Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00076

MỘT TIẾP CẬN THUẬT TOÁN PHÂN CỤM MỜ MỚI DỰA TRÊN MA TRẬN KẾT HỢP MỜ TRUNG LẬP

Lê Hoàng Sơn[1], Nguyễn Văn Căn[2], Hoàng Việt Long[2], Đoàn Ngọc Tú[2]

[1]Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội

[2] Khoa Công nghệ thông tin, Trường Đại học Kỹ thuật - Hậu cần CAND

sonlh@vnu.edu.vn,cannv@truongt36.edu.vn,longhv08@gmail.com,doanngoctu9893@gmail.com

TÓM TẮT: *Trong bài báo này, chúng tôi đề xuất một thuật toán phân cụm mờ mới dựa trên ma trận kết hợp mờ trung lập. Dữ liệu sau khi được mờ hóa về dạng tập mờ trung lập được sử dụng để tính toán ra ma trận kết hợp mờ. Bằng cách xây dựng chuỗi hữu hạn, ma trận tương đương của ma trận mờ trung lập sẽ được tạo ra và sử dụng để tính lát cắt của ma trận kết hợp mờ. Các cụm sẽ được phát hiện từ ma trận lát cắt sử dụng thủ tục kiểm tra. Thuật toán mới sẽ được thử nghiệm đánh giá về chất lượng cụm và thời gian tính toán.*

Từ khóa: *Phân cụm mờ, tập mờ trung lập, ma trận kết hợp mờ, ma trận lát cắt, chất lượng cụm.*

I. GIỚI THIỆU

Phân cụm dữ liệu có nhiều ứng dụng quan trọng trong khai phá dữ liệu, nhận dạng mẫu, truy hồi thông tin và học máy. Tuy nhiên, trong thực tế các dữ liệu thường phức tạp, thiếu hụt hoặc có tính chất mơ hồ, không chắc chắn. Để giải quyết vấn đề này, lý thuyết tập mờ đã được Zadeh [17] đề xuất trong đó thông tin không chắc chắn được mô hình hóa dưới dạng độ thuộc của phần tử vào một tập. Thuật toán phân cụm trên tập mờ của Zadeh là Fuzzy C-Means (FCM) đưa ra bởi Bezdek năm 1984 [2] và cho đến nay đã được ứng dụng trong nhiều lĩnh vực khác nhau với kết quả khả quan so với thuật toán phân cụm rõ.

Một vấn đề cơ bản trong các nghiên cứu liên quan đến tập mờ truyền thống của Zadeh là khả năng biểu diễn các thông tin liên quan đến tính "không thuộc" và tính "do dự". Tuy nhiên, đơn cử như trường hợp khi bỏ phiếu bầu cử, cử tri có thể bỏ 3 loại phiếu: phiếu đồng ý, phiếu không đồng ý và phiếu trung lập (Không có ý kiến). Khi đó, tập mờ truyền thống của Zadeh không phù hợp để mô hình hóa các thông tin về tính "không thuộc" và tính "do dự". Một số mở rộng của tập mờ truyền thống đã được đề xuất như: tập mờ trực cảm của Antanassov [1] hay tập mờ trung lập của Smarandache [9] trong đó tập mờ trung lập là tổng quát hóa của các tập mờ được đưa ra trước đây như tập mờ, tập mờ trực cảm và tập mờ trực cảm dạng khoảng. Tập mờ trung lập đã được nghiên cứu và áp dụng trong các lĩnh vực khác nhau như hệ thống chẩn đoán y tế [7], hệ hỗ trợ quyết định [8], robot [10], phân tích thông tin xã hội và giáo dục, v.v.

Đối với bài toán phân cụm mờ trên tập mờ trung lập, vấn đề quan trọng nhất là xác định các độ đo tương tự để phân chia phần tử vào các cụm. Ye [11] đề xuất ba giải pháp sử dụng độ đo tương tự cho tập mờ trung lập gồm độ đo Jaccard, Dice và Cosine để áp dụng cho hệ ra quyết định đa tiêu chí với dữ liệu mờ trung lập đơn giản. Trong [13] và [12], Ye tiếp tục đề xuất các phương pháp cải tiến tập mờ trung lập cho hệ chuyên gia ra quyết định bằng các độ đo tương tự mở rộng. Mondal và Pramanik [6] nghiên cứu về độ đo tương tự mờ dựa trên hàm tiếp tuyến và ứng dụng trong y học. Các nghiên cứu về độ tương tự mờ trung lập có thể xem trong [14, 15, 16].

Trong bài báo này, chúng tôi đề xuất một thuật toán phân cụm mờ trên tập mờ trung lập mới sử dụng ma trận kết hợp mờ trung lập. Ý tưởng chính của thuật tóan là xây dựng ma trận kết hợp mờ trung lập từ dữ liệu và sau đó thiết lập ma trận lát cắt (λ–cutting) của ma trận kết hợp tương đương. Từ ma trận lát cắt, các cụm được xác định dựa trên các cột phần tử giống nhau trong ma trận.

Phần tiếp theo của bài báo có cấu trúc như sau: trong phần II, chúng tôi đề xuất thuật toán phân cụm mới thông qua phân tích lý thuyết chi tiết. Phần III là một số kết quả thực nghiệm trên bộ dữ liệu thực. Cuối cùng là kết luận và các hướng phát triển.

II. ĐỀ XUẤT THUẬT TOÁN PHÂN CỤM DỰA TRÊN MA TRẬN KẾT HỢP MỜ TRUNG LẬP

A. Tập mờ trung lập

Cho X là một tập khác rỗng, với một phần tử của X ký hiệu là $x \in X$, tập mờ trung lập A xác định trên không gian nền X được đặc trưng bởi ba hàm số: hàm $T_A(x)$ đo độ thuộc chỉ ra rằng sự kiện x sẽ xảy ra, hàm đo độ trung lập $I_A(x)$ tức là không có ý kiến gì về việc sự kiện x có xảy ra hay không và hàm đo độ không thuộc $F_A(x)$ tin rằng sự kiện x sẽ không xảy ra với $x \in X$. Ở đây $T_A(x)$, $I_A(x)$, $F_A(x) \in [0,1]$ và $0 \leq T_A(x) + I_A(x) + F_A(x) \leq 3$. Như vậy về mặt tập hợp, một tập mờ trung lập được biểu diễn như sau:

$$A = \{\langle x; T_A(x); I_A(x); F_A(x)\rangle : x \in X\}.$$

Cho $A_1 = \{\langle x; T_1(x); I_1(x); F_1(x)\rangle | x \in X\}$ và $A_2 = \{\langle x; T_2(x); I_2(x); F_2(x)\rangle | x \in X\}$ là hai tập mờ trung lập. Khi đó ta có các phép toán tập hợp cơ bản:

a) $A_1 \subseteq A_2$ khi và chỉ khi $T_1(x) \le T_2(x); I_1(x) \ge I_2(x); F_1(x) \ge F_2(x)$;
b) $A_1^c = \{\langle x; F_1(x); I_1(x); T_1(x)\rangle | x \in X\}$;
c) $A_1 \cap A_2 = \{\langle x; \min\{T_1(x); T_2(x)\}; \max\{I_1(x); I_2(x)\}; \max\{F_1(x); F_2(x)\}\rangle | x \in X\}$;
d) $A_1 \cup A_2 = \{\langle x; \max\{T_1(x); T_2(x)\}; \min\{I_1(x); I_2(x)\}; \min\{F_1(x); F_2(x)\}\rangle | x \in X\}$.

B. Đề xuất ma trận kết hợp mờ trung lập

Định nghĩa 1. Cho B_j $(j = 1,2,\dots,n)$ là n tập mờ trung lập. Khi đó $M = (m_{ij})_{n\times n}$ được gọi là ma trận kết hợp mờ trung lập nếu các phần tử $m_{ij} = m(B_i, B_j)$ thỏa mãn tính chất như sau:

a) $0 \le m_{ij} \le 1, \forall i, j = 1, 2, \dots, n.$
b) $m_{ij} = 1$ nếu và chỉ nếu $B_i = B_j$.
c) $m_{ij} = m_{ji},\ \forall i, j = 1, 2, \dots, n.$

Định nghĩa 2. Cho $M = (m_{ij})_{n\times n}$ là một ma trận kết hợp. Ma trận $M^2 = M * M = (\overline{m}_{ij})_{n\times n}$được gọi là ma trận hợp thành của M nếu

$$\overline{m}_{ij} = max_p \left\{min\{m_{ip}, m_{pj}\}\right\}, i, j = 1, 2, \dots, n. \quad (1)$$

Bổ đề 1. Cho $M = (m_{ij})_{n\times n}$ là một ma trận kết hợp. Khi đó ma trận hợp thành M^2 cũng là một ma trận kết hợp.

Chứng minh:

(a): Vì M là một ma trận kết hợp, nên với bất kì $i, j = 1, 2, \dots, n$, ta có $0 \le M_{ij} \le 1$. Do đó ta có

$$0 \le \overline{m}_{ij} = max_p \left\{min\{m_{ip}, m_{pj}\}\right\} \le 1.$$

(b): Vì $m_{ij} = 1$ khi và chỉ khi $B_i = B_j, i, j = 1, 2, \dots, n$, do đó

$\overline{m}_{ij} = max_p \left\{min\{m_{ip}, m_{pj}\}\right\} = 1$ khi và chỉ khi $B_i = B_p = B_j, p = 1,2, \dots, n.$

(c): Vì $M_{ij} = M_{ji}, i, j = 1,2,3, \dots, n$, ta suy ra

$$\overline{m}_{ij} = max_p \left\{min\{m_{ip}, m_{pj}\}\right\} = max_p \left\{min\{m_{pi}, m_{jp}\}\right\}$$
$$= max_p \left\{min\{m_{jp}, m_{pi}\}\right\} = \overline{m}_{ji} \quad \blacksquare$$

Dựa trên Bổ đề 1, chúng ta có thể suy ra bổ đề sau:

Bổ đề 2. Cho $M = (m_{ij})_{n\times n}$ là một ma trận kết hợp. Khi đó với bất kì số nguyên dương p, ta có

$$M^{2p+1} = M^{2p} * M^1 \quad (2)$$

Khi đó ma trận hợp thành M^{2p+1} cũng là một ma trận kết hợp.

Định nghĩa 3. Cho $M = (m_{ij})_{n\times n}$ là một ma trận kết hợp. Nếu $M^2 \subseteq M$ tức là với bất kì $i, j = 1,2, \dots, n$,bất đẳng thức dưới đây thỏa mãn

$$max_p \left\{min\{m_{ip}, m_{pj}\}\right\} \le m_{ij} \quad (3)$$

Khi đó, M được gọi là một ma trận kết hợp tương đương.

Sử dụng nguyên lý bắc cầu của ma trận tương đương, theo [18] ta có thể chứng minh bổ đề dưới đây:

Bổ đề 3. Cho $M = (m_{ij})_{n\times n}$ là một ma trận kết hợp. Khi đó chuỗi hữu hạn ma trận kết hợp của M:

$$M \to M^2 \to M^4 \to \cdots \to M^{2p} \to \cdots \quad (4)$$

thỏa mãn tính chất: tồn tại một số nguyên dương p sao cho $M^{2(p)} = M^{2(p+1)}$, hơn nữa ta có $M^{2(p)}$ cũng là một ma trận kết hợp tương đương.

Dựa vào khái niệm ma trận kết hợp tương đương, chúng tôi đưa ra một số khái niệm như sau:

Định nghĩa 4. Cho $M = (m_{ij})_{n\times n}$ là một ma trận kết hợp tương đương. Khi đó $M_\lambda = (\lambda m_{ij})_{n\times n}$ được gọi là ma trận lát cắt (λ–cutting) của M, trong đó

$$\lambda m_{ij} = \begin{cases} 0, & m_{ij} < \lambda, \\ 1, & m_{ij} \ge \lambda, \end{cases} i, j = 1,2, \dots, n \quad (5)$$

và λ được gọi là độ tin cậy mức $\lambda \in [0, 1]$.

Sử dụng các khái niệm về tập mờ trung lập nêu trên, chúng tôi giới thiệu một thuật toán phân cụm như sau.

C. Phân cụm mờ dựa trên ma trận kết hợp mờ trung lập

Bước 1: Cho $U = \{u_1, u_2, \dots, u_p\}$ là không gian nền và cho $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$ là vector trọng số của phần tử $\alpha_l (l = 1,2, \dots, p)$, với $\alpha_l \in [0,1]$, $l = 1,2, \dots, p$, và $\sum_{l=1}^{p} \alpha_l = 1$. Xét một họ tập mờ trung lập B_j $(j = 1,2, \dots, n)$, với

$$B_j = \left\{ \langle y, T_{B_j}(y_l), I_{B_j}(y_l), F_{B_j}(y_l) \rangle \mid y_l \in U \right\} \tag{6}$$

$$\varphi_{B_j}(y_l) = 3 - T_{B_j}(y_l) - I_{B_j}(y_l) - F_{B_j}(y_l), \qquad j = 1,2, \dots, n.$$

Bước 2: Xây dựng ma trận kết hợp mờ trung lập như sau:

$$m_{ij} = m(B_i, B_j) = \frac{\sum_{l=1}^{p} \left(T_{B_i}(y_l)^2 \, T_{B_j}(y_l)^2 + I_{B_i}(y_l)^2 I_{B_j}(y_l)^2 + F_{B_i}(y_l)^2 \, F_{B_j}(y_l)^2 \right)}{\max\left(\sum_{l=1}^{p} \alpha_l \left(T^2{}_{B_i}(y_l) + I^2{}_{B_i}(y_l) + F^2{}_{B_i}(y_l) + \varphi^2{}_{B_i}(y_l) \right), \sum_{l=1}^{p} \alpha_l \left(T^2{}_{B_j}(y_l) + I^2{}_{B_j}(y_l) + F^2{}_{B_j}(y_l) + \varphi^2{}_{B_j}(y_l) \right) \right)}$$

Bước 3: Nếu ma trận kết hợp mờ trung lập $M = (m_{ij})_{p \times p}$ là một ma trận kết hợp tương đương, thì chúng ta có thể xây dựng ma trận lát cắt $M_\lambda = (\lambda m_{ij})_{n \times n}$ theo Định nghĩa 4. Nếu không, chúng ta sẽ xây dựng chuỗi hữu hạn của ma trận kết hợp M theo bổ đề 3 để tạo ra một ma trận kết hợp tương đương $\bar{M}$ và áp dụng Định nghĩa 4 để xây dựng ma trận lát cắt.

Bước 4: Nếu tất cả phần tử của dòng (cột) thứ i trong M_λ (hoặc $\bar{M}_\lambda$) giống với các phần tử của dòng (cột) thứ j trong M_λ (hoặc $\bar{M}_\lambda$), thì các tập mờ trung lập B_i và B_j là cùng cụm. Bằng quy tắc này, chúng ta có thể phân loại tất cả tập mờ trung lập $B_j (j = 1,2, \dots, n)$.

Các bước của thuật toán phân cụm này có thể được mô tả bởi Hình 1. Bằng việc sử dụng ma trận lát cắt cho ma trận kết hợp tương đương, thuật toán phân cụm mờ trung lập phân loại các tập mờ trung lập dựa trên các mức độ tin cậy λ cho trước. Như vậy, mức độ tin cậy có mối quan hệ với các phần tử của ma trận kết hợp tương đương. Trong các ứng dụng thực tế, ta có thể xác định được các mức độ tin cậy theo các phần tử của ma trận kết hợp tương đương và các tình huống cụ thể. Như vậy, thuật toán sẽ trở nên linh hoạt và có tính khả thi cao hơn. Tuy nhiên, trong nhiều trường hợp, ta có thể sử dụng thuật toán phân cụm tự động mà không cần phải xác định số cụm thủ công. Nói cách khác, thuật toán cần phải có thêm khả năng tự động sinh ra tham số λ một cách tối ưu nhất dựa theo cấu trúc của các cụm. Vấn đề này cần nghiên cứu kỹ trong tương lai.

Hình 1. Mô hình phân cụm mờ trung lập dựa trên ma trận kết hợp

III. THỰC NGHIỆM VÀ ĐÁNH GIÁ

A. Môi trường thực nghiệm

1. Công cụ thực nghiệm

Trong phần này chúng tôi cài đặt và thử nghiệm so sánh thuật toán của chúng tôi với các thuật toán của Ye2014 [13] và Ye2016 [14]. Khác với thuật toán đề xuất của chúng tôi, thuật toán của Ye2014 [13] sử dụng ma trận tương tự để tiến hành xác định cụm trong CSDL. Đối với thuật toán Ye2016 [14], thuật toán của tác giả đã sử dụng ma trận kết hợp mờ trung lập tương đương để tiến hành xác định cụm. Do đó, chúng tôi đã lựa chọn 2 thuật toán Ye2014 và Ye2016 để tiến hành so sánh với thuật toán đề xuất.

Thuật toán được thực hiện trên ngôn ngữ lập trình Matlab 2015a, sử dụng máy tính có CPU Intel(R) Core (TM) i5-2520M tốc độ 2.4 GHz, bộ nhớ RAM 4096 MB và sử dụng hệ điều hành Windows 7 Professional 64 bits.

2. Tập dữ liệu EPPO

EPPO [4] là bộ cơ sở dữ liệu toàn cầu do Ban thư ký của Tổ chức Bảo vệ thực vật châu Âu và Địa Trung Hải (EPPO) tạo ra. Cơ sở dữ liệu này vẫn còn đang được phát triển, mục tiêu của EPPO là thu thập các thông tin về dịch bệnh trong nông nghiệp. Bảng 1 dưới đây cung cấp cho chúng ta các thông tin về các dịch bệnh trên thực vật.

Bảng 1. Bảng mô tả dữ liệu thực nghiệm

Bộ dữ liệu	Số phần tử	Số thuộc tính	Số lớp
eppo_standard_pp1	1452	289	4
eppo_standard_pm8	167	3	2
eppo_standard_pm4	555	35	2

3. Độ đo phân cụm

Độ đo Davies-Bouldin (DB) [3]:

$$DB = \frac{1}{k}\sum_{l=1}^{k} D_l, \tag{8}$$

$$D_l = \max_{l\neq m}\{D_{l,m}\};\ D_{l,m} = (\bar{d}_l + \bar{d}_m)/d_{m,l}.$$

Với $\bar{d}_l, \bar{d}_m$ là các khoảng cách trong nhóm trung bình của các cụm thứ l và thứ m tương ứng, còn $d_{m,l}$ là khoảng cách giữa các cụm này. Với công thức tính như sau:

$$\bar{d}_l = \frac{1}{N_l}\sum_{x_i\in C_l}\|x_i - \bar{x}_l\|\,; d_{l,m} = \|\bar{x}_l - \bar{x}_m\|.$$

Khi các thuật toán được cài đặt thực nghiệm, kết quả độ đo DB nhận được càng nhỏ càng tốt.

Độ đo Simplified Silhouete Width Criterion (SSWC):

$$SSWC = \frac{1}{N}\sum_{j=1}^{N} S_{x_j} \tag{9}$$

$$S_{x_j} = \frac{b_{p,j}-a_{p,j}}{\max\{a_{p,j},b_{p,j}\}}$$

Với độ đo SSWC, khi cài đặt thực nghiệm, giá trị độ đo càng lớn thì thuật toán càng hiệu quả.

Độ đo IFV [5]:

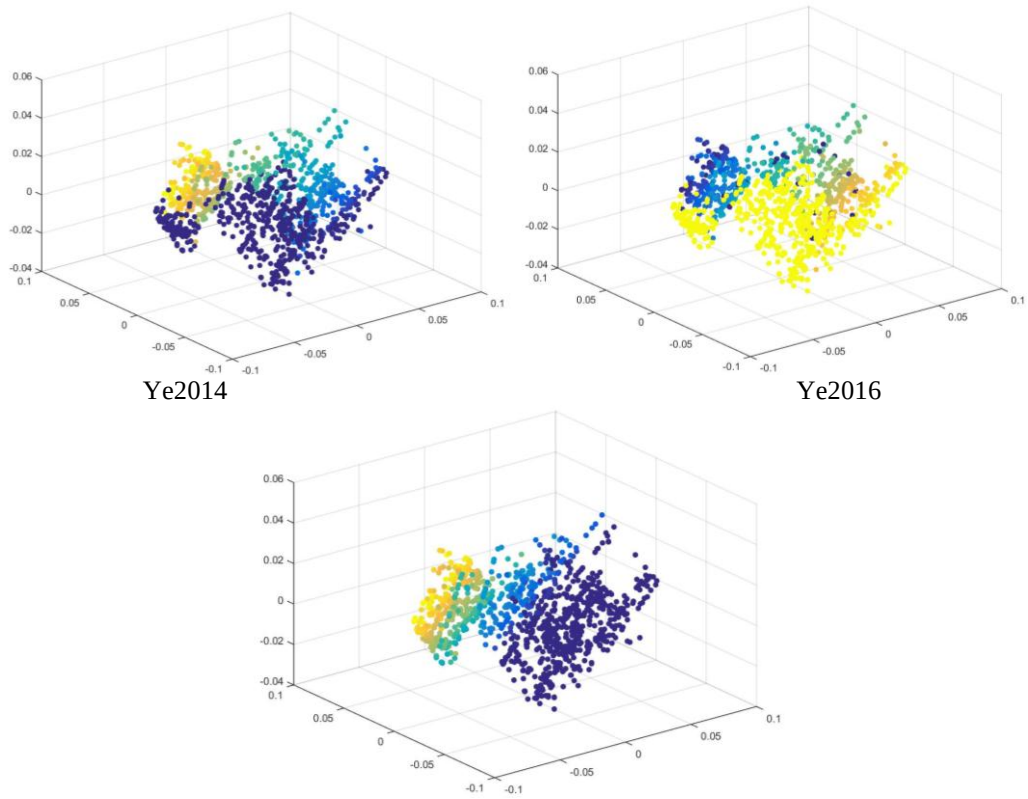
$$IFV = \frac{1}{C}\sum_{j=1}^{C}\left\{\frac{1}{N}\sum_{k=1}^{N} {u_{kj}}^2\left[\log_2 C - \frac{1}{N}\sum_{k=1}^{N}\log_2 u_{kj}\right]^2\right\} \times \frac{SD_{max}}{\sigma_D} \tag{10}$$

$$SD_{max} = \max_{k\neq j}\|V_k - V_j\|^2\ ,\ \overline{\sigma_D} = \frac{1}{C}\sum_{j=1}^{C}\left(\frac{1}{N}\sum_{k=1}^{N}\|X_k - V_j\|^2\right).$$

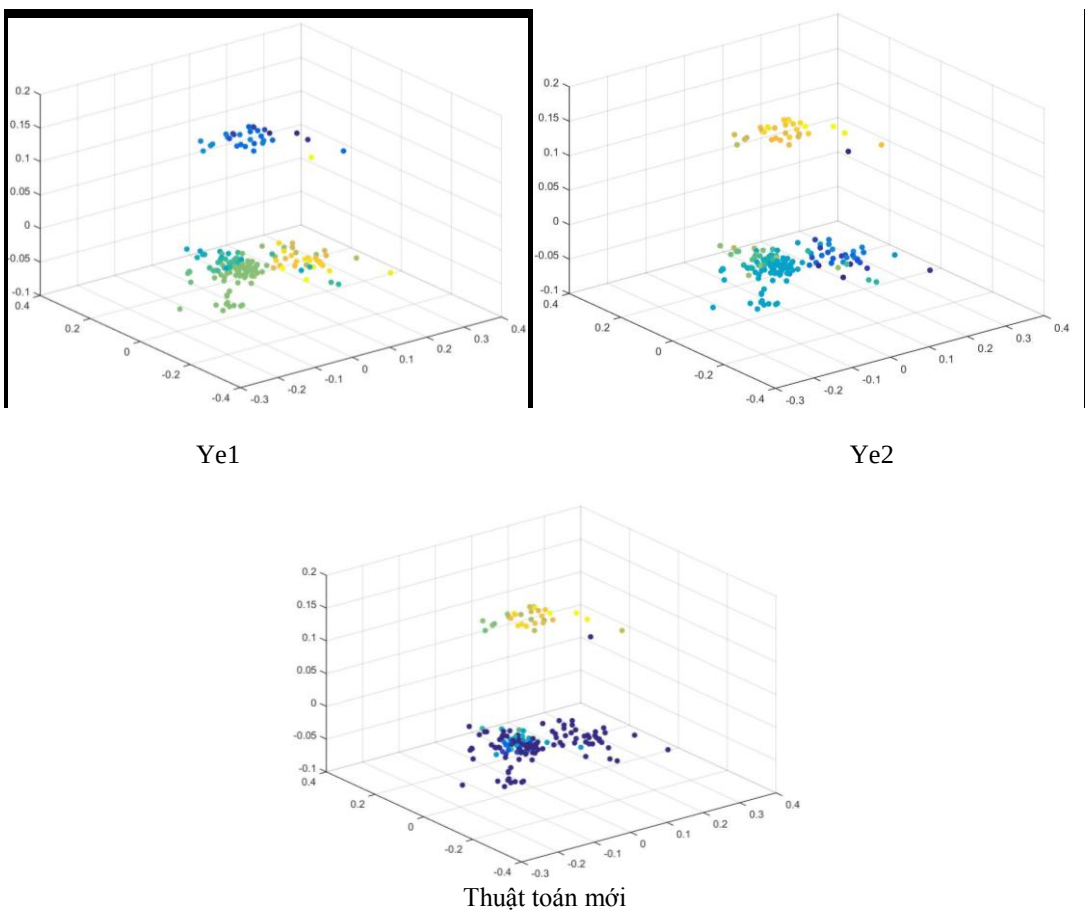
Thuật toán càng tốt thì giá trị IVF càng cao.

B. Đánh giá hiệu suất

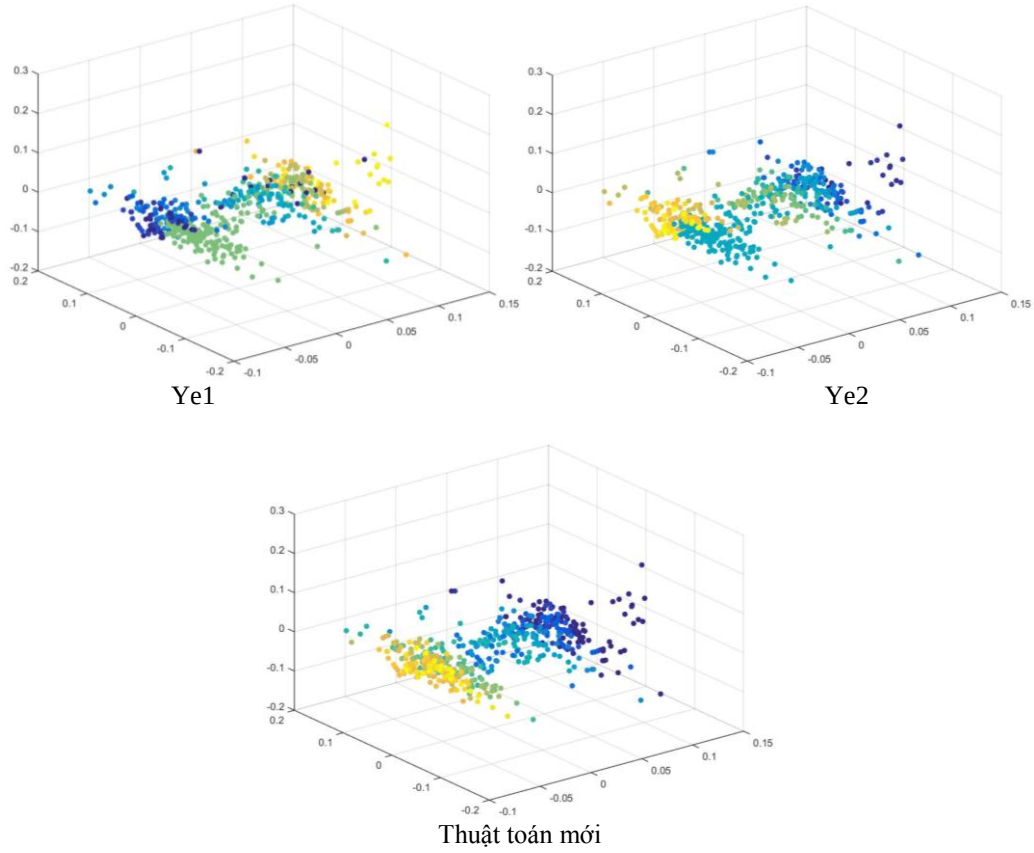
Hình 2, 3, 4 biểu diễn các kết quả của các thuật toán phân cụm. Mỗi màu sắc trong hình đại diện cho một cụm. Số lượng các cụm phụ thuộc vào từng phương pháp và các tham số đầu vào mỗi phương pháp. Ở đây chúng tôi lựa chọn số lượng cụm ở mỗi thuật toán xấp xỉ nhau. Qua mỗi hình và các hình con của nó, ta có thể thấy thuật toán đề xuất biểu diễn rõ ràng các cụm hơn so với những thuật toán khác.



Hình 2. Kết quả phân cụm của 3 thuật toán với bộ dữ liệu eppo_standard_pp1



Hình 3. Kết quả phân cụm của 3 thuật toán với bộ dữ liệu eppo_standard_pm8



Hình 4. Kết quả phân cụm của 3 thuật toán với bộ dữ liệu eppo_standard_pm4

Bảng 2 đưa ra các chỉ số (DB, SSWC, IFV và PBM) sau khi chuẩn hóa các bộ dữ liệu tương ứng với 3 bộ dữ liệu thực nghiệm. Dễ dàng nhận thấy đối với chỉ số DB, thuật toán đề xuất có kết quả không tốt so với Ye2014 và Ye2016, nhưng với chỉ số SSWC và IFV thuật toán của chúng tôi lại cho kết quả tốt hơn.

Bảng 2. Kết quả so sánh giữa thuật toán đề xuất với Ye2014 và Ye2016
(Bôi đậm thể hiện kết quả tốt nhất trong một cột)

Bộ dữ liệu	Thuật toán	DB	SSWC	IFV
eppo_standard_pp1	Ye2014	**28.155232**	0.557163	399.482864
	Ye2016	30.859587	0.712286	559.711
	Thuật toán đề xuất	508.395223	**0.998623**	**589.791505**
eppo_standard_pm4	Ye2014	**5.720429**	0.660127	20036.600822
	Ye2016	**5.720429**	0.712286	559711.195116
	Thuật toán đề xuất	145.697383	**1**	**8769117.402737**
eppo_standard_pm8	Ye2014	38.40345	0.581924	13108.155304
	Ye2016	**3.323803**	0.671745	10186.476571
	Thuật toán đề xuất	72.11886	**1**	**475193.424271**

Bảng 3 đưa ra thời gian phân cụm của các thuật toán ứngvới 3 bộ dữ liệu thực nghiệm.

Bảng 3. So sánh thời gian chạy (giây) giữa các thuật toán

Bộ dữ liệu	Ye21014	Ye2016	Thuật toán đề xuất
eppo_standard_pp1	1501.350482	283.430899	4222.546329
eppo_standard_pm8	41.350550	8.953963	77.954504
eppo_standard_pm4	4535.715869	53.485130	1043.375767

Về thời gian chạy, thuật toán Ye2016 có thời gian thực hiện nhỏ. Nhưng đối với bộ dữ liệu eppo_standard_pm4, thì thuật toán đề xuất chạy nhanh hơn so với thuật toán Ye2014. Về chất lượng cụm, thuật toán đề xuất có chỉ số đánh giá SSWC và IFV cao hơn so với 2 thuật toán Ye2014 và Ye2016. Dựa vào đồ thị biểu diễn có thể nhận thấy kết quả phân cụm của thuật toán đề xuất có tính gom cụm hơn, không rời rạc và ít nhiễu.

IV. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất một thuật toán phân cụm mờ mới dựa trên ma trận kết hợp mờ trung lập. Kết quả thực nghiệm trên bộ dữ liệu EPPO cho thấy chất lượng của thuật toán mới là tốt hơn các thuật toán phân cụm mờ trung lập khác. Các kết quả phân cụm cũng phân bố rõ ràng và it nhiễu và ngoại lệ. Tuy nhiên thời gian tính toán lại lâu hơn các thuật toán khác. Do vậy, trong tương lai chúng tôi sẽ nghiên cứu cải tiến thời gian thực hiện của thuật toán phân cụm mờ trên tập mờ trung lập.

V. TÀI LIỆU THAM KHẢO

[1] Atanassov, K. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems. 20: 87-96.

[2] Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.

[3] Davies, D. L., & Bouldin, D. W. (1979) A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2): 224-227.

[4] EPPO Global Database, https://gd.eppo.int/.

[5] Hu, C., Meng, L., & Shi, W. (2008). Fuzzy clustering validity for spatial data. Geo-spatial information science, 11(3), 191-196.

[6] Mondal, K., & Pramanik, S. (2015). Weighted fuzzy similarity measure based on tangent function and its application to medical diagnosis. International Journal of Innovative Research in Science, Engineering and Technology, 4: 158-164.

[7] Mondaland, K., & Pramanik, S. (2015). Weighted fuzzy similarity measure based on tangent function and its application to medical diagnosis. International Journal of Innovative Research in Science, Engineering and Technology. 4: 158-164.

[8] Pramanik. S., & Chackrabarti, S.N. (2013). A study on problems of construction workers in West Bengal basedon neutrosophic cognitive maps, International Journal of Innovative Research in Science, Engineering and Technology 2, 11: 6387-6394.

[9] Smarandache, F. (1998). Neutrosophy: Neutrosophic probability, set, and logic. American Research Press, Rehoboth.

[10] Smarandache, F., & Vladareanu, L. (2014). Applications of Neutrosophic logic to robotics. Neutrosophic Theoryand Its Applications. 1: 61-66 .

[11] Ye, J,. & Smarandache, F. (2016). Similarity measure of refined single-valued neutrosophic sets and its multicriteria decision making method, Neutrosophic Sets and Systems, 12: 41-44.

[12] Ye, J,. & Zhang, Q.S. (2014) Single valued neutrosophic similarity measures for multiple attribute decision making. Neutrosophic Sets and Systems, 2: 48-54.

[13] Ye, J. (2014). Clustering methods using distance-based similarity measures of single-valued Neutrosophic sets, Journal of Intelligent Systems, 23(4): 379-389.

[14] Ye, J. (2016). A netting method for clustering-simplied neutrosophic information, Soft Computing. Doi:10.1007/s00500-016-2310-z.

[15] Ye, J. (2017). Single-valued neutrosophic similarity measures based on cotangent function and their application in the fault diagnosis of steam turbine, Soft Comput. 21(3): 817-825.

[16] Ye, J., & Fu, J. (2016). Multi-period medical diagnosis method using a single valued neutrosophic similarity measure based on tangent function. Computer Methods and Programs in Biomedicine 123: 142-149.

[17] Zadeh, L. A., (1965). Fuzzy sets. Information and Control, 8(3): 338-353.

[18] P. Z. Wang, (1983) Fuzzy Set Theory and Applications, Shanghai Scientific and Technical Publishers, Shanghai.

A NEW APPROACH FOR FUZZY CLUSTERING BASED ON NEUTROSOPHIC ASSOCIATION MATRIX

Le Hoang Son, Nguyen Van Can, Hoang Viet Long, Doan Ngoc Tu

ABSTRACT*: In this paper, we propose a new fuzzy clutering algorithm based on the neutrosophic association matrix.Data that are fuzzified into neutrosophic sets are used to create the neutrosophic association matrix. By deriving a finite sequence of neutrosophic association matrices, we generate the neutrosophic equivalence matrix and futher its lamda-cutting. Clusters are determined based on the lamda-cutting matrix. We evaluate the algorithms by both the clustering quality and computational time.*

Keywords*: Fuzzy clustering,neutrosophic set, association matrix, cutting matrix, clustering qualiy.*