ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÙI VĂN CHUNG

PHÂN CỤM ĐA MÔ HÌNH VÀ ỨNG DỤNG TRONG PHÂN ĐOẠN ẢNH VIỄN THÁM

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÙI VĂN CHUNG

PHÂN CỤM ĐA MÔ HÌNH VÀ ỨNG DỤNG TRONG PHÂN ĐOẠN ẢNH VIỄN THÁM

Ngành: Công nghệ thông tin

Chuyên ngành: Kỹ thuật phần mềm

Mã số: 60.48.01.03

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. Lê Hoàng Sơn

HÀ NỘI - 2016

PHÂN CỤM ĐA MÔ HÌNH VÀ ỨNG DỤNG TRONG PHÂN ĐOẠN ẢNH VIỄN THÁM

Luận văn thạc sĩ ngành: Công nghệ thông tin - Mã số: 60.48.01.03

Người hướng dẫn khoa học: TS. Lê Hoàng Sơn

Học viên thực hiện luận văn: Bùi Văn Chung

Abstract: 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.

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.

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].

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:

1.2. Tổng quan các thuật toán phân cụm tiêu biểu

1.2.1 Phân cụm cụm phân hoạch

1.2.2 Phân cụm phân cấp

1.2.3 Phân cụm dựa trên mật độ

1.2.4 Phân cụm dựa trên mô hình

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 $z_j = \left\{ z_{j1}, \text{z}_{j2}, ..., x_{jd} \right\}$, với $x_i = \left\{ x_{i1}, x_{i2}, ..., x_{id} \right\}$ thành c cụm mờ $1 < c < N$ với tâm cụm $Z = \left\{ z_1, \text{z}_2, ..., \text{z}_c \right\}$, với $z_j = \left\{ z_{j1}, \text{z}_{j2}, ..., x_{jd} \right\}$. 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.

Thuật toán Fuzzy C-Means

FCM được đề xuất bởi Bezdek năm 1974:

- Input

- $X = \left\{ x_1, x_2, ..., x_N \right\}$

- Số cụm c
- Tham số m

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

- **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 $d_{ij}, i = 1, 2..., N; j = 1, 2..., c$

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

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

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

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.

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:

$K(x,y) = \exp(-\|x-y\|^2 / \sigma^2)$ trong trường hợp

$K(x,x) = 1$ thì công thức (1.9) và (1.10) sẽ được viết lại như sau:

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

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

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

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

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

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

1.3.2 Jaccard Index

1.3.3 Modified Hubert's Γ Index

1.3.4 Dunn's Validity Index

1.3.5 Davies-Bouldin Validity Index

1.3.6 Normalized Mutual Information

1.3.7 Dunn's Index (DI)

1.3.8 Partition Coefficient (PC)

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.

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].

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

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 $s_{v_a,v_b} = e^{-d^2_{v_a,v_b}}$.

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

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.

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].

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.

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

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

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:

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

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

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

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

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

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

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

$$\sum_{q=1}^{3} w_q = 1. \tag{2.8}$$

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:

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

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:

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

Giảm (2.13), ta có:

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

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.

2.5.5 Mã giả

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

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].

3.1.3 Bộ cảm và máy chụp ảnh

3.1.4 Phân loại ảnh viễn thám

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

3.2.1 Nhu cầu thực tế

3.3 Đặc tả dữ liệu

3.4 Các bước phân đoạn ảnh

3.4.1 Tiền xử lý ảnh

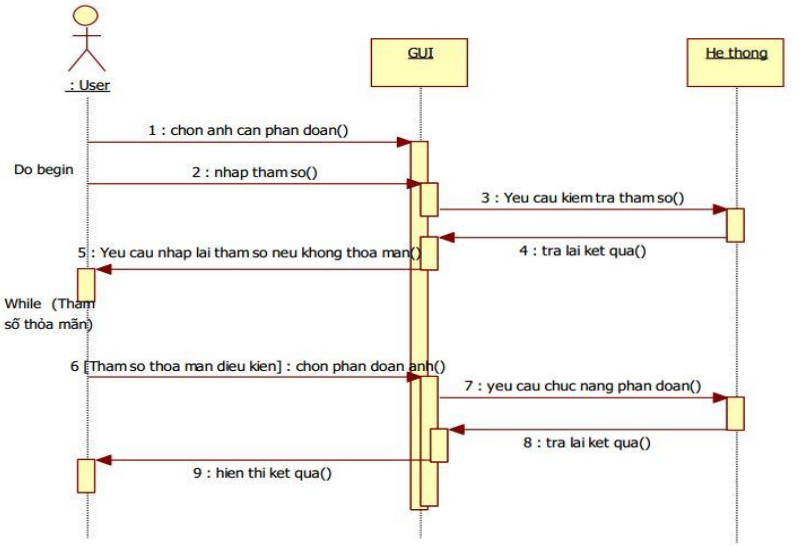
3.4.2 Các bước chính của quá trình phân đoạn ảnh.

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.

3.5.1 Chức năng phân đoạn ảnh viễn thám

- 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

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

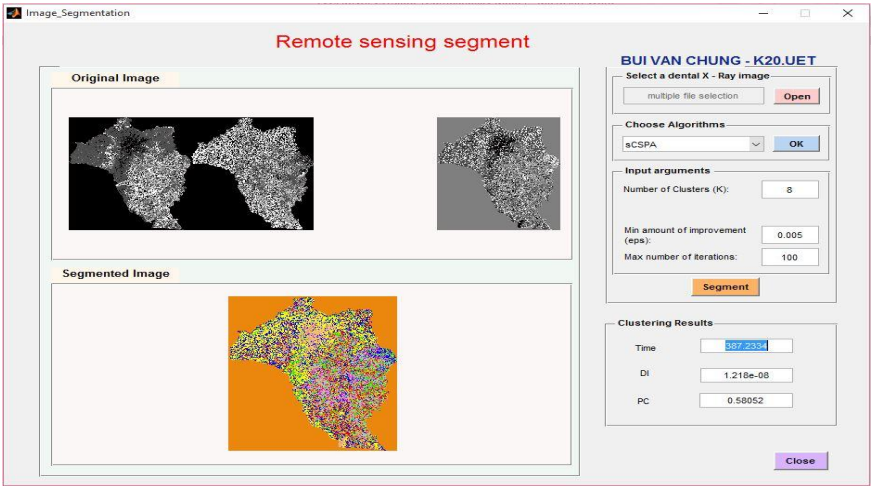
3.6 Minh họa chương trình đánh giá tổng hợp

3.6.1 Giao diện chính của ứng dụng

3.6.2 Chọn ảnh cần phân đoạn

3.6.3 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

3.8 Đánh giá kết quả phân đoạn

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

Luận văn đã trình bày:

- 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.

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, 74*, 133–150.

[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.*