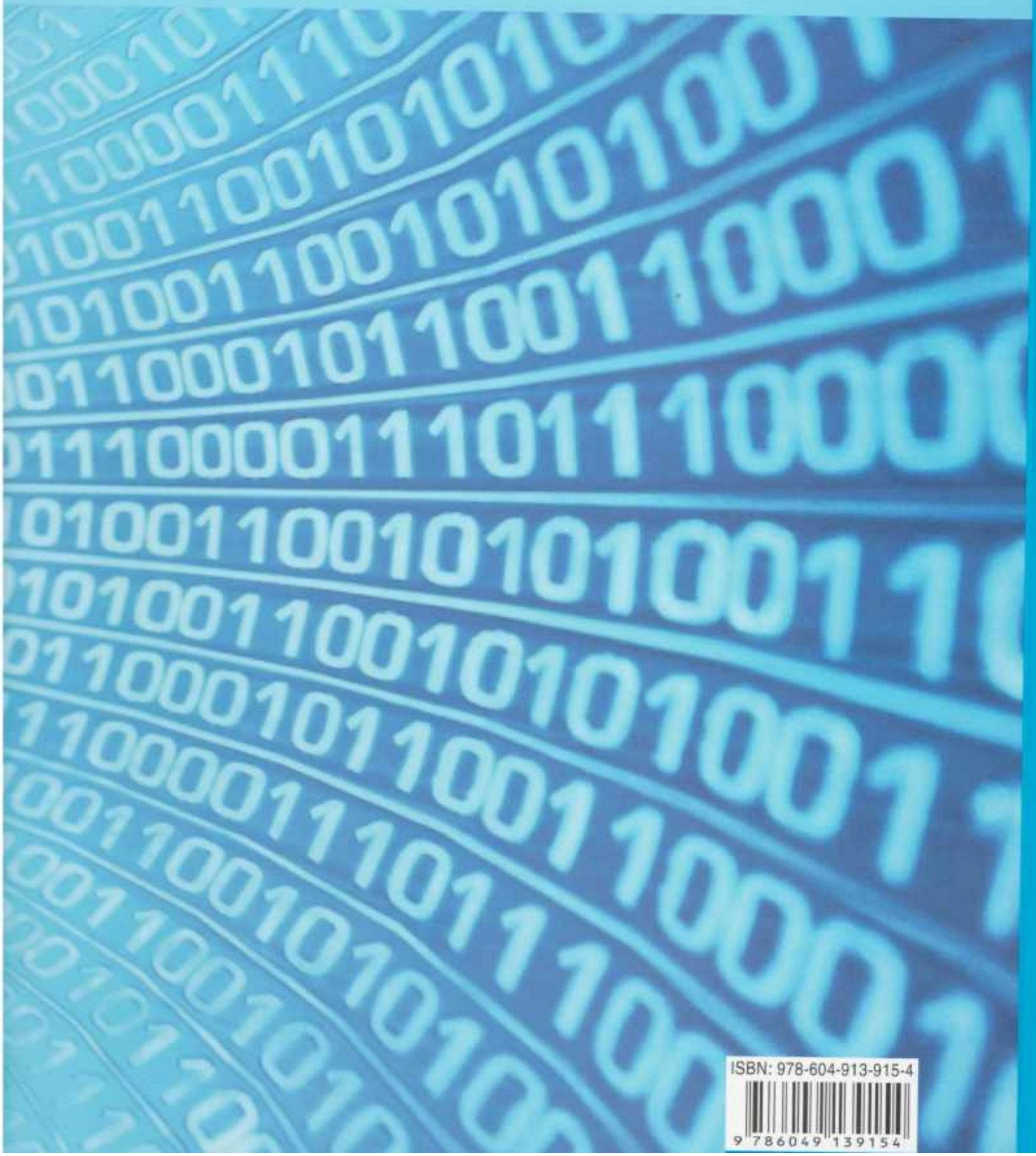
Pham Cong Xuyen, Nguyen Thanh Tung

ABSTRACT: *Clustering problem appears in many different areas. The basic objective of clustering is to group objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. Recently, many researchers have contributed to categorical data clustering, where data objects are made up of categorical attributes. In particular, hierarchical clustering categorical data using the rough set theory has attracted much attention. The key to these approaches is how to select one attribute that is the best to cluster the objects at each time from many candidates of attributes.*

In this paper, we review three rough set based techniques: TR (Total Roughness), MMR (Min-Min Roughness) and MDA (Maximum Dependency Attribute), and propose MAX-MEAN-SU (Maximum Mean of Symmetric Uncertainties), an alternative algorithm for hierarchical clustering attribute selection. MAX-MEAN-SU uses SU (Symmetric Uncertainty), a measure of information theory that allows to quantify the degree of mutual correlation between two attributes, to determine the clustering attribute so that its average correlation with other attributes reaches its maximum value. To evaluate and compare MAX-MEAN-SU with three rough set based techniques, we use the concept of average intra-class similarity to measure the clustering quality of each attribute which is chosen by each method. The experiment results show that the clustering quality of the attribute selected by our method is higher than that of attributes selected by TR, MMR and MDA methods. Therefore, MAX-MEAN-SU can be used as an effective technique to select attributes in hierarchical clustering of categorical data.

Keywords: *Clustering, Categorical Data, Hierarchical Clustering, Rough Set Theory, Clustering Attribute Selection, Symmetric Uncertainty.*

Proceedings of the 12th National Conference on Fundamental and Applied Information Technology Research (FAIR'2019)



ISBN: 978-604-913-915-4