THUẬT TOÁN PHÂN CỤM MỜ CỘNG TÁC VÀ GIẢM CHIỀU DỮ LIỆU CHO BÀI TOÁN PHÂN CỤM ẢNH VỆ TINH SIÊU PHỔ

COLLABORATIVE CLUSTERING ALGORITHM WITH REDUCING DIMENTIONALITY FOR HYPERSPECTRA SATELLINE IMAGES

Đặng Trọng Hợp[1,*], Mai Đình Sinh[2]

TÓM TẮT

Ảnh vệ tinh siêu phổ (Hyperspectral Satelline Images - HSI) gần đây đã nhận được sự quan tâm của nhiều nhà nghiên cứu và được ứng dụng trong nhiều lĩnh vực khác nhau. Phân cụm là một bài toán cơ bản trong xử lý ảnh siêu phổ, đồng thời nó cũng là một trong những bước khó nhất bởi vì hảnh siêu phổ có hàng trăm kênh và đòi hỏi tính toán với hiệu năng cao. Trong bài báo này, chúng tôi đưa ra giải pháp phân cụm ảnh siêu phổ bằng cách sử dụng thuật toán phân cụm mờ cộng tác sau khi đã thực hiện giảm chiều dữ liệu ảnh siêu phổ với phép chiếu ngẫu nhiên dựa trên định lý Johnson Lindenstrauss (Thuật toán C2JL). Các kết quả thử nghiệm với tập dữ liệu ảnh vệ tinh siêu phổ và các chỉ số đánh giá cho thấy phương pháp đề xuất cho kết quả tốt hơn các phương pháp đã có.

Từ khóa: *Hình ảnh siêu phổ; phân cụm mờ; hợp tác phân cụm; giảm tính năng.*

ABSTRACT

Hyperspectral satelline images (HSI) have received popularityand shown their usefulness in various earth observation applications in recent years. Segmentation is the basic problems in HSI processing but it also is one of the most difficult taks because HSI have hundreds of channels and high-performance computing is crucial. In this paper, we proposed solution for HSI segmentation by using collaborative clustering with reducing image dimentionality by random projection based on Johnson Lindenstrauss lemma (C2JL algorithm) which also preserves the relative distance between data points. The experiments which were done on 2 HSI data sets with 5 validity indexes shows that proposed methods give the better performance.

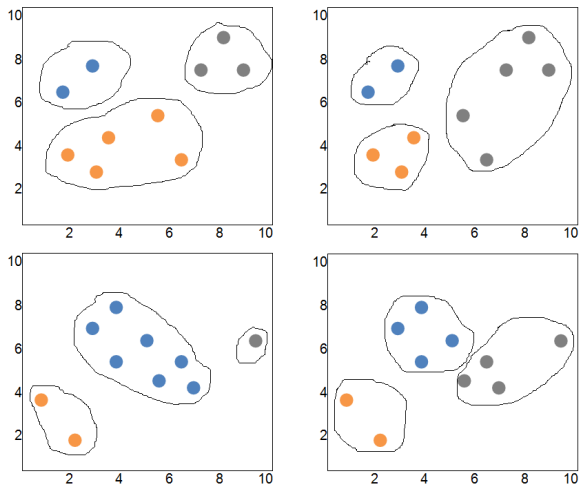
Keywords: *Hyperspectral image; fuzzy clustering; collaborative clustering; feature reduction.*

[1]Trường Đại học Công nghiệp Hà Nội
[2]Viện Kỹ thuật Công trình Đặc biệt, Đại học Kỹ thuật Lê Quý Đôn
[*]Email: hopdt@haui.edu.vn
Ngày nhận bài: 15/01/2021
Ngày nhận bài sửa sau phản biện: 18/6/2021
Ngày chấp nhận đăng: 25/02/2022

1. GIỚI THIỆU

Phân cụm là một công cụ toán học dùng để phát hiện cấu trúc hoặc các mẫu nào đó trong tập dữ liệu, theo đó có đối tượng bên trong cụm dữ liệu thể hiện bậc tương đồng nhất định. Kỹ thuật phân cụm được áp dụng trong rất nhiều lĩnh vực như khai phá dữ liệu, nhận dạng mẫu, xử lý ảnh… Với tư cách là một chức năng khai phá dữ liệu, phân cụm cũng có thể được sử dụng như một công cụ độc lập để quan sát đặc trưng bên trong sự phân bố của dữ liệu. Các thuật toán phân cụm dữ liệu đã được quan tâm nghiên cứu và ứng dụng rộng rãi như: thuật toán phân cụm k-means và các cải tiến của nó [1]; họ các thuật toán phân cụm mờ Fuzzy c-means [2, 3].

Phân cụm cộng tác mờ được giáo sư Pedrycz đề xuất [4] như là công cụ để tìm ra những cấu trúc và nhóm tương đồng của nhiều tập dữ liệu rời rạc có liên quan với nhau. Có hai đặc điểm của phân cụm dữ liệu cộng tác, một là dữ liệu chi tiết ở các tập không thể trao đổi với nhau mà chỉ có thể trao đổi thông tin về cấu trúc, hai là cần xem xét việc phân cụm ở tập dữ liệu này có tác động và chia sẻ tới việc phân cụm ở các tập dữ liệu khác [4, 5, 6, 7]. Một ví dụ về dữ liệu và kết quả của việc phân cụm cộng tác khi có nhiều tập dữ liệu và cấc tập đó có sự cộng tác trong quá trình phân cụm được mô tả trong hình 1. Theo đó, nếu thực hiện phân cụm riêng lẻ từng tập dữ liệu ta sẽ có kết quả như hình (a), nếu thực hiện quá trình phân cụm cộng tác để điều chỉnh sẽ cho kết quả như hình (b). Rõ ràng nếu nhìn toàn bộ dữ liệu của cả hai tập dữ liệu ta sẽ thấy kết quả hình (b) hợp lý hơn do hình dạng của cấu trúc cụm của hai tập có sự tương đồng.



Hình 1. Kết quả phân cụm (a) trước cộng tác, (b) sau khi cộng tác

Tiếp tục các kết quả của Pedrycz, Coletta và cộng sự đã nghiên cứu các phương pháp tối ưu hóa tham số như tính toán mức độ cộng tác giữa các cặp tập dữ liệu, tính toán tối ưu số cụm dữ liệu trong các tập [7]. Ngoài ra, phương pháp phân cụm dữ liệu cộng tác cũng được nghiên cứu và ứng dụng trong trường hợp dữ liệu có nhiều khung nhìn khác nhau tương ứng với các thuộc tính khác nhau, kết quả phân cụm theo từng khung nhìn có thể cộng tác với nhau [8]. Nhiều hướng nghiên cứu mở rộng cũng như ứng dụng phân cụm mờ cộng tác khác đã được nghiên cứu như: Zhou giới thiệu giải thuật phân cụm cộng tác trong mạng phân tán P2P [9]; Thuật toán phân cụm cộng tác lai tính toán hạt cũng được nhóm của Z. Han nghiên cứu ứng dụng trong bài toán xếp hạng các nhà cung cấp gas [10]; Yan Liu trình bày phương pháp phân cụm mờ cộng tác cho dữ liệu khoảng có quy mô lớn [11]; Trong nghiên cứu của mình, Z. Deng cũng đưa ra một hướng tương tự phân cụm cộng tác là phân cụm dựa trên trao đổi mẫu (tâm cụm) [12].

HSI chứa hàng trăm kênh ảnh cho mỗi điểm ảnh và được ứng dụng trong nhiều lĩnh vực khác nhau, đặc biệt là giám sát bề mặt trái đất. Nó có thể được sử dụng để phân loại các chất liệu của bề mặt bằng cách đo bức xạ đến cảm biến với độ phân giải quang phổ cao trên một dải phổ đủ rộng để có thể phân đoạn giữa các lớp phủ đất khác nhau (nước, đất, lâm nghiệp, cây trồng, khu vực đô thị,...).

Trong những năm qua, có một số kỹ thuật đã được đề xuất để giải quyết vấn đề số lượng lớn kênh phổ của HSI. Giảm chiều là cách phổ biến nhất, đây là quá trình tiền xử lý thường được sử dụng trước khi phân đoạn HSI có thể loại bỏ các thành phần thừa HSI và giảm độ phức tạp tính toán, hai cách tiếp cận giảm số chiều phổ biến là trích chọn đặc trưng [13-14] và trích xuất đặc trưng [15]. Mục tiêu cơ bản của mỗi kỹ thuật giảm kích thước này là làm sao giảm kích thước nhưng đồng thời bảo toàn thông tin cấu trúc của dữ liệu [16].

Nhiều nghiên cứu gần đây đã được nghiên cứu để giải quyết bài toán phân đoạn HSI. Có hai cách tiếp cận cơ bản, cách thứ nhất dựa trên việc học có giám sát như SMV, cách thứ hai là dựa trên việc học không giám sát như thuật toán phân cụm, vì thiếu mẫu dữ liệu được gán nhãn, phân cụm là một trong những kỹ thuật được áp dụng phổ biến nhất để phân đoạn ảnh. Một số kỹ thuật lai cũng đã được nhiều nhà nghiên cứu đề xuất và cho kết quả tốt như sử dụng phương pháp nhân [15] hoặc sử dụng thuật toán tiến hóa để tăng độ chính xác của cụm [17-18]. Hầu hết các thuật toán phân cụm sử dụng kết hợp thông tin phổ và thông tin không gian để tối ưu hóa kết quả phân đoạn [19, 20].

Một trong các vấn đề khó nhất trong quá trình xử lý HSI là số lượng lớn chiều dữ liệu và số lượng lớn các mẫu dữ liệu. Vì vậy, trong bài báo này, chúng tôi đề xuất giải pháp cho phân đoạn HSI bằng cách giảm chiều dữ liệu thông qua một phương pháp dựa trên định lý Johnson Lindenstrauss và chia dữ liệu thành các tập nhỏ hơn sau đó phân cụm cục bộ và cuối cùng thực hiện phân cụm cộng tác.

Phần tiếp theo gồm các nội dung sau: Phần 2 giới thiệu ngắn gọn về định lý Johnson-Lindenstrauss, giảm chiều dữ liệu bằng phép chiếu ngẫu nhiên, thuật toán phân cụm cộng tác; Phần 3 đề xuất phương pháp mới để phân đoạn HSI; Phần 4 trình bày kết quả thực nghiệm; Kết luận và các nghiên cứu trong tương lai được đề cập trong Phần 5.

2. CƠ SỞ NGHIÊN CỨU

2.1. Giảm chiều bằng phép chiếu ngẫu nhiên

Định lý Johnson Lindenstrauss [21], trong đó chỉ ra rằng với tập xác định X gồm p chiều: $X \subset R^p$ có n phần tử, tồn tại một phép biến đổi tuyến tính $f: \mathbb{R}^p \to R^m$ thỏa mãn:

$$(1-\varepsilon)\|x-y\|_2 \le \|f(x-y)\|_2 \le (1+\varepsilon)\|x-y\|_2 \quad \text{với mọi } x, y \in R^p$$

Hay với một tập các điểm dữ liệu trong p chiều, tồn tại một phép biến đổi tuyến tính giảm xuống m chiều, trong đó khoảng cách tương đối Euclidean giữa các điểm trong không gian mới gần như không đổi bằng cách sử dụng phép chiếu nhân ma trận dữ liệu đầu vào với một ma trận ngẫu nhiêu m chiều ta được dữ liệu đầu ra m chiều $Y_{nm} = X_{np}R_{pm}$.

Để tính ma trận R, [22] đề xuất phân phối ngẫu nhiên độc lập của R_{ij} như sau:

$$r_{ij} = \begin{cases} -1, \text{pro} = \frac{1}{2} \\ 1, \text{pro} = \frac{1}{2} \end{cases} \quad (1)$$

$$r_{ij} = +\sqrt{3}\begin{cases} +1, \text{pro} = \frac{1}{6} \\ 0, \text{pro} = \frac{2}{3} \\ -1, \text{pro} = \frac{1}{6} \end{cases} \quad (2)$$

Với tham số đầu vào ϵ theo định lý Johnson-Lindenstrauss số chiều đầu ra m được xác định theo công thức sau, pro là phân phối xác suất cho giá trị tương ứng, n là số phần tử của dữ liệu đầu vào:

$$m = \epsilon^{-2}\log n \quad (3)$$

Định lý chỉ ra rằng khoảng cách giữa các đối tượng bị thay đổi không đáng kể trong khoảng $(1 \pm \epsilon)$.

Mở rộng các kết quả của định lý Johnson- Lindenstrauss chỉ ra rằng: với giá trị x cố định $x \in R^n$ và $s < m \in \mathbb{N}$ tồn tại phân phối chuyển đổi tuyến tính $\Psi_s: R^n \to R^m$ thỏa mãn:

$$(0{,}63-\epsilon)\|x\|_{1,2,s} \le \|\psi_s x\|_1 \le (1{,}63+\epsilon)\|x\|_{1,2,s}$$

Trong đó, Ψ_s là ma trận ngẫu nhiên $\Psi_s = \sqrt{\frac{2}{\pi}\frac{s}{m}}\Phi_s$ với Φ_s là phép nhân Hadamard của hai ma trận ngẫu nhiên A_s với ma trận ngẫu nhiên G, trong đó có A_s có d = m/s giá trị 1 trong một cột và các giá trị khác bằng 0, G là ma trận phân phối chuẩn Gaussian. Giá trị của s thể hiện độ rời rạc của x.

2.2. Phân cụm mờ

Thuật toán Fuzzy C-means phân cụm tập dữ liệu X thành c cụm dữ liệu, trong đó mỗi cụm dữ liệu đại diện bằng một tâm cụm dựa trên tối ưu hóa hàm mục tiêu:

$$J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(d_{ik})^2, \quad 1 \le m \le \infty \quad (4)$$

Trong đó, d_{ik} là khoảng cách giữa các phần tử dữ liệu thứ i và k, U là ma trận phân hoạch, v là tập các tâm cụm.

Tối thiểu hóa hàm mục tiêu trên bằng phương pháp Lagrange ta được u và v theo công thức sau:

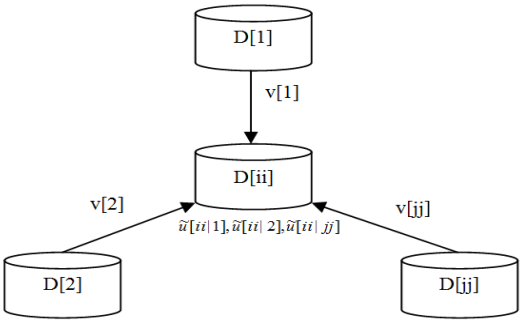
$$u_{ik} = \begin{cases} \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}}, I_k = \emptyset \\ 0, i \notin I_k \\ \left\{\sum_{i\in I_k} u_{ik} = 1, i \in I_k \right. \end{cases} \begin{matrix} 1 \le i \le c, \\ , 1 \le k \le n \\ I_k \ne \emptyset \end{matrix} \quad (5)$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x^k}{\sum_{k=1}^{n}(u_{ik})^m}, \ 1 \le i \le c \quad (6)$$

2.3. Phân cụm cộng tác mờ

Giả sử có P tập dữ liệu D[1], D[2],...,D[P], trong đó chứa N[1], N[2],...,N[P] mẫu dữ liệu trong cùng không gian thuộc tính X. Trong mỗi tập dữ liệu D ta phân thành c cụm. Kết quả phân cụm ở mỗi tập dữ liệu lại tác động tới việc phân cụm ở các khu vực còn lại, chúng ta gọi quá trình này là sự cộng tác giữa và phân cụm cộng tác.

Trong hình 2 các khu vực dữ liệu không trực tiếp trao đổi dữ liệu mà chia sẻ thông tin cấu trúc là ma trận dữ liệu trọng tâm cụm v[jj].



Hình 2. Mô hình phân cụm cộng tác

Bài toán phân cụm dữ liệu cộng tác có hàm mục tiêu cần tối ưu là:

$$Q_{[ii]} = \sum_{k=1}^{N[ii]}\sum_{i=1}^{c} u_{ik}^2[ii]\, d_{ik}^2 + \beta\sum_{jj=1}^{P}\sum_{k=1}^{N[ii]}\sum_{i=1}^{c}(u_{ik} - \tilde{u_{ik}}[ii|jj])^2\, d_{ik}^2 \quad (7)$$

Phần đầu của hàm mục tiêu tương tự như hàm mục tiêu thuật toán C-Means với $u_{ik}[ii]$ là độ thuộc của phần tử thứ k vào cụm i trong tập dữ liệu ii; d_{ik} là khoảng cách từ phần tử thứ i tới tâm cụm i. Phần sau của hàm mục tiêu thể hiện sự tối ưu trong quá trình cộng tác.

Tham số β phản ánh mức độ cộng tác giữa các tập dữ liệu. $\tilde{u}[ii|jj]$ là ma trận độ thuộc tác động của tập dữ liệu jj lên tập ii và được tính theo công thức [16].

$$\tilde{u_{ik}}[ii|jj] = \frac{1}{\sum_{j=1}^{c}\left(\frac{|x_k[ii]-v_i[jj]|}{|x_k[ii]-v_j[jj]|}\right)^2} = \frac{1}{\sum_{j=1}^{c}\frac{d_{ik}^2[ii|jj]}{d_{jk}^2[ii|jj]}} \quad (8)$$

Sử dụng phương pháp Lagrange để tối ưu hàm mục tiêu trên sẽ được công thức tính ma trận phân hoạch và tâm cụm như sau:

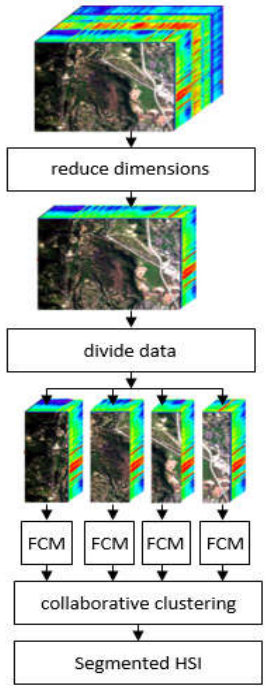
$$u_{rs}[ii] = \frac{1}{\sum_{j=1}^{c}\frac{d_{rs}^2}{d_{js}^2}}\left[1 - \sum_{j=1}^{c}\frac{\beta\sum_{jj=1,jj\ne ii}^{P}\tilde{u_{js}}[ii|jj]}{1+\beta(P-1)}\right] + \frac{\beta\sum_{jj=1,jj\ne ii}^{P}\tilde{u_{rs}}[ii|jj]}{1+\beta(P-1)} \quad (9)$$

$$v_{rt}[ii] = \frac{\sum_{k=1}^{N[ii]} u_{rk}^2[ii]x_{kt} + \beta\sum_{jj=1,jj\ne ii}^{P}\sum_{k=1}^{N[ii]}(u_{rk}[ii]-\tilde{u_{rk}}[ii|jj])^2 x_{kt}}{\sum_{k=1}^{N[ii]} u_{rk}^2[ii] + \beta\sum_{jj=1,jj\ne ii}^{P}\sum_{k=1}^{N[ii]}(u_{rk}[ii]-\tilde{u_{rk}}[ii|jj])^2} \quad (10)$$

Với $1 \le r \le c; 1 \le s \le N[ii]; 1 \le t \le M$.

3. PHƯƠNG PHÁP ĐỀ XUẤT

Bài báo đề xuất giải pháp phân cụm ảnh vệ tinh siêu phổ như hình 3. Mô hình đề xuất gồm 3 pha, Pha 1 thực hiện giảm chiều dữ liệu bằng thuật toán Giám chiều dữ liệu theo phép chiếu ngẫu nhiên, Pha 2 thực hiện chia dữ liệu thành các tập con và phân cụm cục bộ bằng thuật toán FCM, Pha 3 thực hiện thuật toán phân cụm cộng tác để có kết quả cuối cùng.



Hình 3. Mô hình phân cụm ảnh HSI

Algorithm 1: C2JL

Input: Dữ liệu ảnh vệ tinh siêu phổ X, số cụm C

Output: dữ liệu đã phân cụm theo ma trận U

Begin

Phase 1: Giảm chiều dữ liệu

1. Chạy thuật toán **Algorithm 2** để giảm chiều dữ liệu ảnh vệ tinh siêu phổ

Phase 2: Chia dữ liệu và phân cụm cục bộ

1. Chia tập X thành P tập con

2. Chạy thuật toán **Algorithm 3** với mỗi tập dữ liệu con

Phase 3: Phân cụm cộng tác

1. Chạy thuật toán **Algorithm 4** để phân cụm cộng tác

2. Trả kết quả phân cụm theo ma trận U

End

Algorithm 2: Thuật toán giảm chiều dữ liệu

Input: Tập dữ liệu: X, ε và độ thưa dữ liệu s

Output: dữ liệu đã giảm chiều

Tính giá trị m từ ε theo công thức (3)

Tính ma trận G(m,k) theo phân phối Gaussian

Tính ma trận ngẫu nhiên A với giá trị 0 / 1

Tính ma trận nhân Hadamard $\Phi_s = A_s o G$

Tính ma trận ngẫu nhiên $\Psi_s = \sqrt{\frac{2}{\pi}}\frac{s}{m}\Phi_s$

Tính ma trận đầu ra $Y = X_o \Psi_s$

Algorithm 3: FCM

Đầu vào: Tập dữ liệu $X = \{x_1, x_2, \dots, x_n\} \in R^p$, số cụm c (1<c<n), hệ số mờ m (1<m<+∞) và sai số ε, số lần lặp tối đa τ_{max}.

Đầu ra: Kết quả phân cụm

Khởi tạo:

Tâm cụm V;

*T=0. //*Đếm số vòng lặp.

Repeat

Tính toán giá trị tâm cụm v theo công thứ (6);

Cập nhật giá trị ma trận hàm thộc u_{ci} theo công thức (5);

τ=τ+1;

Until ($(|J^{(n)} - J^{(n-1)}| \le \varepsilon)$ hoặc ($\tau \ge \tau_{max}$))

Algorithm 4: Collaborative FCM

Repeat

Trao đổi tâm cụm tới tất cả các tập dữ liệu

For each data site D[ii]

Tính ma trận phân hoạch $\tilde{u}$ theo (8)

Repeat

Tính ma trận phân hoạch u theo (9)

Tính ma trận tâm cụm v theo (10)

Until $|U^t - U^{t-1}| < \varepsilon$ or $t > t_{max}$.

End for

Until $|V^t - V^{t-1}| < \varepsilon 1$ or $t > t_{max}$

4. THỬ NGHIỆM

Trong phần này, giải pháp đề xuất được đánh giá bằng hai thử nghiệm với các dữ liệu ảnh vệ tinh siêu phổ thực tế. Tập dữ liệu và chỉ số Xie-Beni (XB) và chỉ số Partition Coefficient (PC) trong [24] được sử dụng để so sánh với kết quả phân cụm của thuật toán BDC-RPFR-CFCM, FCM. Phương pháp có chỉ số PC nhỏ hơn và chỉ số PC lớn hơn sẽ tốt hơn.

Để thực hiện phân cụm cộng tác, sau khi giảm chiều dữ liệu, tập dữ liệu ảnh vệ tinh siêu phổ được chia thành 3 tập con. Độ chính xác và các chỉ số được cộng cho toàn bộ các tập con. Ngoài ra, hiệu quả phân loại cũng được đánh giá bởi chỉ số True Positive Rate (TPR) và False Positive Rate (FPR) được định nghĩa như sau:

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Trong đó, TP là số điểm thuộc lớp được phân lớp đúng, FN là số điểm không thuộc lớp bị phân lớp sai, FP là số điểm thuộc lớp bị phân lớp sai, TN là số điểm không thuộc lớp được phân lớp đúng. Thuật toán có chỉ số TPR cao hơn và FTR thấp hơn sẽ tốt hơn.

4.1. Thử nghiệm 1

Tập dữ liệu ảnh vệ tinh siêu phổ khu vực Indian Pines ở North-western Indiana gồm ảnh 145 *x* 145 chiều và 200 kênh phổ có bước song từ 0,4μm - 2,4μm. Số điểm mẫu là 10249. Dữ liệu gồm 16 lớp khác nhau bao phủ bề mặt khu vực với số điểm của từng lớp trong bảng 1.

Bảng 1. Số điểm trong từng lớp theo ảnh mẫu đã gán nhãn

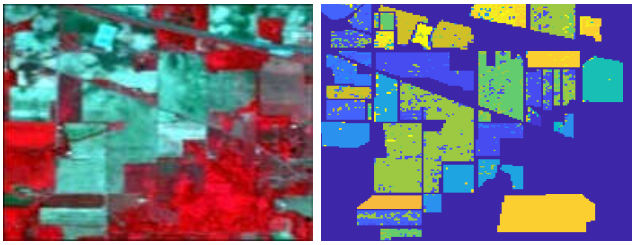
TT	Lớp	Số điểm
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

Theo công thức (3), ta xác định được mối quan hệ giữa chỉ số ϵ và số chiều dữ liệu sau khi giảm theo bảng 2 với giá trị ϵ từ 0,1 đến 0,9. Chạy với các giá trị này được giá trị tối ưu là ϵ = 0,3

Kết quả phân cụm theo từng thuật toán được trình bày trong hình 4. Các chỉ số đánh giá trong bảng 3 cho thấy thuật toán đề xuất CFCM-C2JL cho kết quả tốt hơn với 3/5 chỉ số.

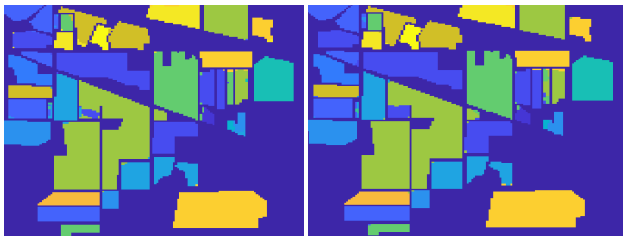
Bảng 2. Giá trị ϵ và số chiều của dữ liệu Indian Pines

ε	0,1	0,2	**0,3**	0,4	0,5	0,6	0,7	0,8	0,9
Dimension	230	58	**26**	14	9	6	5	4	3



a) Dataset b) FCM

c) CFCM d) IT2FCM



e) IT2FCM* f) CFCM-C2JL

Hình 4. Dữ liệu Indian Pines và kết quả phân đoạn ảnh

Bảng 3. Chỉ số đánh giá kết quả phân cụm với dữ liệu Indian Pines

Chỉ số	FCM	CFCM	IT2FCM	IT2FCM*	CFCM-C2JL
PC	0,234	0,376	**0,456**	0,455	0,453
XB	1,094	0,967	0,785	**0,659**	0,698
TPR	90,53%	95,08%	98,04%	98,35%	**98,38%**
FPR	5,29%	5,06%	3,28%	2,99%	**1,63%**
Accuracy	86,34%	94,12%	97,43%	97,45%	**97,66%**

4.2. Thử nghiệm 2

Ảnh vệ tinh siêu phổ chụp bởi vệ tinh ROSIS vùng Pavia, bắc Italia. Số kênh ảnh là 103, số điểm ảnh là 42776, ảnh kích thước 610x610 với độ phân giải 1 điểm ảnh 1,3m^2. Ảnh đã gán nhãn gồm 9 lớp với số điểm của từng lớp trong bảng 4.

Bảng 4. Số điểm trong từng lớp theo ảnh mẫu đã gán nhãn

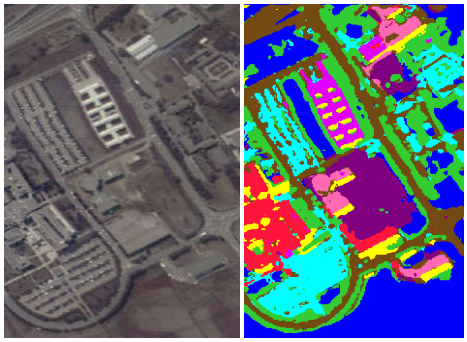
TT	Lớp	Số điểm
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947

Theo công thức (3), ta xác định được mối quan hệ giữa chỉ số ϵ và số chiều dữ liệu Pavia University sau khi giảm theo bảng 5 với giá trị ϵ từ 0,1 đến 0,9 của tập dữ liệu. Chạy với các giá trị này được giá trị tối ưu là ϵ = 0,4.

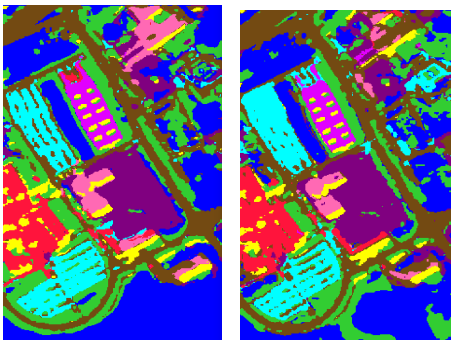
Kết quả phân cụm theo từng thuật toán được trình bày trong hình 5. Các chỉ số đánh giá trong bảng 6 cho thấy thuật toán đề xuất CFCM-C2JL cho kết quả tốt hơn với 3/5 chỉ số.

Bảng 5. Giá trị ϵ và số chiều của dữ liệu Pavia University

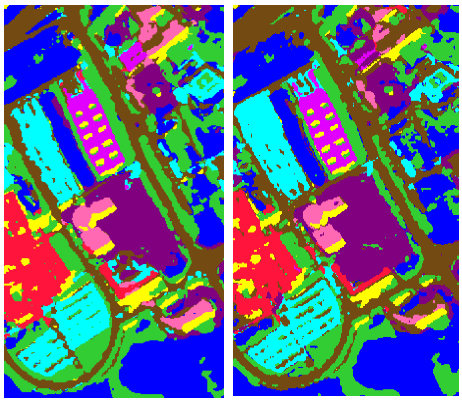
ε	0,1	0,2	0,3	**0,4**	0,5	0,6	0,7	0,8	0,9
Dimension	201	50	22	**13**	8	6	5	4	3



a) Dataset b) FCM



c) CFCM d) IT2FCM



e) IT2FCM* f) CFCM-C2JL

Hình 5. Dữ liệu Pavia University và kết quả phân đoạn ảnh

Bảng 6. Chỉ số đánh giá kết quả phân cụm với dữ liệu Pavia University

Chỉ số	FCM	CFCM	IT2FCM	IT2FCM*	CFCM-C2JL
PC	0,481	0,583	0,672	**0,674**	0,653
XB	0,895	0,799	0,647	**0,628**	0,677
TPR	88,56%	92,95%	95,03%	96,17%	**96,89%**
FPR	4,27%	3,61%	3,52%	2,26%	**1,88%**
Accuracy	86,62%	93,78%	94,11%	96,29%	**96,41%**

5. KẾT LUẬN

Trong bài báo này, chúng tôi đã nghiên cứu các vấn đề gặp phải trong bài toán phân đoạn ảnh vệ tinh siêu phổ, đặc biệt là vấn đề số chiều dữ liệu và khối lượng dữ liệu phải xử lý rất lớn. Từ đó đưa ra giải pháp phân đoạn ảnh vệ tinh siêu phổ bằng thuật toán phân cụm mờ cộng tác và giảm chiều dữ liệu theo phép chiếu ngẫu nhiên dựa trên định lý Johnson Lindenstrauss. Các kết quả thử nghiệm với ảnh HSI thực tế đã cho thấy giải pháp đề xuất cho kết quả tốt trong phần lớn các chỉ số đánh giá.

Trong tương lai chúng tôi sẽ tiếp tục cải tiến hiệu năng của thuật toán bằng cách triển khai mô hình tính toán song song cũng như sử dụng một số thuật toán tốt hơn FCM trong pha xử lí thứ 2 thực hiện phân cụm cục bộ.

TÀI LIỆU THAM KHẢO

[1]. J. B. MacQueen, 1967. *Some Methods for classification and Analysis of Multivariate Observations*. Proc. 5th Berkeley Symp. Math. Stat. Probab.

[2]. Dunn, 1973. *A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters*. J. Cybern., pp. 32–57.

[3]. J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press.

[4]. Witold Pedrycz, 2007. *Collaborative and knowledge based Fuzzy Clustering*. International Journal of Innovative Computing, Information and Control ICIC International, ISSN 1349-4198 Volume 3, pp. 1-12.

[5]. Witold Pedrycz, 2008. *Collaborative fuzzy clustering.* Pattern Recognition Letters 23, pp.1675–1686.

[6]. Witold Pedrycz, 2008. *Collaborative clustering with the use of Fuzzy C-Means and its quantification.* Fuzzy Sets and Systems, pp. 2399 -2427.

[7]. Luiz F. S. Coletta, Lucas Vendramin, Eduardo Raul Hruschka, Ricardo J. G. B. Campello, Witold Pedrycz, 2012. *Collaborative Fuzzy Clustering Algorithms: Some Refinements and Design Guidelines*. IEEE Transaction on Fuzzy Systems, Vol. 20, No. 3, pp. 444-462.

[8]. Yizhang Jiang, Fu-Lai Chung, Shitong Wang, Zhaohong Deng, Jun Wang, Pengjiang Qian, 2014. *Collaborative Fuzzy Clustering From Multiple Weighted Views*. IEEE Trans.on Cybernetics, pp. 1-13.

[9]. Jin Zhou, C. L. Philip Chen, Long Chen, Han-Xiong Li, 2014. *Collaborative Fuzzy Clustering Algorithm in Distributed Network Environments.* IEEE Trans. on Fuzzy Systems, pp. 1-14.

[10]. Z. Han, J. Zhao, Q. Liu, W. Wang, 2016. *Granular-computing based hybrid collaborative fuzzy clustering for long-term prediction of multiple gas holders levels.* Inf. Sci. (Ny)., vol. 330, pp. 175–185.

[11]. Y. Liu, F. Yu, 2016. *Collaborative Fuzzy Clustering Method for Large Scale Interval Data*. Control Decis. Conf., pp. 3906–3911.

[12]. Z. Deng, S. Member, Y. Jiang, F. Chung, 2016. *Transfer Prototype-Based Fuzzy Clustering*. Trans. FUZZY Syst., vol. 24, no. 5, pp. 1210–1232.

[13]. Yuan Yuan, Jianzhe Lin, Qi Wang, 2016. *Dual-Clustering-Based Hyperspectral Band Selection by Contextual Analysis*. IEEE Transactions on Geoscience and Remote Sensing, Vol. 54.

[14]. Sen Jia, Guihua Tang, Jiasong Zhu, Qingquan Li, 2016. *A Novel Ranking-Based Clustering Approach for Hyperspectral Band Selection*. IEEE Transactions on Geoscience and Remote Sensing, Vol 54.

[15]. M. Imani, H. Ghassemian, 2014. *Band Clustering-Based Feature Extraction for Classification of Hyperspectral Images Using Limited Training Samples.* IEEE Geoscience and Remote Sensing Letters, Vol. 11.

[16]. A. A, B. E., J. A. J, A. . G. B. H, 2016. *Comparison of Dimensionality Reduction Techniques for Clustering and Visualization of Load Profiles*. in IEEE PES Transmission & Distribution Conference and Exposition, Jalisco.

[17]. J. Senthilnath, Sushant Kulkarni, J. A. Benediktsson, X. S. Yang, 2016. *A Novel Approach for Multispectral Satellite Image Classification Based on the Bat Algorithm.* IEEE Geoscience and Remote Sensing Letters.

[18]. Pedram Ghamisi, Abder-Rahman Ali, Micael S. Couceiro, Jón Atli Benediktsson, 2015. *A Novel Evolutionary Swarm Fuzzy Clustering Approach for Hyperspectral Imagery*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol 8.

[19]. Kévin Bernard, Yuliya Tarabalka, Jesús Angulo, Jocelyn Chanussot, Jón Atli Benediktsson, 2012. *Spectral-Spatial Classification of Hyperspectral Data Based on a Stochastic Minimum Spanning Forest Approach.* IEEE Transactions on image processing, Vol 21.

[20]. Bing Tu, Xiaofei Zhang, Xudong Kang, Jinping Wang, Jón Atli Benediktsson, 2019. *Spatial Density Peak Clustering for Hyperspectral Image Classification With Noisy Labels*. IEEE Transactions on Geoscience and Remote sensing.

[21]. W. Johnson, J. Lindenstraus, 1984. *Extensions of Lipshotz mapping into Hilbert space*. in Contemporary Mathematics, Texas.

[22]. A. Dimitris, 2003. *Database-friendly random projections: Johnson-Lindenstrauss with binary coins.* Journal of Computer and System Sciences, pp. 271–678.

[23]. F. Krahmer, R. Ward, 2016. *A Unified Framework for Linear Dimensionality Reduction in L1*. Results in Mathematics, vol. 70, no. 1-2, pp. 209-231.

[24]. Jin Zhou, Chiman Kwan, Bulent Ayhan, Michael T. Eismann, 2017. *A Novel Cluster Kernel RX Algorithm for Anomaly and Change Detection Using Hyperspectral Images*. IEEE Transactions on Geoscience and Remote sensing.

AUTHORS INFORMATION

Dang Trong Hop[1], Mai Dinh Sinh[2]

[1]Faculty of Information Technology, Hanoi University of Industry

[2]Institute of Special Engineering, Le Quy Don Technical University