

A SEMI-SUPERVISED FUZZY CLUSTERING METHOD FOR DATA PARTITION WITH CONFIDENCE PROBLEM BASED ON PICTURE FUZZY CLUSTERING

Phung The Huan¹, Hoang Thi Canh^{1*}, Pham Huy Thong²

¹TNU - University of Information and Communication Technology

²VNU - Information Technology Institute

ARTICLE INFO	ABSTRACT
Received: 21/02/2022 Revised: 20/4/2022 Published: 21/4/2022	Data clustering and applications have received much research attention in recent years. During data collection, it is possible that some data with lower confidence (wrong value, incorrect attribute, etc.). This will reduce the clustering performance with possible outliers and noises. Several research directions have been proposed to solve this problem. First, for data elements with wrong values or wrong attributes can use Safe semi-supervised fuzzy clustering methods. Secondly, for noisy data elements, the concept of Picture Fuzzy Set can be used, although there are some related studies to reduce noises and increase the quality of clustering, it is only on the traditional fuzzy set. In this paper, we propose a new algorithm named as PT2FCM, to handle the problem of data partition with confidence problem. The proposed method is implemented and experimentally compared against the related methods, including the standard Picture fuzzy clustering (FCPFS), and the Confidence-weighted safe semi-supervised clustering (CS3FCM), etc. The experimental results show that the proposed method has better performance comparing to selected methods on the same datasets.
KEYWORDS Fuzzy clustering Semi-supervised fuzzy clustering Safe clustering Confidence weight Picture fuzzy set	

MỘT PHƯƠNG PHÁP PHÂN VÙNG DỮ LIỆU THEO ĐỘ TIN CẬY DỰA TRÊN PHÂN CỤM MỜ VIỄN CẢNH

Phùng Thế Huân¹, Hoàng Thị Canh^{1*}, Phạm Huy Thông²

¹Trường Đại học Công nghệ Thông tin và Truyền thông - ĐHTT Nguyễn

²Viện Công nghệ Thông tin - ĐHQG Hà Nội

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 21/02/2022 Ngày hoàn thiện: 20/4/2022 Ngày đăng: 21/4/2022	Phân cụm dữ liệu và các lĩnh vực ứng dụng là một trong những hướng nghiên cứu nhận được nhiều sự quan tâm từ các nhà khoa học trong những năm gần đây. Trong quá trình thu thập dữ liệu, có thể một số dữ liệu có độ tin cậy thấp hơn (sai giá trị, thuộc tính không chính xác, v.v.) tồn tại trong toàn bộ tập dữ liệu. Điều này sẽ làm giảm hiệu suất phân cụm với các nhiễu và ngoại lệ có thể xảy ra. Một số hướng nghiên cứu đã được đưa ra để giải quyết vấn đề này. Thứ nhất, đối với các dữ liệu sai giá trị, sai thuộc tính có thể sử dụng các phương pháp phân cụm bán giám sát mờ an toàn. Thứ hai, đối với các điểm dữ liệu nhiễu có thể sử dụng khái niệm tập mờ viễn cảnh, cho dù đã có một số nghiên cứu liên quan nhằm tăng chất lượng phân cụm, tuy nhiên chỉ dừng lại ở tập mờ truyền thống. Trong bài báo này, chúng tôi đề xuất một phương pháp mới trong phân vùng dữ liệu theo độ tin cậy dựa trên phân cụm mờ viễn cảnh có tên gọi PT2FCM. Thuật toán đề xuất được so sánh thực nghiệm với một số phương pháp liên quan như phân cụm bán giám sát mờ trên tập mờ viễn cảnh (FCPFS), phân cụm bán giám sát mờ an toàn (CS3FCM), v.v. Các kết quả thực nghiệm cho thấy, phương pháp đề xuất có chất lượng phân cụm tốt so với các phương pháp liên quan trong cùng tập dữ liệu.
TỪ KHÓA Phân cụm mờ Phân cụm bán giám sát mờ Phân cụm an toàn Trọng số an toàn Tập mờ viễn cảnh	

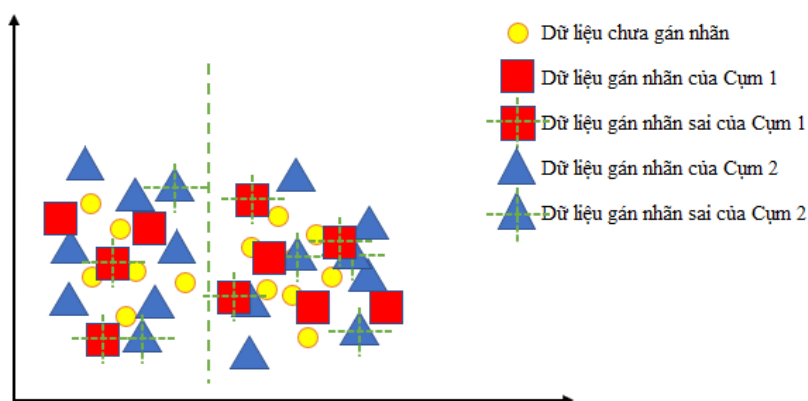
DOI: <https://doi.org/10.34238/tnu-jst.5563>

* Corresponding author. Email: htcanh@ictu.edu.vn

1. Giới thiệu

Ngày nay, do sự gia tăng của các nguồn dữ liệu đến từ các mạng xã hội, kênh đa phương tiện như Facebook, Amazon, Youtube, v.v. nên nhu cầu cần thiết được đặt ra là cấu trúc phân cấp dữ liệu lớn để truy cập và truy xuất một cách nhanh chóng, dễ dàng. Dữ liệu có nhiều định dạng khác nhau như văn bản, hình ảnh, âm thanh, video,... Kỹ thuật phân cụm dữ liệu đã được áp dụng trong nhiều lĩnh vực như ứng dụng trong sinh học [1], ứng dụng trong thư viện [2], bảo hiểm [3], tài chính [4], phân đoạn hình ảnh y tế [5], [6],... Tuy nhiên, trong quá trình thu thập dữ liệu có thể tồn tại một số dữ liệu có độ tin cậy thấp hơn trong toàn bộ tập dữ liệu. Điều này sẽ ảnh hưởng đến hiệu suất của các kết quả phân cụm dữ liệu.

Vấn đề phân cụm dữ liệu theo độ tin cậy được coi là việc giải quyết bài toán phân cụm trong trường hợp một số dữ liệu được gán nhãn không chính xác. Như được minh họa trong Hình 1, giả sử tập dữ liệu bao gồm 2 cụm và có một số dữ liệu đã được gán nhãn (hình vuông thể hiện dữ liệu đã được gán nhãn thuộc Cụm 1 và các hình tam giác thể hiện dữ liệu đã được gán nhãn thuộc Cụm 2), hình tròn thể hiện các điểm dữ liệu chưa được gán nhãn. Đường nét đứt ngầm hiển thị ranh giới giữa hai cụm. Ngoài ra còn một số dữ liệu được gán nhãn không chính xác, được biểu diễn bằng dấu thập phía trên các ký hiệu của dữ liệu được gán nhãn. Mục tiêu của bài toán này là tìm ra đường ranh giới “tốt nhất” giữa hai cụm với các dữ liệu được gán nhãn chính xác và không chính xác.



Hình 1. Phân cụm dữ liệu theo độ tin cậy

Một trong những cách tiếp cận điển hình để giải quyết vấn đề phân cụm dữ liệu theo độ tin cậy là phân cụm bán giám sát mờ. Được khởi xướng bởi Pedrycz và Waletzky [7], ý tưởng chính là sử dụng các kết quả phân loại có sẵn như một phần của quá trình phân cụm với việc bổ sung vector thể hiện hai giá trị giữa dữ liệu được gán nhãn và không được gán nhãn. Ngoài ra, phân cụm bán giám sát mờ an toàn cũng đã được sử dụng để giải quyết vấn đề phân cụm dữ liệu theo độ tin cậy. Ý tưởng chính của cách tiếp cận này như được mô tả trong nghiên cứu của Gan và đồng nghiệp [8]-[11] bao gồm 2 bước chính: i) tính toán trọng số tin cậy của dữ liệu được gán nhãn bằng một đồ thị cục bộ; ii) xây dựng và xác định các trung tâm cụm và các giá trị phân tử mờ theo dữ liệu được gán nhãn có trọng số tin cậy cao. Một trong những cách tiếp cận khác đó là sử dụng tập mờ viền cảnh [12], là một sự khái quát và mở rộng của tập mờ truyền thống với tập mờ trực cảm. Các mô hình dựa trên mờ viền cảnh có thể được áp dụng cho nhiều tình huống liên quan đến các lựa chọn như: đồng ý, do dự, không đồng ý và từ chối trả lời. Các tình huống này có thể cho kết quả tốt hơn trên các thuật toán phân cụm dựa trên tập mờ trực cảm hoặc tập mờ truyền thống. Trong bài báo này, chúng tôi đề xuất một phương pháp mới trong phân vùng dữ liệu theo độ tin cậy dựa trên phân cụm mờ viền cảnh có tên gọi PT2FCM. Phương pháp đề xuất được thực nghiệm và so sánh đem lại hiệu suất tốt so với các phương pháp liên quan.

Các phần tiếp theo của bài báo được trình bày như sau: Trong phần 2 chúng tôi trình bày một số phương pháp tiếp cận trong phân cụm bán giám sát mờ theo độ tin cậy và các ưu nhược điểm.

Chi tiết của phương pháp đề xuất sẽ được trình bày trong phần 3. Trong phần 4 chúng tôi tiến hành đánh giá thực nghiệm, so sánh hiệu suất phân cụm của phương pháp PT2FCM và các phương pháp liên quan. Cuối cùng, phần 5 đưa ra các kết luận và đóng góp của bài báo, đề xuất hướng phát triển trong thời gian tới.

2. Một số phương pháp tiếp cận trong phân cụm bán giám sát mờ theo độ tin cậy

2.1. Phương pháp phân cụm bán giám sát mờ tiêu chuẩn (SSFCM)

SSFCM là phương pháp phân cụm bán giám sát mờ tiêu chuẩn được đề xuất bởi Pedrycz và Waletzky [7]. Trong hàm mục tiêu của phương pháp SSFCM gồm 2 thành phần: thành phần học không giám sát và thành phần học có giám sát. Với hàm mục tiêu được biểu diễn như sau:

$$J_s = \sum_{k=1}^N \sum_{i=1}^C u_{ik}^m d_{ik}^2 + \alpha \sum_{k=1}^N \sum_{i=1}^C (u_{ik} - f_{ik} b_k)^m d_{ik}^2 \rightarrow \text{Min} \quad (1)$$

Trong đó, tập dữ liệu $X = \{X_1, X_2, \dots, X_k, \dots, X_N\}$ với số điểm dữ liệu N , số cụm: (C) , độ thuộc của điểm dữ liệu k đối với cụm i : (u_{ki}) , khoảng cách d_{ki} từ điểm dữ liệu k đến tâm cụm V_i , độ thuộc f_{ki} được gán nhãn k trong cụm i , số mờ m .

Tham số α được sử dụng để cân bằng giữa các thành phần học giám sát và học không giám sát, trong đó b_k được sử dụng để phân biệt giữa các phần tử được gán nhãn và không được gán nhãn.

$$f_{ik} = \begin{cases} 1 & \text{neá } X_k \text{ ñöôc gán nhaõ} \\ 0 & \text{neá ngöôc laï} \end{cases} \quad (2)$$

$$b_k = \begin{cases} 1 & \text{neá } X_k \text{ ñöôc gán nhaõ} \\ 0 & \text{neá ngöôc laï} \end{cases} \quad (3)$$

Các tâm cụm V_i và độ thuộc u_{ik} được tính toán như sau:

$$V_i = \frac{\sum_{k=1}^N u_{ik}^2 x_k + \alpha \sum_{k=1}^N (u_{ik} - f_{ik} b_k)^2 x_k}{\sum_{k=1}^N u_{ik}^2 + \alpha \sum_{k=1}^N (u_{ik} - f_{ik} b_k)^2}, \quad \forall i = 1, \dots, C \quad (4)$$

$$u_{ik} = \frac{1}{1 + \alpha} \left\{ \frac{1 + \alpha \left(1 - b_k \sum_{j=1}^C f_{ij} \right)}{\sum_{j=1}^C \frac{d_{ik}^2}{d_{jk}^2}} + \alpha f_{ik} b_k \right\}, \quad \forall i = 1, \dots, C, \quad k = 1, \dots, L \quad (5)$$

Phương pháp phân cụm bán giám sát mờ tiêu chuẩn SSFCM có hiệu suất phân cụm tốt do trong phương pháp phân cụm bán giám sát mờ có sử dụng các thông tin bổ trợ, là các thông tin do người dùng đưa vào để hướng dẫn quá trình phân cụm. Phương pháp này có thể đạt hiệu suất cao với không nhiều số lượng thông tin bổ trợ được gán nhãn cho các điểm dữ liệu. Tuy nhiên, hiệu suất của phương pháp phụ thuộc vào số cụm và việc khởi tạo ngẫu nhiên các giá trị tâm cụm ban đầu.

2.2. Phương pháp phân cụm bán giám sát an toàn có trọng số độ tin cậy (CS3FCM)

Trong phương pháp CS3FCM [8] mỗi một phần tử khác nhau thì có một ảnh hưởng khác nhau đến hiệu suất phân cụm. Về mặt hình thức, có 2 tập dữ liệu: tập thứ nhất là $X = [x_1, x_2, \dots, x_l]$ là tập dữ liệu được gán nhãn và tập thứ hai $X_u = [x_{l+1}, x_{l+2}, \dots, x_n]$ là tập dữ liệu không được gán nhãn. Trong đó C là số cụm, phần tử x_k có nhãn $y_k \in \{1, \dots, c\}$. Trong phương pháp CS3FCM,

các tác giả đã sử dụng FCM để chia tất cả các điểm dữ liệu thành các cụm sau đó tính toán ma trận phân hoạch $U = [u]_{c \times n}$ và ước lượng các nhãn đầu ra $Y = [y_1, \dots, y_l, y_{l+1}, \dots, y_n]$ sử dụng thuật toán đối sánh Kuhn–Munkres [13], nhãn ước lượng $y_k \mid_{k=1}^l$ và nhãn chính xác $y_k \mid_{k=1}^l$ được so sánh để có được ma trận N_c bao gồm các phần tử p_{ij} , trong đó p_{ij} đo lường khả năng của nhãn i^{th} được phân lớp vào lớp j^{th} .

$$N_c = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & \cdots & p_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ p_{c1} & p_{c2} & \cdots & p_{cc} \end{bmatrix} \quad (6)$$

$$\text{Điều kiện } \sum_{i=1}^c u_{ij} = 1 \text{ và } 0 \leq p_{ij} \leq 1$$

Đối với một phần tử được gán nhãn x_k , nếu $y_k = y_k$ và p_{y_k, y_k} là cao thì độ an toàn, tin cậy của phần tử đó là cao. Trọng số s_k của x_k có thể được tính toán như sau:

$$s_k = \begin{cases} p_{y_k, y_k} \times u_{y, k} & \text{if } y_k = y_k \\ p_{y_k, y_k} \times (1 - u_{y, k}) & \text{otherwise} \end{cases} \quad (7)$$

Gan và cộng sự đã xây dựng một biểu đồ cục bộ $W = [w_{kr}]_{m \times n}$ để định nghĩa các hàng xóm cho các phần tử được gán nhãn và xác định trọng số của biểu đồ:

$$w_{kr} = \begin{cases} \exp\left\{-\frac{\|x_k - x_r\|_2^2}{\delta^2}\right\} & \text{nếu } x_r \in N_p(x_k) \text{ và } y_k = y_r \\ 0 & \text{nếu ngược lại} \end{cases} \quad (8)$$

Trong đó, $N_p(x_k)$ đại diện cho tập dữ liệu của p hàng xóm gần nhất của các phần tử được gán nhãn x_k , trong khi x_r đại diện cho các phần tử không được gán nhãn. Hàm mục tiêu của CS3FCM được tính như sau:

$$J_c = \sum_{k=1}^N \sum_{i=1}^C u_{ik}^m d_{ik}^2 + \lambda_1 \sum_{k=1}^l s_k \sum_{i=1}^C (u_{ik} - f_{ik})^2 d_{ik}^2 + \lambda_2 \sum_{k=1}^l \frac{1}{s_k} \sum_{r=l+1}^N w_{kr} \sum_{i=1}^C (u_{ik} - u_{ir})^2 \rightarrow \text{Min} \quad (9)$$

$$\sum_{i=1}^C u_{ik} = 1, \forall k = 1, \dots, N$$

Độ thuộc của u_{ik} đối với dữ liệu được gán nhãn x_k được tính toán như sau:

$$u_{ik} = \frac{p_{ik} + \frac{1 - \sum_{j=1}^C p_{jk}}{\sum_{j=1}^C q_{jk}}}{q_{ik}} \quad (10)$$

Độ thuộc của u_{ir} đối với dữ liệu không được gán nhãn x_r được tính toán như sau:

$$z_{ir} + \frac{1 - \sum_{j=1}^C \frac{z_{ir}}{t_{ir}}}{\sum_{j=1}^C \frac{1}{t_{ir}}} \quad (11)$$

$$u_{ir} = \frac{z_{ir}}{t_{ir}}$$

Tâm cụm v_i được tính bởi:

$$v_i = \frac{\sum_{k=1}^N u_{ik}^2 x_k + \lambda_1 \sum_{k=1}^l s_k (u_{ik} - f_{ik})^2 x_k}{\sum_{k=1}^N u_{ik}^2 + \lambda_1 \sum_{k=1}^l s_k (u_{ik} - f_{ik})^2} \quad (12)$$

Phương pháp phân cụm bán giám sát mờ an toàn có trọng số tin cậy CS3FCM có hiệu suất phân cụm tốt hơn so với các phương pháp SSFCM. Trong phương pháp phân cụm bán giám sát mờ an toàn có trọng số CS3FCM, thuật toán tiến hành so sánh và tính toán các mức độ ảnh hưởng của các điểm dữ liệu. Các điểm dữ liệu nếu có mức độ ảnh hưởng lớn tới các điểm còn lại thì mang trọng số lớn, ngược lại các điểm dữ liệu có mức độ ảnh hưởng nhỏ thì có trọng số nhỏ. Tuy nhiên, phương pháp này có bất lợi về thời gian chạy do quá trình xây dựng các đồ thị cục bộ, trải qua quá trình kiểm tra và cập nhật các nhãn của các điểm dữ liệu.

2.3. Phương pháp phân cụm mờ viễn cảnh (FCPFS)

Khái niệm tập mờ viễn cảnh [12] được đưa ra lần đầu vào năm 2014 bởi Bùi Công Cường trên cơ sở mở rộng và tổng quát hóa tập mờ [14] và tập mờ trực cảm [15] để đề xuất khái niệm tập mờ viễn cảnh. Một tập mờ viễn cảnh trong một tập nền không rỗng X được định nghĩa như sau:

$$A = \{ (x, \mu A(x), \eta A(x), \xi A(x)) \mid x \in X \} \quad (13)$$

Trong đó, $\mu A(x)$ là độ khẳng định của mỗi phần tử $x \in N$, $\eta A(x)$ là độ trung lập (độ do dự) và $\xi A(x)$ là độ phủ định, thỏa mãn các ràng buộc:

$$0 \leq \mu_{ij}, \eta_{ij}, \xi_{ij} \leq 1, 0 \leq \mu A(x) + \eta A(x) + \xi A(x) \leq 1 \quad (14)$$

Độ từ chối của một phần tử ký hiệu là $\vartheta A(x)$ được tính như sau:

$$\vartheta A(x) = 1 - (\mu A(x) + \eta A(x) + \xi A(x)) \quad (15)$$

Phạm Huy Thông và Lê Hoàng Sơn đã đề xuất một thuật toán phân cụm mờ viễn cảnh (FCPFS) [16], với hàm mục tiêu được mô tả như sau:

$$J_m(U, \eta, \xi, V) = \sum_{i=1}^N \sum_{j=1}^C (\mu_{ij} (2 - \xi_{ij}))^m \|x_i - v_j\|^2 + \sum_{i=1}^N \sum_{j=1}^C \eta_{ij} (\log \eta_{ij} + \xi_{ij}) \rightarrow \text{Min} \quad (16)$$

Với các ràng buộc được định nghĩa như sau:

$$0 \leq \mu_{ij}, \eta_{ij}, \xi_{ij} \leq 1, 0 \leq \mu_{ij} + \eta_{ij} + \xi_{ij} \leq 1, 1 \leq i \leq N, 1 \leq j \leq C \quad (17)$$

$$\sum_{j=1}^C \mu_{ij} (2 - \xi_{ij}) = 1, 1 \leq i \leq N \quad (18)$$

$$\sum_{j=1}^C (\eta_{ij} + \frac{1}{C} \xi_{ij}) = 1, 1 \leq i \leq N \quad (19)$$

Sử dụng phương pháp Lagrange, Phạm Huy Thông và cộng sự đã tính toán được nghiệm tối ưu của hàm mục tiêu (16) là:

$$v_j = \frac{\sum_{i=1}^N (\mu_{ij}(2 - \xi_{ij}))^m x_i}{\sum_{i=1}^N (\mu_{ij}(2 - \xi_{ij}))^m} \quad (20)$$

$$\mu_{ij} = \frac{1}{(2 - \xi_{ij}) \sum_{k=1}^C \left(\frac{\|x_i - v_j\|^2}{\|x_i - v_k\|^2} \right)^{1/(m-1)}} \quad (21)$$

$$\eta_{ij} = \frac{\exp(-\xi_{ij})}{\sum_{k=1}^C \exp(-\xi_{ik})} \left(1 - \frac{1}{C} \sum_{k=1}^C \xi_{ik} \right) \quad (22)$$

$$\xi_{ij} = 1 - (\mu_{ij} + \eta_{ij}) - (1 - (\mu_{ij} + \eta_{ij})^\alpha)^{1/\alpha} \quad (23)$$

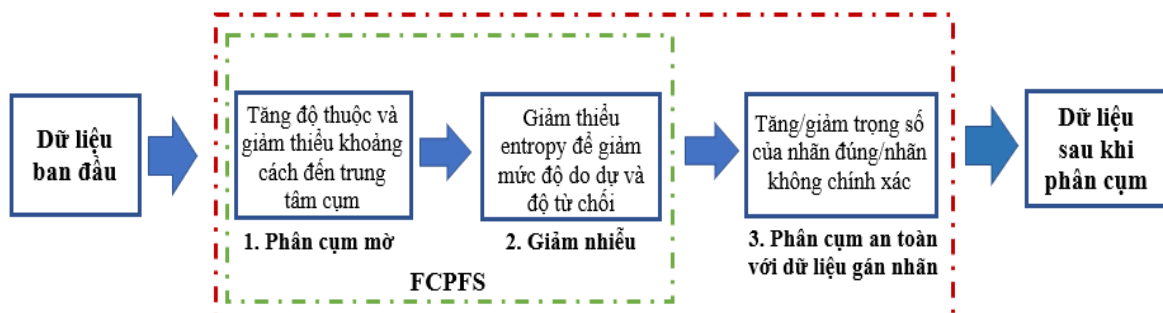
Trong đó, $\alpha \in (0,1)$ được gọi là hệ số mũ được sử dụng để điều khiển độ từ chối của các tập mờ viễn cảnh.

Phương pháp phân cụm mờ viễn cảnh FCPFS có hiệu suất phân cụm tốt hơn so với các phương pháp đã trình bày ở trên như SSFCM, CS3FCM, đặc biệt đối với các dạng dữ liệu có nhiễu và dữ liệu bị gán nhãn sai. Do trong phương pháp phân cụm mờ viễn cảnh FCPFS có sử dụng khái niệm tập mờ viễn cảnh, trong đó bao gồm các thuộc tính như độ thuộc, độ trung lập và độ từ chối của một điểm dữ liệu đối với các cụm. Các thuộc tính này có thể cho kết quả phân cụm tốt hơn hiệu suất phân cụm đối với tập mờ và mờ trực cảm.

3. Phương pháp đề xuất

3.1. Ý tưởng thuật toán

Phương pháp phân cụm bán giám sát mờ an toàn viễn cảnh được đề xuất PT2FCM dựa trên ý tưởng kết hợp phân cụm bán giám sát mờ an toàn và tập mờ viễn cảnh, được biểu diễn trong hàm mục tiêu mới bao gồm 3 thành phần chính. Ý tưởng và các bước thực hiện được mô tả trong Hình 2. Giai đoạn đầu tiên của PT2FCM là giảm thiểu khoảng cách giữa các phần tử dữ liệu và trung tâm cụm thông qua độ mờ viễn cảnh. Giai đoạn thứ hai là để xử lý "dữ liệu nhiễu" bằng cách tích hợp đại lượng entropy giữa các mức độ do dự và độ từ chối của mô hình tập mờ viễn cảnh. Giai đoạn thứ ba nhằm mục đích phối hợp phân cụm bán giám sát mờ an toàn với dữ liệu đã gán nhãn để xử lý dữ liệu chưa chính xác. Kết quả đầu ra của thuật toán là các cụm cuối cùng với độ tin cậy và mức độ chính xác. Chi tiết thuật toán và các công thức theo ý tưởng này sẽ được trình bày trong phần tiếp theo.



Hình 2. Mô hình thuật toán đề xuất PT2FCM

3.2. Chi tiết thuật toán

3.2.1. Mô hình hóa

Từ các ý tưởng chính ở trên, phần này sẽ hình thành mô hình hoá phương pháp đề xuất. Hàm mục tiêu bao gồm 3 thành phần, được hiển thị như sau:

$$J = \sum_{k=1}^N \sum_{j=1}^C (\mu_{kj} (2 - \xi_{kj}))^2 \|X_k - V_j\|^2 + \sum_{k=1}^N \sum_{j=1}^C \eta_{kj} (\log \eta_{kj} + \xi_{kj}) + \sum_{k=1}^L \sum_{j=1}^C \frac{(\mu_{kj} (2 - \xi_{kj}) f_{kj})^2}{1 + (\mu_{kj} (2 - \xi_{kj}) f_{kj})^2} \|X_k - V_j\|^2 \rightarrow \text{Min} \quad (24)$$

Với các ràng buộc:

$$\mu_{kj} + \eta_{kj} + \xi_{kj} \leq 1 \quad (25)$$

$$\sum_{j=1}^C (\eta_{kj} + \frac{\xi_{kj}}{C}) = 1 \quad (26)$$

$$\sum_{j=1}^C (\mu_{kj} (2 - \xi_{kj})) = 1 \quad (27)$$

$$k = 1..N, j = 1..C$$

Trong đó tập dữ liệu $X = \{X_1, X_2, \dots, X_N\}$ với số lượng phần tử N , số các phần tử được gán nhãn trong X : $L < N$; số lượng cụm (C); μ_{kj} , η_{kj} , ξ_{kj} lần lượt là độ thuộc, độ do dự, độ từ chối của phần tử X_k đối với cụm j .

Trong hàm mục tiêu (24), 2 thành phần đầu tiên là các thành phần của phân cụm mờ viễn cảnh gốc (FCPFS) [16], thành phần thứ 3 tương ứng với phân cụm bán giám sát mờ an toàn trên phân cụm mờ viễn cảnh.

- Phần đầu tiên $\sum_{k=1}^N \sum_{j=1}^C (\mu_{kj} (2 - \xi_{kj}))^2 \|X_k - V_j\|^2$ đại diện cho phân cụm mờ viễn cảnh.
- Phần thứ hai $\sum_{k=1}^N \sum_{j=1}^C \eta_{kj} (\log \eta_{kj} + \xi_{kj})$ đại diện cho thông tin entropy giúp giảm nhiễu dữ liệu

thông qua các mức độ do dự và từ chối của một điểm dữ liệu.

- Phần thứ ba $\sum_{k=1}^L \sum_{j=1}^C \frac{(\mu_{kj} (2 - \xi_{kj}) f_{kj})^2}{1 + (\mu_{kj} (2 - \xi_{kj}) f_{kj})^2} \|X_k - V_j\|^2$ là thành phần cho các phần tử dữ liệu

được gán nhãn, trong đó $k = 1..L$ và L là số phần tử dữ liệu được gán nhãn. Phần tử số $(\mu_{kj} (2 - \xi_{kj}) f_{kj})^2$ mô tả phân cụm bán giám sát mờ, trong đó f_{kj} là một giá trị hằng số, có giá trị bằng 1 hoặc 0.

$$f_{kj} = \begin{cases} 1 & \text{nếu phần tử nằm trong cụm } j \\ 0 & \text{nếu ngược lại} \end{cases} \quad (28)$$

Phần mẫu số $1 + (\mu_{kj} (2 - \xi_{kj}) f_{kj})^2$ mô tả phân cụm bán giám sát an toàn. Ý nghĩa của thành phần này như sau: Sau khi phân cụm, nếu bất kỳ điểm dữ liệu nào được gán đúng nhãn, trọng số sẽ được tăng lên; nếu không, trọng số sẽ được giảm xuống.

• Thông tin bổ trợ cho phân cụm bán giám sát mờ là các cấp độ thành viên trước đó. Sử dụng thuật toán FCPFS ban đầu để phân cụm tất cả dữ liệu bao gồm cả dữ liệu được gán nhãn và không được gán nhãn. Từ đó tính được 4 giá trị $(\bar{\mu}_{kj}, \bar{\eta}_{kj}, \bar{\xi}_{kj}, \bar{V})$ để làm cơ sở tính toán cho tất cả các phần tử dữ liệu.

3.2.2. Giải nghiệm

Các phương án tối ưu của mô hình được xác định trong các công thức (24–27) thông qua phương pháp Lagrange.

$$V_j = \frac{\sum_{k=1}^N (\mu_{kj} (2 - \xi_{kj}))^2 X_k + \sum_{k=1}^L \frac{(\mu_{kj} (2 - \xi_{kj}) f_{kj})^2}{1 + (\mu_{kj} (2 - \xi_{kj}) f_{kj})^2} X_k}{\sum_{k=1}^N (\mu_{kj} (2 - \xi_{kj}))^2 + \sum_{k=1}^L \frac{(\mu_{kj} (2 - \xi_{kj}) f_{kj})^2}{1 + (\mu_{kj} (2 - \xi_{kj}) f_{kj})^2}} \quad (29)$$

Độ thuộc u của dữ liệu được gán nhãn như sau:

$$\mu_{kj} = \frac{1}{(2 - \xi_{kj}) \sum_{i=1}^C \left(\frac{(f_{ki})^2}{1 + \frac{(f_{ki})^2}{(\bar{\mu}_{ki} (2 - \bar{\xi}_{ki}) f_{ki})^2}} \right) \|X_k - V_i\|^2} \quad (30)$$

Độ thuộc u của dữ liệu chưa được gán nhãn như sau: $\mu_{kj} = \frac{1}{(2 - \xi_{kj}) \sum_{i=1}^C \frac{\|X_k - V_j\|^2}{\|X_k - V_i\|^2}}$ (31)

Các độ thuộc khác được tính toán như sau: $\eta_{kj} = (1 - \frac{1}{C} \sum_{i=1}^C \xi_{ki}) \frac{e^{-\xi_{kj}}}{\sum_{i=1}^C e^{-\xi_{ki}}}$ (32)

$$\xi_{kj} = 1 - (\mu_{kj} + \eta_{kj}) - (1 - (\mu_{kj} + \eta_{kj})^\alpha)^\frac{1}{\alpha} \quad (33)$$

3.2.3. Chi tiết thuật toán PT2FCM

Thuật toán của phân cụm bán giám sát mờ an toàn trên tập mờ viễn cảnh (PT2FCM) trước hết sử dụng FCPFS để phân vùng tất cả dữ liệu với kết quả đóng vai trò là đầu vào của phương pháp được đề xuất. Phương pháp PT2FCM mở rộng thuật toán FCPFS với việc thêm phần tử bán giám sát an toàn cho dữ liệu được gán nhãn. Phương pháp phân vùng dữ liệu này không những đem lại hiệu quả cao đối với dữ liệu được gán nhãn mà còn giảm số lượng nhãn nghi ngờ. Thuật toán đề xuất được trình bày trong Thuật toán 1 dưới đây:

Thuật toán 1. Các bước chính của thuật toán PT2FCM

Đầu vào	Tập dữ liệu X với số lượng phần tử N trong d chiều, số lượng phần tử được gán nhãn trong X : $L < N$; số lượng cụm (C); ngưỡng ε ; tham số mờ m ; số mũ $\alpha \in (0,1]$ và số lần lặp tối đa $Maxsteps > 0$
Đầu ra	Các ma trận độ thuộc μ, η, ξ và các tâm cụm V
1:	Chạy thuật toán FCPFS với tất cả các phần tử dữ liệu để lấy $(\bar{\mu}_{kj}, \bar{\eta}_{kj}, \bar{\xi}_{kj}, \bar{V})$
2:	Khởi tạo: $t = 0$
3:	$\mu_{kj}^t \leftarrow random; \eta_{kj}^t \leftarrow random; \xi_{kj}^t \leftarrow random$ ($k = 1..N, j = 1..C$) thỏa mãn các ràng buộc (25-27)
4:	Repeat
5:	$t = t + 1$

6:	Tính V_j ($j = 1, \dots, C$) bằng công thức (29)
7:	Tính $\mu_{kj}^{(t)}$ cho dữ liệu đã được gán nhãn ($k = 1, \dots, L; j = 1, \dots, C$) bằng công thức (30)
8:	Tính $\mu_{kj}^{(t)}$ cho dữ liệu chưa được gán nhãn ($k = L + 1, \dots, C; j = 1, \dots, C$) bằng công thức (31)
9:	Tính $\eta_{kj}^{(t)}$ ($k = 1, \dots, N; j = 1, \dots, C$) bằng công thức (32)
10:	Tính $\xi_{kj}^{(t)}$ ($k = 1, \dots, N; j = 1, \dots, C$) bằng công thức (33)
11:	Until $\ \mu^t - \mu^{t-1}\ + \ \eta^t - \eta^{t-1}\ + \ \xi^t - \xi^{t-1}\ \leq \varepsilon$ hoặc $Maxsteps \geq$ số lần lặp tối đa

4. Đánh giá thực nghiệm

4.1. Các điều kiện thực nghiệm

Các thuật toán được thực hiện trên máy tính để bàn với bộ vi xử lý Intel(R) Xeon(R), 16GB RAM; ngôn ngữ lập trình C với Dev-C++ IDE. Các bộ dữ liệu thử nghiệm được lấy từ Kho lưu trữ học máy UCI [17] được trình bày trong Bảng 1. Trong phần này, chúng tôi cài đặt thực nghiệm để so sánh, đánh giá hiệu năng của phương pháp đề xuất PT2FCM với các phương pháp SSFCM, CS3FCM và FCPFS đã được trình bày trong phần 2.

Các tiêu chí để so sánh, đánh giá giữa các thuật toán như sau: i). Độ chính xác phân cụm: Giá trị cao hơn của độ chính xác phân cụm CA (*Clustering Accuracy*) [8] thể hiện cho hiệu suất phân cụm tốt hơn. ii). Chất lượng phân cụm: Giá trị nhỏ hơn của độ đo DB (*Davies–Bouldin index*) [18] đem lại chất lượng phân cụm tốt hơn.

Bảng 1. Dữ liệu UCI dùng cho thực nghiệm

STT	Bộ dữ liệu	Số lượng mẫu	Số thuộc tính	Số cụm
1	Australian	690	14	2
2	Balance-scale	625	4	3
3	Dermatology	366	34	6
4	Heart	270	13	2
5	Iris	150	4	3
6	Spambase	4601	57	2
7	Tae	151	5	3
8	Waweform	5000	40	3
9	Wdbc	569	30	2
10	Wine	178	13	3

4.2. Kết quả thực nghiệm

4.2.1. Đánh giá theo độ chính xác phân cụm

Bảng 2. Giá trị độ chính xác phân cụm (CA) đối với 10 bộ dữ liệu

Phương pháp	PT2FCM		FCPFS		CS3FCM		SSFCM	
	Trung bình	Phương sai	Trung bình	Phương sai	Trung bình	Phương sai	Trung bình	Phương sai
Australian	0,5576	0,0274	0,5312	0,0150	0,5186	0,0153	0,5227	0,0162
Balance-scale	0,5212	0,0417	0,5944	0,0388	0,5404	0,0218	0,5490	0,0305
Dermatology	0,6872	0,0274	0,7155	0,0203	0,6728	0,0154	0,6719	0,0162
Heart	0,8331	0,0306	0,7233	0,0188	0,8081	0,0164	0,5414	0,0194
Iris	0,8103	0,0287	0,7542	0,0327	0,8069	0,0833	0,6411	0,0175
Spambase	0,5940	0,0255	0,5447	0,0146	0,5674	0,0478	0,5144	0,0143
Tae	0,4821	0,0293	0,4618	0,0461	0,4844	0,0194	0,4580	0,0181
Waweform	0,5519	0,0365	0,5174	0,0229	0,5134	0,0312	0,5711	0,0253
Wdbc	0,6835	0,0265	0,7471	0,0161	0,6041	0,0286	0,5683	0,0153
Wine	0,8405	0,0301	0,7922	0,0586	0,7579	0,0329	0,5889	0,0189

Độ chính xác phân cụm của phương pháp đề xuất PT2FCM được so sánh với các phương pháp SSFCM, CS3FCM và FCPFS được trình bày trong Bảng 2.

Trong Bảng 2, phương pháp đề xuất PT2FCM có 5/10 giá trị tốt nhất (Australian, Heart, Iris, Spambase Wine); phương pháp FCPFS có 3/10 giá trị tốt nhất (Balance-scale, Dermatology, Wdbc); phương pháp CS3FCM chỉ có 1/10 giá trị tốt nhất (Tae); còn lại phương pháp SSFCM cũng chỉ có 1/10 giá trị tốt nhất (Waweform). Mặt khác, tỉ số trung bình của độ chính xác phân cụm theo phương pháp PT2FCM so với các phương pháp FCPFS, CS3FCM và SSFCM lần lượt là 1,03, 1,05 và 1,17. Do đó, một cách tổng thể độ chính xác phân cụm theo phương pháp PT2FCM tốt hơn các phương pháp FCPFS, CS3FCM và SSFCM.

4.2.2. Đánh giá theo chất lượng cụm

Chất lượng phân cụm theo độ đo DB của phương pháp đề xuất PT2FCM được so sánh với các phương pháp SSFCM, CS3FCM và FCPFS được trình bày trong Bảng 3.

Bảng 3. Giá trị chất lượng phân cụm theo độ đo DB đối với 10 bộ dữ liệu

Phương pháp	PT2FCM		FCPFS		CS3FCM		SSFCM	
	Trung bình	Phương sai	Trung bình	Phương sai	Trung bình	Phương sai	Trung bình	Phương sai
Australian	2,3610	0,1678	3,3491	0,0182	6,0265	0,0152	2,2527	0,0377
Balance-scale	3,0984	0,2812	3,1854	0,1062	4,9964	0,1919	3,9196	0,1511
Dermatology	12,5601	1,1231	9,9297	1,8434	15,5687	2,0770	19,5570	0,9930
Heart	2,1627	1,0649	2,2254	0,0261	2,4040	0,1521	2,5519	0,9348
Iris	3,0312	0,1798	3,0838	0,0539	3,4747	0,2629	3,0464	0,0497
Spambase	2,6472	0,2487	3,2775	0,0157	4,4607	0,5125	2,7543	0,1186
Tae	2,8903	0,1694	2,7060	0,1394	2,7640	0,0443	2,8909	0,0393
Waweform	8,3421	1,2704	9,4595	3,0705	5,1157	0,1207	22,9106	1,9370
Wdbc	1,9235	0,0591	2,6836	0,0983	2,2870	0,0151	2,0328	0,0299
Wine	3,1089	0,0465	2,7263	0,0400	3,0655	0,0648	2,8765	0,0191

Trong Bảng 3, phương pháp PT2FCM có 5/10 giá trị tốt nhất (Balance-scale, Heart, Iris, Spambase, Wdbc); phương pháp FCPFS có 3/10 giá trị tốt nhất (Dermatology, Tae, Wine); phương pháp CS3FCM chỉ có 1/10 giá trị tốt nhất (Waveform); còn lại phương pháp SSFCM cũng chỉ có 1/10 giá trị tốt nhất (Australian). Tỉ số trung bình của chất lượng phân cụm của phương pháp PT2FCM so với các phương pháp FCPFS, CS3FCM và SSFCM lần lượt là 0,98, 0,84 và 0,65. Do đó, chất lượng phân cụm theo phương pháp PT2FCM tốt hơn các phương pháp liên quan.

5. Kết luận

Trong bài báo này, chúng tôi đề xuất một phương pháp mới trong phân vùng dữ liệu theo độ tin cậy dựa trên phân cụm mờ viễn cảnh có tên gọi PT2FCM. Thuật toán đề xuất được so sánh thực nghiệm các thuật toán phân cụm bán giám sát mờ tiêu chuẩn (SSFCM), phân cụm bán giám sát mờ an toàn với trọng số tin cậy (CS3FCM) và phân cụm bán giám sát mờ trên tập mờ viễn cảnh (FCPFS) theo các tiêu chí về độ chính xác phân cụm và chất lượng cụm. Các kết quả thực nghiệm cho thấy, phương pháp đề xuất có độ chính xác phân cụm và chất lượng phân cụm tốt so với các phương pháp liên quan. Cụ thể như sau: i.) Về độ chính xác phân cụm: Tỉ số trung bình của độ chính xác phân cụm theo phương pháp PT2FCM so với các phương pháp FCPFS, CS3FCM và SSFCM lần lượt là 1,03, 1,05 và 1,17.; ii). Về chất lượng cụm: Tỉ số trung bình của chất lượng phân cụm theo phương pháp PT2FCM so với các phương pháp FCPFS, CS3FCM và SSFCM lần lượt là 0,98, 0,84 và 0,65.

Thuật toán được đề xuất có khả năng loại bỏ hoặc giảm bớt mức độ ảnh hưởng của các phần tử dữ liệu bị nhiễu. Tuy nhiên, phương pháp này vẫn còn một số hạn chế như mô hình chứa nhiều tham số dẫn đến thời gian tính toán cao. Trong những nghiên cứu tiếp theo, nhóm tác giả sẽ phát triển các thuật toán mới để khắc phục những nhược điểm này.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên; trong đề tài mã số T2022-07-03.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] Y. Wu, Y. Guo, Y. Xiao, and S. Lao, "AAE-SC: A scRNA-seq clustering framework based on adversarial autoencoder," *IEEE Access*, vol. 8, pp. 178962-178975, 2020.
- [2] J. T. Santoso, S. Jumini, and Bhawika, *Unsupervised Data Mining Technique for Clustering Library in Indonesia*, Library Philosophy and Practice, pp. 1-9, 2021.
- [3] W. Shi, W. N. Chen, T. Gu, H. Jin, and J. Zhang, "Handling Uncertainty in Financial Decision Making: A Clustering Estimation of Distribution Algorithm With Simplified Simulation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 1, pp. 42-56, 2020.
- [4] S. Majumdar and Laha, *Clustering and classification of time series using topological data analysis with applications to finance*, Expert Systems with Applications, vol. 162, 2020.
- [5] A. Kumar, H. S. Bhaduria, and A. Singh, "Semi-supervised OTSU based hyperbolic tangent Gaussian kernel fuzzy C-mean clustering for dental radiographs segmentation," *Multimedia Tools and Applications*, vol. 79, no. 3, pp. 2745-2768, 2020.
- [6] K. Zhao, Y. Jiang, K. Xia, L. Zhou, Y. Chen, K. Xu, and P. Qian, "View-collaborative fuzzy soft subspace clustering for automatic medical image segmentation," *Multimedia Tools and Applications*, vol. 79, no. 13, pp. 9523-9542, 2020.
- [7] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 27, no. 5, pp. 787-795, 1997.
- [8] H. Gan, Y. Fan, Z. Luo, R. Huang, and Z. Yang, "Confidence-weighted safe semi-supervised clustering," *Engineering Applications of Artificial Intelligence*, vol. 81, pp. 107-116, 2019.
- [9] H. Gan, Z. Li, W. Wu, Z. Luo, and R. Huang, "Safety-aware graph-based semi-supervised learning," *Expert Systems with Applications*, vol. 107, pp. 243-254, 2018.
- [10] H. Gan, Y. Fan, Z. Luo, and Q. Zhang, "Local homogeneous consistent safe semi-supervised clustering," *Expert Systems with Applications*, vol. 97, pp. 384-393, 2018.
- [11] H. Gan, "Safe Semi-Supervised Fuzzy C-Means Clustering," *IEEE Access*, vol. 7, pp. 95659-95664, 2019.
- [12] B. C. Cuong and V. Kreinovich, "Picture fuzzy sets," *Journal of Computer Science and Cybernetics*, vol. 30, no. 4, pp. 409-420, 2014.
- [13] L. Lovász and M. D. Plummer, *Matching theory*, American Mathematical Soc, vol. 378, 2009.
- [14] L. A. Zadeh, "Fuzzy sets," *In Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*, pp. 394-432, 1996.
- [15] K. Atanassov, "Intuitionistic fuzzy sets," *International Journal Bioautomation*, vol. 20, no. 1, pp. 1-6, 2016.
- [16] P. H. Thong and L. H. Son, "Picture fuzzy clustering: a new computational intelligence method," *Soft Comput.*, vol. 20, no. 9, pp. 3549-3562, 2016.
- [17] D. Dua and C. Graff, "UCI Machine Learning Repository," 2019. [Online]. Available <http://archive.ics.uci.edu/ml>. [Accessed Jan. 10, 2022].
- [18] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 224-227, 1979.