Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), TP. HCM, ngày 23-24/12/2021
DOI: 10.15625/vap.2021.0036

MỘT TIẾP CẬN PHÂN VÙNG ẢNH VIỄN THÁM DỰA TRÊN MAPREDUCE VÀ PHÂN CỤM MỜ

Nguyễn Tư Trung, Nguyễn Ngọc Quỳnh Châu, Trần Mạnh Tuấn
Khoa Công nghệ thông tin, Trường Đại học Thủy lợi, Hà Nội, Việt Nam
trungnt@tlu.edu.vn, chaunnq@tlu.edu.vn, tmtuan@tlu.edu.vn

Tóm tắt: *Phân vùng ảnh viễn thám là bước quan trọng trong quá trình xử lý ảnh viễn thám. Quá trình phân vùng ảnh có các phương pháp cơ bản như sau: dựa trên điểm ảnh, dựa trên biên, dựa trên vùng. Trong đó phương pháp phân vùng ảnh dựa trên điểm ảnh có các phương pháp lấy ngưỡng và phân cụm dữ liệu. Kỹ thuật phân cụm dữ liệu cũng đã có nhiều nghiên cứu chỉ ra hiệu quả trong quá trình phân đoạn ảnh nói chung và ảnh viễn thám nói riêng. Tuy nhiên, với phương pháp này gặp khó khăn trong các bài toán dữ liệu lớn. Với bài toán phân vùng ảnh viễn thám thường kích thước lớn do vậy tốc độ thực thi chậm. Trong nghiên cứu này, nhóm nghiên cứu đưa ra mô hình Mapreduce_Fuzzy sẽ khắc phục được các nhược điểm trong bài toán phân vùng ảnh viễn thám. Trong thực nghiệm cũng chỉ rằng thời gian thực thi của mô hình đề xuất cải thiện hơn cho bài toán phân vùng ảnh viên thám.*

Từ khóa: *Ảnh viễn thám, phân vùng ảnh, phân cụm mờ, MapReduce_Fuzzy.*

I. GIỚI THIỆU

Phân vùng ảnh viễn thám là vấn đề được nghiên cứu từ rất lâu và hiện tại vẫn đang được quan tâm. Ảnh viễn thám ngày càng phức tạp cả về kích thước, số lượng kênh phổ cũng như mức độ chi tiết của ảnh. Có nhiều phương pháp phân vùng khác nhau như: phương pháp dựa trên biên, phương pháp hình thái, phương pháp dựa trên điểm ảnh,... Trong đó một số phương pháp chỉ sử dụng cường độ của mỗi điểm ảnh để phân vùng. Trong [1], Balaji và cộng sự trình bày một phương pháp phân đoạn ảnh mới dựa trên đặc trưng màu từ ảnh với việc chuyển điểm ảnh từ không gian RGB sang không gian L*a*b* và phân cụm trên không gian này. Trong [3], các tác giả đã đề xuất thuật toán KMeans mới sử dụng thay thế tâm cụm. Trong [6], các tác giả cũng kết hợp giữa thuật toán phân cụm mờ và các biểu thức điều chỉnh mức xám khác để tăng cường độ tương phản ảnh y tế. Trong [14], các tác giả đã sử dụng Wavelet để giảm nhiễu cho ảnh y tế. Hiện nay, một số thuật toán sử dụng thêm thông tin ngữ cảnh trong quy trình để giảm bớt tính hỗn tạp của các phân đoạn [10]. Trong [16], các tác giả đã sử dụng tiếp cận cục bộ dựa trên thuật toán phân cụm Fuzzy C-Means để tăng cường độ tương phản của ảnh viễn thám.

Trong một số nghiên cứu về bài toán phân vùng ảnh dưa trên thuật toán KMeans đã cho thấy khi phân vùng ảnh viễn thám kích thước lớn, tốc độ hội tụ của thuật toán vẫn rất chậm. Trong [3], các tác giả đề xuất thuật toán khởi tạo tâm CCEA để tăng tốc độ thuật toán KMeans. Theo [8], KMeans đánh mất đặc trưng ngữ cảnh (thông tin lân cận) của mỗi điểm ảnh khi chỉ xem xét đặc trưng cường độ. Do đó, các tác giả đã đề xuất thuật toán 2D-KMeans với việc bổ sung giá trị trung vị như tham số không gian (thông tin ngữ cảnh cục bộ) để tăng hiệu quả phân cụm [8]. Tuy nhiên, cải tiến này làm tăng dữ liệu lên gấp đôi. Điều này làm giảm tốc độ xử lý dữ liệu nói chung,… và phân cụm so với KMeans gốc rất nhiều nói riêng. Nhược điểm của thuật toán KMeans chỉ xem xét điểm ảnh thuộc về cụm gần nhất. Để cải tiến điều này, thuật toán Fuzzy C-Means [2] khắc phục được vấn đề của thuật toán KMeans. Tuy nhiên, thuật toán Fuzzy C-Means thực thi tốt với các ảnh thường có kích thước nhỏ, với các ảnh viễn thám thì kích thước lớn thì thuật toán Fuzzy C-Means gặp vấn đề liên quan đến bộ nhớ, tốc độ thực thi khi thực hiện phân vùng ảnh kích thước lớn. Kích thước, độ phức tạp của ảnh nói chung và ảnh viễn thám nói riêng ngày càng lớn sẽ là thách thức đối với các phương pháp xử lý dữ liệu truyền thống. Sẽ hiệu quả hơn nếu chúng ta áp dụng các phương pháp xử lý dữ liệu lớn.

Hiện nay, với sự phát triển của công nghệ thông tin, cuộc cách mạng công nghiệp 4.0 dẫn đến sự bùng nổ về dữ liệu (Big Data). Dữ liệu lớn và phân tích của nó đóng một vai trò quan trọng trong thế giới Công nghệ thông tin với các ứng dụng của Công nghệ đám mây, Khai thác dữ liệu, Hadoop và MapReduce [11]. Các công nghệ truyền thống chỉ áp dụng cho dữ liệu có cấu trúc trong khi dữ liệu lớn bao gồm cả dữ liệu có cấu trúc, bán cấu trúc và không có cấu trúc. Làm thế nào để xử lý hiệu quả dữ liệu lớn đã trở thành thách thức lớn trong thời đại mới và cần có những phương pháp xử lý mới. MapReduce là mô hình xử lý dữ liệu phân tán rất hiệu quả, đã và đang được ứng dụng rộng rãi trong xử lý dữ liệu lớn [5]. Trong các công trình [9, 13, 17], các tác giả cũng trình bày một số cải tiến của thuật toán FCM cho việc xử lý dữ liệu lớn.

Trong nghiên cứu này, chúng tôi sẽ trình bày mô hình MapReduce_Fuzzy với việc sử dụng mô hình MapReduce với Fuzzy C-Mean nhằm cải tiến về thời gian tính toán của Fuzzy C-KMeans trong bài toán phân vùng ảnh viễn thám mà không làm giảm chất lượng phân vùng ảnh. Việc cải tiến này được nhóm nghiên cứu dựa trên cơ sở thực nghiệm chỉ ra ở phần IV của bài báo. Nhóm nghiên cứu áp dụng mô hình MapReduce_Fuzzy có thể cải tiếp áp dụng trong một số bài toán big data.

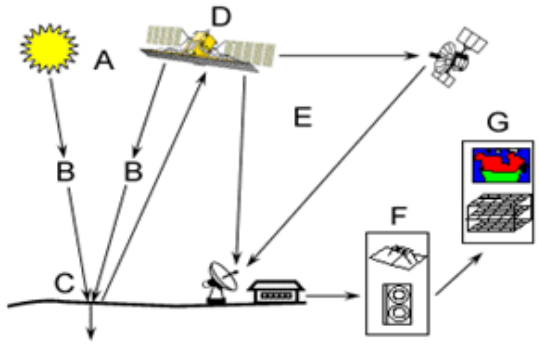
Phần tiếp theo của bài báo được tổ chức như sau: trong phần II, chúng tôi trình bày các nghiên cứu liên quan đến ảnh viễn thám, mô hình MapRedure, phân cụm Fuzzy C-Mean. Phần III, trình bày mô hình đề xuất MapReduce_Fuzzy dựa trên mô hình MapReduce kết hợp phân cụm Fuzzy C-Mean. Trong phần IV, trình bày một số kết quả thực nghiệm trên dữ liệu ảnh viễn thám và đánh giá so sánh mô hình mới với mô hình đã có. Phần cuối cùng là kết luận và hướng phát triển.

II. NGHIÊN CỨU LIÊN QUAN

Trong phần này, Mục 2.1 sẽ trình bày tổng quan về ảnh viễn thám. Mục 2.2 trình trình bày tổng quan về MapReduce và Mục 2.3 trình bày về thuật toán Fuzzy C-Mean.

Tổng quan về viễn thám

Viễn thám là ngành khoa học thu thập từ xa các thông tin trên bề mặt Trái đất [4], nó bao gồm cảm nhận và ghi lại năng lượng phát ra, xử lý, phân tích dữ liệu và ứng dụng các thông tin sau phân tích. Phần lớn các hệ thống thu nhận và xử lý ảnh viễn thám có quy trình 7 bước như trên hình 1 [4].



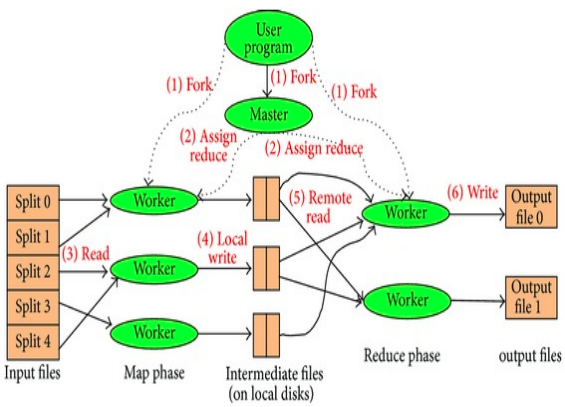
Hình 1. Tiến trình thu thập và xử lý ảnh viễn thám

Trong hình 1, A là nguồn năng lượng hay nguồn sáng, B là Bức xạ và khí quyển, C là Tương tác với đối tượng đích, D là Thu nhận năng lượng bằng đầu cảm biến, E là Truyền, nhận và xử lý năng lượng, F là Diễn giải và phân tích, G là Ứng dụng. Ảnh viễn thám có các đặc trưng: kênh ảnh, độ phân giải không gian, độ phân giải phổ, độ phân giải bức xạ, độ phân giải thời gian. Có nhiều loại ảnh/vệ tinh viễn thám khác nhau như: Vệ tinh Landsat, SPOT, MOS, IRS, IKONOS, WORLD VIEW – 2, COSMOS [12] ... Trong đó, ảnh Landsat 7 ETM+ gồm 8 kênh: Chàm, Lục, Đỏ, Cận hồng ngoại, Hồng ngoại trung (sóng ngắn), Hồng ngoại nhiệt, Hồng ngoại trung (sóng ngắn) và kênh toàn sắc [12]. Ảnh SPOT 5 gồm 5 kênh: Lục, Đỏ, Cận hồng ngoại, Hồng ngoại trung (sóng ngắn) và kênh toàn sắc [12]. Ngoài ra, hiện nay còn có ảnh Quickbird, gồm 5 kênh: Lam, Lục, Đỏ và cận hồng ngoại và kênh toàn sắc.

Với những ưu điểm nổi bật so với các phương pháp truyền thống, công nghệ viễn thám đã được sử dụng rộng rãi và mang lại hiệu quả to lớn trong nông nghiệp, lâm nghiệp…, quản lý tài nguyên thiên nhiên và giám sát môi trường [12]. Tuy nhiên để có được các ứng dụng trên ảnh viễn thám cần phân tích và xử lý ảnh để xác định các đối tượng, thành phần từ ảnh.

Tổng quan về mô hình MapReduce

MapReduce là mô hình xử lý tính toán song song và phân tán do google đề xuất (hình 2) [5]. Nó bao gồm hai chức năng cơ bản: "Map" và "Reduce" được xác định bởi người dùng. Thông qua mô hình MapReduce bao gồm các máy: master và worker. Ứng dụng với từng ngôn ngữ, chương trình có nhiệm vụ phân mảnh tệp dữ liệu đầu vào. Trong đó máy master làm nhiệm vụ điều phối sự hoạt động của quá trình thực hiện MapReduce trên các máy worker, các máy worker làm nhiệm vụ thực hiện Map và Reduce với dữ liệu mà nó nhận được. Dữ liệu được cấu trúc theo dạng key, value.



Hình 2. Sơ đồ mô hình MapReduce

Biểu diễn hình thức mô hình MapReduce: Theo [7, 15], ta có biểu diễn hình thức của mô hình MapReduce như sau:

$$\text{map: (K1 k1, V1 v1)} \rightarrow \text{list(K2 k2, V2 v2)} \quad (1)$$

$$\text{reduce: (K2 k2, list(V2 v2))} \rightarrow \text{list(K3 k3, V3 v3)} \quad (2)$$

Thuật toán Fuzzy C-Means

Thuật toán phân cụm mờ được Bezdek [2] đề xuất dựa trên độ thuộc μ_{ik} của phần tử dữ liệu x_k từ cụm i. Hàm mục tiêu được xác định như sau:

$$J_m = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m d^2(x_k, V_i) \quad (3)$$

Trong đó:

- c: số cụm;
- n: số phần tử dữ liệu;
- μ_{ik}: ma trận độ thuộc của phần tử x_k vào cụm thứ i;
- m: là số mờ hóa;
- x_k: phần tử thứ k;
- V_i: Vector trung tâm của cụm thứ i;
- $d^2(x_k,V_i)$ = Khoảng cách giữa x_k và V_i;

Thành viên (μ_{ik}) được ước lượng với khoảng cách giữa điểm ảnh thứ *k* và tâm cụm thứ i và bị ràng buộc như sau:

$$\begin{cases} 0 \le \mu_{ik} \le 1 \\ \sum_{i=1}^{c}\mu_{ik} = 1 \end{cases} \quad (4)$$

Sử dụng phương pháp Lagrange, xác định được tâm của cụm dựa vào (5) và độ thuộc dựa vào (6) từ hàm mục tiêu (3) và điều kiện rằng buộc (8):

$$V_i = \frac{\sum_{k=1}^{n}(\mu_{ik})^m x_k}{\sum_{k=1}^{n}(\mu_{ik})^m},\ 1 \le i \le c \quad (5)$$

$$\mu_{ik} = \left[\left[\sum_{j=1}^{c}\left(\frac{d(x_k,V_i)}{d(x_k,V_j)}\right)^{\frac{2}{m-1}}\right]\right]^{-1},\ 1 \le i \le c,\ 1 \le k \le n \quad (6)$$

Khi đó thuật toán Fuzzy C-means như sau (Bảng 1).

Bảng 1. Thuật toán Fuzzy C-means

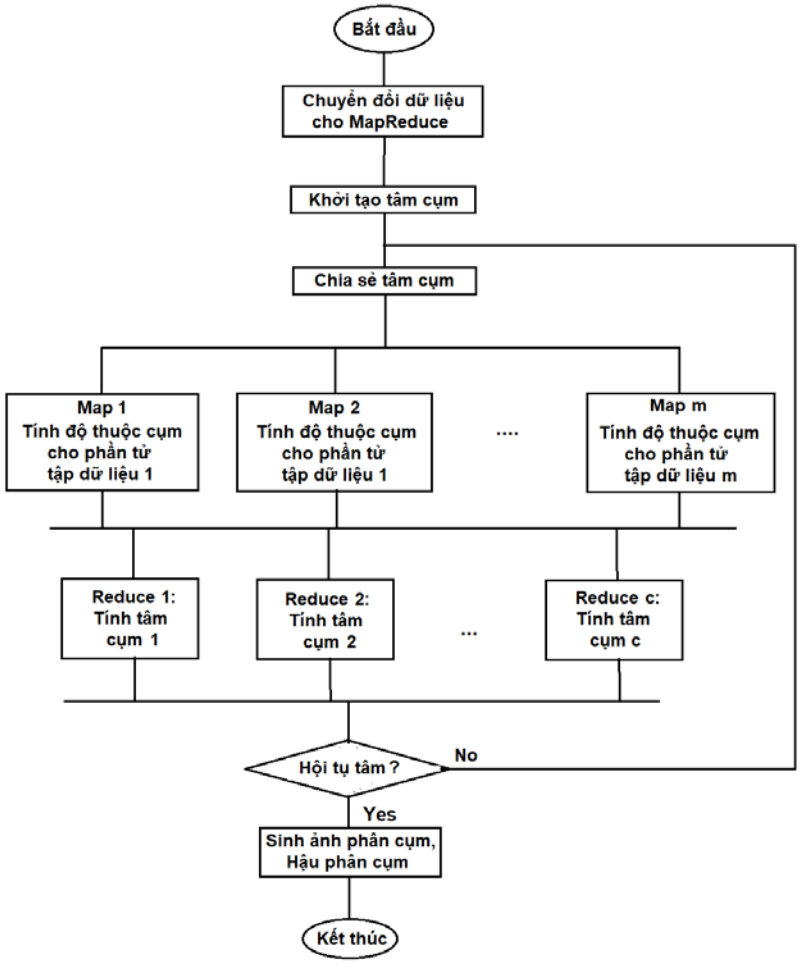
Input	Tập dữ liệu X gồm N phần tử trong không gian r chiều; số cụm C; mờ hóa m; ngưỡng ε; số lần lặp lớn nhất MaxStep>0.
Output	Ma trận μ và tâm cụm V.
FCM	
1	t=0
2	$V_i^{(t)} \leftarrow random; \quad (i = \overline{1,C})$
3	Repeat
4	t=t+1
5	Tính $\mu_{ik}^{(t)}; (k = \overline{1,N}; i = \overline{1,C})$ bởi công thức (6)
6	Tính $V_i^{(t)}; (i = \overline{1,C})$ bởi công thức (5)
7	Until $\left\|U^{(t)} - U^{(t-1)}\right\| \le \varepsilon$ hoặc t > MaxStep

III. ĐỀ XUẤT MÔ HÌNH MAPREDUCE_FUZZY

Trong phần này, chúng tôi trình bày mô hình MapReduce_Fuzzy đề xuất ở Mục 3.1. Trong Mục 3.2 là chứng minh chất lượng của thuật toán Fuzzy C-Means và MapReduce_Fuzzy là tương đương nhau.

A. Mô hình MapReduce_Fuzzy

Trong phần này, chúng tôi đề xuất mô hình MapReduce_Fuzzy cho phân vùng ảnh. Mô hình đề xuất MapReduce_ Fuzzy (hình 4) đầu tiên ảnh đầu vào được được chuyển thành sang dạng list cho xử lý MapReduce. Tiếp đó, các tâm cụm khởi tạo ngẫu nhiên. Tiếp theo, hệ thống sẽ thực hiện phân chia dữ liệu có được thành nhiều mảnh, mỗi mảnh được xử lý song song bởi MapTask (thực hiện việc tính độ thuộc của mỗi điểm dữ liệu thuộc mảnh dữ liệu tương ứng với các tâm cụm theo công thức (6) để thu được dữ liệu trung gian. Sau khi, tất cả các mảnh được xử lý xong bởi MapTask, dữ liệu trung gian sẽ được sắp xếp, trộn, gom nhóm theo từng cụm. Tiếp đó, dữ liệu đã được gom cụm sẽ được xử lý bởi ReduceTask để tính lại các tâm cụm theo công thức (5). Hệ thống thực hiện kiểm tra độ hội tụ của các tâm cụm. Nếu chưa hội tụ thì quay lại tiếp tục thực hiện MapTask và ReduceTask. Nếu đã hội tụ thì sinh ảnh phân cụm và thực hiện các công việc hậu phân cụm.



Hình 4. Sơ đồ thuật toán MapReduce_Fuzzy

1. Chuyển đổi dữ liệu list cho xử lý MapReduce

Chuyển đổi dữ liệu điểm ảnh (thành phần cường độ x_{ij}) thành list các hàng. Mỗi hàng bao gồm: thông tin vị trí (chỉ số dòng, cột) và danh sách giá trị là các thành phần của vector biểu diễn cho một điểm ảnh (thành phần cường độ x_{ij}). Lý do cần thông tin vị trí là để khôi phục lại ảnh phân cụm và thực hiện các việc hậu phân cụm về sau,... Như vậy, đầu ra của việc phân cụm, các phần tử dữ liệu phải bao gồm thông tin cường độ, trung vị và vị trí tương ứng.

2. Biểu diễn hình thức cho các thủ tục map_Fuzzy và reduce_Fuzzy

Input: Mỗi phần tử dữ liệu x_{ij} là một bộ gồm: thông tin vị trí dòng, cột, thành phần cường độ: (i, j, x_{ij}).

Output: Kết quả sau khi hội tụ bao gồm một bộ: chỉ số cụm c và danh sách các phần tử thuộc cụm c là bộ dữ liệu (i, j, x_{ij}).

Khi này, ta xác định được các cặp k1, v1 và k3, v3 như sau:

- k1 là offset, v1 là nội dung dòng dữ liệu (ứng với đối tượng x_{ij}), tức là (i, j, x_{ij});
- k3 là thông tin cụm mới (sau khi tính lại) c_{New}, v3 là danh sách các bộ (i, j, x_{ij}) của các phần tử thuộc về cụm lưu trong k3;

Hàm Map thực hiện việc gán dữ liệu về cụm gần nhất nên suy ra k2, v2 như sau:

- k2 là chỉ số cụm center_ind gần nhất với x_{ij}, v2 là bộ (i,j,x_{ij})

Khi này, ta có biểu diễn hình thức của các thủ tục Map và Reduce như sau:

$$\text{map\_Fuzzy: (offset, } x_{ij}) \rightarrow \text{list(center\_ind, (i,j,}x_{ij}, \mu_{ij}^{c})) \quad (7)$$

$$\text{reduce\_Fuzzy: (center\_ind, list((i,j,}x_{ij}, \mu_{ij}^{c}))) \rightarrow \text{list(}c_{New}\text{, list}(x_{ij}^{cNew})) \quad (8)$$

3. Thủ tục map_Fuzzy

Bảng 2 mô tả thuật toán cho thủ tục map_ Fuzzy. Mục đích thuật toán map_ Fuzzy là tính danh sách độ thuộc với mỗi tâm (trong tập tâm đã chia sẻ) của đối tượng dữ liệu đầu vào.

Bảng 2. Thuật toán cho hàm map_Fuzzy

Input:	Tập tâm chia sẻ lstCenter, key k1 là giá trị offset, value v1 là thông tin đối tượng x_{ij}: info(x_{ij}), tức là (i,j,x_{ij})
Output:	lstK2V2 là list cặp (k2,v2): k2 là chỉ số cụm, v2 là bộ info(i,j,x_{ij}, μ_{ij}^{c})
map_Fuzzy	
1	Trích các phần thông tin cường độ và trung vị: x_{ij}
2	Duyệt cen_ind = 0 đến lstCenter.length
3	Tính μ_{ij}^{c}
4	Khởi tạo bộ (k2,v2)
5	Gán k2 = cen_ind
6	Gán v2 = v1
7	Add bộ (k2,v2) vào lstK2V2
8	return lstK2V2

4. Thủ tục reduce_Fuzzy

Bảng 3 mô tả thuật toán cho thủ tục reduce_Fuzzy. Mục đích thuật toán reduce_Fuzzy là tính lại giá trị tâm cụm mới từ danh sách các đối tượng và độ thuộc về cụm tương ứng.

Bảng 3. Thuật toán cho hàm reduce_Fuzzy

Input:	key là chỉ số cụm cen_ind, value là danh sách thông tin các đối tượng x_{ij} và độ thuộc về cụm có chỉ số cen_ind, tức là list(info(i,j,$x_{ij}, \mu_{ij}^{cen\_ind}$))
Output:	Cặp (k3,v3): k3 là tâm cụm mới tính lại c_{New} (chỉ số cụm cen_ind), v3 là list(info(i,j,$x_{ij}, x_{ij}^{cen\_ind}$))
map_Fuzzy	
1	Khởi tạo mảng c_{New} có số phần tử bằng số chiều của các đối tượng x_{ij}
2	Khởi tạo biến *totalM* = 0
3	Duyệt list((i,j,$x_{ij}, \mu_{ij}^{cen\_ind}$)
4	Trích các phần thông tin cường độ và độ thuộc: $x_{ij}, \mu_{ij}^{cen\_ind}$
5	Tính c_{New} += $x_{ij} * (\mu_{ij}^{cen\_ind})^m$
6	Tính *totalM* += $(\mu_{ij}^{cen\_ind})^m$
7	Chia c_{New} cho *totalM* (từng thành phần) để thu được giá trị tâm mới
8	Gán k3 = c_{New}
9	Gán v3 = list((info(i,j,$x_{ij}, \mu_{ij}^{cen\_ind}$))
10	Return (k3,v3)

5. Sinh ảnh kết quả phân cụm và giai đoạn hậu phân cụm

Từ dữ liệu đầu ra của của hàm reduce_Fuzzy, đơn giản nhất, có thể khôi phục lại ảnh kết quả phân cụm từ thông tin vị trí và giá trị cường độ của tâm cụm,… Ngoài ra, sau đó, chúng ta có thể thực hiện những việc khác đánh giá dữ liệu, phân tích dữ liệu, nhận dạng, phân lớp, ra quyết định,… về sau.

B. Chứng minh chất lượng của thuật toán Fuzzy C-Means và MapReduce_Fuzzy là giống nhau

Mệnh đề: Nếu cùng tập dữ liệu đầu vào, tập tâm khởi tạo thì hai thuật toán FCM (Fuzzy C-Means) và MRF (MapReduce_Fuzzy) cho cùng kết quả phân cụm.

Input: Tập dữ liệu X = {x_1, x_2, …, x_n}, tập tâm khởi tạo C = {c_1, c_k, …, c_m}
Chứng minh:

- Tại vòng lặp đầu tiên: Với mỗi điểm dữ liệu x_i, xét độ thuộc với mỗi tâm c_j là μ_i^j (trong tập C= {c_1, c_k, …, c_m})
  - ✓ Nhận xét (a): Vì khoảng cách d(x_i,c_j) từ điểm dữ liệu x_i đến c_j là như nhau dù được tính trong thuật toán FCM hay MRF nên theo công thức (10), độ thuộc μ_i^j là như nhau dù được tính trong thuật toán FCM hay MRF.
  - ✓ Nhận xét (b): Tất cả các điểm dữ liệu (trong tập dữ liệu đầu vào) và độ thuộc tương ứng sẽ được gom theo mỗi cụm.
  - ✓ Nhận xét (c): Từ nhận xét (b), với mỗi cụm đại diện bởi tâm c_s, c_{s_new} (tính lại) được tính bởi thuật toán FCM hay MRF là giống nhau.

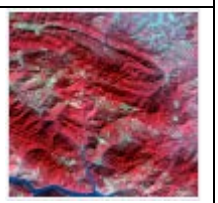
- ✓ Nhận xét (d): Từ nhận xét (c), tập tâm cụm đầu ra của vòng lặp đầu tiên là giống nhau với thuật toán FCM hay MRF.

- Tại vòng lặp thứ 2: Tập tâm được lấy từ đầu ra của vòng lặp đầu tiên nên giống nhau với thuật toán FCM hay MRF.
  - ✓ Với lập luận tương tự như trong vòng lặp đầu tiên thì tập tâm đầu ra của vòng lặp thứ 2 cũng giống nhau với thuật toán FCM hay MRF.
- Các vòng lặp tiếp theo lập luận tương tự như vòng lặp 2 ta đều có kết quả phân cụm giống nhau sau mỗi vòng lặp với thuật toán FCM hay MRF.
- Như vậy, kết quả phân cụm sau khi hội tụ là giống nhau với thuật FCM hay MRF. Đây là ***điều phải chứng minh***.

IV. THỰC NGHIỆM

Chúng tôi tiến hành thử nghiệm thuật toán MRF (MapReduce_Fuzzy) đề xuất và so sánh với kết quả của thuật toán FCM (Fuzzy C-Means). Tập dữ liệu phục vụ cho thử nghiệm gồm ba loại. Một là, loại ảnh Landsat 7 ETM+ chụp khu vực Hòa Bình, bao gồm 11 ảnh ranh giới từng huyện và một ảnh theo ranh giới tỉnh của tỉnh Hòa Bình. Hai là, loại ảnh SPOT 4, chụp khu vực Hòa Bình và Sơn La với 35 ảnh. Ba là, loại ảnh Quickbird, được tải từ dữ liệu mẫu trên trang http://opticks.org. Trong khuôn khổ bài báo có hạn, nhóm tác giả trình việc thử nghiệm với ba mẫu ảnh đầu vào khác nhau như được minh họa trong bảng 3. Trong đó, thực nghiệm 1 là ảnh là ảnh Lansat huyện Đà Bắc, tỉnh Hòa Bình, kích thước 1596×1333; thực nghiệm 2 với ảnh đầu vào là ảnh Quickbird, kích thước 2056×2065; thực nghiệm 3 với anh đầu vào là ảnh SPOT, kích thước 2201×2101. Trong thử nghiệm này, chúng tôi sử dụng công cụ Spark để cài đặt thuật toán MRF sử dụng mô hình MapReduce.

Bảng 4. Các ảnh đầu vào trong các thực nghiệm

Thử nghiệm 1	Thử nghiệm 2	Thử nghiệm 3
		

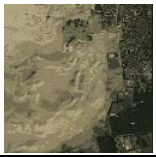
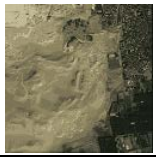
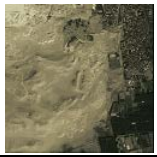
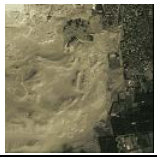
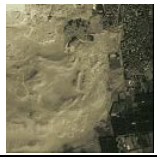
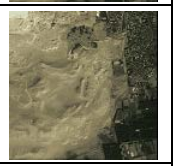
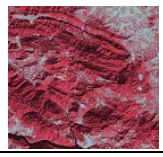
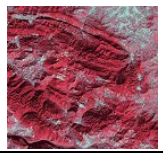
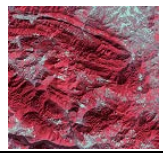
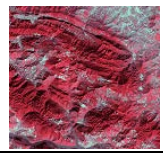
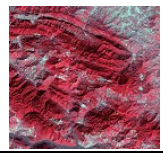
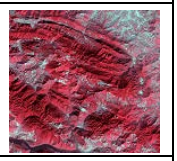
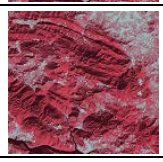
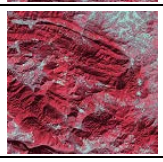
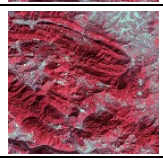
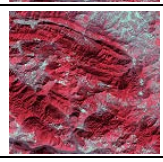
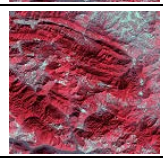
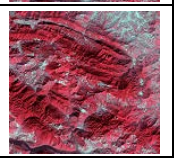
Kết quả thực nghiệm về phân vùng ảnh bằng 2 phương pháp FCM và MRF với số cụm khác nhau thể hiện ở bảng 5.

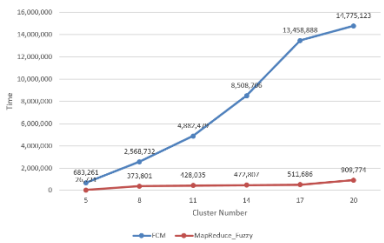
Bảng 5. Thống kê thời gian phân cụm ảnh Lansat huyện Đà Bắc (ms)

	Số cụm	**5**	**8**	**11**	**14**	**17**	**20**
Thực nghiệm 1	**FCM**	683,261	2,568,732	4,882,479	8,508,706	13,458,888	14,775,123
	MRF	26,721	373,801	428,035	472,807	511,686	909,774
Thực nghiệm 2	**FCM**	681,186	2,873,355	5,809,545	15,960,209	18,042,748	26,700,984
	MRF	127,773	336,061	641,885	940,179	1,202,135	2,062,376
Thực nghiệm 3	**FCM**	3,441,597	4,786,402	5,797,639	11,689,740	13,393,833	35,209,072
	MRF	179,329	819,436	854,393	1,004,062	1,110,313	2,744,031

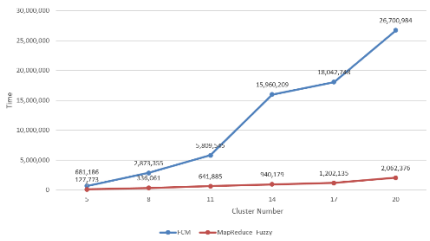
Bảng 6. Ảnh kết quả phân vùng ảnh của 2 phương pháp FCM và MRF trên các thực nghiệm

	Số cụm	**5**	**8**	**11**	**14**	**17**	**20**
Thực nghiệm 1	**FCM**						
	MRF						

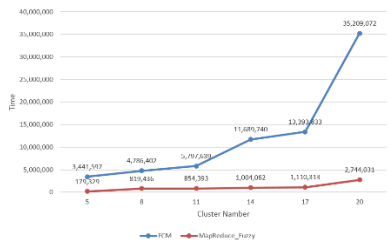
Thực nghiệm 2	**FCM**						
	MRF						
Thực nghiệm 3	**FCM**						
	MRF						



(a) Thực nghiệm 1



(b) Thực nghiệm 2



(c) Thực nghiệm 3

Hình 6. Biểu đồ thời gian phân vùng ảnh dựa trên các thực nghiệm

Kết quả thực nghiệm về thời gian thực thi của 2 thuật toán FCM và MRF được thể hiện ở bảng 6 và hình 6. Với số cụm nhỏ thì thời gian thực thi của 2 phương pháp chênh lệch nhau không nhiều, nhưng số cụm tăng lên thì mức độ chênh lệch về thời gian thực thi của 2 thuật toán chênh lệch nhau rất nhiều. Với thực nghiệm 1, khi số cụm càng tăng thì thời gian thực thi của thuật toán MapReduce_Fuzzy tăng rất chậm (từ 26721 ms đến 909774 ms); thời gian thực thi của thuật toán FCM tăng rất nhanh (từ 683261 ms đến 14775123 ms). Với thực nghiệm 2, khi số cụm càng tăng thì thời gian thực thi của thuật toán MapReduce_Fuzzy tăng rất chậm (từ 127773 ms đến 2062376 ms); thời gian thực thi của thuật toán FCM tăng rất nhanh (từ 681186 ms đến 26700984 ms). Với thực nghiệm 3, khi số cụm càng tăng thì thời gian thực thi của thuật toán MapReduce_Fuzzy tăng rất chậm (từ 179 ms đến 1110313 ms); thời gian thực thi của thuật toán FCM tăng rất nhanh (từ 3441579 ms đến 35209072 ms). Do vậy ở thuật toán MapReduce_Fuzzy cho thấy tốt hơn rất nhiều so với thuật toán FCM về thời gian thực thi nhưng vẫn đảm bảo chất lượng cụm của 2 phương pháp này là như nhau.

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi tập trung vào việc cải tiến thời gian thực thi của phân vùng ảnh viễn thám. Đóng góp chính của nhóm tác giả là đề xuất thuật toán phân cụm ảnh MapReduce_Fuzzy sử dụng mô hình MapReduce để cải thiện tốc độ phân cụm từ thuật toán Fuzzy C-Means, nhưng đảm bảo chất lượng cụm vẫn không bị thay đổi. Đồng thời, chúng tôi đã đưa ra biểu diễn hình thức và thuật toán chi tiết cho thủ tục map_Fuzzy và reduce_Fuzzy. Ngoài ra, chúng tôi đã cài đặt thực nghiệm để đánh giá so sánh giữa 2 thuật toán FCM và MRF. Các kết quả thử nghiệm cho thấy thuật toán MRF cho thời gian phân cụm tốt rất nhiều so với thuật toán FCM mà không làm giảm chất lượng phân cụm.

Trong nghiên cứu tiếp theo, chúng tôi dự kiến áp dụng mô hình MapReduce cho những thuật toán học máy khác để có thể khai thác, phân tích và xử lý dữ liệu lớn hiệu quả hơn.

TÀI LIỆU THAM KHẢO

[1] Balaji T., Sumathi M., Relational Features of Remote Sensing Image classification using Effective KMeans Clustering, International Journal of Advancements in Research & Technology, Volume 2, Issue 8, August-2013, pp. 103-107.

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

[3] Chih-Tang Chang và cộng sự, A Fuzzy KMeans Clustering Algorithm Using Cluster Center Displacement, Journal of Information science and Engineering 27, 2011, pp. 995-1009.

[4] Canada Center for Remote Sensing, Fundamentals of Remote Sensing, http://www.ccrs.nrcan.gc.ca, 2008.

[5] Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, USENIX Association OSDI '04: 6th Symposium on Operating Systems Design and Implementation, 2004.

[6] Hasanien A.E., A. Badr, A Comparative Study on Digital Mamography Enhancement Algorithms Based on Fuzzy Theory, Studies in Informatics and Control, Vol.12, No.1, March 2003.

[7] Herodotos Herodotou, Business Intelligence and Analytics: Big Systems for Big Data, Cyprus University of Technology, 2016.

[8] Intan aidha yusoff, Nor ashidi mat isa, Two-Dimensional Clustering Algorithms for Image Segmentation, WSEAS Transactions on Computers, Issue 10, Volume 10, October 2011.

[9] Ludwig, S.A. MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability. Int. J. Mach. Learn. & Cyber. 6, 923-934 (2015). https://doi.org/10.1007/s13042-015-0367-0.

[10] Meritxell Bach Cuadra, Jean-Philippe Thiran, Satellite Image Segmentation and Classification, Diploma project, Fall 2004.

[11] Nandhini.P, A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce, Int. Journal of Engineering Research and Application, 2018.

[12] Nguyễn Khắc Thời và cộng sự, Giáo trình Viễn thám, Trường Đại học Nông nghiệp Hà Nội, 2011.

[13] Razavi, S. M., Kahani, M. & Paydar, S. Big data fuzzy C-means algorithm based on bee colony optimization using an Apache Hbase. J Big Data 8, 64 (2021). https://doi.org/10.1186/s40537-021-00450-w.

[14] Shruti Dalmiya và cộng sự, Application of Wavelet based KMeans Algorithm in Mammogram Segmentatio, International Journal of Computer Application, Volume 52 - No.15, August 2012.

[15] Tom White, Hadoop: The Definitive Guide: The Definitive Guide, 2009.

[16] Trung Nguyen Tu và cộng sự, Enhancing Remote Sensing Image Contrast based on Combination of Fuzzy Logic and Local Approach, Journal of Information Hiding and Multimedia Signal Processing, Volume 10, Number 4, December 2019.

[17] Q. Yu and Z. Ding, An improved Fuzzy C-Means algorithm based on MapReduce, 2015 8th International Conference on Biomedical Engineering and Informatics (BMEI), 2015, pp. 634-638, doi: 10.1109/BMEI.2015.7401581.

A MAPREDUCE AND FUZZY CLUSTERING BASED ON APPROACH FOR CLUSTERING REMOTE SENSING IAMGES

Nguyen Tu Trung, Nguyen Ngoc Quynh Chau, Tran Manh Tuan

Abtracts: *Clustering is a important step in processing remote sensing images. The basic approachs in the clustering include: the pixel based aproach and the object based approach. Clustering techniques are very effective in clustering images in general and remote sensing images in particular. Some techniques have difficulty in clustering large-sized images such as remote sensing images. In this paper, the authors improve the fuzzy clustering algorithm using Mapreduce model to support clustering of remote sensing images. Experiments show that the execution speed of the proposed algorithm is better than the original algorithm.*