Tạp chí Khoa học Đại học Cần Thơ
Số chuyên đề: Khoa học tự nhiên
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.320

THUẬT TOÁN XÂY DỰNG CHÙM ẢNH DỰA TRÊN CÁC PIXEL MÀU ĐƯỢC TRÍCH XUẤT

Trương Minh Lượng, Nguyễn Kim Ngân, Trần Nam Hưng, Nguyễn Hồng Chi, Phan Như Huỳnh và Võ Văn Tài*
Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ
**Tác giả liên hệ (Corresponding author): vvtai@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 03/04/2024
Sửa bài (Revised): 31/05/2024
Duyệt đăng (Accepted): 03/07/2024

Title: *The clustering algorithm for images based on extracted color pixels*

Author(s): *Truong Minh Luong, Nguyen Kim Ngan, Tran Nam Hung, Nguyen Hong Chi, Phan Nhu Huynh and Vo Van Tai**

Affiliation(s): *Can Tho University*

TÓM TẮT

Trong nhiều lĩnh vực, việc phân chia hình ảnh thành các chùm có thể giúp chúng ta phân loại, nhận dạng các đối tượng trong ảnh cũng như phát hiện được những yếu tố bất thường. Nghiên cứu này đề xuất một thuật toán phân tích chùm cho ảnh dựa vào hàm mật độ xác suất (PDF) được ước lượng từ đặc trưng trích xuất. Đầu tiên, ta đưa một ảnh bất kỳ về 4 màu cơ bản (đỏ, xanh lục, xanh lam, xám) để trích xuất đặc trưng pixel tại mỗi điểm ảnh. Tiếp theo, các PDF đại diện cho đặc trưng trích xuất sẽ được ước lượng để đại diện cho ảnh trong nhận dạng. Cuối cùng, một thuật toán phân tích chùm mờ cho các PDF được đề xuất. Thuật toán đề nghị được trình bày từng bước và được áp dụng trên những tập ảnh cụ thể. Các kết quả số cho thấy thuật toán đề nghị hiệu quả và ổn định, có thể ứng dụng cho nhiều lĩnh vực khác nhau của thực tế.

Từ khoá: *Hàm mật độ xác suất, khoảng cách, phân tích chùm, trích xuất ảnh*

ABSTRACT

In many fields, dividing images into clusters can help us classify, identify objects in the images, and detect anomalies. This study proposes a clustering analysis algorithm for images based on the probability density function (PDF) estimated from extracted features. First, any image is converted into four basic colors (red, green, blue, and gray) to extract pixel features at each point. Next, the PDFs representing the extracted features will be estimated to represent the image in recognition. Finally, a fuzzy clustering algorithm for the PDFs is proposed. The algorithm is presented step by step and applied to specific image sets. Numerical results showed that the proposed algorithm is effective and stable, applicable to various fields of reality.

Keywords: *Clustering, distance, extracting images, probability density function*

1. GIỚI THIỆU

Phân tích chùm là quá trình phân chia dữ liệu thành các nhóm, gọi là "chùm" dựa trên các đặc điểm của dữ liệu. Khi đó, những phần tử trong cùng một chùm có sự tương tự nhiều hơn so với những phần tử thuộc các chùm khác (Strehl & Ghosh, 2002; Chen et al. 2018). Trong thời đại bùng nổ

thông tin hiện nay, mỗi ngày chúng ta đều phải lưu trữ và trích xuất một số lượng dữ liệu khổng lồ, do đó phân tích chùm cần phải được thực hiện. Cũng chính vì thế mà phân tích chùm trở thành một hướng phát triển quan trọng của thống kê và khoa học dữ liệu ngày nay (Wu et al., 2013, Nguyen-Trang et al., 2023a).

Đối tượng phân tích chùm có thể được chia thành hai nhóm: dữ liệu số và dữ liệu ảnh. Phân tích chùm cho dữ liệu số đã được đề xuất đầu tiên và được áp dụng rộng rãi ngày nay (Hung & Yang, 2015; Rao & Liu, 2020; Vo-Van & Nguyen-Trang, 2023b). Với sự phát triển mạnh mẽ của các thiết bị ghi hình, dữ liệu ảnh ngày càng trở nên phổ biến. Ảnh đã trở thành dữ liệu đầu vào không thể thiếu cho nhiều cải tiến công nghệ, đặc biệt liên quan đến trí tuệ nhân tạo và tự động hoá. Theo đó, việc phân chia ảnh thành các chùm có ý nghĩa thực tiễn trong nhiều lĩnh vực khác nhau. Trong lĩnh vực y học, phân chia chùm ảnh đóng vai trò quan trọng trong việc phát hiện các bất thường trong hình ảnh được chụp từ các thiết bị y tế như máy siêu âm, máy MRI và CT scanner. Bên cạnh đó, trong các ứng dụng về quản lý và an ninh, việc phân chùm hình ảnh có thể giúp phát hiện và nhận dạng đối tượng hoặc hoạt động đáng chú ý từ các hình ảnh giám sát. Ngoài ra, việc phân chia ảnh từ vệ tinh có thể được sử dụng để theo dõi, quản lý tài nguyên tự nhiên như rừng và đất, nhằm đánh giá sự thay đổi của rừng, phân biệt giữa các loại đất, hay theo dõi mức độ ô nhiễm môi trường, phân tích mật độ dân số, kết cấu đô thị, và các yếu tố khác.

So với dữ liệu số, phân tích chùm cho ảnh có sự phức tạp hơn khá nhiều. Trích xuất ảnh là bước đầu tiên để thực hiện quá trình phân tích chùm. Thông thường, ảnh sẽ được trích xuất thông qua màu sắc, kết cấu và hình dạng (Nguyen-Trang et al., 2023b). Chúng ta không thể khẳng định sử dụng đặc trưng nào để trích xuất ảnh là tốt nhất cho tất cả trường hợp bởi vì nó phụ thuộc vào tập ảnh và mục đích nhận dạng. Khi đặc trưng của ảnh được trích xuất, chúng phải được biểu diễn thành một đối tượng nào đó để nhận diện. Trong đa số các nghiên cứu, đối tượng nhận diện ảnh là ma trận đặc trưng (Izakian et al., 2016; Qi et al., 2016; Rao & Liu, 2020). Nhận dạng ảnh từ ma trận đặc trưng làm cho bài toán phân tích chùm tốn nhiều thời gian để thực hiện và không hiệu quả cho nhiều trường hợp (Xu et al., 2015, Wang et al., 2020). Nhận dạng ảnh từ các vector hoặc khoảng nhiều chiều từ đặc trưng được trích xuất là một sự cải tiến đáng kể so với ma trận số. Những thuật toán phân tích chùm theo hướng này đã cải tiến được tốc độ tính toán và hiệu quả trong một số trường hợp. Tuy nhiên kết quả không hữu ích cho mọi trường hợp (Montanari & Calò, 2013; Zhu et al., 2021). Nhận dạng ảnh từ hàm mật độ xác suất (PDF) đại diện cho đặc trưng được trích xuất là một sự đột phá, mang đến nhiều lợi ích cho bài toán phân tích chùm. Do đó, hướng nghiên cứu này được sự quan tâm của nhiều nhà thống kê trong thời gian gần đây (Hung & Yang, 2015; Chen and Hung, 2021; Nguyen-Trang et al., 2023b).

Phương pháp phân tích chùm dựa trên hàm mật độ xác suất là một trong những cách tiếp cận để phân chia ảnh thành các nhóm có ý nghĩa. Bằng cách xác định các điểm dữ liệu có mật độ cao trong không gian đặc trưng của ảnh, chúng ta có thể nhận ra các vùng có sự tương đồng và phân chia chúng thành các nhóm riêng biệt. Trong quá trình này, hàm mật độ xác suất được sử dụng để đo lường mức độ "tập trung" của các điểm dữ liệu trong không gian. Các vùng có mật độ cao hơn được coi là các vùng quan trọng và có thể tương ứng với các đối tượng hoặc cấu trúc quan trọng trong ảnh. Bằng cách phân chia ảnh dựa trên các điểm dữ liệu có mật độ cao, chúng ta có thể xác định các vùng có ý nghĩa trong ảnh và nhóm chúng lại với nhau (Vo-Van & Nguyen-Trang, 2018).

Tuy nhiên, tất cả các nghiên cứu về phân tích chùm ảnh dựa trên PDF đại diện đã liệt kê ở trên đều chỉ dựa vào một PDF mà thông thường là PDF đại diện cho màu xám. Việc chỉ dựa vào một màu để trích xuất đặc trưng có thể không phân biệt được sự khác biệt của các ảnh. Chúng ta biết rằng một màu nào đó trong thực tế là sự kết hợp chủ yếu của 3 màu cơ bản RGB (đỏ: Red (R), xanh lục: Green (G), xanh da trời: Blue (B)). Do đó, nếu trích xuất đặc trưng của ảnh cùng lúc từ các màu này có thể cải thiện quá trình nhận dạng, từ đó nâng cao hiệu quả trong xây dựng chùm. Nghiên cứu này khảo sát hiệu quả của thuật toán phân tích chùm ảnh dựa trên đặc trưng được trích xuất từ những không gian màu khác nhau và biểu diễn các đặc trưng này thành các PDF đại diện.

2. CÁC VẤN ĐỀ LIÊN QUAN ĐẾN THUẬT TOÁN ĐỀ NGHỊ

2.1. Ước lượng hàm mật độ xác suất

Có nhiều phương pháp để ước lượng PDF từ dữ liệu rời rạc, bao gồm phương pháp tham số và phi tham số. Trong hai phương pháp trên, phi tham số được xem có nhiều ưu điểm hơn vì không đòi hỏi điều kiện của dữ liệu (Bowman & Azzalini, 1997). Trong các phương pháp phi tham số, hàm hạt nhân được xem là một phương pháp hiệu quả, được sử

dụng phổ biến nhất hiện nay (Nguyen-Trang et al., 2023a). Phương pháp này rất linh hoạt và mạnh mẽ trong việc xử lý dữ liệu không đồng nhất và phức tạp. Về cơ bản, phương pháp này hoạt động bằng cách tạo ra hàm hạt nhân cho từng biến, sau đó kết hợp nhiều hàm hạt nhân để ước lượng PDF cho toàn bộ tập dữ liệu nhiều chiều. Giả sử ta có dữ liệu rời rạc d-chiều, khi đó PDF được ước lượng theo phương pháp hàm hạt nhân có dạng:

$$f(x) = \frac{1}{N}.\frac{1}{h_1 h_2 \dots h_d}\sum_{i=1}^{N}\prod_{j=1}^{d} K_j\left(\frac{x_j - x_{ij}}{h_j}\right),$$

trong đó

N là số phần tử của dữ liệu,

d là số chiều,

h_j là tham số trơn của biến thức j,

x_j là biến thứ $j, j = 1,2, \dots, d$,

x_{ij} là dữ liệu thứ i của biến thứ $j, i = 1,2, \dots, N$,

$K_j(\cdot)$ là hàm hạt nhân của biến thứ j. Hàm hạt nhân phải thỏa mãn hai điều kiện $K_j(x) \geq 0$ và $\int K_j(x)\,dx = 1$.

Một số hàm hạt nhân phổ biến bao gồm hàm tam giác, chữ nhật, Epanechnikov, song lượng và chuẩn tắc. Hiện tại, chưa có hàm hạt nhân nào được xem là tối ưu cho tất cả các loại dữ liệu. Hầu hết các nghiên cứu hiện tại sử dụng hàm hạt nhân dạng chuẩn:

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

Ngoài ra, việc chọn tham số trơn h cũng là một vấn đề trong quá trình ước lượng PDF. Khi tham số trơn h nhỏ thì PDF ước lượng sẽ kém trơn. Ngược lại, khi h càng lớn thì tính trơn của PDF sẽ tăng lên, nhưng sẽ kém chính xác trong ước lượng. Các nhà toán học khẳng định việc chọn tham số trơn quan trọng hơn việc chọn hàm hạt nhân. Nghiên cứu này sử dụng hàm hạt nhân dạng chuẩn và tham số trơn theo Bowman & Azzalini (1997). Cụ thể tham số trơn của biến thứ j được tính như sau:

$$h_j = \left(\frac{4}{N(d+2)}\right)^{\frac{1}{d+4}} \times \sigma_j,$$

trong đó N là số phần tử của dữ liệu, d là số chiều và σ_j là độ lệch chuẩn của biến thứ j.

2.2. Hàm mật độ xác suất đại diện chùm

Cho $\mathcal{F} = \{f_1(x), f_2(x), \dots, f_n(x)\}$ là tập hợp các PDF với $n \geq 2$. Các PDF cần được xếp vào k chùm $\mathcal{C} = \{C_1, C_2, \dots, C_k\}, k \geq 2$. Hàm đại diện của mỗi chùm được xác định bởi công thức:

$$f_{v_i} = \frac{1}{\sum_{j=1}^{n}(\mu_{ij})^2} \times \sum_{j=1}^{n}(\mu_{ij})^2 f_j, i = 1,2, \dots, k,$$

trong đó μ_{ij} là xác suất để hàm mật độ f_j xếp vào chùm C_i. Khi đó, ta nhận thấy hàm đại diện của nhóm f_{v_i} không âm và $\int_{\mathbb{R}^d} f_{v_i}(x)\,dx = 1$. Do đó, hàm đại diện của một nhóm cũng là một PDF.

2.3. Độ đo đánh giá sự tương tự của hai hàm mật độ xác suất

Trong không gian xác suất $(\Omega, \mathcal{F}, \mathbb{P})$, với Ω là không gian mẫu, họ $\mathcal{F}$ các tập con đo được của σ-đại số trong Ω và hàm xác suất $\mathbb{P}: \mathcal{F} \to [0;1]$ gồm hai PDF f và g. Khi đó, ta có các khoảng cách sau:

Khoảng cách $\mathcal{L}^1$:

$$\|f,g\|_1 = \int_{\mathbb{R}^n} f_{max}(x)\,dx - \int_{\mathbb{R}^n} f_{min}(x)\,dx = 2\left(\int_{\mathbb{R}^n} f_{max}(x)\,dx - 1\right).$$

Khoảng cách Divergence:

$$\mathcal{D}_{DVG}(f \parallel g) = 2\int_{-\infty}^{+\infty}\frac{(f(x)-g(x))^2}{(f(x)+g(x))^2}dx.$$

Khoảng cách Kullback-Leibler:

$$\mathcal{D}_{KL}(f \parallel g) = \int_{-\infty}^{+\infty} f(x).\ln\left(\frac{f(x)}{g(x)}\right)dx.$$

Khoảng cách Jensen-Shannon:

$$\mathcal{D}_{JS}(f \parallel g) = \frac{1}{2}\mathcal{D}_{KL}(f \parallel h) + \frac{1}{2}\mathcal{D}_{KL}(g \parallel h),$$

trong đó $h = \frac{1}{2}(f+g)$.

Khoảng cách Bhattacharyya:

$$\mathcal{D}_B(f,g) = \int_{\mathbb{R}^n}[f(x)g(x)]^{\frac{1}{2}}dx.$$

Các khoảng cách trên được sử dụng để đánh giá sự tương tự của hai PDF. Khi khoảng cách càng nhỏ thì sự tương tự của chúng càng lớn và ngược lại, khoảng cách càng lớn thì sự tương tự của chúng càng nhỏ. Có nhiều nghiên cứu việc chọn khoảng cách tối ưu cho bài toán nhận dạng thống kê. Tuy nhiên, chưa có khoảng cách nào được xem là tốt nhất

cho mọi trường hợp. Nghiên cứu này sử dụng khoảng cách $\mathcal{L}^1$ để xây dựng thuật toán phân loại.

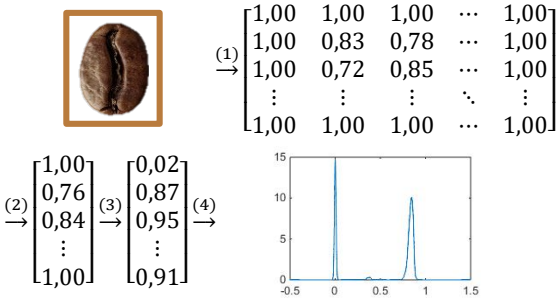
2.4. Vấn đề trích xuất hình ảnh

Trong lĩnh vực thị giác máy tính, có ba yếu tố chính để phân biệt và nhận dạng một hình ảnh. Chúng là màu sắc, kết cấu và hình dạng. Trong số đó, trích xuất đặc trưng màu sắc của hình ảnh được xem là phổ biến, thu hút nhiều sự quan tâm từ cộng đồng nghiên cứu (Nguyen-Trang et al., 2023b).

Về màu sắc, chúng ta có nhiều thang đo khác nhau, trong đó RGB (R: red; G: Green; B: Blue) được xem là phổ biến nhất hiện nay. Mỗi màu có giá trị trong khoảng $[0; 255]$ và một màu thực tế mà chúng ta quan sát được là sự kết hợp của ba màu trên. Như vậy thực tế chúng ta có $256 \times 256 \times 256 = 16777216$ màu khác nhau. Ảnh xám (Gr) cũng là một sự kết hợp của ba ảnh màu RGB.

Điểm ảnh là đơn vị cơ bản của một hình ảnh số. Mỗi điểm ảnh đại diện cho một điểm cụ thể trên bề mặt hình ảnh và chứa thông tin về màu sắc và độ sáng tại điểm đó. Trong hình ảnh số, điểm ảnh thường được biểu diễn bằng các pixel. Mỗi pixel có thể chứa thông tin về màu sắc theo các kênh màu (ảnh màu) hoặc độ sáng (ảnh trắng đen). Một ảnh có pixel càng lớn thì chất lượng ảnh càng tốt.

Trong nghiên cứu này, đặc trưng pixel được sử dụng để nhận dạng cho ảnh và PDF ước lượng từ pixel trích xuất được sử dụng làm phần tử đại diện. Cụ thể, thuật toán trích xuất đặc trưng của ảnh được đề nghị như sau:



Hình 1. Minh họa các bước trích xuất một hình ảnh thành PDF

Thuật toán 1:

Bước 1. Chuyển mỗi ảnh về ảnh R, G, B và Gr.

Bước 2. Trích xuất pixel cho mỗi ảnh từ Bước 1 để có ma trận điểm ảnh cho mỗi trường hợp.

Bước 3. Chuyển ma trận điểm ảnh của mỗi trường hợp thành vector cột.

Bước 4. Ước lượng PDF cho mỗi trường hợp bằng phương pháp hàm hạt nhân.

Như vậy sau thuật toán này, mỗi ảnh có thể được nhận diện bởi 4 PDF. Hình 1 minh hoạ cho các bước của Thuật toán 1 đề nghị.

2.5. Tiêu chuẩn đánh giá chất lượng của một thuật toán phân tích chùm

Khi có nhiều phương pháp phân tích chùm trên cùng một tập dữ liệu, chúng ta cần phải có sự lựa chọn. Trong nghiên cứu này, chỉ số hạng điều chỉnh (ARI:Adjusted Rand Index) được sử dụng. Đây là tiêu chuẩn được xem phổ biến nhất hiện nay được sử dụng trong đánh giá chất lượng của thuật toán xây dựng chùm (Rand, 1971).

Gọi P là kết quả phân vùng thực tế dựa trên các nhãn đã có sẵn, và Q là kết quả phân vùng của thuật toán, chỉ số ARI được tính bởi công thức:

$$ARI = \frac{a - (a+c)(a+b)/(a+b+c+d)}{((a+c)+(a+b))/2 - (a+c)(a+b)/(a+b+c+d)}$$

Trong công thức trên, a là số các cặp phần tử được xếp vào cùng một chùm ở cả P và Q, b là số các cặp phần tử thuộc cùng một chùm trong P nhưng thuộc vào hai chùm khác nhau trong Q, c là số các cặp phần tử thuộc hai chùm khác nhau trong P nhưng thuộc vào cùng một chùm trong Q, và d là số các cặp phần tử thuộc vào hai chùm khác nhau trong cả P lẫn Q.

Chỉ số ARI có giá trị thuộc [0,1]. Nếu giá trị của nó càng lớn thì thuật toán phân tích chùm càng tốt. Nếu giá trị $ARI = 1$ thì kết quả xây chùm là tối ưu.

3. THUẬT TOÁN ĐỀ NGHỊ

Cho N ảnh $\{I_1, I_2, \dots, I_N\}$, khi đó thuật toán phân tích chùm mờ cho các ảnh gồm các bước sau:

Thuật toán 2:

Bước 1. Áp dụng Thuật toán 1, tìm 4 PDF $\{g_{1i}, g_{2i}, g_{3i}, g_{4i}\}$ đại diện cho đặc trưng được trích xuất của ảnh $I_i, i = 1, 2, \dots, N$.

Bước 2. Đặt $f_i = g_{1i} + g_{2i} + g_{3i} + g_{4i}, i = 1, 2, \dots, N$. Khởi tạo các phần tử đại diện của chùm tại vòng lặp $t = 0$.

$$\{f_1, f_2, \dots, f_N\} \equiv \left\{f_1^{(0)}, f_2^{(0)}, \dots, f_N^{(0)}\right\} = f^{(0)}.$$

Bước 3. Cập nhật các PDF đại diện theo công thức sau:

$$f_i^{(t+1)} = \frac{\sum_{j=1}^{N} K_\lambda\left(f_i^{(t)}, f_j^{(t)}\right) \cdot f_j^{(t)}}{\sum_{j=1}^{N} K_\lambda\left(f_i^{(t)}, f_j^{(t)}\right)}, 1 \le i \le N, \quad (1)$$

trong đó

$$K_\lambda\left(f_i^{(t)}, f_j^{(t)}\right) = \begin{cases} \exp\left(-\frac{\left\|f_i^{(t)}, f_j^{(t)}\right\|_1}{\lambda}\right) & \text{nếu } \left\|f_i^{(t)}, f_j^{(t)}\right\|_1 \le w_s, \\ 0 & \text{nếu } \left\|f_i^{(t)}, f_j^{(t)}\right\|_1 > w_s, \end{cases} \quad (2)$$

với $\lambda = \frac{w_s}{5}, w_s = \frac{2}{N(N-1)} \sum_{i<j} \left\|f_i^{(t)}, f_j^{(t)}\right\|_1$.

Bước 4. Lặp lại Bước 3 cho đến khi

$$\max_i \left\|f_i^{(t+1)}, f_i^{(t)}\right\|_1 < \varepsilon.$$

Sau mỗi vòng lặp, các PDF $f_i^{(t+1)}$ sẽ tự cập nhật gần về các PDF đại diện của nó theo công thức (1). Khi Bước 4 dừng, nếu ta có bao nhiêu PDF đại diện thì tập ảnh đã cho sẽ được chia thành bấy nhiêu chùm.

Bước 5. Giả sử sau Bước 4 ta có được k chùm, khi ma trận phân vùng ban đầu $U^{(0)} = [\mu_{ij}^{(0)}]_{k\times N}$ được thiết lập, trong đó các phần tử của ma trận này được tính theo công thức:

$$\mu_{ij}^{(0)} = \begin{cases} 1 \text{ nếu } f_j \in C_i \\ 0 \text{ nếu } f_j \notin C_i \end{cases}; 1 \le i \le k, 1 \le j \le N.$$

Bước 6. Tìm phần tử đại diện f_{v_i} của chùm và tính độ rộng chùm giữa mỗi hàm f_{v_i} và các hàm mật độ f_j, trong đó

$$f_{v_i} = \frac{\sum_{j=1}^N \left(\mu_{ij}^{(0)}\right)^2 f_j}{\sum_{j=1}^N \left(\mu_{ij}^{(0)}\right)^2}, 1 \le i \le k.$$

Bước 7. Cập nhật ma trận phân vùng mờ $U^{(1)}$ theo công thức:

$$u_{ij}^{(1)} = \begin{cases} \frac{1}{\sum_{m=1}^k \left(\left\|f_j, f_{v_i}\right\|_1 / \left\|f_j, f_{v_m}\right\|_1\right)^2} & \text{nếu } \left\|f_j, f_{v_i}\right\|_1 > 0, \\ 0 & \text{ngược lại,} \end{cases}$$

trong đó $1 \le i \le k, 1 \le j \le N$.

Bước 8. Tính giá trị $\left\|U^{(1)} - U^{(0)}\right\| = \max_{i,j}\left\{\left|\mu_{ij}^{(1)} - \mu_{ij}^{(0)}\right|\right\}$. Lặp lại Bước 7 t lần cho đến khi điều kiện dừng sau được thỏa mãn:

$$\left\|U^{(t+1)} - U^{(t)}\right\| < \varepsilon.$$

Khi thuật toán kết thúc ta nhận được ma trận $U^{(t)} = [\mu_{ij}^{(t)}]_{k\times N}$ chính là xác suất để xếp ảnh thứ j vào chùm thứ $i, j = 1,2,\dots,N;\ i = 1,2,\dots,k$.

Trong Bước 4 và Bước 8, ε đo sự khác biệt giá trị của hai vòng lặp một cách có ý nghĩa. Nghiên cứu này chọn $\varepsilon = 0{,}0001$ trong ví dụ số và các áp dụng.

4. VÍ DỤ MINH HỌA

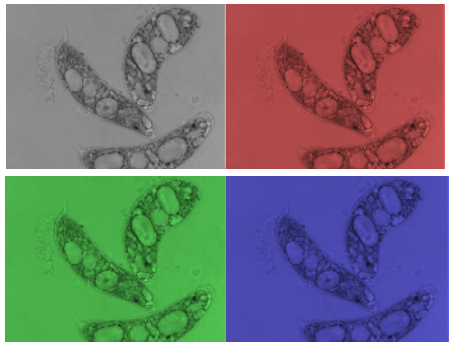
Phần này sử dụng 3 ảnh trùng roi xanh (C_1), 2 ảnh trùng giày (C_2) và 2 ảnh kí sinh trùng sốt rét (C_3) để minh hoạ các bước của Thuật toán 2. Những loại trùng này đều có vai trò quan trọng trong hệ sinh thái và sức khỏe con người, từ việc là nguồn thức ăn cho các sinh vật khác đến gây bệnh và tác động đến sự phát triển của các quần thể sinh vật. Ảnh mẫu của ba nhóm sinh vật trên được thể hiện ở Hình 2.

C_1 C_2 C_3

Hình 2. Hình ảnh minh họa ba nhóm sinh vật kí sinh trùng sốt rét, trùng giày và trùng roi xanh

Các bước của thuật toán được thực hiện như sau:

Bước 1*:* Đưa mỗi ảnh về bốn màu cơ bản (xem Hình 3).



Hình 3. Đưa ảnh C1 về bốn màu cơ bản

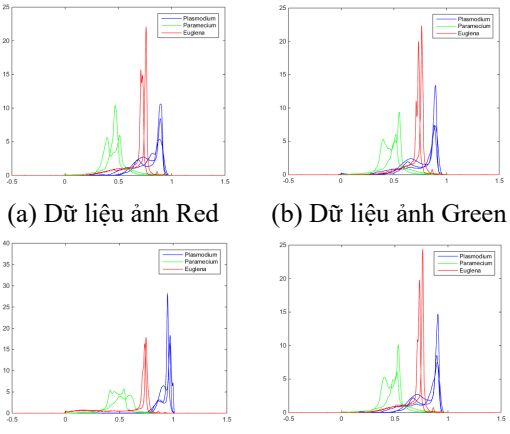
Trích xuất đặc trưng từng màu và ước lượng các PDF $g_{1i}, g_{2i}, g_{3i}, g_{4i}, 1 \le i \le 7$ tương ứng với bốn màu cho mỗi hình ảnh. Trong đó $\{g_{1i}, g_{2i}, g_{3i}, g_{4i}, i = 1,2,3\}$ là các PDF đại diện cho C_1, $\{g_{1i}, g_{2i}, g_{3i}, g_{4i}, i = 4,5\}$ là các PDF đại diện cho C_2, $\{g_{1i}, g_{2i}, g_{3i}, g_{4i}, i = 6,7\}$ là các PDF đại diện cho C_3. Hình 4 thể hiện 7 PDF được trích xuất từ mỗi màu.

Bước 2: Khởi tạo các phần tử đại diện của chùm tại vòng lặp $t = 0$ bằng cách lấy tổng bốn PDF tương ứng với bốn màu.

$$f_j^{(0)} \equiv g_{1j} + g_{2j} + g_{3j} + g_{4j}, 1 \le j \le N.$$

$$\{f_1, f_2, \dots, f_N\} \equiv \left\{f_1^{(0)}, f_2^{(0)}, \dots, f_N^{(0)}\right\} = f^{(0)}.$$

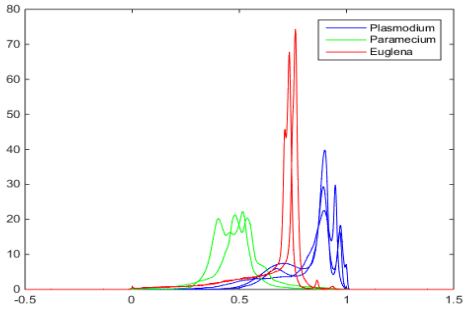
Đồ thị các PDF cho bởi Hình 5.



(a) Dữ liệu ảnh Red (b) Dữ liệu ảnh Green

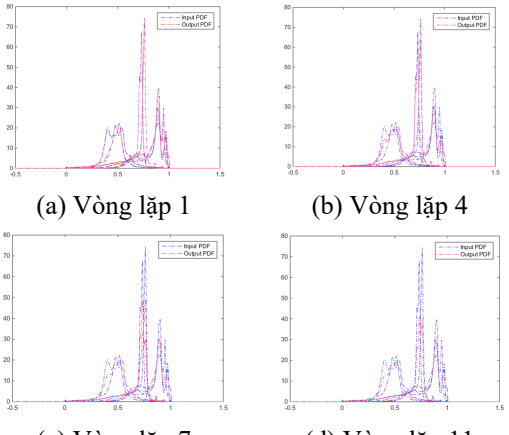
(c) Dữ liệu ảnh Blue (d) Dữ liệu ảnh Gray

Hình 4. PDF tương ứng với bốn màu



Hình 5. Các PDF tại vòng lặp $t = 0$

Bước 3: Cập nhật các PDF đại diện, kết quả thực hiện của bước này được mô tả trong Hình 6.



(a) Vòng lặp 1 (b) Vòng lặp 4

(c) Vòng lặp 7 (d) Vòng lặp 11

Hình 6. Sự hội tụ của 7 PDF vào 3 PDF đại diện

Bước 4: Sau 11 vòng lặp, *Bước 3* dừng. Kết quả mô tả các PDF trong 11 vòng lặp của thuật toán được biểu diễn trong Hình 6. Trong vòng lặp cuối cùng, 7 PDF (các đường màu xanh) hội tụ vào 3 PDF đại diện (các đường màu đỏ). Do đó, 7 PDF được chia thành 3 chùm:

$$C_1 = \{f_1, f_2, f_3\}, C_2 = \{f_4, f_5\}, C_3 = \{f_6, f_7\}.$$

Kết quả phân tích chùm này đúng như dữ liệu gốc ban đầu.

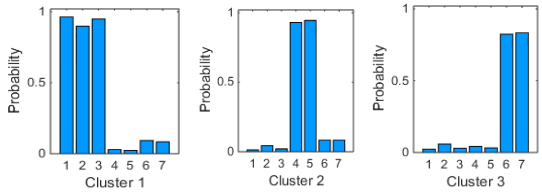
Bước 5: Từ kết quả của *Bước 4*, thuật toán thiết lập ma trận phân vùng ban đầu như sau:

$$U^{(0)} = \left[\mu_{ij}^{(0)}\right] = \begin{bmatrix} 1 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Bước 6, 7, 8: Tìm phần tử đại diện của chùm và cập nhật ma trận phân vùng mờ đến khi thỏa mãn điều kiện dừng. Khi đó ta sẽ có xác suất thuộc mỗi chùm của 7 PDF được trình bày chi tiết trong Bảng 1 và minh hoạ bởi Hình 7.

Bảng 1. Mối quan hệ mờ giữa 7 PDF và 3 chùm

f	C_1	C_2	C_3
f_1	0,9630	0,0133	0,0237
f_2	0,8982	0,0432	0,0586
f_3	0,9494	0,0206	0,0300
f_4	0,0287	0,9288	0,0425
f_5	0,0239	0,9440	0,0321
f_6	0,0930	0,0832	0,8238
f_7	0,0838	0,0832	0,8331



Hình 7. Xác suất thuộc vào 3 chùm của 7 PDF

Chỉ số đánh giá chất lượng của kết quả phân tích chùm đối với 4 màu Gr, R, G, B là $ARI = 1$.

Thực hiện phân tích chùm tương tự cho 2 màu và 3 màu, ta có Bảng 2.

Bảng 2 cho thấy việc sử dụng các đặc trưng từ các thang màu (Gr, R), (Gr, G), (R, G) và (Gr, R, G) không cho kết quả phân tích chùm tốt, tất cả các trường hợp còn lại đều cho kết quả tối ưu với ARI đều bằng 1.

Bảng 2. Bảng tổng hợp chỉ số ARI trong phân tích chùm cho 7 ảnh

Màu	ARI
Gr	1,0
R	1,0
G	1,0
B	1,0
Gr, R	0,4425
Gr, G	0,4425
Gr, B	1,0
R, G	0,4425
R, B	1,0
G, B	1,0
Gr, R, G	0,4425
Gr, R, B	1,0
Gr, G, B	1,0
R, G, B	1,0
Gr, R, G, B	1,0

5. ỨNG DỤNG

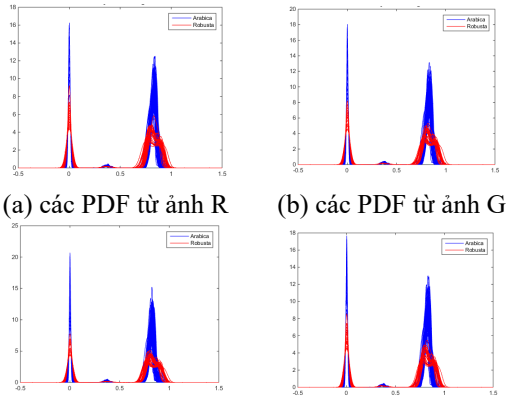
5.1. Áp dụng cho dữ liệu ảnh hạt cà phê

Để minh họa cho thuật toán đề nghị, ta thực hiện phân tích chùm cho hai loại hạt cà phê Arabica và Robusta. Trên thế giới cũng như tại Việt Nam, Arabica và Robusta là hai trong các loại hạt cà phê được trồng phổ biến nhất. Hai loại hạt cà phê này cũng khó phân biệt bằng mắt thường đối với những người không phải chuyên gia. Thuật toán được thực hiện trên tập dữ liệu *Robusta or Arabica* gồm 112 ảnh với 82 ảnh hạt cà phê Arabica và 30 hạt ảnh cà phê Robusta được lấy miễn phí từ website universe.roboflow.com. Mẫu ảnh của hai nhóm được cho bởi Hình 8.

(a) Cà phê Arabica (b) Cà phê Robusta

Hình 8. Hình ảnh minh họa hai loại hạt cà phê

Trích xuất mỗi ảnh thành 4 PDF dựa trên 4 không gian màu cơ bản và phương pháp ước lượng hàm hạt nhân, ta có Hình 9.

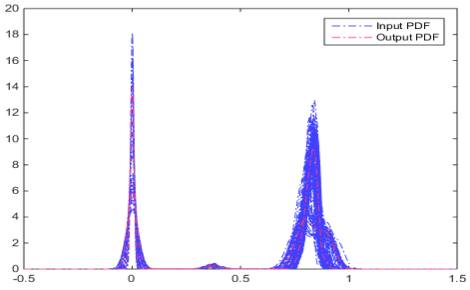


(a) các PDF từ ảnh R (b) các PDF từ ảnh G

(c) các PDF từ ảnh B (d) các PDF từ ảnh Gr

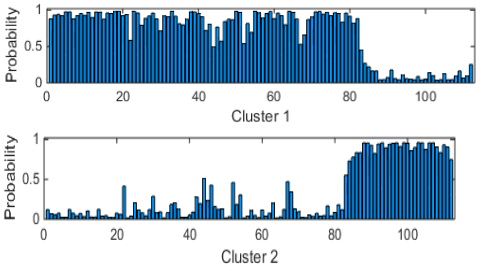
Hình 9. PDF cho toàn bộ dữ liệu ảnh RGBGr

Sử dụng hàm đại diện từ 4 PDF cho mỗi ảnh và tìm hàm đại diện cho hai nhóm, ta nhận được Hình 10.



Hình 10. Minh họa hai PDF đại diện

Sau 6 vòng lặp, thuật toán dừng, khi đó ta có hai chùm như tập dữ liệu ảnh gốc ban đầu và xác suất thuộc vào hai chùm của các ảnh được cho bởi Hình 11.



Hình 11. Xác suất thuộc vào hai chùm của 112 ảnh hạt cà phê

Hình 11 một lần nữa cho thấy khả năng phân tích chùm hợp lý của các ảnh khi xác suất thuộc vào chùm đúng của các ảnh đều cao.

Thực hiện phân tích chùm cho trường hợp trích xuất đặc trưng từ 1 màu, 2 màu, 3 màu và 4 màu và

tính chỉ số *ARI* cho các trường hợp, ta nhận được Bảng 3.

Bảng 3. Chỉ số ARI trong các trường hợp cho tập dữ liệu ảnh cà phê

Màu	ARI
Gr	0,9625
R	0,9625
G	1,0
B	1,0
Gr, R	0,9625
Gr, G	1,0
Gr, B	1,0
R, G	0,9625
R, B	1,0
G, B	1,0
Gr, R, G	0,9625
Gr, R, B	1,0
Gr, G, B	1,0
R, G, B	1,0
Gr, R, G, B	1,0

Từ Bảng 3, ta thấy kết quả phân loại dữ liệu ảnh cà phê của các phương pháp đều rất tốt vì giá trị *ARI* của nó bằng 0,9625 hoặc bằng 1, trong đó trường hợp sử dụng bốn màu Gr, R, G, B vẫn cho kết quả ổn định tối ưu với giá trị $ARI = 1$.

5.2. Áp dụng cho tập ảnh bệnh đậu mùa khỉ, thủy đậu và sởi

Bộ dữ liệu được trích lọc gồm 60 hình ảnh thuộc các nhóm bệnh về da gồm 30 ảnh đậu mùa khỉ (C_1), 20 ảnh thủy đậu (C_2) và 10 ảnh bệnh Sởi (C_3) từ các tập dữ liệu *Monkeypox*, *Chickenpox*, *Measles* được lấy miễn phí từ website universe.roboflow.com. Tất cả các hình ảnh đều có kích thước 128×128 pixels và ở định dạng kiểu file jpg. Ảnh mẫu của ba nhóm bệnh được thể hiện ở Hình 12.

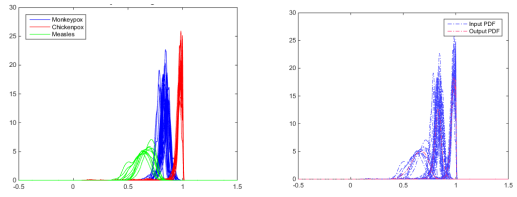
C_1 C_2 C_3

Hình 12. Hình ảnh minh họa ba loại bệnh đậu mùa khỉ, thủy đậu và bệnh sởi

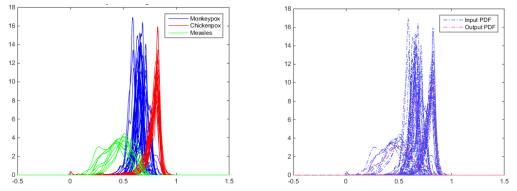
Sau 8 vòng lặp, thuật toán dừng. Khi đó ta có kết quả phân tích chùm giống như dữ liệu gốc ban đầu và xác suất thuộc vào chùm của mỗi ảnh được cho bởi Hình 14.

Ba loại bệnh Đậu mùa khỉ, Thủy đậu, Sởi đều là các bệnh truyền nhiễm phổ biến gây ra bởi các loại virus khác nhau. Các nhóm bệnh này có thể gây ra các triệu chứng đa dạng, bao gồm sự xuất hiện của mụn đỏ trên. Chúng có thể xuất hiện ở nhiều phần khác nhau của cơ thể và thường đi kèm với các triệu chứng khác như sốt, đau đầu và mệt mỏi. Về mặt hình ảnh biểu hiện bên ngoài, người dân cũng có thể nhầm lẫn 3 bệnh này nên cũng cần thực hiện bài toán phân tích chùm.

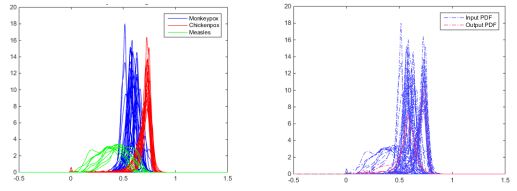
Tiến hành trích xuất tập ảnh thành các PDF đối với ảnh G, R, B và Gr, ta có Hình 13.



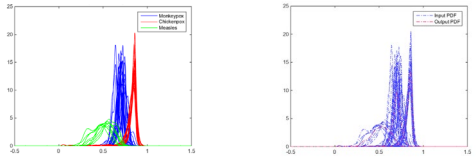
(a) PDF được trích xuất và đại diện của 3 nhóm từ ảnh R



(b) PDF được trích xuất và đại diện của 3 nhóm từ ảnh G

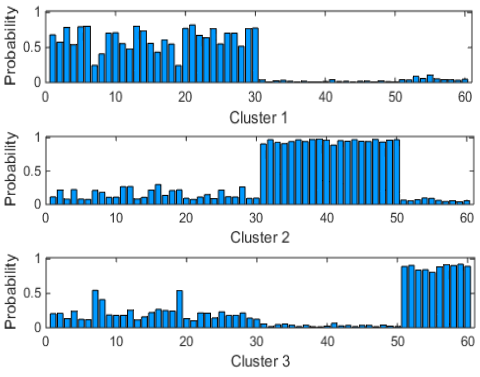


(c) PDF được trích xuất và đại diện của 3 nhóm từ ảnh B



(d) PDF được trích xuất và đại diện của 3 nhóm từ ảnh Gr

Hình 13. PDF đại diện cho các ảnh và đại diện của mỗi chùm được trích xuất từ các màu RGBGr



Hình 14. Xác suất thuộc vào 3 chùm của 60 ảnh

Hình 14 cũng cho thấy kết quả phân tích chùm khá tốt khi xác suất của ảnh thuộc vào chùm đúng đều cao.

Tính chỉ số *ARI* của các kết quả phân tích chùm trong các trường hợp khác nhau của không gian màu, ta nhận được Bảng 4.

Bảng 4 cho thấy khi sử dụng đặc trưng trích xuất từ một màu Gr, R, hoặc 2 màu (Gr, G), (Gr, B), (R, G), (G, B), hoặc 3 màu (Gr, R, G), (Gr,g, B), hoặc cả 4 màu sẽ cho kết quả phân tích chùm tối ưu vì chỉ số ARI có giá trị bằng 1.

Từ ví dụ minh hoạ và hai ứng dụng trên, chúng ta cũng thấy rằng việc sử dụng 1 màu, hoặc 2 màu hoặc 3 màu chúng ta cũng có thể nhận được kết quả phân tích chùm tốt, nhưng không ổn định. Việc sử dụng cùng lúc đặc trưng trích xuất từ 4 màu có thể mang lại kết quả phân tích chùm ổn định hơn.

Bảng 4. Bảng tổng hợp chỉ số ARI trong phân tích của 60 ảnh

Màu	ARI
Gr	1,0
R	1,0
G	0,6748
B	0,6748
Gr, R	0,3792
Gr, G	1,0
Gr, B	1,0
R, G	1,0
R, B	0,4479
G, B	1,0
Gr, R, G	1,0
Gr, R, B	0,3792
Gr, G, B	1,0
R, G, B	0,4479
Gr, R, G, B	1,0

6. KẾT LUẬN

Nghiên cứu này đã đề xuất phương pháp trích xuất đặc trưng pixel màu sắc của ảnh thành các PDF để đại diện. Các PDF này sau đó được sử dụng để xây dựng thuật toán phân tích chùm cho dữ liệu ảnh. Ví dụ số và ứng dụng cho thấy thuật toán phân tích chùm đề nghị ổn định và hiệu quả, có tiềm năng trong các áp dụng thực tế.

Trong thời gian tới, nghiên cứu sẽ kiểm tra thêm sự hiệu quả của thuật toán trên những tập dữ liệu lớn của thực tế trong những lĩnh vực khác nhau để cải tiến các tham số của mô hình đề nghị. Nghiên cứu cũng sẽ mở rộng việc trích xuất các đặc trưng khác của ảnh để nhận dạng, từ đó cải thiện thuật toán phân tích chùm.

TÀI LIỆU THAM KHẢO

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-plus illustrations* (Vol. 18). Oxford, England: OUP.

Chen, J., Chang, Y., & Hung, W. (2018). A robust automatic clustering algorithm for probability density functions with application to categorizing color images. *Communications in Statistics-Simulation and Computation*, *47*(7), 2152–2168.

Chen, J. H., & Hung, W. L. (2021). A jackknife entropy-based clustering algorithm for probability density functions. *Journal of Statistical Computation and Simulation*, *91*(5), 861–875.

Hung, W. L., & Yang, J. H. (2015). Automatic clustering algorithm for fuzzy data. *Journal of Applied Statistics*, *42*(7), 1503–1518.

Izakian, Z., Mesgari, M. S., & Abraham, A. (2016). Automated clustering of trajectory data using a particle swarm optimization. *Computers, Environment and Urban Systems*, *55*, 55–65.

Montanari, A., & Calò, D. G. (2013). Model-based clustering of probability density functions. *Advances in Data Analysis and Classification*, *7*(3), 301–319.

Nguyen-Trang, T., Nguyen-Thoi, T., Nguyen-Thi, K. N., & Vo-Van, T. (2023). Balance-driven automatic clustering for probability density functions using metaheuristic optimization. *International Journal of Machine Learning and Cybernetics*, *14*(4), 1063–1078.

Nguyen-Trang, T., Nguyen-Thoi, T., & Vo-Van, T. (2023). Globally automatic fuzzy clustering for probability density functions and its application

for image data. *Applied Intelligence*, *53*, 18381–18397.

Qi, X., Li, C. G., Zhao, G., Hong, X., & Pietikäinen, M. (2016). Dynamic texture and scene classification by transferring deep image features. *Neurocomputing*, *171*, 1230–1241.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rao, C., & Liu, Y. (2020). Three-dimensional convolutional neural network (3D-CNN) for heterogeneous material homogenization. *Computational Materials Science*, *184*, 109850.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*(Dec), 583–617.

Vo-Van, T., & Nguyen-Trang, T. (2018). Similar coefficient for cluster of probability density functions. *Communica- tions in Statistics-Theory and Methods*, *47*(8), 1792–1811.

Vo-Van, T., & Nguyen-Trang, T. (2018). Similar coefficient of cluster for discrete elements. *Sankhya B*, *80*(1), 19–36.

Wang, X., Zhao, Y., & Pourpanah, F. (2020). Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics*, *11*, 747–750.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107.

Xu, L., Hu, Q., Hung, E., Chen, B., Tan, X., & Liao, C. (2015). Large margin clustering on uncertain data by considering probability distribution similarity. *Neurocomputing*, *158*, 81–89.

Zhu, Y., Deng, Q., Huang, D., Jing, B., & Zhang, B. (2021). Clustering based on Kolmogorov–Smirnov statistic with application to bank card transaction data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *70*(3), 558–578.