LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu độc lập của riêng tôi, không sao chép ở bất kỳ một công trình hoặc một luận văn, luận án của các tác giả khác. Các số liệu, kết quả nêu trong luận văn này là trung thực và chưa được công bố trong bất kỳ công trình nào khác. Các trích dẫn, các số liệu và kết quả tham khảo dùng để so sánh đều có nguồn trích dẫn rõ ràng.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Hà Nội, tháng 04 năm 2016

Tác giả luận văn

Bùi Văn Chung

LỜI CẢM ƠN

Để hoàn thành tốt luận văn này, đầu tiên em xin bày tỏ lòng biết ơn chân thành và sâu sắc đến Tiến sĩ Lê Hoàng Sơn, người đã tận tình và trực tiếp hướng dẫn em trong suốt quá trình triển khai và nghiên cứu đề tài, tạo điều kiện để em hoàn thành luận văn này.

Thứ hai, em xin bày tỏ lòng biết ơn chân thành tới toàn thể các thầy cô giáo trong khoa Công nghệ thông tin, trường Đại học Công nghệ Hà Nội, Đại học Quốc gia Hà Nội đã dạy bảo tận tình em trong suốt quá trình em học tập tại khoa.

Thứ ba, em xin được gửi lời cảm ơn tới các thầy cô, các anh chị và các bạn trong Trung tâm Tính toán Hiệu năng cao, trường Đại học Khoa học tự nhiên đã giúp đỡ tôi trong suốt thời gian làm luận văn này.

Cuối cùng tôi xin chân thành cảm ơn tới gia đình, bạn bè, đồng nghiệp đã luôn bên em cổ vũ, động viên, giúp đỡ em trong suốt quá trình học tập và thực hiện luận văn.

Mặc dù đã cố gắng hoàn thành luận văn trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em rất mong được sự góp ý chân thành của thầy cô và các bạn để em hoàn thiện luận văn của mình.

Luận văn này được thực hiện dưới sự tài trợ của đề tài NAFOSTED, mã số: 102.05-2014.01.

Xin chân thành cảm ơn!

Hà Nội, ngày 20 tháng 4 năm 2016

Học viên

Bùi Văn Chung

MỤC LỤC

DANH SÁCH HÌNH VẼ

DANH MỤC CÁC KÝ HIỆU VIẾT TẮT

Từ hoặc cụm từ	**Từ viết tắt**	**Từ Tiếng Anh**
Tập mờ	FS	Fuzzy Set
Phân cụm mờ C - Means	FCM	Fuzzy C – Means
Phân cụm mờ K-Means	KFCM	Kernel fuzzy C-means
Thuật toán phân cụm	GK	Gustafson–Kessel
Hệ thống thông tin địa lý	GIS	Geographic Information System
Thuật toán phân cụm đa mô hình	MCLA	Meta-CLustering Algorithm
Thuật toán phân cụm đa mô hình dựa trên sự tương đồng	CSPA	Cluster-based Similarity Partitioning Algorithm
Thuật toán xây dựng biểu đồ hỗn hợp.	HBGF	Hybrid Bipartite Graph Formulation
Chỉ số thực vật	NDVI	Normalized difference vegetation index
Tỷ số chỉ số thực vật	RVI	Ratio vegetion index
Chỉ số sai khác thực vật	DVI	Difference vegetion index
Chỉ số màu xanh thực vật	GVI	Green vegetation index
Chỉ số màu sáng thực vật	LVI	Light vegetation index
Chỉ số úa vàng thực vật	YVI	Yellow vegetation index
Chỉ số màu nâu thực vật	BVI	Brown vegetation index
Chỉ số thực vật cây trồng	CVI	Crop vegetion index

LỜI MỞ ĐẦU

1. ĐẶT VẤN ĐỀ

Trong những năm gần đây, công nghệ thông tin đã có những chuyển biến mạnh mẽ, tác động lớn đến sự phát triển của xã hội. Sự bùng nổ thông tin đã đem đến lượng dữ liệu khổng lồ. Chúng ta càng có nhu cầu khám phá kho dữ liệu đó phục vụ cho nhu cầu con người, điều đó đòi hỏi con người phải biết khai thác dữ liệu và xử lý thông tin đó thành tri thức có ích.

Một trong những kỹ thuật quan trọng trong quá trình khai phá dữ liệu và xử lý dữ liệu lớn là kỹ thuật phân cụm dữ liệu. Phân cụm đặc biệt hiệu quả khi ta không biết về thông tin của các cụm, hoặc khi ta quan tâm tới những thuộc tính của cụm mà chưa biết hoặc biết rất ít về những thông tin đó. Phân cụm được coi như một công cụ độc lập để xem xét phân bố dữ liệu, làm bước tiền xử lý cho các thuật toán khác. Việc phân cụm dữ liệu có rất nhiều ứng dụng như trong lập quy hoạch đô thị, nghiên cứu trái đất, địa lý, khai phá Web v.v.

Ngày nay, cùng với kỹ thuật phân cụm kết hợp với lý thuyết mờ của Zadeh phương pháp phân cụm mờ đã và đang phát triển và được ứng dụng rộng rãi trong thực thực tiễn, phân đoạn ảnh, phân đoạn ảnh viễn thám, nhận dạng mặt người, nhận dạng cử chỉ và điệu bộ, phân tích rủi ro, dự báo nguy cơ phá sản cho ngân hàng và nhiều bài toán khác. Những vấn đề chính được quan tâm nhiều trong phân cụm nói chung và phân mờ nói riêng là nâng cao chất lượng phân cụm, tính toán thông qua một số độ đo chất lượng cụ thể. v.v. được áp dụng trong phân đoạn ảnh viễn thám đa mô hình. Và trong khuôn khổ luận văn này sẽ tìm hiểu vấn đề đó trên cơ sở khảo sát một số thuật toán phân cụm đa mô hình cho bài toán phân cụm ảnh viễn thám, cụ thể là thuật toán SCPA, MG.

2. MỤC ĐÍCH CỦA LUẬN VĂN

Trong luận văn này chúng tôi khảo sát môt số thuật toán phân cụm mờ, cụ thể là thuật toán FCM, KFCM, MG, SCPA. Các thuật toán này sẽ được áp dụng cho bài toán phân cụm ảnh viễn thám đa mô hình.

Cụ thể với một cơ sở dữ liệu mẫu là bộ ảnh vệ tinh của một số khu vực được khảo sát khu vực Bảo Lâm và Thanh Hóa. Qua đây, tính hiệu quả của các thuật toán đa mô hình cho bài toán phân cụm ảnh viễn thám theo các tiêu chí về chất lượng và độ đo.

3. BỐ CỤC CỦA LUẬN VĂN

Luận văn gồm 3 chương, có phần mở đầu, phần kết luận, phần mục lục, phần tài liệu tham khảo. Các nội dung cơ bản của luận văn được trình bày theo cấu trúc như sau:

Chương 1: Tổng quan về phân cụm

Trong chương này, luận văn sẽ trình bày tổng quan về tập mờ, bài toán phân cụm và phân cụm mờ và thuật toán cơ bản giải quyết vấn đề phân cụm trên tập mờ đó là thuật toán Fuzzy C – Means (FCM), KFCM. Từ thuật toán này đưa ra thuật toán đa mô hình cho bài toán phân cụm ảnh viễn thám.

Chương 2: Phân cụm đa mô hình

Trong chương này, tổng quan về học đa mô hình và phân cụm đa mô hình. Tiếp theo, giới thiệu về thuật toán đa mô hình SCPA, MCLA, HBGF và MG.

Chương 3: Ứng dụng phân đoạn ảnh viễn thám

Trong chương này, chúng tôi cài đặt và đánh giá hiệu năng các thuật toán đa mô hình: MG và SCPA từ đây thấy hiệu quả của các thuật toán phân cụm đa mô hình cho ảnh viễn thám được khẳng định.

CHƯƠNG 1: TỔNG QUAN VỀ PHÂN CỤM

1.1. Khái quát phân cụm

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp học không giám sát trong học máy, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn và quan trọng trong tập dữ liệu lớn để từ đó cung cấp thông tin, tri thức cho việc ra quyết định.

Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm, sao cho các đối tượng trong cùng một cụm tương tự nhau và các đối tượng khác cụm thì không tương tự nhau [1].

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm nội tại bên trong của bộ dữ liệu không có nhãn. Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh giá hiệu quả của phân tích phân cụm, điều này phụ thuộc vào mục đích cuối cùng của phân cụm dữ liệu. Do đó, người sử dụng phải cung cấp tiêu chuẩn, theo cách như vậy mà kết quả của phân cụm sẽ phù hợp với nhu cầu của người sử dụng cần.

Định nghĩa 1.1

Cho X là một tập dữ liệu gồm N vector: $\{x_1, x_2, ..., x_N\}$. Bài toán phân cụm là chia tập dữ liệu X , c cụm dữ liệu c.

Thỏa mãn 3 điều kiện sau:

- $z_i \neq \varnothing, \quad i = 1, 2, ..., c$
- $X = \bigcup_{i=1}^{c} z_i$
- $z_i \cap z_j = \varnothing$ với $i \neq j$; $i, j = 1, 2, ..., c$

Phân cụm được đóng vai trò quan trọng trong các nghành khoa học:

Thương mại: Phân cụm dữ liệu giúp các nhà cung cấp biết được nhóm khác hàng quan trọng có các đặc trưng tương đồng nhau và đặc tả họ từ các mẫu trong cơ sở dữ liệu khách hàng.

- Sinh học: Phân cụm dữ liệu được sử dụng để xác định các loại sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.

- Phân tích dữ liệu không gian: Do sự đồ sộ của dữ liệu không gian như dữ liệu thu được từ các hình ảnh chụp từ vệ tinh, các thiết bị y học hoặc hệ thống thông tin địa lý (GIS), v.v, làm cho người dùng rất khó để kiểm tra các dữ liệu không gian một cách chi tiết. Phân cụm dữ liệu có thể trợ giúp người dùng tự động phân tích và xử lý các dữ liêu không gian như nhận dạng và chiết xuất các đặc tính hoặc các mẫu dữ liệu quan tâm có thể tồn tại trong cơ sở dữ liệu không gian.

- Lập quy hoạch đô thị: Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý, v.v, nhằm cung cấp thông tin cho quy hoạch đô thị.

- Nghiên cứu trái đất: Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm.

- Địa lý: Phân lớp các động vật, thực vật và đưa ra đặc trưng của chúng.

- Khai phá Web: Phân cụm dữ liệu có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu Web, khám phá ra các mẫu truy cập của khách hàng đặc biệt hay khám phá ra cộng đồng Web, v.v.

1.2. Tổng quan các thuật toán phân cụm tiêu biểu

Các kỹ thuật phân cụm có rất nhiều cách tiếp cận và các ứng dụng trong thực tế, nó đều hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán [1]. Hiện nay, các kỹ thuật phân cụm có thể phân loại theo các cách tiếp cận chính sau:

1.2.1 Phân cụm cụm phân hoạch

Kỹ thuật này phân hoạch một tập hợp dữ liệu có n phần tử thành k nhóm cho đến khi xác định số các cụm được thiết lập. Số các cụm được thiết lập là các đặc trưng được lựa chọn trước. Phương pháp này là tốt cho việc tìm các cụm

hình cầu trong không gian Euclidean. Ngoài ra, phương pháp này cũng phụ thuộc vào khoảng cách cơ bản giữa các điểm, để lựa chọn các điểm dữ liệu nào có quan hệ là gần nhau với mỗi điểm khác và các điểm dữ liệu nào không có quan hệ hoặc có quan hệ là xa nhau so với mỗi điểm khác. Tuy nhiên, phương pháp này không thể xử lý các cụm có hình dạng kỳ quặc hoặc các cụm có mật độ các điểm dầy đặc. Các thuật toán phân hoạch dữ liệu có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề phân cụm dữ liệu, do nó phải tìm kiếm tất cả các phân hoạch có thể được. Chính vì vậy, trên thực tế thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Như vậy, ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược ăn tham để tìm kiếm nghiệm.

Một số thuật toán phân cụm theo tiếp cận phân hoạch: Thuật toán K-Means, thuật toán K-Medoids

Thuật toán K-Means: Cho k là số cụm sau khi phân hoạch. ($1 \leq k \leq n$, với n là số điểm trong không gian giữ liệu)

Thuật toán k-means gồm 4 bước:

B1. Chọn ngẫu nhiên k điểm làm trọng tâm ban đầu của k cụm.

B2. Gán (hoặc gán lại) từng điểm vào cụm có trọng tâm gần điểm đang xét nhất. Nếu không có phép gán nào thì dừng. Vì không có phép gán nào có nghĩa là các cụm đã ổn định và thuật toán không thể cải thiện làm giảm độ phân biệt hơn được nữa.

B3. Tính lại trọng tâm cho từng cụm.

B4. Quay lại bước 2. Minh họa thuật toán với k=2

Ưu điểm của phương pháp phân cụm k-means

- Độ phức tạp của thuật toán là O (tkn) với t là số lần lặp (t khá nhỏ so với n), k là số cụm cần phân hoạch, n là số điểm trong không gian dữ liệu.
- K-means phù hợp với các cụm có dạng hình cầu.

Nhược điểm của phương pháp k-mean

- Không đảm bảo đạt được tối ưu toàn cục và kết quả đầu ra phụ thuộc nhiều vào việc chọn k điểm khởi đầu. Do đó có thể phải chạy lại thuật toán với nhiều bộ khởi đầu khác nhau để có được kết quả đủ tốt. Trong thực tế có thể áp dụng thuật giải di truyền để phát sinh các bộ khởi đầu.
- Cần phải xác định trước số cụm.
- Khó xác định số cụm thực sự mà không gian dữ liệu có. Do đó có thể phải thử với các giá trị k khác nhau.
- Khó phát hiện các loại cụm có hình dạng phức tạp và nhất là các dạng cụm không lồi.
- Không thể xử lý nhiễu và mẫu cá biệt.
- Chỉ có thể áp dụng khi tính được trọng tâm.

Thuật toán K-Medoids

Thuật toán K-Medoids là cải tiến của thuật toán k-means, k-medoids khác k-means:

- Chiến lược cho k trọng tâm đầu tiên.
- Phương pháp tính độ phân biệt
- Phương pháp tính trọng tâm trong cụm

Thuật toán K-Medoids được thực hiện qua các bước sau:

B1: Chọn ngẫu nhiên k điểm $O_i (i = 1,..., k)$ làm trung tâm (medoids) ban đầu của k cụm.

B2: Gán (hoặc gán lại) từng điểm vào cụm có trung tâm gần điểm đang xét nhất

B3: Với mỗi điểm trung tâm $O_i (i = 1,...,k)$:

B3.1. Lần lượt xét các điểm không là trung tâm x.

B3.2. Tính S là độ lợi khi hoán đổi O_i bởi x. S được xác định như sau:

$$S = E_x - E_{Oi} \quad (1.1)$$

Với E_{oi}, E_x lần lượt là giá trị hàm mục tiêu trước và sau khi thay bởi x.

$$E = \sum_{i=1}^{k} \sum d\left(p, O_i\right)^2 \quad (1.2)$$

B3.3. Nếu S là âm thì thay thế O_i trong bộ k trung tâm bởi x (chọn trung tâm mới tốt hơn).

B4. Nếu có ít nhất 1 sự thay đổi trong B3 thì tiếp tục quay lại B2. Ngược lại thì kết thúc thuật toán.

Ưu điểm: Thuật toán K-medoids làm việc được với nhiễu và biệt lệ.

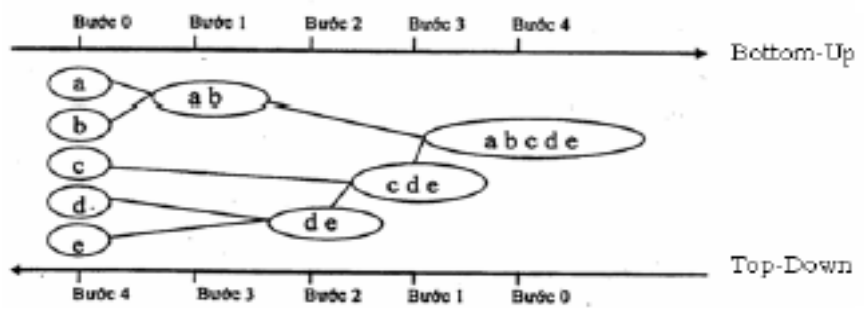
Nhược điểm: Thuật toán K-medoids chỉ hiệu quả khi tập dữ liệu không quá lớn vì có độ phức tạp là $O(k(n-k)^{2t})$. Trong đó: n là số điểm trong không gian dữ liệu, k là số cụm cần phân hoạch, t là số lần lặp.

1.2.2 Phân cụm phân cấp

Phương pháp này xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Có hai cách tiếp cận phổ biến của kỹ thuật này đó là:

+ Hoà nhập nhóm, thường được gọi là tiếp cận Bottom-Up

+ Phân chia nhóm, thường được gọi là tiếp cận Top-Down



Hình 1.1 Các chiến lược phân cụm phân cấp

1.2.3 Phân cụm dựa trên mật độ

Kỹ thuật này nhóm các đối tượng dữ liệu dựa trên hàm mật độ xác định, mật độ là số các đối tượng lân cận của một đối tượng dữ liệu theo một nghĩa nào đó. Trong cách tiếp cận này, khi một dữ liệu đã xác định thì nó tiếp tục được phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa trên mật độ của các đối tượng, để xác định các cụm dữ liệu có thể phát hiện ra các cụm dữ liệu với hình thù bất kỳ. Kỹ thuật này có thể khắc phục được các phần tử ngoại lai hoặc giá trị nhiễu rất tốt, tuy nhiên việc xác định các tham số mật độ của thuật toán là rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả phân cụm.

Một số thuật toán PCDL dựa trên mật độ điển hình như: DBSCAN, OPTICS, DENCLUE, SNN, v.v.

Thuật toán DENCLUE

Thuật toán DENCLUE (DENsity - Based CLUstEring) được đề xuất bởi [19], đây là thuật toán phân cụm dữ liệu dựa trên một tập các hàm phân phối mật độ. Ý tưởng chính của thuật toán này như sau:

- Ảnh hưởng của một đối tượng tới láng giềng của nó được xác định bởi hàm ảnh hưởng.

- Mật độ toàn cục của không gian dữ liệu được mô hình phân tích như là tổng tất cả các hàm ảnh hưởng của các đối tượng.

- Các cụm được xác định bởi các đối tượng mật độ cao trong đó mật độ cao là các điểm cực đại của hàm mật độ toàn cục.

Định nghĩa hàm ảnh hưởng: Cho x, y là hai đối tượng trong không gian d chiều ký hiệu là F^d, hàm ảnh hưởng của y lên x được xác định: $f_B^y : F^d \to R_0^+$, mà được định nghĩa dưới dạng một hàm ảnh hưởng cơ bản : $f_b : f_B^y(x) = f_b(x, y)$. Hàm ảnh hưởng là hàm tuỳ chọn, miễn là nó được xác định bởi khoảng cách d(x,y) của các đối tượng, thí dụ như khoảng cách Euclide.

1.2.4 Phân cụm dựa trên mô hình

Phương pháp này cố gắng khám phá các phép xấp xỉ tốt của các tham số mô hình sao cho khớp với dữ liệu một cách tốt nhất. Chúng có thể sử dụng chiến lược phân cụm phân hoạch hoặc phân cụm phân cấp, dựa trên cấu trúc hoặc mô hình này để nhận dạng ra các phân hoạch. Phương pháp phân cụm dựa trên mô hình cố gắng khớp giữa các dữ liệu với mô hình toán học, nó dựa trên giả định rằng dữ liệu được tạo ra bằng hỗn hợp phân phối xác suất cơ bản. Các thuật toán phân cụm dựa trên mô hình có hai cách tiếp cận chính: mô hình thống kê và mạng nơron. Phương pháp này gần giống với phương pháp phân cụm dựa trên mật độ, vì chúng phát triển các cụm riêng biệt nhằm cải tiến các mô hình đã được xác định trước đó, nhưng đôi khi nó không bắt đầu với một số cụm cố định và không sử dụng cùng một khái niệm mật độ cho các cụm.

Phương pháp phân cụm dữ liệu dựa trên mô hình cố gắng khớp giữa dữ liệu với mô hình toán học, nó dựa trên giả định rằng dữ liệu được tạo ra bằng hỗn hợp phân phối xác suất cơ bản. Các thuật toán phân cụm dựa trên mô hình có hai

tiếp cận chính: Mô hình thống kê và Mạng Nơron. Một số thuật toán điển hình như EM, COBWEB, v..v.

Thuật toán EM được nghiên cứu từ 1958 bởi Hartley và được nghiên cứu đầy đủ bởi Dempster, Laird và Rubin công bố năm 1977. Thuật toán này nhằm tìm ra sự ước lượng về khả năng lớn nhất của các tham số trong mô hình xác suất (các mô hình phụ thuộc vào các biến tiềm ẩn chưa được quan sát), nó được xem như là thuật toán dựa trên mô hình hoặc là mở rộng của thuật toán k-means. EM gán các đối tượng cho các cụm đã cho theo xác suất phân phối thành phần của đối tượng đó. Phân phối xác suất thường được sử dụng là phân phối xác suất Gaussian với mục đích là khám phá lặp các giá trị tốt cho các tham số của nó bằng hàm tiêu chuẩn là hàm logarit khả năng của đối tượng dữ liệu, đây là hàm tốt để mô hình xác suất cho các đối tượng dữ liệu.

Thuật toán gồm 2 bước xử lý: Đánh giá dữ liệu chưa được gán nhãn

(bước E) và đánh giá các tham số của mô hình, khả năng lớn nhất có thể xẩy ra (bước M).

Cụ thể thuật toán EM ở bước lặp thứ t thực hiện các công việc sau:

1) Bước E: Tính toán để xác định giá trị của các biến chỉ thị dựa trên mô hình hiện tại và dữ liệu:

$$z_{ij}^{(t)} = E_{\psi}\left(z_{ij} \mid x\right) = \Pr_{\psi}\left(z_{ij} = 1 \mid x\right) = \frac{f_j\left(x_i\right)\pi_j^{(t)}}{\sum_g^k 1 f g^{(x_i)\pi} g} \qquad (1.3)$$

2) Bước M: Đánh giá xác suất π

$$\pi_j^{(t+1)} = \sum_{i=1}^{n} z_{ij}^{(t)} \Big/ n \qquad (1.4)$$

EM có thể khám phá ra nhiều hình dạng cụm khác nhau, tuy nhiên do thời gian lặp của thuật toán khá nhiều nhằm xác định các tham số tốt nên chí phí tính toán của thuật toán là khá cao. Đã có một số cải tiến được đề xuất cho EM dựa trên các tính chất của dữ liệu: có thể nén, có thể sao lưu trong bộ nhớ và có thể huỷ bỏ. Trong các cải tiến này, các đối tượng bị huỷ bỏ khi biết chắc chắn được nhãn phân cụm của nó, chúng được nén khi không bị loại bỏ và thuộc về một cụm quá lớn so với bộ nhớ và chúng sẽ được lưu lại trong các trường hợp còn lại.

1.2.5 Phân cụm mờ

Phân cụm dữ liệu đóng vai trò quan trọng trong giải quyết bài toán nhân biết mẫu và xác định mô hình mờ. Thuật toán FCM phù hợp hơn với dữ liệu lớn hoặc nhỏ phân bố quanh tâm cụm.

Fuzzy C – Means là một phương pháp phân nhóm cho phép một phần dữ liệu thuộc hai hay nhiều cụm.

Phân cụm N vector $X = \{x_1, x_2, ..., x_N\}$ thành c cụm dựa trên tính toán tối thiểu hóa hàm mục tiêu để đo chất lượng của cụm và tìm tâm cụm sao cho hàm độ đo không tương tự là nhỏ nhất. Một phân cụm mờ vector $X = \{x_1, x_2, ..., x_N\}$ được biểu diễn bởi ma trận $U = [U_{ki}]_{N \times c}$ sao cho một điểm dữ liệu có thể thuộc về nhiều nhóm và được xác định bằng giá trị hàm thuộc u. Ma trận giá trị hàm thuộc có dạng như sau:

$$U = \begin{bmatrix} u_{11} & \text{L} & u_{1c} \\ \text{M} & \text{O} & \text{M} \\ u_{N1} & \text{L} & u_{Nc} \end{bmatrix}$$

Thuật toán phân cụm mờ đã được xuất phát từ việc cực tiểu giá trị hàm mục tiêu:

$$J_m = \sum_{k=1}^{c} \sum_{j=1}^{N} u_{kj}^m d(x_k, \text{z}_j) \qquad (1.5)$$

$d(x_k, \text{z}_j)$: là một độ đo không tương tự.

Giải bài toán $J_m(u,z) \to \min$ với ràng buộc sau:

$$\begin{cases} 0 \le u_{kj} \le 1 & \forall j = 1,2,..,c \\ & \forall k = 1,2,..,N \\ \sum_{j=1}^{c} u_{kj} = 1 & \\ 0 \le \sum_{k=1}^{N} u_{kj} \le N & \end{cases}$$

Thuật toán Fuzzy C – Means phân tập N đối tượng trong không gian R^d chiều $z_j = \{z_{j1}, z_{j2}, ..., x_{jd}\}$, với $x_i = \{x_{i1}, x_{i2}, ..., x_{id}\}$ thành c cụm mờ $1 < c < N$ với tâm cụm $Z = \{z_1, z_2, ..., z_c\}$, với $z_j = \{z_{j1}, z_{j2}, ..., x_{jd}\}$. Cụm mờ của N đối tượng được biểu diễn bằng ma trận mờ μ có N hàng và c cột với N là số các đối tượng và c là số cụm.

Có thể tổng quát bài toán bằng công thức (p) như sau:

$$(p) \begin{cases} \left(\min_{\mu,Z}\right) J_m(\mu, Z) = \sum_{i=1}^{N}\sum_{j=1}^{c} \mu_{ij}^m \left\|x_i - z_j\right\|^2 \\ \sum_{j=1}^{c} \mu_{ij} = 1, \forall i = 1,..., N \\ \mu_{ij} \ge 0, \forall i = 1,..., N; j = 1,..., c \end{cases} \quad (1.6)$$

Trong đó:

- $d_{ij} = \left\|x_i - z_j\right\|$ là khoảng cách Euclide
- $m (m > 1)$ tham số mờ (Đối với $m = 1$ thì Fuzzy C – Means trở thành thuật toán rõ. Giá trị thường sử dụng là $m = 2$)
- Tâm cụm z_j của cụm thứ j được tính theo công thức:

$$z_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m x_i}{\sum_{i=1}^{N} \mu_{ij}^m} \quad (1.7)$$

Thuật toán Fuzzy C-Means

FCM được đề xuất bởi Bezdek năm 1974:

- Input

- $X = \{x_1, x_2, ..., x_N\}$
- Số cụm c
- Tham số m

- Output
- Tâm cụm $Z = \{z_1, z_2, ..., z_c\}$
- Giá trị hàm thuộc $\mu = \left[\mu_{ij}\right]_{N \times c}$

- **Thuật toán**

Bước 1: Lựa chọn $m(m > 1)$; Khởi tạo các giá trị hàm thuộc $\mu_{ij}, i = 1, 2, ..., N; j = 1, 2, ..., c$

Bước 2: Tính toán tâm cụm $z_j; j = 1, 2, ..., c$ theo công thức (1.7)

$$z_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m \, x_i}{\sum_{i=1}^{N} \mu_{ij}^m}$$

Bước 3: Tính khoảng cách Euclide $d_{ij}, i = 1, 2..., N; j = 1, 2..., c$

$$d_{ij}(x_i, z_j) = \sqrt{\left(x_{i1} - z_{j1}\right)^2 + \left(x_{i2} - z_{j2}\right)^2 + ... + \left(x_{id} - z_{jd}\right)^2}$$

Bước 4: Cập nhật các giá trị hàm thuộc $\mu_{ij}, i = 1, 2, ..., N; j = 1, 2, ..., c$ theo công thức (1.8):

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \qquad (1.8)$$

Bước 5: Nếu không hội tụ, lặp lại bước 2.

Một vài luật dừng có thể được sử dụng. Thứ nhất các giá trị đầu và giá trị cuối nhận giá trị nhỏ hơn khi thay đổi giá trị tâm cụm. Hoặc hàm mục tiêu (1.6) $J_m(\mu, Z) = \sum_{i=1}^{N} \sum_{j=1}^{c} \mu_{ij}^m \left\| x_i - z_j \right\|^2$ không thể cực tiểu hơn nữa. Thuật toán FCM nhạy cảm với giá trị khởi tạo và có thể sảy ra tối ưu cục bộ.

*** Ưu và nhược điểm:**

Ưu điểm:

- Cho kết quả tốt nhất cho dữ liệu chồng chéo.
- Dữ liệu điểm duy nhất có thể không thuộc về một cụm duy nhất, ở mỗi điểm được phân vào cụm dựa trên kết quả tính hàm thuộc. Vì vậy, một điểm có thể thuộc về nhiều hơn một cụm.

Nhược điểm:

- Cần tiên nghiệm số lượng các cụm.
- ε càng thấp kết quả nhận được càng tốt nhưng chi phí tính toán càng nhiều.
- Khoảng cách Euclide các yếu tố cơ bản có thể không đồng đều.

Thuật toán KFCM

Từ thuật toán FCM đề xuất thuật toán Kernel fuzzy C-means (KFCM). Xác định giá trị phi tuyến: $\Phi: x \to \Phi(x) \in F$ ở đây $x \in X$. X là không gian dữ liệu và F không gian đặc trưng biến đổi với kích thước vô hạn cao hơn. KFCM giảm thiểu hàm mục tiêu sau đây:

$$J_m(\mathrm{U},\mathrm{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{jk}^{m} \left\|\Phi(x_k) - \Phi(v_j)\right\|^2 \qquad (1.9)$$

ở đây

$$\left\|\Phi(x_k) - \Phi(v_i)\right\|^2 = K(x_k, x_k) + K(v_i, v_i) - 2K(x_k, v_i) \qquad (1.10)$$

Trong đó $K(x,y) = \Phi(x)^T \Phi(y)$ là hàm nhân. Nếu ta tính toán theo hàm Gaussian thì hàm nhân sẽ là: $K(x,y) = \exp(-\left\|x-y\right\|^2 / \sigma^2)$ trong trường hợp $K(x,x) = 1$ thì công thức (1.9) và (1.10) sẽ được viết lại như sau:

$$J_m(U,V) = 2\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m (1 - K(x_k, v_i)) \quad (1.11)$$

Tương tự như FCM xây dựng hàm Lagrange giải (1.11) ta có:

$$u_{ik} = \frac{(1/(1-K(x_k,v_i)))^{1/(m-1)}}{\sum_{j=1}^{c}(1/(1-K(x_k,v_j)))^{1/(m-1)}} \quad (1.12)$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m K(x_k, v_i) x_k}{\sum_{k=1}^{n} u_{ik}^m K(x_k, v_i)} \quad (1.13)$$

$$d(x,y) \overset{\Delta}{=} \|\Phi(x) - \Phi(y)\| = \sqrt{2(1-K(x,y))} \quad (1.14)$$

Các bước thuật toán KFCM

1. Khởi tạo ma trận phân hoạch $U=[u_{jk}], U^{(0)}$.
2. Gán cho c , tmax , m > 1 and ε > 0 là các hằng số dương
3. Tại bước thứ t: Tính vecto tâm cụm v_i^t theo công thức (1.13)
4. Cập nhật lại u_{ik}^t tính theo công thức (1.12)
5. Nếu $E^t = \max_{i,k} \left| u_{ik}^t - u_{ik}^{t-1} \right| < \varepsilon$ thì dừng, sai thì quay lại bước 3.

1.3 Độ đo phân cụm

Nhiều độ đo phân cụm tương đối khác nhau tồn tại mà rất hữu ích trong thực tế là biện pháp định lượng để đánh giá chất lượng của phân cụm dữ liệu, các tiêu chí mới vẫn được đề xuất. Những tiêu chí có được các tính năng riêng biệt mà có thể làm tốt hơn những trường hợp cụ thể của độ đo phân cụm. Ngoài ra, có thể có yêu cầu tính toán hoàn toàn khác nhau. Khó khăn cho người dùng

chọn lựa một tiêu chí cụ thể khi phải đối mặt với hàng loạt các khả năng. Vì vậy trong vấn đề liên quan đến phân cụm ta phải so sánh các độ đo hiện có đã tồn tại trước đó với các tiêu chí mới của độ đo được đề xuất.

Các giải pháp khác có liên quan với các kỹ thuật xác nhận phân cụm, để chất lượng truy cập phân nhóm dựa trên ba nhóm chỉ số giá trị phân cụm [6-8] đã phát triển cho đánh giá định lượng của các kết quả phân nhóm dựa vào bên ngoài, các biện pháp bên trong, và tương đối [9] tương ứng. Cả hai phương pháp xác nhận bên ngoài và bên trong dựa trên kiểm tra thống kê đòi hỏi chi phí tính toán cao. Tuy nhiên, ý tưởng chính của cách tiếp cận thứ ba, dựa trên các tiêu chí tương đối, là để xác định kết quả phân cụm tốt nhất tạo ra từ các thuật toán phân cụm tương tự nhưng với tham số khác nhau.

1.3.1 Adjusted Rand Index

Adjusted Rand Index [10] được xác định bởi:

$$ARI\left(P^*,P\right)=\frac{\sum_{i,j}\binom{N_{ij}}{2}-\left[\sum_i\binom{N_i}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i\binom{N_i}{2}+\sum_j\binom{N_j}{2}\right]-\left[\sum_i\binom{N_i}{2}\sum_j\binom{N_j}{2}\right]/\binom{N}{2}}. \quad (1.15)$$

Ở đây, N là số điểm dữ liệu trong một tập dữ liệu cho trước và N_{ij} là số điểm dữ liệu của các nhãn lớp $C_j^* \in P^*$. N_i là số điểm dữ liệu trong một tập dữ liệu cho trước gán cho cụm C_i trong phân vùng P. N_i là số điểm dữ liệu trong cụm C_i. Giá trị ARI nằm giữa 0 và 1 các chỉ số giá trị tương đương với 1 chỉ khi một phân vùng là hoàn toàn giống với cấu trúc nội tại và gần 0 cho một phân vùng ngẫu nhiên.

1.3.2 Jaccard Index

Hệ số tương tự Jaccard [10] được xác định bởi:

$$J\left(P^*,P\right)=\frac{\sum_{i,j}\binom{N_{ij}}{2}}{\sum_i\binom{N_i}{2}+\sum_j\binom{N_j}{2}-\sum_i\binom{N_{ij}}{2}}. \quad (1.16)$$

Ở đây N_{ij} là số điểm dữ liệu của các nhãn lớp $C_j^* \in P^*$ được gán cho cụm C_i trong phân vùng P . N_i là số điểm dữ liệu trong cụm C_i của phân vùng P và N_i là số điểm dữ liệu trong lớp C_j^*.

1.3.3 Modified Hubert's Γ Index

Modified Hubert's Γ Index [11] được cho bởi phương trình:

$$MHT\left(P\right)=\frac{2}{n(\text{n}-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}PM_{ij}Q_{ij} \quad (1.17)$$

Ở đây PM_{ij} là ma trận khoảng cách, và Q là $n\times n$ là cụm khoảng cách dựa trên ma trận trên phân vùng P , Q_{ij} là khoảng cách giữa các trung tâm cụm mà x_i và x_j thuộc về. Trong Modified Γ Index Hubert của (MHΓ), giá trị cao đại diện cho chất lượng phân cụm tốt hơn.

1.3.4 Dunn's Validity Index

Dunn's Validity Index [12] được cho bởi phương trình sau:

$$DVI(P)=\min\left\{\frac{d\left(c_i,c_j\right)}{\max\limits_{k=1\ldots K_m}\left\{diam\left(c_k\right)\right\}}\right\} \quad (1.18)$$

Trong đó $\left[c_i,c_j,c_k\right]\in P$, $d\left(c_i,c_j\right)$ là các liên kết không giống nhau duy nhất giữa 2 cụm và $diam\left(c_k\right)$ là đường kính của cụm c_k dựa trên đánh giá phân vùng P. Giống như MHΓ, Dunn's Validity Index (DVI) cũng đánh giá chất lượng phân nhóm dựa trên chia và tách các cụm trong phân vùng đó, giá trị cao đại diện cho chất lượng phân cụm tốt hơn.

1.3.5 Davies-Bouldin Validity Index

Chỉ số Davis-Bouldin Validity [14] là một hàm của các tỷ lệ của tổng số trong cụm phân tán và giữa các cụm phân tách.

$$DB(P) = \frac{1}{K}\sum_{i,j=1}^{K} \max_{i \neq j} \left\{ \frac{Dist(Q_i) + Dist(Q_j)}{Dist(Q_i, Q_j)} \right\} \quad (1.19)$$

Trong đó K là số cụm. *Dist(Q_i)* là khoảng cách trung bình của tất cả các các đối tượng từ các cụm trung tâm cụm Q_i trong phân vùng P, *Dist*(Q_i , Q_j) là khoảng cách giữa các tâm cụm *(Q_i,Q_j)*. Do đó, chỉ số Davies-Bouldin sẽ có giá trị nhỏ thì kết quả phân cụm tốt hơn.

1.3.6 Normalized Mutual Information

Cho một tập hợp các phân vùng $\{P_t\}_{t=1}^{T}$ thu được từ một tập dữ liệu mục tiêu, *NMI* tiêu chí dựa trên giá trị phân cụm của phân vùng đánh giá P_a được xác định bằng tổng của *NMI* giữa các phân vùng đánh giá P_a và mỗi P_m phân vùng. Do đó, giá trị NMI cao cho chất lượng phân cụm tốt hơn, hàm *NMI* được tính như sau:

$$NMI\left(P_a, P_b\right) = \frac{\sum_{i=1}^{K_a}\sum_{j=1}^{K_b} N_{ij}^{ab} \log\left(\frac{NN_{ij}^{ab}}{N_i^a N_j^b}\right)}{\sum_{i=1}^{K_a} N_i^a \log(\frac{N_i^a}{N}) + \sum_j^b N_j^b \log(\frac{N_j^b}{N})} \quad (1.20)$$

$$NMI(\mathrm{P}) = \sum_{t=1}^{T} NMI(\mathrm{P}, \mathrm{P}_t)$$

Ở đây P_a và P_b là dán nhãn cho 2 phân vùng để phân chia một tập dữ liệu của các đối tượng N vào K_a và K_b cụm tương ứng. N_{ij}^{ab} là số đối tượng được chia sẻ giữa các cụm $C_i^a \in P_a$ và $C_j^b \in P_b$ trong đó N_i^a và N_j^b là đối tượng trong C_i^a và C_j^b.

1.3.7 Dunn's Index (DI)

DI:

$$V_{DI} = \min_{1 \le i \le C} \min_{\substack{1 \le j \le C \\ i \neq j}} \frac{\delta(C_i, C_j)}{\max_{1 \le k \le C} \Delta_k}, \quad (1.21)$$

$$\delta(C_i, C_j) = \min\{d(X_i, X_j) | X_i \in C_i, X_j \in C_j\}. \quad (1.22)$$

Trong những phương trình, $\delta(C_i, C_j)$ là khoảng cách cụm C_i và C_j, Δ_k là khoảng cách trung bình giữa các phần tử cụm đến tâm cụm thứ k^{th} . Giá trị lớn hơn của chỉ số DI có nghĩa là kết quả phân cụm tốt hơn.

1.3.8 Partition Coefficient (PC)

- PC:

$$V_{PC} = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{C}(u_{kj})^2 . \quad (1.23)$$

Giá trị lớn hơn của chỉ số PC có nghĩa là kết quả phân cụm tốt hơn.

1.4 Kết luận chương

Chương này tập trung giới thiệu hai vấn đề chính. Vấn đề đầu tiên, giới thiệu tổng quan về phân cụm, tổng quan về các thuật toán phân cụm mờ tiêu biểu như FCM, KFCM và độ đo phân cụm. Vấn đề tiếp theo, trình bày về khái niệm độ đo phân cụm và một số độ đo tiêu biểu.

Trong chương 2 luận văn sẽ trình bày các thuật toán phân cụm đa mô hình.

CHƯƠNG II: PHÂN CỤM ĐA MÔ HÌNH

2.1. Tổng quan về học đa mô hình và phân cụm đa mô hình

2.1.1 Học đa mô hình

Học đa mô hình là một phương pháp học máy sử dụng nhiều nhóm học để giải quyết cùng một vấn đề. Ngược với cách tiếp cận của các phương pháp học thông thường là cố gắng tìm hiểu một giả thuyết từ dữ liệu huấn luyện, phương pháp học tập hợp xây dựng một tập các giả thuyết và kết hợp chúng để sử dụng [18]. Phương pháp này dùng để cải thiện hiệu xuất và độ chính xác phân loại. Hệ thống phân loại được chia làm nhiều lớp dựa trên sự kết hợp của một tập các phân loại và sự hợp nhất của chúng để đạt được hiệu suất cao hơn. Ý tưởng chính của hầu hết các phương pháp học tập hợp là sẽ sửa đổi các tập dữ liệu huấn luyện , xây dựng n tập đào tạo mới. Trong các mô hình học tập hợp các lỗi và sai lệch của một bộ phận được bù đắp bởi các thành viên khác trong toàn tập hợp. Khả năng tổng quát hóa của phương pháp tập hợp thường mạnh hơn nhiều so với một phân loại đơn. Dietterich [30] đã đưa ra ba lý do bằng cách xem bản chất của máy học như tìm kiếm một không gian cho giả thuyết chính xác nhất.

Lý do đầu tiên là dữ liệu huấn luyện có thể không cung cấp đủ thông tin lựa chọn một bộ phân loại tốt nhất.

Lý do thứ hai là các quá trình tìm kiếm của các thuật toán phân lớp có thể là không hoàn hảo.

Lý do thứ ba là không gian giả thuyết đang được tìm kiếm có thể không chứa hàm đích thực.

Như vậy học đa mô hình là tập hợp các phương pháp có thể bù đắp cho những điều không hoàn hảo trong quá trình tìm kiếm quy luật.

2.1.2 Phân cụm đa mô hình

Phân cụm đa mô hình đã được chứng minh là một lựa chọn tốt khi phải xử lý vấn đề phân tích cụm bao gồm việc tạo ra một tập hợp các cụm từ các số liệu tương tự và kết hợp chúng thành một cụm đồng nhất. Mục tiêu của quá trình kết hợp này là để nâng cao chất lượng phân cụm dữ liệu riêng lẻ. Có nhiều phương pháp phân cụm khác nhau được sử dụng như: phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dựa trên lưới, v.v. Tuy nhiên, mỗi phương pháp có đặc trưng và cách thức thực hiện khác nhau; do vậy không thuật toán nào có thể làm việc hiệu quả trên mọi tập dữ liệu. Phân cụm đa mô hình là cách tiếp cận trong đó kết hợp các giải pháp của các thuật toán phân cụm đơn nhằm thu được nghiệm có chất lượng tốt hơn nghiệm của các thuật toán đơn đó và phản ánh chính xác hơn phân bố của các điểm dữ liệu. Các thuật toán phân cụm đa mô hình được xây dựng theo nhiều tiếp cận khác. Các thuật toán phân cụm đa mô hình có tính ổn định, độ tin cậy, khả năng song song hóa và tính co giãn tốt hơn các thuật toán phân cụm đơn [18].

Vững mạnh: Quá trình kết hợp phải có hiệu suất tốt hơn so với trung bình các thuật toán phân cụm đơn.

Tính nhất quán: Các kết quả của sự kết hợp nên bằng cách nào đó, rất giống với tất cả các kết quả kết hợp thuật toán phân nhóm duy nhất.

Mới lạ: Phân cụm đa mô hình phải cho phép tìm kiếm các giải pháp không thể đạt được bằng thuật toán phân cụm đơn.

Tính ổn định: Kết quả với độ nhạy nhiễu thấp hơn và sự chênh lệch.

2.2 Thuật toán phân cụm đa mô hình CSPA (sCSPA)

Các thuật toán CSPA được [18] đề xuất hoạt động bằng cách đầu tiên tạo ra một ma trận đồng kết hợp của tất cả các đối tượng, và sau đó sử dụng Metis [24] để phân vùng không gian tương tự này để tạo ra số lượng mong muốn của các cụm.

sCSPA mở rộng CSPA bằng cách sử dụng các giá trị trong S để tính toán ma trận tương đồng. Nếu chúng ta hình dung từng đối tượng như là một điểm trong $\sum_{q=1}^{r} k^{(q)}$ chiều không gian, với mỗi chiều tương ứng với xác suất của nó thuộc về một cụm, sau đó SS^T là giống như việc tìm kiếm các điểm trong không gian mới này. Như vậy kỹ thuật đầu tiên biến đổi các đối tượng vào một không gian gán nhãn và sau đó giải thích những điểm giữa các vectơ biểu diễn các đối tượng. Sử dụng khoảng cách Euclide trong không gian gán nhãn để có được độ đo tương tự. Các điểm chấm tìm được là rất cao cùng liên quan với đo Euclide, nhưng khoảng cách Euclide cung cấp đối với ngữ nghĩa tốt hơn. Khoảng cách Euclide giữa v_a và v_b được tính như:

$$d_{v_a,v_b} = \sqrt{\sum_{q=1}^{r}\sum_{i=1}^{k(q)}\left(S_{v_a i}^{(q)} - S_{v_b i}^{(q)}\right)^2} \qquad (2.1)$$

Điều này có thể được giải thích như là một độ đo của sự khác biệt trong các thành viên của các đối tượng cho mỗi cụm. Khác biệt này được chuyển đổi thành một độ đo tương tự bằng cách sử dụng $s_{v_a,v_b} = e^{-d_{v_a,v_b}^2}$.

$$sim(v_a, v_b) = \frac{1}{r}\sum_{i=1}^{k^{(q)}} S_{v_a i}^{(q)} \times S_{v_b i}^{(q)} \qquad (2.2)$$

Thuật giải:

```
function sim_mat=CSPA_SimMat(Um)
    q=size(Um,1);
    data_n=size(Um{1},1);
    cluster_n=size(Um{1},2);
    sim_mat=zeros(data_n,data_n);
    for i=1:data_n-1
        for j=i+1:data_n
            for n=1:q
                for k=1:cluster_n
```

```
                    sim_mat(i,j)=sim_mat(i,j)+(Um{n}(i,k)-
Um{n}(j,k))^2;
                end
            end
        end
    end
    for i=1:data_n-1
        for j=i+1:data_n
            sim_mat(i,j)=exp(-sqrt(sim_mat(i,j)));
            sim_mat(j,i)=sim_mat(i,j);
        end
    end
    for i=1:data_n
        sim_mat(i,i)=1;
    end
end
```

2.3. Thuật toán phân cụm đa mô hình MCLA (sMCLA)

Trong MCLA mỗi cụm được đại diện bởi một vector n-chiều kết hợp. Ý tưởng là để nhóm và thu gọn cụm vào siêu cụm, và sau đó gán từng đối tượng để các siêu cụm trong đó nó tốt nhất. Các cụm được chia nhóm theo phân vùng đồ thị dựa phân cụm. sMCLA là mở rộng MCLA bằng cách chấp nhận phân cụm mềm như đầu vào. sMCLA có thể được chia thành các bước sau:

Xây dựng Meta-Graph của cụm: Tất cả các $\sum_{q=1}^{r} k^{(q)}$ theo từng cụm hoặc chỉ số vector s_i (với trọng số), các siêu cạnh của S, có thể được xem như là đỉnh của một đồ thị vô hướng. Các trọng số cạnh giữa hai cụm s_a và s_b được thiết lập như là $W_{a,b} = Euclidean\_dist(s_a, s_b)$. Khoảng cách Euclide là một thước đo của sự khác biệt về thành viên của tất cả các đối tượng đến hai cụm này. Như trong các

thuật toán SCSPA, khoảng cách Euclid được chuyển đổi thành một giá trị tương tự.

Nhóm các cụm vào siêu cụm: Các Meta-graph xây dựng trong bước trước được phân chia sử dụng để tạo ra METIS k cân bằng siêu cụm. Vì mỗi đỉnh trong Meta - graph đại diện cho một nhãn cụm riêng biệt, một cụm Meta đại diện cho một nhóm các các nhãn cụm tương ứng.

Thu gọn Meta-clusters sử dụng trọng số: Thu gọn tất cả các cụm chứa trong mỗi meta-cluster để tạo thành vector liên kết của nó. Mỗi meta-clusters chứa một giá trị cho mọi đối tượng của nó. Vector liên kết này được tính là trung bình của các vectơ liên kết để mỗi cụm được nhóm lại thành các meta-cluster. Đây là một hình thức có trọng số của các bước thực hiện trong MCLA.

Mã giả:

Input: Data set $X = \{x_1, x_2, ..., x_m\}$;

$C = \left\{ C_j \middle| 1 \le j \le k^* \right\}$;

Process

1. $V = C$;
2. $E = \phi$;
3. for $i = 1, ..., k^*$:
4. for $j = 1, ..., k^*$:
5. if C_i và C_j thuộc về cụm khác nhau
6. then $E = E \cup \{e_{ij}\}$ % thêm cạnh
7. $w_{ij} = \left| C_i \cap C_j \right| / \left(\left| C_i \right| + \left| C_j \right| - \left| C_i \cap C_j \right| \right)$;
8. end
9. end

10. $G=(V,E)$;

11. $\left(C_1^{(M)}+C_2^{(M)}+...C_k^{(M)}\right)$ = METIS(G) ;

12. for $p=1...k$:

13. for $i=1...m$:

14. $h_{pi}^{(M)}=\sum_{C\in C_P^{(M)}}\prod(x_i\in C)\Big/\left|C_P^{(M)}\right|$;

15. end

16. end

17. for $i=1,...,m$:

18. $\lambda_i=\arg\max_{p\in\{1,...,k\}}h_{pi}^{(M)}$;

19. end

20. Output: Phân cụm đa mô hình λ ;

2.4. Thuật toán phân cụm đa mô hình HBGF (sHBGF)

Xét một tập dữ liệu $X=\{x_1,x_2,...,x_n\}$. Phân cụm đa mô hình là tập hợp các giải pháp S phân cụm: $C=\{c_1,c_2,...,c_s\}$. Mỗi giải pháp phân cụm C_l trong đó $l=1,...,S$ là một phân vùng của tập X, tức là $C_l=\{C_l^1,C_l^2,...,C_l^{K_l}\}$ trong đó $\cup_K C_l^K=X$. Với tập hợp các giải pháp phân nhóm C và số cụm K. Mục tiêu là để kết hợp các phân nhóm khác nhau giải pháp là tính toán một phân vùng mới của X vào K cụm rời nhau.

Một phân vùng đồ thị có đầu vào một đồ thị có trọng số và một số nguyên K. Một đồ thị có trọng số G được định nghĩa như là một cặp $G=(V,E)$, trong đó V là một tập hợp các đỉnh và E là một ma trận $|V|\times|V|$ tương tự. Mỗi phần tử E_{ij} của E giống nhau giữa đỉnh V_i và V_j, với $E_{ij}=E_{ji}$ và $E_{ij}>=0\forall i,j$. Cho G và K, các vấn đề về phân vùng G vào đồ thị con K bao

gồm trong tính toán một phân vùng của V thành các K nhóm của đỉnh $V=\{V_1,V_2,...,V_K\}$. Đề xuất phương pháp HBGF để tìm ra một phân vùng K trong đó có sự giống nhau của các trường và cụm. Cụ thể với một cụm $C_l=\{C_1,C_2,...,C_s\}$. HBGF xây dựng một đồ thị hai phía $G=(V,E)$ như sau: $V=V^c\cup V^I$ trong đó mỗi đỉnh của V^c đại diện cho một cụm của tập C và V^I chứa N đỉnh đại diện cho một thể hiện của tập dữ liệu X. Nếu đỉnh i và j đại diện cho từng cụm hoặc các trường hớp $E_{ij}=0$; nếu không i thuộc về cụm j, $E_{ij}=E_{ji}=1$ và 0 nếu ngược lại sử dụng thuật toán đa chiều phân vùng đồ thị để tìm một phân vùng K của đồ thị hai phía [28].

Mã giả

Input: Data set $X=\{x_1,x_2,...,x_m\}$;

$C=\{C_j|1\le j\le k^*\}$; % Biểu đồ phân vùng gói L hoặc METIS

Process:

1. $V=X\cup C$; % Thiết lập đỉnh v_i như trường hợp x_i trong D hoặc cụm C_i trong C ;

2. $E=\phi$;

3. for $i=1,...,k^*$:

4. for $j=1,...,k^*$:

5. if $v_i\in v_j$ % v_i là một trường hợp X ; v_j là một cụm trong C;

6. then $E=E\cup\{e_{ij}\}$ % thêm cạnh $e_{ij}=(v_i,v_j)$

7. $w_{ij}=1$;

8. end

9. end

```
10. G = (V, E);
11.  λ = L(G); %Gọi các gói phân vùng đồ thị trên G
12. Output: Phân cụm đa mô hình λ ;
```

2.5 Thuật toán MG

2.5.1 Phân cụm bởi các thuật toán đơn

Cho một tập dữ liệu X gồm N điểm dữ liệu trong kích thước r. Chia các số liệu vào các cụm C với một số tham số xác định trước như số m và số lượng tối đa các bước lặp. Bước đầu tiên của thuật toán mới được sử dụng một số thuật toán phân cụm mờ đơn lẻ như FCM [5] và KFCM [23] để tạo ra các giải pháp phân cụm khác nhau.

2.5.2 Tổng hợp các kết quả phân cụm đơn

Sau khi nhận được các giải pháp phân cụm đơn tập hợp chúng thành một trong những cách thức như sau. Hãy xem xét các khoảng cách Euclide giữa hai điểm dữ liệu của chương trình đa phân cụm như sau.

$$d_{ij}^{(q)} = d^{(q)}\left(X_i, X_j\right) = \left(\sum_{l=1}^{C(q)} \left(u_{il}^{(q)} - u_{jl}^{(q)}\right)^2\right)^{1/2}, \quad (2.3)$$

$$i, j = \overline{1, N}; i \neq j,$$

Trong đó $U_{il}^{(q)}$ là độ thuộc của các điểm dữ liệu i^{th} đến cụm l^{th} ($i = \overline{1, N}$, $l = \overline{1, C(q)}$) trong kết quả phân cụm q^{th}. Nó có thể là khác nhau $C(q)$ cho kết quả phân cụm khác nhau, nhưng trong trường hợp này $C(q) = C$, $\forall q = 1,2,3$. Ma trận thành viên cho mỗi kết quả phân cụm thỏa mãn các ràng buộc (2.3) sau:

$$\begin{cases} u_{kj}^{(q)} \in [0,1] \\ \sum_{j=1}^{C(q)} u_{kj}^{(q)} = 1 \\ k = \overline{1,N};\ j = \overline{1,C(q)} \end{cases}. \quad (2.4)$$

Ma trận tương tự $S^{(q)}$ cho kết quả phân cụm q^{th} với ($\forall q = 1,2,3$) là tính toán như:

$$S^{(q)} = \sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}^{(q)}, \quad (2.5)$$

$$S_{ij}^{(q)} = e^{-\left(d_{ij}^{(q)}\right)^2}. \quad (2.6)$$

Ma trận tương tự cuối cùng được tổng hợp bởi các tổng trực tiếp của các vector trọng số như sau.

$$S = F\left(S^{(1)}, S^{(2)}, S^{(3)}\right) = \sum_{q=1}^{3} w_q \times S^{(q)}, \quad (2.7)$$

Trong đó w_q là trọng số của các ma trận tương tự $S^{(q)}$ thỏa mãn,

$$\sum_{q=1}^{3} w_q = 1. \quad (2.8)$$

2.5.3 Đi tìm trọng số thích hợp

Theo phương trình (2.7), các trọng số của ma trận tương tự phải được xác định để tính toán ma trận tương tự cuối cùng. Ý tưởng sử dụng một số biện pháp xác định phân cụm bên trong như chỉ số Dunn's (DI) và Partition Coefficient (PC) [22] để tạo ra những trọng số và định nghĩa độ đo.

Từ phương trình (2.7-2.8), kết hợp với độ đo DI, PC công thức sau đây được sử dụng để tạo ra các trọng số:

$$w_q^h = \frac{V_h^{(q)}}{\sum_{q=1}^{3} V_h^{(q)}}, \tag{2.9}$$

$$w_q' = \left(\sum_{h=1}^{2} w_q^h \right) / 2, \tag{2.10}$$

$$w_q = \frac{w_q'}{\sum_{q=1}^{3} w_q'}, \tag{2.11}$$

Trong đó $V_h^{(q)}$ là giá trị của độ đo được xác thực h^{th} (h = 1(DI) or 2 (PC)) cho kết quả phân cụm ($\forall q = 1,2,3$). Bằng cách sử dụng các biện pháp xác thực phân cụm bên trong, các ma trận tương tự cuối cùng nghiêng vào kết quả phân cụm có hiệu quả tốt nhất trong số đó.

2.5.4 Xác định kết quả cuối cùng

Bây giờ, ta có các ma trận tương tự cuối cùng S. Để xác định ma trận thành viên cuối cùng từ S, nó là cần thiết để giải quyết các phương trình:

$$S_{kl} = \sum_{j=1}^{C} u_{kj} u_{lj} + \varepsilon_{kl}, \tag{2.12}$$

Trong đó ε_{kl} là một sai số giữa 2 điểm dữ liệu X_k và X_l.

Các phương pháp Gradient được áp dụng để giải quyết các phương trình (2.12) bằng cách giảm thiểu các tổng sau đây của ô lỗi:

$$\varphi^2 = \frac{\sum_{k=1}^{N} \sum_{l=1}^{N} \left(S_{kl} - \sum_{j=1}^{C} u_{kj} u_{lj} \right)^2}{\sum_{k=1}^{N} \sum_{l=1}^{N} \left(S_{kl} - \overline{S} \right)^2} \to \min. \tag{2.13}$$

Giảm (2.13), ta có:

$$J = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(S_{kl} - \alpha \sum_{j=1}^{C} u_{kj}u_{lj} \right)^2 \rightarrow \min . \qquad (2.14)$$

Lấy đạo hàm của J đối với α, ta được

$$\alpha = \frac{\sum_{k=1}^{N}\sum_{l=1}^{N} S_{kl} \sum_{j=1}^{C} u_{kj}u_{lj}}{\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\sum_{j=1}^{C} u_{kj}u_{lj} \right)^2} . \qquad (2.15)$$

Các vectơ gốc được xác định như sau.

$$\frac{\partial J}{\partial u_{kj}} = -2\alpha \sum_{\substack{l=1 \\ l\neq k}}^{N} u_{lj}\left(S_{kl} - \alpha \sum_{j=1}^{C} u_{kj}u_{lj} \right). \qquad (2.16)$$

Từ (2.15-2.16), các phương pháp sau đây được sử dụng để tìm ra giải pháp cuối cùng.

Input	Data X và số cụm - C
Output	Ma trận thành viên U
MG:	
1	Khởi tạo $U(0)$. Cài đặt số bước: p = 0
2	Thiết lập p = p + 1 và tính toán theo phương trình (2.15)
3	Tính $\frac{\partial J}{\partial u_{kj}(p-1)}$ bởi phương trình (2.19) và tìm ra giải pháp tối ưu đối với hướng các vector gốc bằng bằng cách sử dụng một tìm kiếm trực tiếp.
4	Cập nhật: $u_{kj}(p) = u_{kj}(p-1) - \lambda \frac{\partial J}{\partial u_{kj}(p-1)}$ với $\lambda > 0$ là kích thước

bước.

5 Nếu $\|U(p)-U(p-1)\| < \varepsilon$ thì dừng; Nếu không thì quay về bước 2

Để xác định các cụm và các tâm cụm từ ma trận thành viên tính toán bởi chương trình lặp đi lặp lại ở trên.

2.5.5 Mã giả

Sau đây là mã giả cho thấy các hoạt động của thuật toán, việc xây dựng các giải pháp đơn của FCM, KFCM và GK thuật toán này được thể hiện trong dòng 2, 5 và 8. Các ma trận tương tự cuối cùng được xây dựng trong dòng 14 bằng cách sử dụng các phương pháp tổng hợp phân cụm đơn.

```
1.    U1 = FCM(data,number_of_cluster);
2.    S1 = createSimilarity(U1);
3.    V1 = validity(U1,measure_name);
4.    U2 = KFCM(data,number_of_cluster);
5.    S2 = createSimilarity(U2);
6.    V2 = validity(U2,measure_name);
7.    U3 = GK(data,number_of_cluster);
8.    S3 = createSimilarity(U3);
9.    V3 = validity(U3,measure_name);
10.   SumV=V1+V2+V3;
11.   w1 = V1 / SumV;
12.   w2 = V2 / SumV;
13.   w3 = V3 / SumV;
14.   S=w1*S1+w2*S2+w3*S3;
15.   U=U1;
16.   ax=calculatealpha(U,S);
17.   do {
18.   Uold=U;
```

```
19.   for i=1:n do
      a. for j=i:n do
      b. for k=1:K do
      c. Q(i,j)=Q(i,j) + U(i,k)*U(j,k);
      d. endfor;
      e. Q(j,i)=Q(i,j);
      f. endfor;
20.   endfor;
21.   for a=1:n do
      a. for l=1:K do
      b. for i=1:n do
      c. if (a != i) then
      d. G(a,l) = G(a,l)-2*ax*(S(a,i)-ax*Q(a,i))*U(i,l);
      e. endif;
      f. endfor;
      g. U(a,l)=U(a,l)-step_size*G(a,l);
      h. endfor;
22.   endfor;
23.   ax=calculatealpha(U,S);
24.   } while( max(abs(U-Uold)) > threshold);
```

2.6 Kết luận chương

Trong chương 2 giới thiệu một số thuật toán phân cụm đa mô hình tiêu biểu. Tiếp theo chương 3 xây dựng ứng dụng phân đoạn ảnh viễn thám và kết quả thực nghiệm.

CHƯƠNG III: ỨNG DỤNG PHÂN ĐOẠN ẢNH VIỄN THÁM

3.1 Tổng quan về ảnh viễn thám

3.1.1 Tổng quan

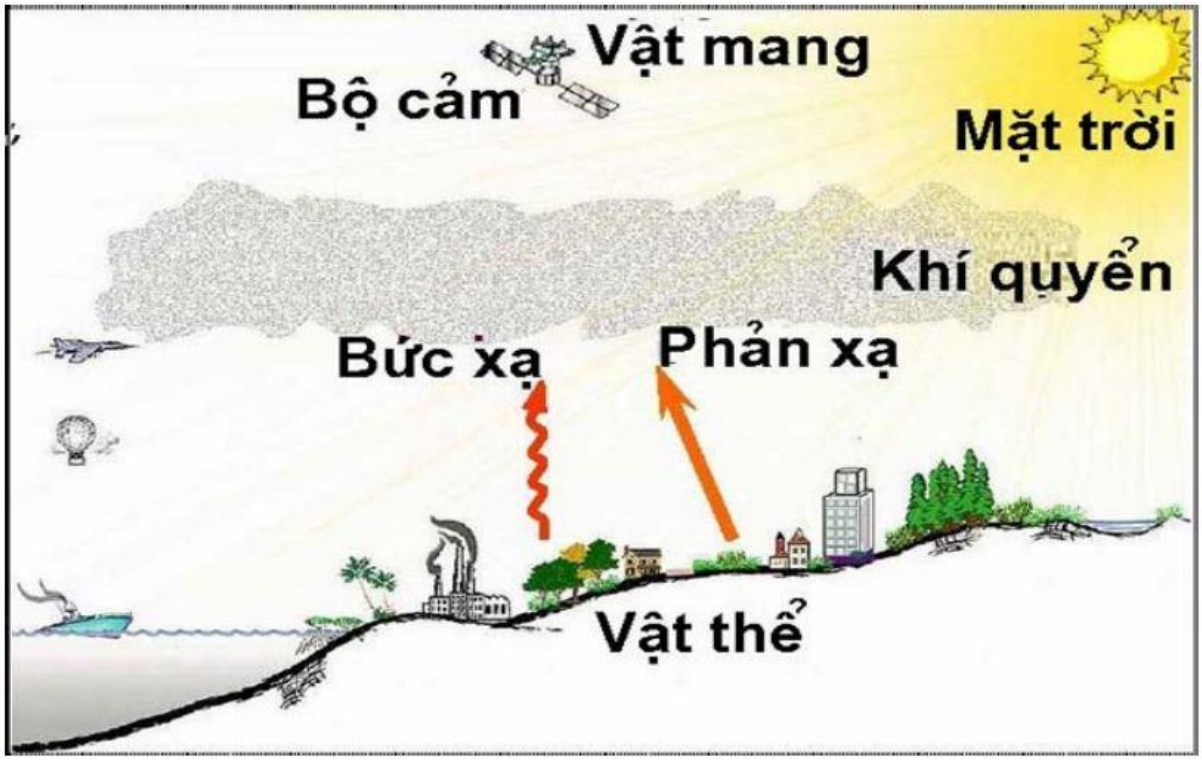
Viễn thám được hiểu là một khoa học để thu nhận thông tin về một đối tượng, một khu vực hoặc một hiện tượng thông qua việc phân tích. Những phương tiện này không có sự tiếp xúc trực tiếp với đối tượng, khu vực hoặc với hiện tượng được nghiên cứu. Thực hiện được những công việc đó chính là thực hiện viễn thám - hay hiểu đơn giản**:** Viễn thám là thăm dò từ xa về một đối tượng hoặc một hiện tượng mà không có sự tiếp xúc trực tiếp với đối tượng hoặc hiện tượng đó. Mặc dù có rất nhiều định nghĩa khác nhau về viễn thám, nhưng mọi định nghĩa đều có nét chung, nhấn mạnh "viễn thám là khoa học thu nhận từ xa các thông tin về các đối tượng, hiện tượng trên trái đất" [2].

3.1.2 Nguyên lý cơ bản của viễn thám

Sóng điện từ được phản xạ hoặc bức xạ từ vật thể là nguồn cung cấp thông tin chủ yếu về đặc tính của đối tượng. Ảnh viễn thám cung cấp thông tin về các vật thể tương ứng với năng lượng bức xạ ứng với từng bước sóng đã xác định. Đo lường và phân tích năng lượng phản xạ phổ ghi nhận bởi ảnh viễn thám, cho phép tách thông tin hữu ích về từng lớp phủ mặt đất khác nhau do sự tương tác giữa bức xạ điện từ và vật thể. Thiết bị dùng để cảm nhận sóng điện từ phản xạ hay bức xạ từ vật thể được gọi là bộ cảm biến. Bộ cảm biến có thể là các máy chụp ảnh hoặc máy quét. Phương tiện mang các bộ cảm biến được gọi là vật mang (máy bay, khinh khí cầu, tàu con thoi hoặc vệ tinh, v.v.) [3].

Nguồn năng lượng chính thường sử dụng trong viễn thám là bức xạ mặt trời, năng lượng của sóng điện từ do các vật thể phản xạ hay bức xạ được bộ cảm biến đặt trên vật mang thu nhận. Thông tin về năng lượng phản xạ của các

vật thể được ảnh viễn thám thu nhận và xử lí tự động trên máy hoặc giải đoán trực tiếp từ ảnh dựa trên kinh nghiệm của chuyên gia. Cuối cùng, các dữ liệu hoặc thông tin liên quan đến các vật thể và hiện thượng khác nhau trên mặt đất sẽ được ứng dụng vào trong nhiều lĩnh vực khác nhau như: nông lâm nghiệp, địa chất, khí tượng, môi trường, v.v. [3].



Hình 3.1. Thể hiện sơ đồ nguyên lý thu nhận ảnh viễn thám [3]

3.1.3 Bộ cảm và máy chụp ảnh

Một thiết bị dùng để cảm nhận sóng điện từ phản xạ hoặc bức xạ từ vật thể được gọi là bộ viễn cảm, thường gọi tắt là bộ cảm. Máy chụp ảnh hoặc máy quét là những bộ viễn cảm.

Các loại máy chụp ảnh sử dụng thông dụng trong viễn thám là máy chụp ảnh hàng không, máy chụp ảnh đa phổ, máy chụp toàn cảnh. Các máy chụp ảnh hàng không được lắp trên máy bay, trên tầu vệ tinh dùng vào mục đích chụp ảnh đo đạc địa hình. Các tư liệu của máy chụp ảnh sử dụng vào mục đích đo đạc nên

cấu tạo của máy chụp ảnh phải thoả mãn các điều kiện quang học và hình học cơ bản như sau:

+ Quang sai của máy chụp ảnh phải nhỏ.

+ Độ phân giải kính vật phải cao và độ nét ảnh phải được bảo đảm trong toàn bộ trường ảnh.

+ Các yếu tố định hướng trong như chiều dài tiêu cự, toạ độ điểm chính ảnh phải được xác định chính xác.

+ Trục quang của ống kính phải vuông góc với mặt phẳng phim.

+ Hệ thống chống nhoè phải đủ khả năng loại trừ ảnh hưởng của chuyển động tương đối giữa vật mang và Trái đất, nhất là khi chụp ảnh từ vệ tinh [3].

3.1.4 Phân loại ảnh viễn thám

Phân loại Phân loại ảnh viễn thám theo nguồn năng lượng và chiều dài bước sóng, ta có thể chia ảnh vệ tinh thành 3 loại cơ bản:

- Ảnh quang học là loại ảnh được tạo ra bởi việc thu nhận các bước sóng ánh sáng nhìn thấy (bước sóng 0.4 – 0.76 micromet). Nguồn năng lượng chính là bức xạ mặt trời

- Ảnh hồng ngoại (ảnh nhiệt) là loại ảnh được tạo ra bởi việc thu nhận các bước sóng hồng ngoại phát ra từ vật thể (bước sóng 8 – 14 micromet). Nguồn năng lượng chính là bức xạ nhiệt của các vật thể.

- Ảnh radar là loại ảnh được tạo ra bởi việc thu nhận các bước sóng trong dải sóng cao tần (bước sóng từ 1mm – 1m). Nguồn năng lượng chính là sóng rada phản xạ từ các vật thể do vệ tinh tự phát xuống theo những bước sóng đã được xác định.

3.2 Nhu cầu thực tế và bài toán phân đoạn ảnh viễn thám

3.2.1 Nhu cầu thực tế

Như phần trên đã chỉ ra, phân cụm được ứng dụng trong nhiều lĩnh vực khác nhau, và một trong số các lĩnh vực đang được quan tâm nhiều hiện nay là phân đoạn ảnh viễn thám. Ngày nay, việc xử lý các hình ảnh viễn thám có vai trò vô cùng quan trọng trong khí tượng, bản đồ, nông – lâm nghiệp, địa chất, môi trường, dự báo thời tiết, dự báo thiên tai liên quan đến biến đổi khí hậu. Đây là công cụ hữu hiệu cho ngành bản đồ, theo dõi biến đổi thảm phủ thực vật, độ che phủ rừng, theo dõi tốc độ sa mạc hóa, phân tích cấu trúc địa chất trên bề mặt cũng như bên trong lòng đất mà trong đó, quá trình phân đoạn thường được yêu cầu như là giai đoạn sơ bộ. Tuy nhiên các phân vùng trong ảnh viễn thám rất phức tạp nên việc phân đoạn chính xác là rất quan trọng [2]. Chính vì thế, thuật toán phân cụm đa mô hình sẽ được ứng dụng cho bài toán phân đoạn ảnh viễn thám nhằm thu được kết quả tốt nhất.

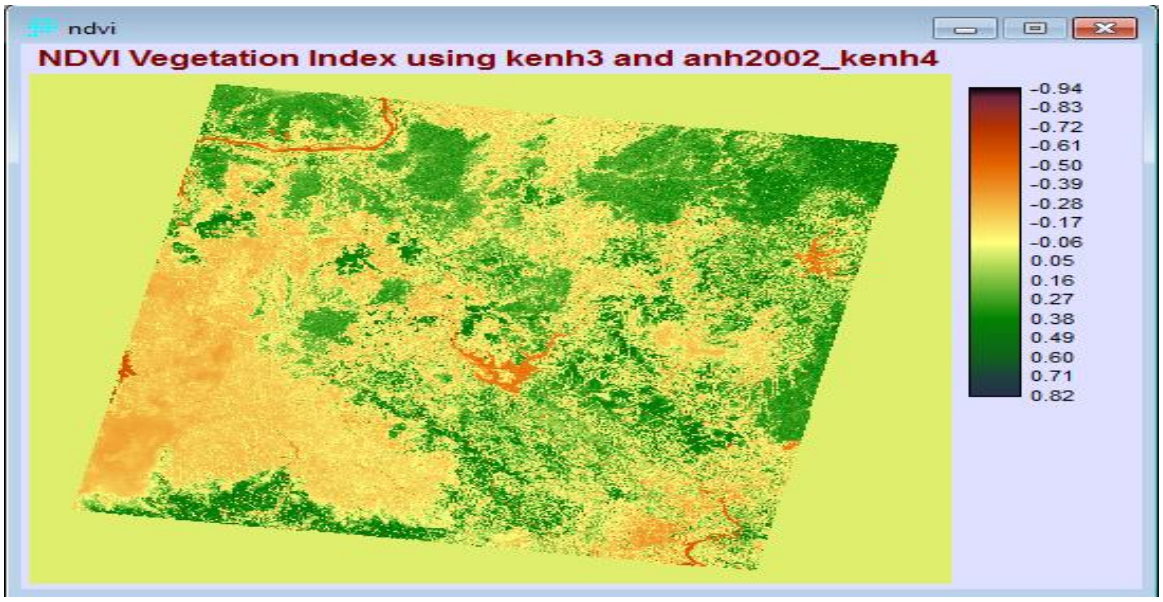
Chính vì thế từ yêu cầu thực tế đặt ra, ứng dụng này nhằm mục đích phân đoạn ảnh viễn thám dựa vào thuật toán phân cụm đa mô hình. Ứng dụng cho phép phân đoạn ảnh viễn thám thành các vùng khác nhau trong đó có vùng cần trích lọc, nhằm loại bỏ những vùng nền không hữu ích trong các quá trình xử lý tiếp theo của hệ thống nhận dạng, phân tích ảnh viễn thám theo yêu cầu thực tế.

3.2.1 Mục đích ứng dụng

Từ yêu cầu thực tế đặt ra viễn thám là một kỹ thuật và phương pháp thu nhận thông tin về các đối tượng từ một khoảng cách nhất định mà không có những tiếp xúc trực tiếp với đối tượng, ứng dụng này nhằm mục đích phân đoạn ảnh viễn thám dựa vào thuật toán phân cụm đa mô hình mờ mà cụ thể là các thuật toán sCSPA, MG. Mục tiêu sử dụng giải thuật phân cụm đa mô hình để thực hiện phân đoạn ảnh viễn thám sử dụng tiêu chí đánh giá theo các chỉ số NDIV, RVI để đưa ra các thông tin cho vùng được khảo sát theo một tiêu chí được đề ra.

3.2.2 Tiêu chí đánh giá theo chỉ số thực vật

Bất kỳ vật thể nào trên bề mặt đất đều có tác dụng điện từ. Đồng thời bất kỳ vật thể nào có nhiệt độ cao hơn nhiệt độ không tuyệt đối (nhiệt độ k =-273,16^{0}C) đều liên tục phát ra sóng điện từ (nhiệt bức xạ). Do thành phần cấu tạo của các vật thể trên bề mặt trái đất khác nhau nên sự hấp thu hoặc phát xạ các sóng điện từ là khác nhau, ngay như thảm thực vật mỗi loại thực vật khác nhau cũng hấp thu và phát xạ các sóng điện từ cũng khác nhau. Vì vậy trên cơ sở các dữ liệu viễn thám ta có thể xác định được các đặc trưng quang phổ khác nhau của của bề mặt trái đất. Trong đó một trong những đặc trưng quang phổ quan trọng nhất của viễn thám là quang phổ thực vật. Từ những đặc trưng này làm cơ sở để xây dựng lên các chỉ số thực vật, là những thông tin quan trọng trong nghiên cứu và phục vụ khí tượng nông nghiệp.



Hình 3.2: Bản đồ chỉ số thực vật (NDVI) bề mặt trái đất theo MODIS [2].

Các chỉ số phổ thực vật được phân tách từ các băng cận hồng ngoại, hồng ngoại và dải đỏ là các tham số trung gian mà từ đó có thể thấy được các đặc tính khác nhau của thảm thực vật như: sinh khối, chỉ số diện tích lá, khả năng quang hợp, tổng các sản phẩm sinh khối theo mùa mà thực vật có thể tạo ra. Những đặc

tính đó có liên quan và phụ thuộc rất lớn vào dạng thực vật bao phủ và thời tiết, đặc tính sinh lý, sinh hoá và sâu bệnh…Công nghệ gần đúng để giám sát đặc tính các hệ sinh thái khác nhau là phép nhận dạng chuẩn và phép so sánh giữa chúng. Đặc trưng cho bề mặt trái đất bao gồm các chỉ số thực vật như sau:

+ Chỉ số thực vật NDVI

$$NDIV = (IR - R) / (IR + R) \quad (3.1)$$

Trong đó IR là giá trị bức xạ của bước sóng cận hồng ngoại, R là giá trị bức xạ của bước sóng nhìn thấy. Chỉ số thực vật được dùng rất rộng rãi để xác định mật độ phân bố của thảm thực vật, đánh giá trạng thái sinh trưởng và phát triển của cây trồng, làm cơ sở số liệu để dự báo sâu bệnh, hạn hán, diện tích năng suất và sản lượng cây trồng.

Để thuận tiện cho việc xử lý ảnh NDVI ta sử dụng công thức chuyển ảnh:

$$Pixel_{value} = (NDIV + 1) \times 127 \quad (3.2)$$

+ Tỷ số chỉ số thực vật RVI

$$RIV = IR / R \quad (3.3)$$

RVI thường dùng để xác định chỉ số diện tích lá, sinh khối khô của lá và hàm lượng chất diệp lục trong lá. Vì vậy chỉ số RVI được dùng để đánh giá mức độ che phủ và phân biệt các lớp thảm thực vật khác nhau nhất là những thảm thực vật có độ che phủ cao.

+ Chỉ số sai khác thực vật DVI hay còn gọi là chỉ số thực vật môi trường EVI, chỉ số thực vật cây trồng CVI.

$$DVI = IR - R \quad (3.4)$$

+ Chỉ số màu xanh thực vật GVI: $GVI=1.6225CH_2 - 2.2978CH_1 + 11.0656$. Trong đó CH_2 và CH_1 là quang phổ của các bước sóng cận hồng ngoại

và bước sóng nhìn thấy của vệ tinh NOAA/AVHRR. Hệ số GVI có ưu điểm là giảm được mức tối thiểu sự ảnh hưởng của đất đai đến chỉ số thực vật.

+ Chỉ số màu sáng thực vật LVI: Năm 1976 R. J. Kauth và G. S Thomas đã tìm được mối liên hệ giữa chỉ số hạn hán thực vật và số liệu vệ tinh TM: LVI=0.3037b1+0.2793b2+0.4743b3+0.5585b4+0.5082b5+0 .1863b7. Trong đó b1-b7 là quang phổ của các bước sóng khác nhau của ảnh vệ tinh TM.

+ Chỉ số úa vàng thực vật YVI:

$$YVI = (R+G)/2 \quad (3.5)$$

Trong đó R là quang phổ bước sóng nhìn thấy (0.63-0.69), G bước sóng xanh (0.52-0.60). Chỉ số này chỉ mức độ hạn hán của thực vật.

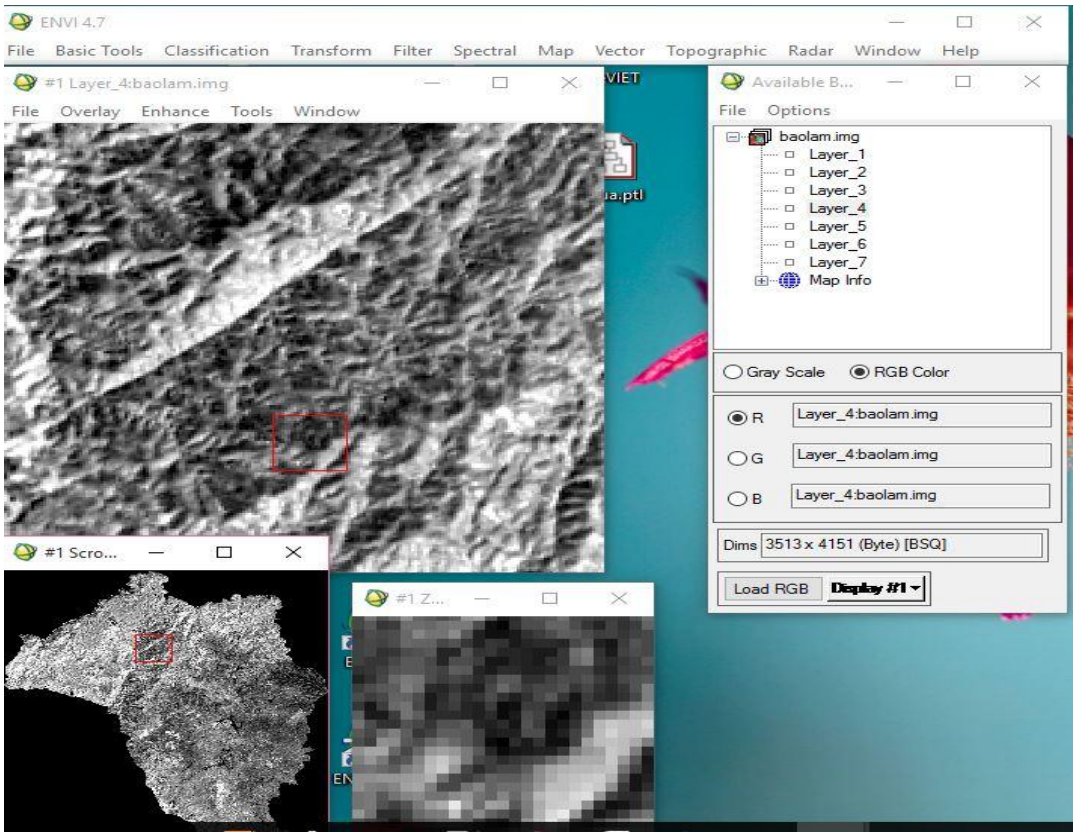
+ Chỉ số màu nâu thực vật BVI:

$$BVI = (b5+b7)/2 \quad (3.6)$$

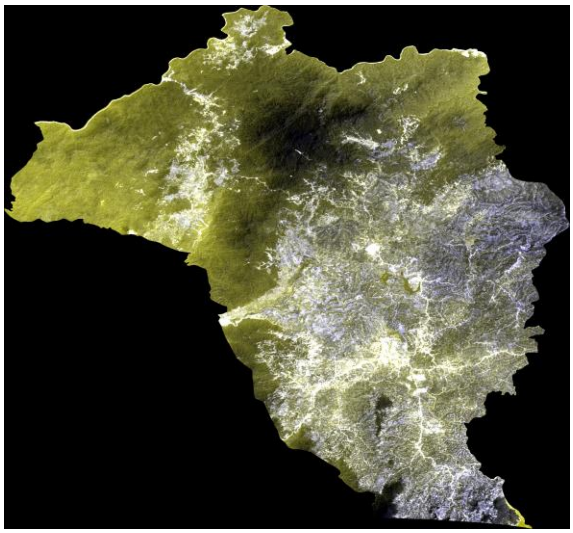
Chỉ số này phản ánh mức độ thiếu nước của thực vật. Chỉ số này còn được dùng để đánh giá tác hại của sâu bệnh đối với cây trồng. Do các chỉ số viễn thám thực vật rất phong phú vì vậy hoàn toàn có khả năng sử dụng các thông tin viễn thám để giải quyết nhiều vấn đề khác nhau trong sản xuất nông nghiệp.

3.3 Đặc tả dữ liệu

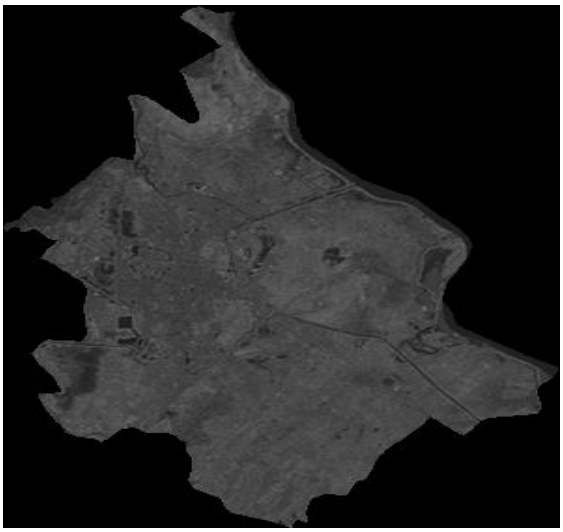
Dữ liệu là bộ ảnh phân lớp thuộc 2 vùng của khu vực huyện Bảo Lâm – tỉnh Lâm Đồng 3.2a và khu vực tỉnh Thanh Hóa 3.2b. Từ ảnh ban đầu của 2 khu vực trên ta sử dụng phần mềm ENVI đọc ảnh và chia ra các kênh khác nhau.



Hình 4: Ảnh sử dụng phần mềm ENVI chia kênh



Hình 5.a



Hình 5.b

Hình 5.a Ảnh là khu huyện Bảo Lâm với diện tích tự nhiên 146.344 ha. Đây là khu vực được bao phủ bởi 7 lớp bao gồm như nước, đá, đất, rừng nguyên sinh, rừng tự nhiên, đất canh tác.

Hình 5.b Ảnh khu vực tỉnh Thanh Hóa với diện tích tự nhiên 11.130,2 km^2 được bao phủ bởi 7 lớp bao gồm như nước, đá, đất, rừng nguyên sinh, rừng tự nhiên, đất canh tác.

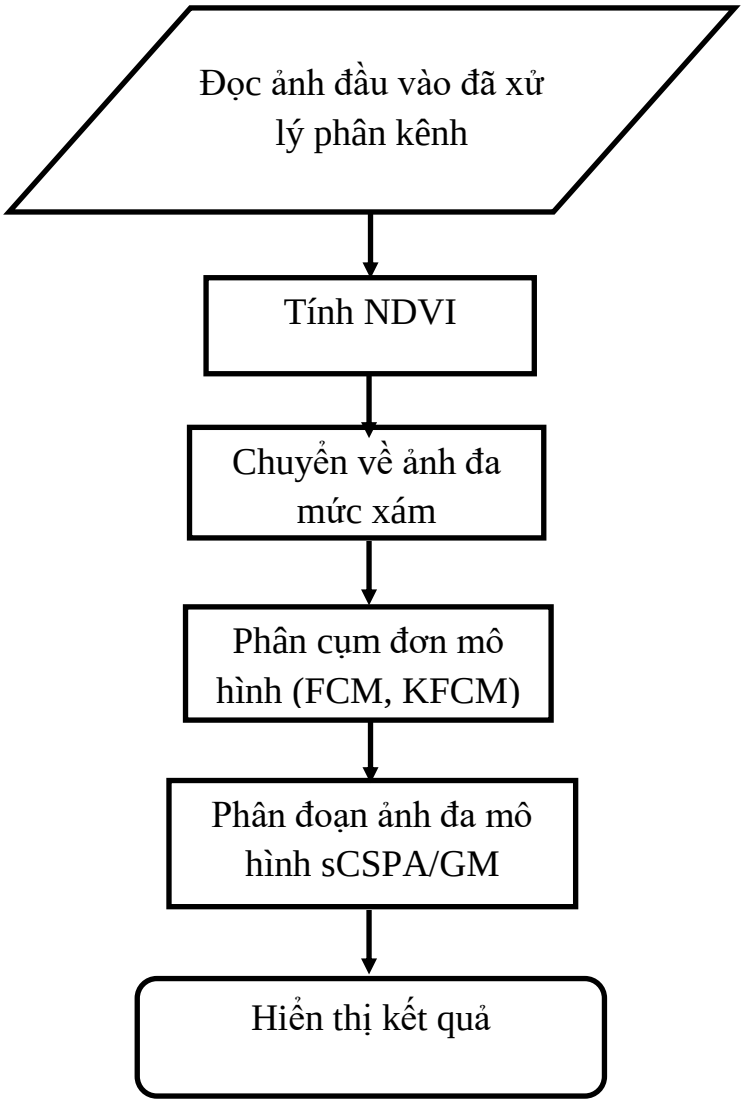
3.4 Các bước phân đoạn ảnh

3.4.1 Tiền xử lý ảnh

Sử dụng phần mềm ENVI là một hệ thống xử lý ảnh khá mạnh. ENVI được thiết kế để đáp ứng yêu cầu của các nhà nghiên cứu có nhu cầu sử dụng dữ liệu ảnh viễn thám, bao gồm các loại ảnh vệ tinh và ảnh hàng không. ENVI hỗ trợ hiển thị dữ liệu và phân tích các dữ liệu ảnh ở mọi kích thước và ở nhiều kiểu định dạng khác nhau. Cho phép làm việc với từng kênh phổ riêng lẻ hoặc toàn bộ ảnh. Khi một file ảnh được mở mỗi kênh phổ của ảnh đó có thể thao tác với tất cả các chức năng hiện có của hệ thống. Với file dữ liệu được mở ta dễ dàng lựa chọn các kênh từ các file ảnh để xử lý cùng nhau.

Từ dữ liệu ảnh ban đa là ảnh đa kênh bao gồm 7 kênh mô tả các phân lớp của ảnh ta sử dụng hai kênh 3 và 4 để thực hiện việc phân đoạn ảnh viễn thám.

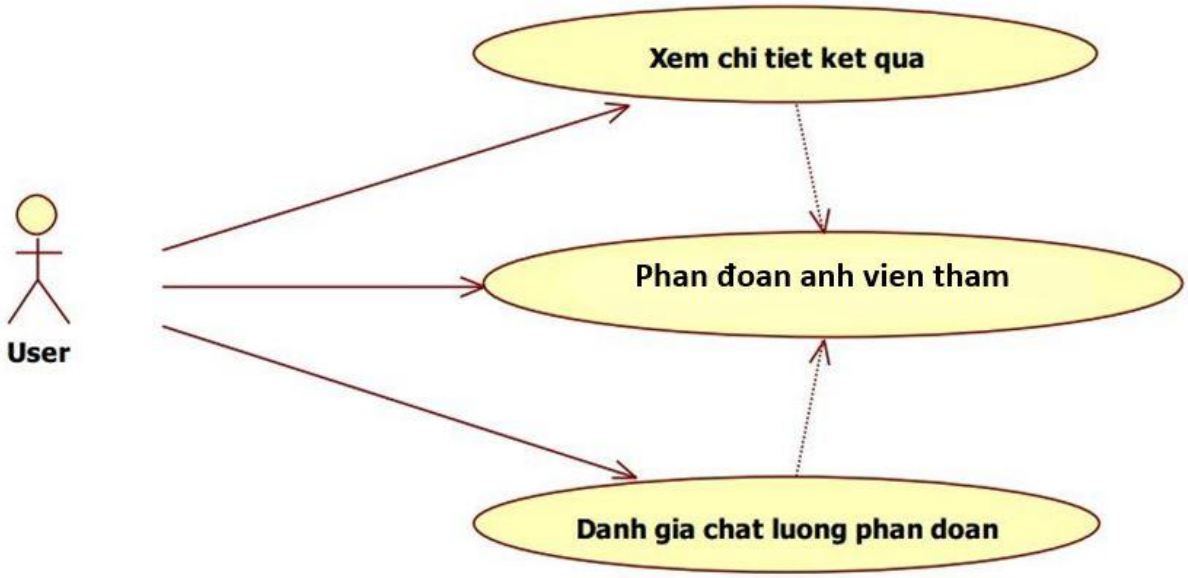
3.4.2 Các bước chính của quá trình phân đoạn ảnh.



Hình 6: Các bước của quá trình phân đoạn ảnh

3.5 Thiết kế hệ thống

Hệ thống cho phép người dùng phân đoạn ảnh viễn thám, xem chi tiết kết quả cũng như thời gian chạy và các độ đo đánh giá chất lượng phân cụm.



Hình 7: Biểu diễn Usecase mô tả chức năng ứng dụng

3.5.1 Chức năng phân đoạn ảnh viễn thám

- Tác nhân: Người dùng

- Input: Ảnh viễn thám cần phân đoạn

- Output: ảnh đã được phân đoạn

- Mô tả chi tiết:

+ Người dùng chọn ảnh cần phân đoạn

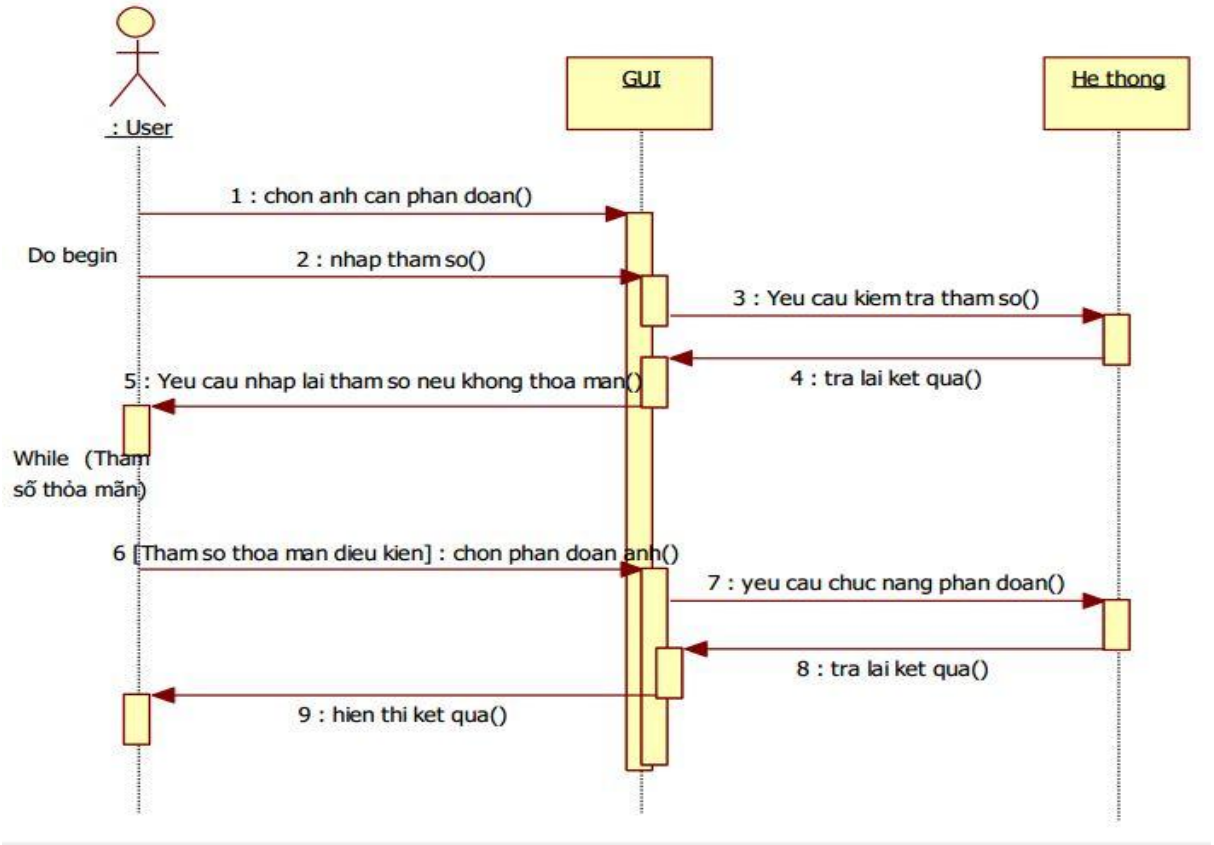
+ Người dùng nhập các tham số

+ Hệ thống kiểm tra tham số và yêu cầu nhập lại cho đến khi thỏa mãn

+ Người dùng chọn phân đoạn ảnh

+ Hệ thống thực hiện phân đoạn đa mô hình sCSPA/GM và trả lại kết quả.

- Biểu đồ trình tự:



Hình 8: Biểu đồ trình tự chức năng phân đoạn ảnh

3.5.2 Chức năng xem chi tiết kết quả

- Tác nhân: Người dùng

- Input: Ảnh đã được phân đoạn và người dùng chọn xem chi tiết kết quả phân cụm (phân đoạn).

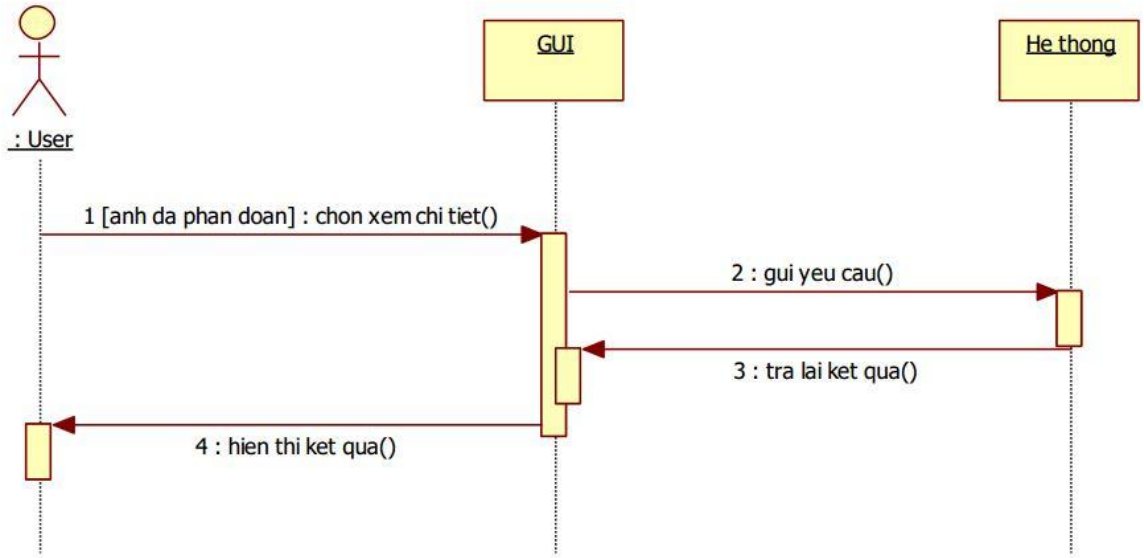
- Output: Kết quả chi tiết được hiển thị

- Mô tả chi tiết:

+ Người dùng chọn chức năng xem chi tiết kết quả

+ Hệ thống hiển thị kết quả chi tiết

- Biểu đồ trình tự:



Hình 9: Biểu đồ trình tự chức năng xem kết quả

3.5.3 Chức năng đánh giá chất lượng phân đoạn ảnh viễn thám

- Tác nhân: Người dùng

- Input: Ảnh đã được phân đoạn và người dùng chọn các độ đo đánh giá kết quả phân cụm (phân đoạn).

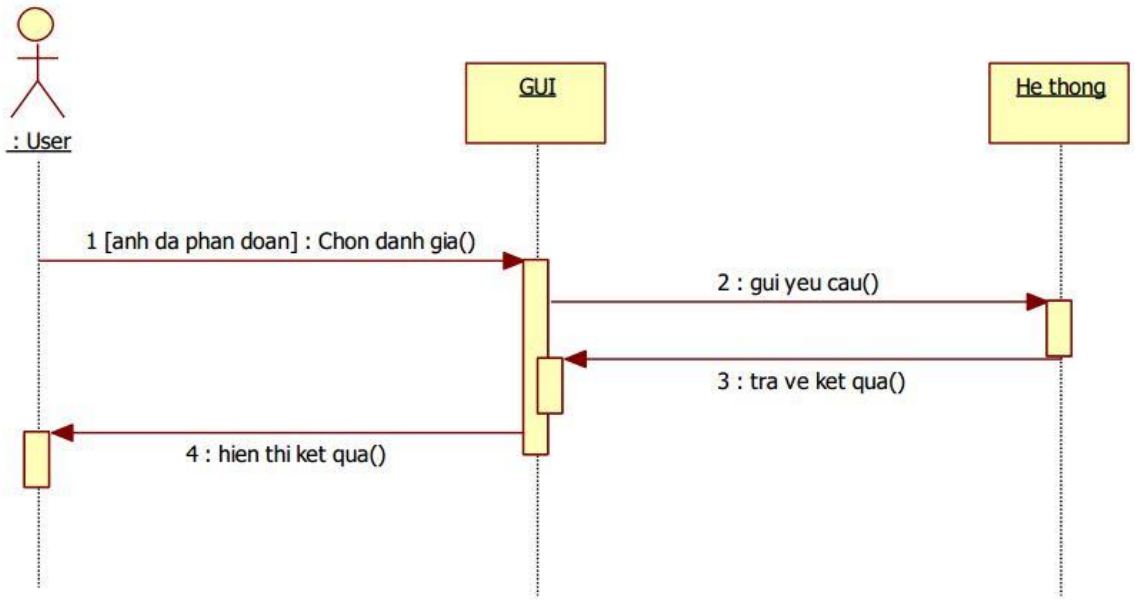
- Output: Kết quả đánh giá được hiển thị

- Mô tả chi tiết:

+ Người dùng chọn chức năng đánh giá kết quả

+ Hệ thống hiển thị kết quả đánh giá.

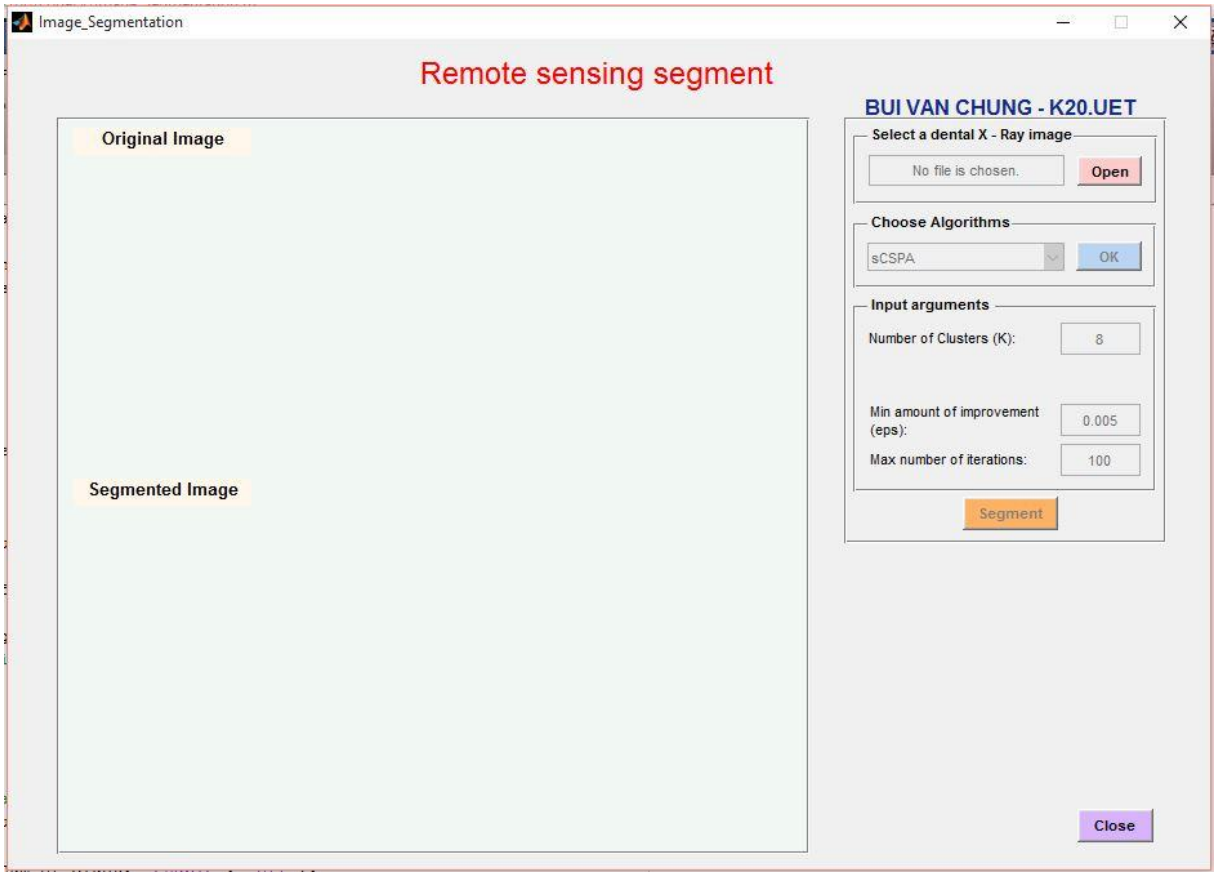
- Biểu đồ trình tự:



Hình 10: Biểu đồ trình tự chức năng đánh giá kết quả

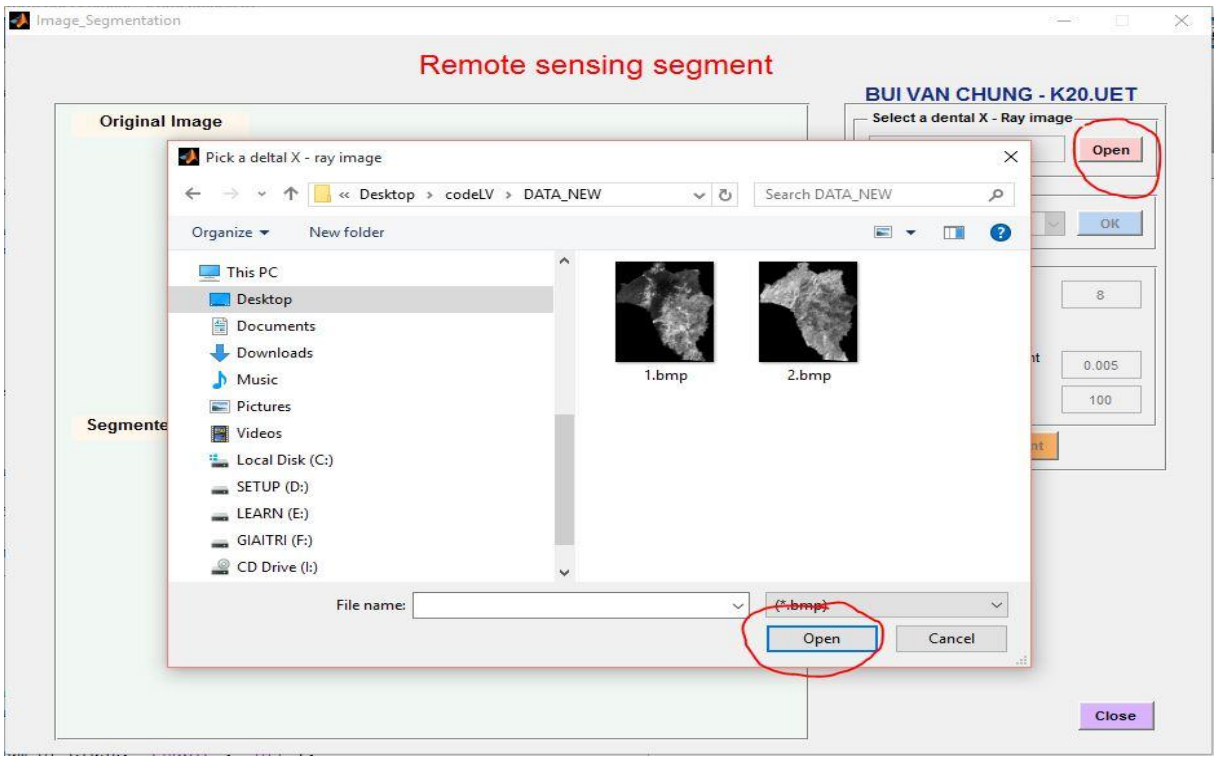
3.6 Minh họa chương trình đánh giá tổng hợp

3.6.1 Giao diện chính của ứng dụng



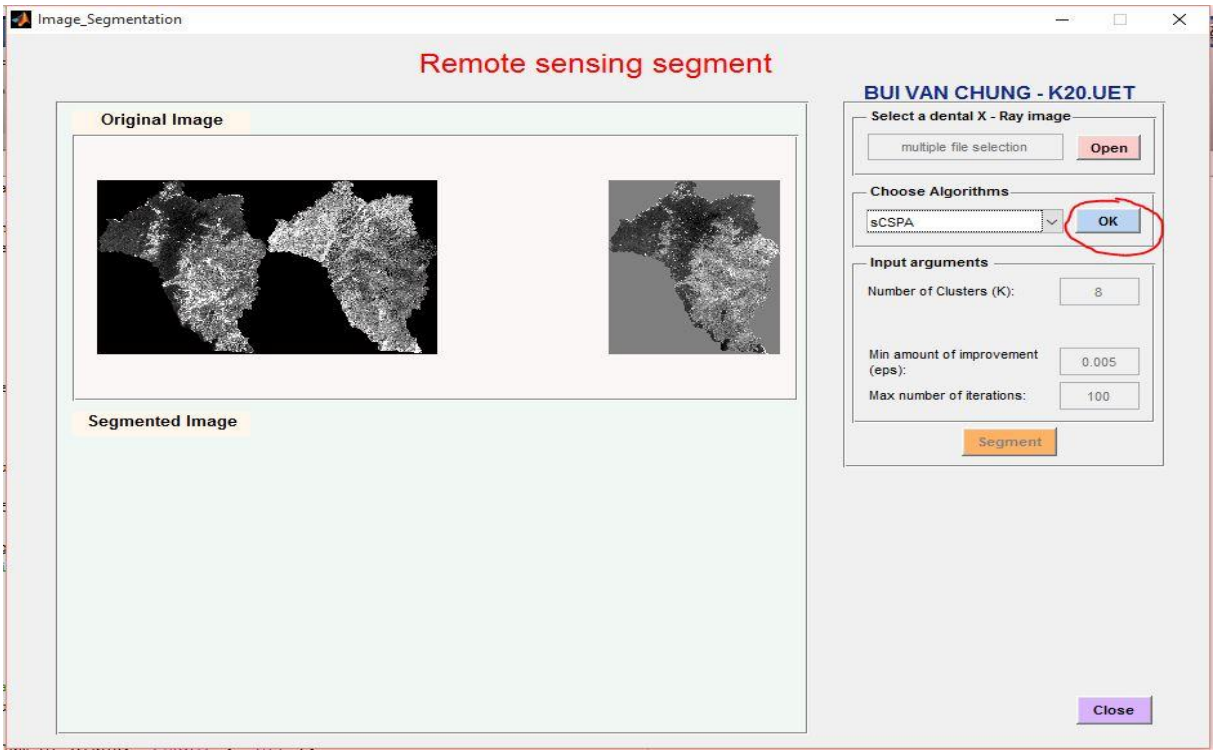
Hình 11: Giao diện chính của phần mềm ứng dụng

3.6.2 Chọn ảnh cần phân đoạn



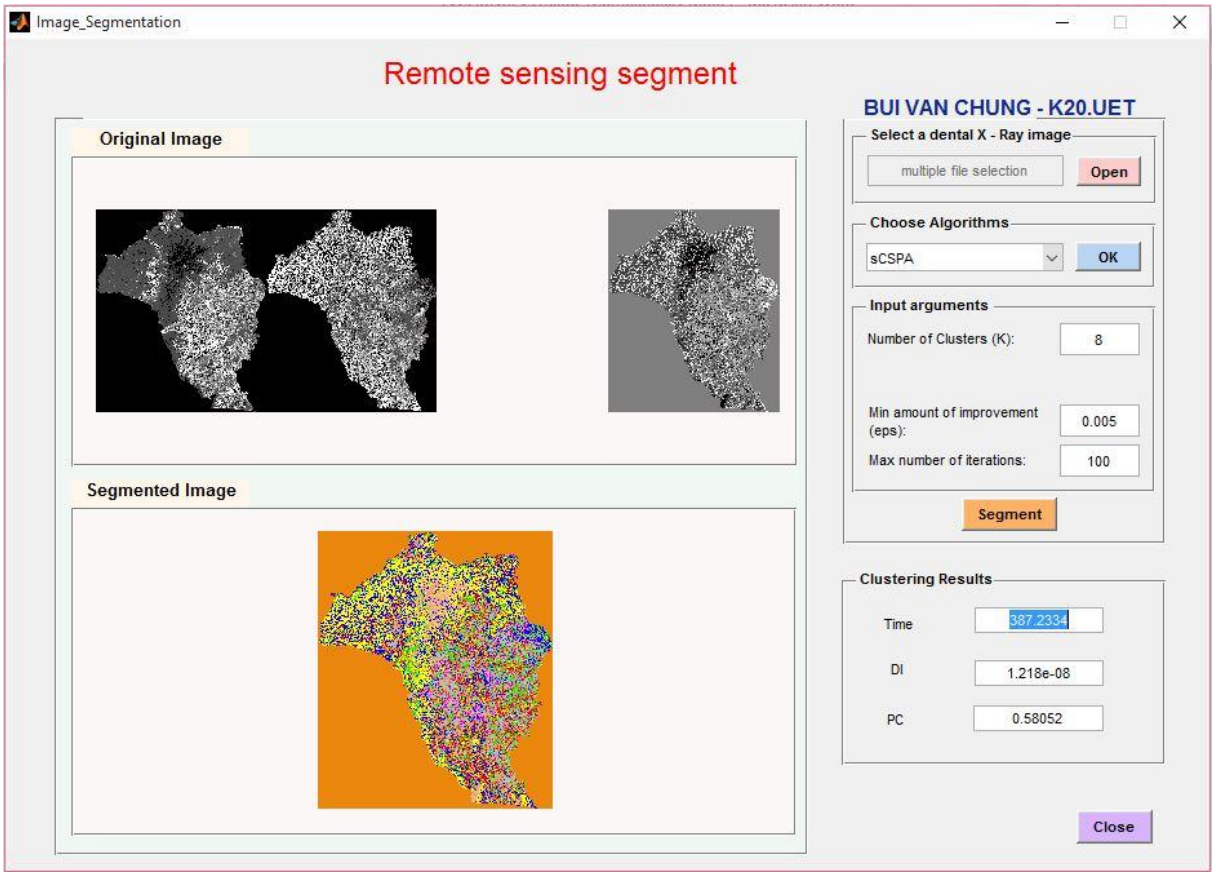
Hình 12: Chọn ảnh cần phân đoạn

3.6.3 Chọn tham số và thuật toán phân đoạn ảnh



Hình 13: Chọn tham số và thuật toán phân đoạn ảnh

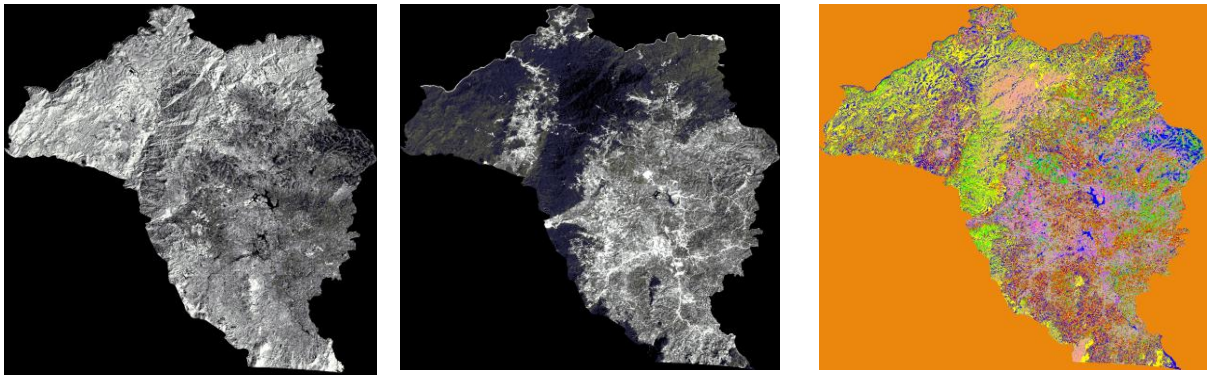
3.6.4 Kết quả phân đoạn ảnh và độ đo



Hình 14: Kết quả phân đoạn ảnh và độ đo

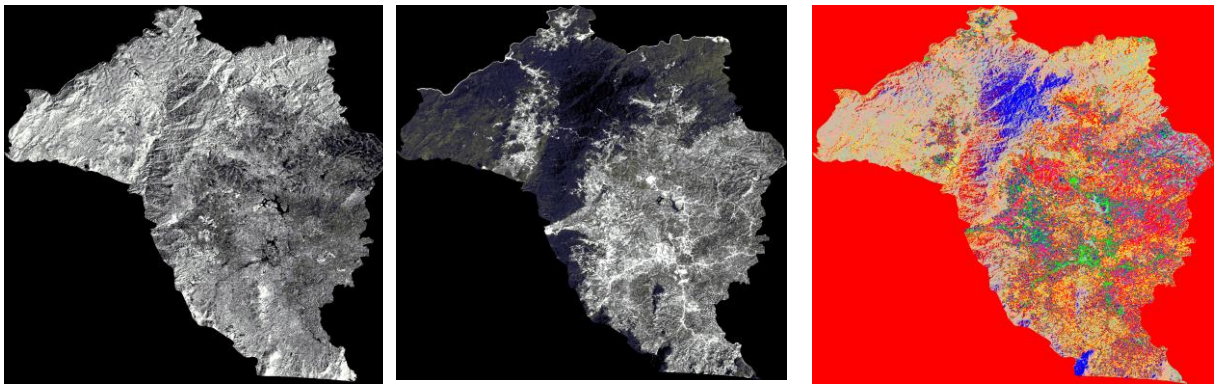
3.7 Kết quả ảnh thu được

3.7.1 Ảnh baolam.img



Ảnh ban đầu (kênh 3) Ảnh ban đầu (kênh 4) Ảnh sau khi phân đoạn sử dụng thuật toán sCSPA

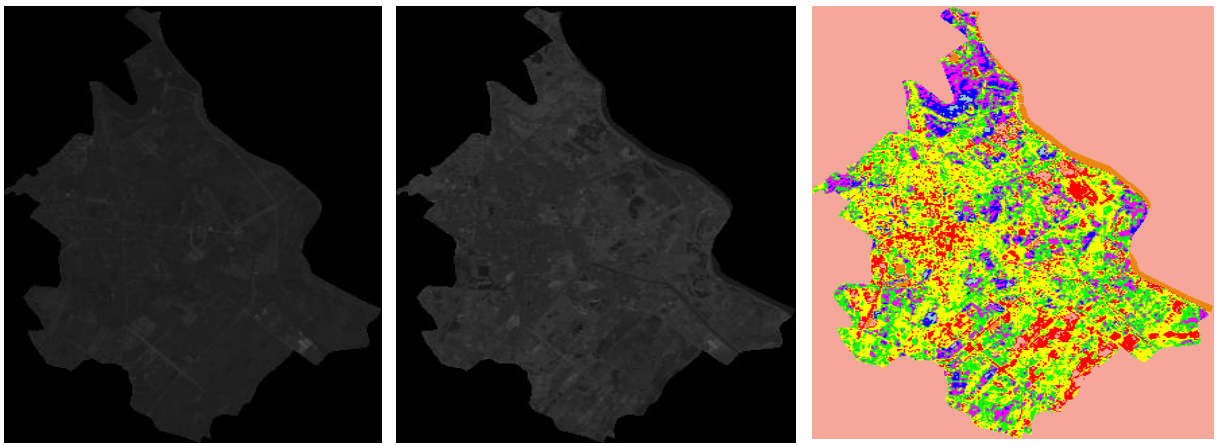
Hình 15: Ảnh baolam.img trước và sau khi phân đoạn sử dụng sCSPA



Ảnh ban đầu (kênh 3) Ảnh ban đầu (kênh 4) Ảnh sau khi phân đoạn sử dụng thuật toán GM

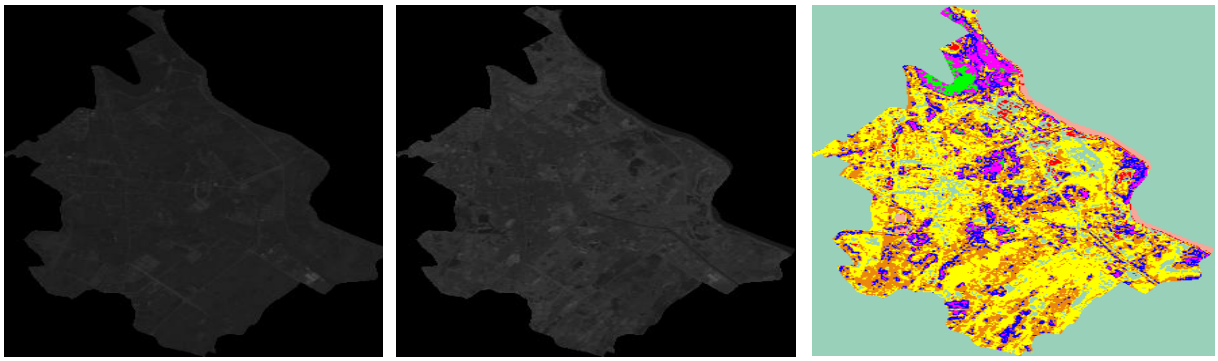
Hình 16: Ảnh baolam.img trước và sau khi phân đoạn sử dụng GM

3.7.2 Ảnh thanhhoa.img



Ảnh ban đầu (kênh 3) Ảnh ban đầu (kênh 4) Ảnh sau khi phân đoạn

Hình 17: Ảnh baolam.img trước và sau khi phân đoạn GM



Ảnh ban đầu (kênh 3) Ảnh ban đầu (kênh 4) Ảnh sau khi phân đoạn

Hình 18: Ảnh baolam.img trước và sau khi phân đoạn sCSPA

3.8 Đánh giá kết quả phân đoạn

Chương trình được cài đặt bằng Matlap Chương trình được chạy thực nghiệm trên máy tính Laptop với thông số kỹ thuật: Intel(R) Core(TM) i3-2330M CPU @ 2.2GHz DDRam3 4Gb.

Kết quả phân đoạn ảnh bởi thuật toán phân cụm đa mô hình sử dụng sCSPA, GM được đánh giá bằng cách so sánh thời gian tính toán, độ đo PC, DI với cùng số cụm đầu vào trên các ảnh.

Ảnh	Số cụm	PC	
		GM	sCSPA

Thanhhoa1993	8	0.49957	0.32681
Thanhhoa2000	9	0.72774	0.33549
Thanhhoa2003	8	0.51785	0.46461
Thanhhoa2009	8	0.68921	0.35549
Thanhhoa2013	8	0.50017	0.32584

Bảng 3.1: Bảng giá trị PC

Từ bảng so sánh trên ta thấy được qua chỉ số độ đo PC ta thấy ở thuật toán MG có giá trị luôn lớn hơn thuật toán sCSPA chứng tỏ thuật toán MG phân cụm tốt hơn.

3.9 Tổng kết chương

Chương III đã mô tả quá trình xây dựng ứng dụng phân đoạn ảnh viễn thám bằng phương pháp phân cụm phân cụm đa mô hình, cụ thể là thuật toán sCSPA, GM: từ đặc tả yêu cầu, thiết kế hệ thống đến triển khai cài đặt chương trình. Từ đó minh họa một cách rõ ràng cách hoạt động, ứng dụng cũng như hiệu quả của thuật toán phân cụm đa mô hình trong phân đoạn ảnh viễn thám. Một số kết quả của các ảnh phân đoạn cũng được đưa ra. Đặc biệt có sự so sánh tính hiệu quả của quá trình phân đoạn giữa thuật toán sCSPA, GM từ đó cho thấy tính giá trị của phân cụm đa mô hình trong ứng dụng phân đoạn ảnh viễn thám.

KẾT LUẬN

Với rất nhiều ý nghĩa trong thực tế, xử lý ảnh ngày càng thu hút sự quan tâm đặc biệt từ các nhà khoa học trên thế giới, đặc biệt là trong xử lý ảnh viễn thám. Trong đó, phân đoạn ảnh được coi như bước cơ bản và thiết yếu đầu tiên trước khi áp dụng các thao tác xử lý ảnh mức cao hơn. Đóng góp chính luận văn:

- Tìm hiểu được những kiến thức tổng quan phân cụm, phân cụm đa mô hình.

- Tổng hợp các phương pháp phân đoạn ảnh đa mô hình, với mỗi phương pháp đều đưa ra thuật toán, đánh giá trực quan về từng thuật toán. Từ đó cho chúng ta có cái nhìn từ tổng thể đến chi tiết các thuật toán đa mô hình trong phân đoạn ảnh viễn thám.

- Cài đặt thuật toán phân cụm mờ đơn FCM, KFCM và thuật toán phân cụm đa mô hình sCSPA, GM để phân đoạn ảnh viễn thám. Trong đó có đưa ra độ đo PC và thời gian chạy để đánh giá chất lượng của kết quả thu được. Từ đó cho thấy tính hiệu quả của thuật toán phân cụm đa mô hình mờ ứng dụng trong việc phân đoạn ảnh viễn thám.

Dựa trên những kết quả bước đầu đã đạt được, trong tương lai, đề tài có thể được phát triển theo các hướng như sau:

- Tiếp tục cải tiến, xây dựng một phương pháp phân cụm đa mô hình mờ mới để đạt được hiệu quả phân đoạn ảnh cao hơn.

- Phát triển hệ thống hỗ trợ, trong đó phân đoạn ảnh viễn thám phục vụ quan trọng trong khí tượng, bản đồ, nông – lâm nghiệp, địa chất, môi trường, dự báo thời tiết, dự báo thiên tai liên quan đến biến đổi khí hậu. Đây là công cụ hữu hiệu cho ngành bản đồ, theo dõi biến đổi thảm phủ thực vật, độ che phủ rừng, theo dõi tốc độ sa mạc hóa, phân tích cấu trúc địa chất trên bề mặt.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt

[1] Bùi Công Cường, Nguyễn Doãn Phước (2006). Hệ mờ, mạng nơron và ứng dụng, *Nhà xuất bản Khoa học kỹ thuật.*

[2] Nguyễn Đình Dương (1998). Bài giảng: Kỹ thuật và các phương pháp viễn thám. *Trường ĐH Mỏ Địa Chất.*

[3] Nguyễn Khắc Thời (2011) Giáo trình: Ảnh viễn thám. *Trường ĐH Nông nghiệp Hà Nội – 2011.*

Tài liệu tiếng Anh

[4] Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers.

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

[6] Dunn, J. C. (1974). "Well-separated clusters and optimal fuzzy partitions." Cybernetics and Systems **4**(1): 95-104.

[7] Davies, D. L. and Bouldin, D. W. (1979). "A cluster separation measure." IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2): 95-104.

[8] Halkidi, M., Batistakis, Y., et al. (2001). "On clustering validation techniques." Journal of Intelligent Information Systems **17**(2): 107-145.

[9] Theodoridis, S., Koutroumbas, K., et al. (1999). Pattern Recognition, Academic Press.

[10] Halkidi, M., Batistakis, Y., et al. (2002). "Cluster validity methods: part I." ACM SIGMOD Record **31**(2): 40-45.

[11] Zhi-Hua Zhou: “Ensemble Methods Foundations and Algorithms”, pages 135–155.Ensemble.

[12] Dunn, J. C. (1974). "Well-separated clusters and optimal fuzzy partitions." Cybernetics and Systems **4**(1): 95-104.

[13] Lesot, M. J., & Kruse, R. (2006). Gustafson-Kessel-like clustering algorithm based on typicality degrees. *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU* (pp. 1300-1307).

[14] Davies, D. L. and Bouldin, D. W. (1979). "A cluster separation measure." IEEE Transactions on Pattern Analysis and Machine Intelligence **1**(2): 95-104.

[15] Vinh, N., Epps, J., et al. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? in the Proceedings of the 26th International Conference on Machine Learning (ICML'09).

[16] Son, L. H., Thong, N. T. (2015). Intuitionistic Fuzzy Recommender Systems: An Effective Tool for Medical Diagnosis. *Knowledge-Based Systems.*

[17] Srivastava, V., Tripathi, B. K., & Pathak, V. K. (2013). Evolutionary fuzzy clustering and functional modular neural network-based human recognition. *Neural Computing and Applications*, *22*(1), 411-419.

[18] Strehl, A., & Ghosh, J. (2003). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, *3*, 583-617.

[19] Alexander Hinneburg, Daniel A. Keim (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. *Knowledge-Based Systems.*

[20] UC Irvine (2015). UCI Machine Learning Repository. Available at: http://archive.ics.uci.edu/ml.

[21] Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, *25*(03), 337-372.

[22] Vendramin, L., Campello, RJ, & Hruschka, ER. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *3*(4), 209-235.

[23] Zhang, D., & Chen, S. (2002). Fuzzy clustering using kernel method. *2002 International Conference on Control and Automation, 2002. ICCA, 2002*.

[24] Karypis G and Kumar V 1998 A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing* **20(1)**, 359–392.

[25] D. E. Gustafson and W. C. Kessel: in *Proc. IEEE CDC*, Vol.2, pp.761-766(1979).

[26] Le Hoang Son, Pham Van Hai (2016). A novel multiple fuzzy clustering method based on internal clustering validation measures with gradient descent. *Inernational Journal of Fuzzy Systems*.

[27] J. Valente de Oliveira and W. Pedrycz: Advances in Fuzzy Clustering and Its Applications. *IEEE Press, Piscataway, NJ*

[28] Bojun Yan and Carlotta Domeniconi. Subspace Metric Ensembles for Semi- supervised Clustering of High Dimensional Data. *IEEE Trans Pattern Anal Mach Intell (TPAMI).*

[29] Fern XZ and Brodley CE 2003 Random projection for high dimensional clustering: A cluster ensemble approach *Proceedings of the Twentieth International Conference on Machine Learning*. ACM Press.

[30] Thomas G Dietterich: Ensemble Methods in Machine Learning. *Oregon State University Corvallis Oregon USA.*