Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2026.114

PHÂN TÍCH CHÙM MỜ CHO DỮ LIỆU ẢNH DỰA VÀO HÀM MẬT ĐỘ XÁC SUẤT ĐẠI DIỆN ĐƯỢC TRÍCH XUẤT TỪ MÔ HÌNH RESNET50

Nguyễn Kim Ngân, Võ Thành Vinh, Trương Hiếu Thành, Nguyễn Duy Tân, Hồ Phúc Thiện, Châu Ngọc Thơ, Bao Nhã Linh và Võ Văn Tài*
Trường Khoa học Tự nhiên, Đại học Cần Thơ, Việt Nam
**Tác giả liên hệ (Corresponding author): vvtai@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 05/09/2025
Sửa bài (Revised): 27/09/2025
Duyệt đăng (Accepted): 04/05/2026

Title: *Fuzzy cluster analysis for image data based on representative probability density function extracted from Resnet50 model*

Author(s): *Nguyen Kim Ngan, Vo Thanh Vinh, Truong Hieu Thanh, Nguyen Duy Tan, Ho Phuc Thien, Chau Ngoc Tho, Bao Nha Linh and Vo Van Tai**

Affiliation(s): *College of Natural Science, Can Tho University, Viet Nam*

TÓM TẮT

Thuật toán phân tích chùm mờ cho dữ liệu ảnh được xây dựng với những giai đoạn chính sau: đầu tiên là việc trích xuất đặc trưng của ảnh bằng phương pháp học sâu Resnet50 với hai khối tích chập đầu tiên và quá trình gộp đặc trưng toàn cục. Các đặc trưng này sau đó được chuẩn hóa và ước lượng thành hàm mật độ xác suất đại diện. Tiếp theo là quá trình xác định số chùm thích hợp dựa vào hàm tự cập nhật chùm phù hợp. Cuối cùng là phương pháp xác định xác suất thuộc vào chùm của mỗi ảnh. Thuật toán đề nghị đã được trình bày chi tiết các bước thực hiện về lý thuyết và được minh họa trên tập ảnh cụ thể. Ứng dụng trên một tập ảnh cụ thể cũng cho thấy thuật toán đề nghị cho kết quả hợp lý và thuận lợi hơn một số thuật toán phổ biến khác qua các tham số đánh giá.

Từ khóa: *Dữ liệu ảnh, đặc trưng trích xuất, hàm mật độ xác suất, phân tích chùm*

ABSTRACT

This study develops a fuzzy clustering algorithm for image data with the following main stages. First, image features are extracted using the deep learning method ResNet50, specifically the first two convolutional blocks combined with global feature pooling. The extracted features are then normalized and transformed into representative probability density functions. Next, the appropriate number of clusters is determined through a self-updating suitable cluster function. Finally, the method determines the probability of each image belonging to a cluster. The proposed algorithm has been presented in detail with theoretical implementation steps and illustrated on a specific image dataset. Applications on several specific image datasets also demonstrate that the proposed algorithm yields reasonable and more advantageous results compared to some other popular algorithms through evaluation parameters.

Keywords: *Cluster analysis, Extracted features, Image data, Probability density function*

1. GIỚI THIỆU

Phân tích chùm là một kỹ thuật trong khai phá dữ liệu và học máy dùng để nhóm các đối tượng có đặc điểm tương đồng vào cùng một chùm, sao cho các đối tượng trong cùng một chùm thì "gần nhau" hơn so với các đối tượng thuộc chùm khác (Vo et al., 2023). Phân tích chùm được áp dụng trong nhiều lĩnh vực khác nhau, được xem là nền tảng không thể thiếu liên quan đến khoa học dữ liệu nên đã và đang nhận được rất nhiều sự quan tâm của các nhà khoa học trong lĩnh vực thống kê và công nghệ thông tin (Chen & Shiu, 2007; Hung & Yang, 2015; Pham & Vo, 2024).

Phân tích chùm có thể được thực hiện cho nhiều loại dữ liệu khác nhau, bao gồm dữ liệu điểm, dữ liệu khoảng và dữ liệu hàm mật độ xác suất (PDF). Đối với dữ liệu điểm, loại dữ liệu phổ biến nhất, đã có rất nhiều thuật toán, trong đó có những thuật toán nổi tiếng như k-means (Zhou & Zhu, 2019) và FCM (Bezdek et al., 1984). Phát triển từ các thuật toán nổi tiếng này, hàng loạt các thuật toán sau đó đã được đề nghị (Hung et al., 2015; Vo & Nguyen, 2018; Hung et al., 2021). Khi dữ liệu phức tạp hơn, chẳng hạn như nhiệt độ dao động trong ngày hay khoảng thời gian một người dùng một ứng dụng trong tháng. Lúc này, việc phân chùm cần tính đến toàn bộ phạm vi giá trị. Khi đó, phân tích chùm cho dữ liệu khoảng đã được đề xuất (Hung et al., 2016; Chang & Jeng, 2025; Le et al., 2023). Ở cấp độ cao hơn, khi một đối tượng cần biểu thị cả một phân phối, thì phân tích chùm cho PDF cần được thiết lập (Chen & Hung, 2015; Nguyen et al., 2023a).

Ngày nay, dữ liệu ảnh trở thành một nguồn thông tin vô cùng quan trọng, xuất hiện ở hầu hết các lĩnh vực của khoa học và đời sống. Với sự phát triển mạnh mẽ của công nghệ số, ảnh số được tạo ra và lưu trữ với khối lượng khổng lồ từ camera giám sát, vệ tinh, thiết bị y tế, điện thoại thông minh đến các hệ thống mạng xã hội (Vo, 2025). Phân tích chùm cho ảnh là một nhánh quan trọng của khai phá dữ liệu và thị giác máy tính. Nó tập trung vào việc nhóm các ảnh có nội dung hoặc đặc trưng tương đồng vào cùng một nhóm mà không cần nhãn dữ liệu ban đầu. Ứng dụng của phân tích chùm trong ảnh rất đa dạng cho nhiều lĩnh vực như môi trường, kinh tế và y học.

Trích xuất đặc trưng ảnh là bước quan trọng trong xử lý và phân tích ảnh, nhằm chuyển đổi dữ liệu ảnh thô có kích thước lớn thành những thông tin gọn nhẹ nhưng giàu ý nghĩa, phản ánh bản chất của đối tượng trong ảnh. Các đặc trưng này có thể là màu sắc, kết cấu, hình dạng hoặc các đặc trưng học sâu từ mạng nơ-ron (Conze et al., 2025). Sau khi trích xuất, đối tượng ảnh được biểu diễn dưới nhiều dạng khác nhau: ma trận, vector hoặc PDF. Nhờ vậy, dữ liệu ảnh trở nên có cấu trúc hơn, dễ dàng áp dụng các thuật toán học máy và phân tích nhằm phục vụ các bài toán nhận dạng, phân loại hay phân chùm. Phân tích chùm cho PDF, lần đầu tiên được đề xuất bởi Vo & Pham (2010). Hướng nghiên cứu này sau đó nhận được sự quan tâm của nhiều nhà khoa học (Chen & Hung, 2015; Nguyen et al., 2023a). Yếu điểm chung của các nghiên cứu này là cần kinh nghiệm của người sử dụng thông qua việc phải chọn các tham số trong các thuật toán. Việc áp dụng phân tích chùm cho ảnh từ các nghiên cứu này cũng còn hạn chế bởi giai đoạn trích xuất đặc trưng của ảnh. Các nghiên cứu hầu như chỉ sử dụng các phương pháp thống kê như ma trận đồng hiện mức xám (Vo et al., 2023), không gian màu RGB (Vo, 2025). Việc kết hợp trích xuất đặc trưng của ảnh từ các phương pháp học sâu và thống kê để biểu diễn thành PDF như trong nghiên cứu này hầu như chưa được quan tâm nhiều.

Phân tích chùm được phát triển theo hai hướng phân tích chùm mờ và không mờ. Trong phân tích chùm không mờ, mỗi phần tử chỉ được xếp vào một chùm duy nhất, trong khi phân tích chùm mờ một phần tử có thể được xếp vào nhiều chùm với một xác suất nhất định. Do đó, việc phân tích chùm mờ được xem là sự mở rộng của phân tích chùm không mờ (Nguyen et al., 2023a; Pham & Vo, 2024). Các phương pháp phân tích chùm mờ đã và đang được nghiên cứu, phát triển nhằm giải quyết hạn chế của phân tích chùm không mờ khi dữ liệu có tính chất chồng lấn hoặc không ranh giới rõ ràng. Phương pháp phổ biến nhất là FCM, trong đó mỗi điểm dữ liệu được gán mức độ thành viên thuộc về từng chùm thay vì chỉ nằm trong một chùm duy nhất. Điều này giúp phản ánh tốt hơn sự mơ hồ vốn có của dữ liệu thực. Phát triển từ FCM dành cho dữ liệu điểm, một số thuật toán phân tích chùm mờ cho PDF đã được đề xuất (Pham & Vo, 2024). Trong các thuật toán này, việc tìm xác suất thuộc vào chùm của mỗi phần tử dựa vào mối quan hệ của phần tử đại diện chùm, trong đó phần tử đại diện chùm là một PDF duy nhất. Vì số lượng PDF trong mỗi chùm là khác nhau, nên việc mỗi chùm có duy nhất 1 PDF đại diện có thể tạo ra những mối quan hệ mờ không hợp lý. Nghiên cứu này cải tiến việc tìm phần tử đại diện chùm, căn cứ vào mức độ gần nhau của các phần tử trong mỗi chùm.

2. THUẬT TOÁN TRÍCH XUẤT ẢNH THÀNH HÀM MẬT ĐỘ XÁC SUẤT ĐẠI DIỆN

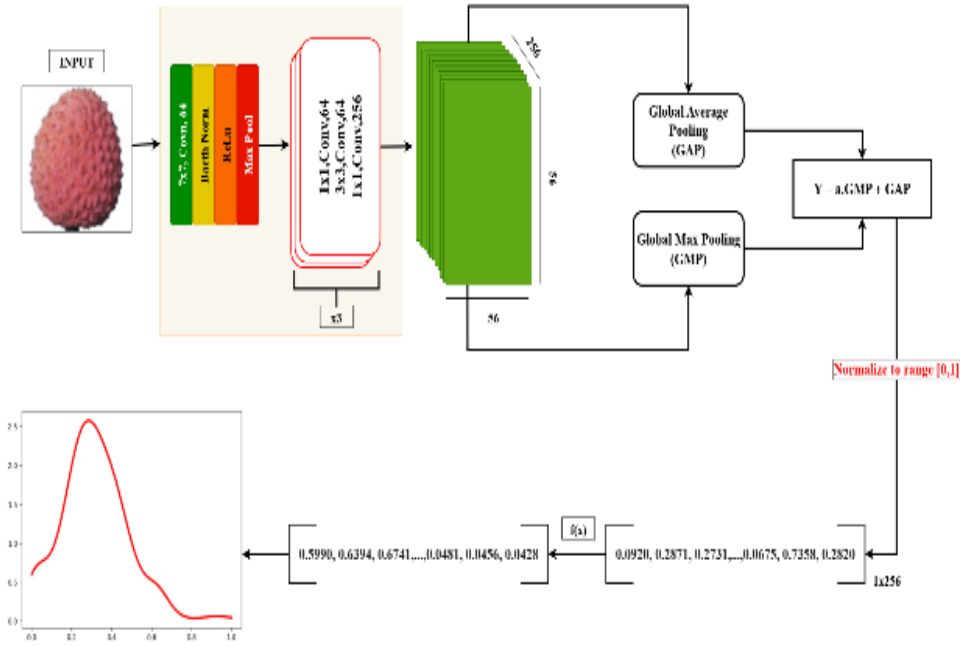
Trong nghiên cứu này, phương pháp học sâu Resnet50 (Bussa & Boppana, 2025) được sử dụng như một bộ trích xuất đặc trưng thay vì khai thác toàn bộ mạng cho nhiệm vụ phân loại. Cụ thể, chỉ hai khối tích chập đầu tiên, gồm Conv1 và Conv2_x được khai thác. Đây được xem là giai đoạn nền tảng trong quá trình học biểu diễn thị giác. Conv1 chủ yếu phát hiện các đặc trưng cơ bản của ảnh như biên, góc, cạnh và đường thẳng, trong khi Conv2_x tiếp tục kết hợp các tín hiệu này để hình thành những mẫu phức tạp hơn, bao gồm họa tiết, kết cấu cục bộ và các mô hình hình học giàu thông tin. Việc chỉ khai thác đặc trưng ở giai đoạn sớm mang lại hai lợi ích quan trọng sau:

(i) Bảo toàn mức độ chi tiết cao của ảnh, tránh sự trừu tượng hóa quá mạnh ở các tầng sâu.

(ii) Tạo ra biểu diễn tổng quát, ít phụ thuộc vào nhãn gốc mà Resnet50 được huấn luyện ban đầu.

Các đặc trưng từ Conv1 và Conv2_x do đó được xem như biểu diễn trung gian giàu thông tin, đóng vai trò cơ sở cho các bước phân tích và mô hình hóa tiếp theo.

Quá trình trích xuất đặc trưng của một ảnh thành PDF được minh họa bởi Hình 1.



Hình 1. Sơ đồ minh hoạ trích xuất mỗi ảnh thành 1 PDF đại diện

Sơ đồ trong Hình 1 được mô tả cụ thể bởi các bước trong Thuật toán 1 sau đây:

Bước 1. Tiền xử lý ảnh:

Giả sử $I \in \mathbb{R}^{W\times H\times 3}$ là ảnh màu cần trích xuất đặc trưng. Để phù hợp với đầu vào của mô hình Resnet50, ảnh I được resize về kích thước chuẩn

$$I' \in \mathbb{R}^{224\times 224\times 3}.$$

Sau đó, ảnh này được đưa vào mô hình để tiến hành trích xuất.

Bước 2. *Trích xuất đặc trưng từ Resnet-50:*

Ảnh I' được đưa qua khối Stem, được minh hoạ như sau:

$$I' \xrightarrow{ConV\ 7\times 7, stride=2, pad=3} \xrightarrow{BN} \xrightarrow{ReLU} \xrightarrow{MaxPool\ 3\times 3, stride=2}$$

Sau quá trình này ta được Tensor $T_1 \in \mathbb{R}^{56\times 56\times 64}$. Tensor T_1 sau đó đi qua khối conv2_x, gồm 3 bottleneck. Mỗi bottleneck thực hiện như sau:

$$T \xrightarrow{Conv\ 1\times 1} \xrightarrow{Conv\ 3\times 3} \xrightarrow{Conv\ 1\times 1} + skip\ connection \Rightarrow T',$$

đồng thời giữ nguyên kích thước không gian. Đầu ra cuối cùng là

$$T_2 \in \mathbb{R}^{56\times 56\times 256}$$

Bước 3. *Gộp đặc trưng toàn cục (Pooling):*

Với mỗi kênh $c = 1, 2,\ldots, 256$, tính toán Global Average Pooling (GAP) và Global Max Pooling (GMP) như sau:

$$GAP_c = \frac{1}{56 \times 56}\sum_{i=1}^{56}\sum_{j=1}^{56} T_2(i,j,c),$$

$$GMP_c = \max_{i,j} T_2(i,j,c).$$

Kết hợp hai giá trị trên theo công thức

$$Y_c = \alpha \cdot GMP_c + GAP_c, c = 1, \ldots, 256,$$

ta thu được vector đặc trưng:

$$Y \in \mathbb{R}^{256},$$

với $\alpha \in [0;1]$ biểu diễn mức độ đóng góp của vector GMP_c. Trong nghiên cứu này, $\alpha = 0{,}1$ được chọn.

Bước 4. *Chuẩn hóa vector đặc trưng:*

Áp dụng chuẩn hóa min-max cho *Y* theo công thức sau:

$$\tilde{Y}_c = \frac{Y_c - \min(Y)}{max(Y) - \min(Y)}, c = 1, 2, \ldots, 256,$$

nhằm đưa toàn bộ giá trị về khoảng $[0;1]$. Khi đó, $\tilde{Y} \in [0;1]^{256}$.

Bước 5. Biểu diễn đặc trưng dưới dạng PDF:

Xem $\tilde{Y}$ như một tập gồm $N = 256$ mẫu ngẫu nhiên trong miền $[0;1]$. Sử dụng phương pháp hàm hạt nhân để ước lượng:

$$\hat{f}(x) = \frac{1}{256 \cdot h}\sum_{c=1}^{256} K\left(\frac{x - \tilde{Y}_c}{h}\right),$$

với $K(\cdot)$ là hàm hạt nhân được chọn dạng chuẩn và h là tham số trơn được xác định theo Nguyentrang et al. (2023).

Như vậy, sau khi kết thúc thuật toán này, các đặc trưng của mỗi ảnh sẽ được đại diện bởi một PDF.

3. THUẬT TOÁN PHÂN TÍCH CHÙM ẢNH ĐỀ NGHỊ

Giả sử $\mathcal{P} = \{I_1, I_2, \ldots, I_n\}$ là tập gồm *n* ảnh cần phân chùm. Khi đó, thuật toán phân tích chùm mờ đề nghị (Thuật toán 2) gồm các bước sau:

Bước 1. *Trích xuất mỗi ảnh thành PDF đại diện*

Áp dụng Thuật toán 1, để trích xuất mỗi ảnh thành 1 PDF đại diện. Sau bước này, ta có $\mathcal{P} = \{p_1(x), p_2(x), \ldots, p_n(x)\}$ là tập gồm *n* PDF cần phân chùm và $\mathcal{C}^{(t)}$ là tập các PDF đại diện của các chùm tại vòng lặp *t.*

Bước 2. *Xác định số chùm và phần tử đại diện ban đầu:*

Xem mỗi PDF ban đầu là một chùm, nghĩa là ta có *n* chùm cũng như *n* PDF đại diện:

$$\mathcal{C}^{(0)} = \{c_1{}^{(0)}, c_2{}^{(0)}, \ldots, c_n{}^{(0)}\}$$
$$= \{p_1(x), p_2(x), \ldots, p(x)\}. \quad (1)$$

Bước 3. *Cập nhật trọng tâm cho phần tử đại diện chùm:*

Cập nhật trọng tâm của chùm $\mathcal{C}^{(t+1)}, t = 0,1, \ldots$, trong đó mỗi phần tử $c_i^{(t+1)}$ được tính bằng cách lấy trung bình có trọng số của các $c_j^{(t)}$ tại bước *t*. Cụ thể, công thức cập nhật được xác định như sau:

$$c_i^{(t+1)} = \frac{\sum_{j=1}^{n} w_{ij}^{(t)} c_j^{(t)}}{\sum_{j=1}^{n} w_{ij}^{(t)}}, i = 1,2, \ldots, n \quad (2)$$

trong đó

$$w_{ij}^{(t)} = \begin{cases} \frac{1}{1 + d_{ij}^{(t)}/\sigma} & \text{nếu } d_{ij}^{(t)} \le \mu\alpha_{ij}(t), \\ 0 & \text{nếu } d_{ij}^{(t)} > \mu\alpha_{ij}(t) \end{cases} \quad (3)$$

với

$$\mu = \frac{1}{2.\binom{n}{2}}\sum_{i<j} d_{ij}^{(t)},$$

$$\sigma = \sqrt{\frac{1}{\binom{n}{2}}\sum_{i<j}\left(d_{ij}^{(t)} - \mu\right)^2}. \quad (4)$$

Trong (4), chúng ta có

* $d_{ij}^{(t)}$ là khoảng cách L_2 giữa $c_i^{(t)}$ và $c_j^{(t)}$, được xác định bởi

$$d_{ij}^{(t)} = \left(\int_{R^n}\left[c_i^{(t)}(x) - c_j^{(t)}(x)\right]^2 dx\right)^2,$$

* μ là trung bình mẫu được điều chỉnh của từng đôi khoảng cách $d_{ij}^{(t)}$,

* σ là độ lệch chuẩn của từng đôi khoảng cách $d_{ij}^{(t)}$,

* $\alpha_{ij}(t)$ ngưỡng hội tụ, xác định bởi

$$\alpha_{ij}(t) = \frac{\alpha_{ij}(t-1)}{1 + \alpha_{ij}(t-1) w_{ij}^{(t)}}, \alpha_{ij}(0) = 1. \quad (5)$$

Bước 4. *Xác định số chùm và phần tử đại diện chùm*:

Lặp lại **Bước** 3 cho đến khi điều kiện sau được thỏa:

$$\left\|c_i^{(t+1)} - c_i^{(t)}\right\|_2 = \left(\int_{R^n}\left[c_i^{(t+1)}(x) - c_j^{(t)}(x)\right]^2 dx\right)^{\frac{1}{2}} < \epsilon, \qquad (6)$$

với ϵ là một số dương nhỏ tùy ý. Nghiên cứu này chọn $\epsilon = 0{,}0001$.

Khi (6) thỏa mãn, gọi

$$\mathcal{C}^* = \left\{c_k^{(t+1)} \in \mathcal{C}^{(t+1)} \middle| c_k^{(t+1)} \neq c_m^{(t+1)}\right\},$$

với $\forall m < k, m, k \in \{1,2,\dots,n\}$.

Mỗi phần tử $\mathcal{C}^*$ là một PDF đại diện cho các PDFs ban đầu trong tập $\mathcal{P}$ mà chúng đã hội tụ đến chính nó khi điều kiện (6) được thỏa. Khi đó, số chùm thích hợp của tập $\mathcal{P}$ chính là số phần tử của $\mathcal{C}^*$, được kí hiệu $|\mathcal{C}^*|$. Giả sử ta có c = $|\mathcal{C}^*|$.

Bước 5. *Thiết lập ma trận phân vùng:*

***Bước** 5.1*. Xây dựng ma trận phân vùng ban đầu:

Với mỗi PDF $p_j \in \mathcal{P}$, tính khoảng cách đến từng phần tử đại diện của mỗi chùm c_i, *i* = 1, 2, …, *c*:

$$d_{ij} = \|p_j - c_i^*\|_2. \qquad (7)$$

Dựa vào (7), xây dựng các ma trận phân vùng $U = \left[u_{ij}\right]_{c\times n}$, với các phần tử được tính theo công thức sau:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^2} \qquad (8)$$

***Bước** 5.2*. Tính độ lệch chuẩn trong mỗi chùm:

Với mỗi $p_j \in \mathcal{P}$, (*j* =1, 2,…, *n*), gán nhãn:

$$y_j = \arg\max_{i\in\{1,2,\dots,c\}}\{u_{ij}\},\ j = 1,2,\dots \qquad (9)$$

Nhãn y_j là chỉ số chùm mà mẫu *j* thuộc về mạnh nhất.

Khi đó, với mỗi chùm $i \in \{1,2,\dots,c\}$, tập các PDF thuộc cụm đó là:

$$C_i = \{p_j \in \mathcal{P} | y_j = i, j = 1,2,\dots,n\}.$$

Độ lệch chuẩn của mật độ theo chùm thứ *i* được tính như sau:

$$s_{i,q} = \sqrt{\frac{1}{n_i - 1}\sum_{h=1}^{n_i}\left(c_{ih,q} - \bar{c}_{i,k}\right)^2} \quad . \qquad (10)$$

trong đó

$$\bar{c}_{i,q} = \frac{1}{n_i}\sum_{h=1}^{n_i} c_{ih,q},$$

q = 1,2,…, Q là chỉ số thành phần của PDF sau khi rời rạc hóa,

$c_{ih,q}$ là giá trị mật độ tại vị trí lưới k của mẫu PDF thứ h trong chùm i.

***Bước** 5.3*. Xây dựng đường biên và ma trận phân vùng tương ứng:

Với độ lệch vừa tìm được ở ***Bước*** 5.2, tiến hành xây dựng biên trên và biên dưới cho các phần tử đại diện như sau:

$$c_i^{*(\text{upper})} = c_i^* + s_i,\ c_i^{*(\text{lower})} = c_i^* - s_i \qquad (11)$$

Các ma trận phân vùng tương ứng cho các phần tử đại diện được tính như sau:

$$u_{ij}^{(\text{upper})} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^{(upper)}}{d_{kj}^{(upper)}}\right)^2},$$

$$u_{ij}^{(\text{lower})} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^{(lower)}}{d_{kj}^{(lower)}}\right)^2}, \qquad (12)$$

với

$$d_{ij}^{(upper)} = \left\|p_j - c_i^{*(upper)}\right\|_2,$$

$$d_{ij}^{(lower)} = \left\|p_j - c_i^{*(lower)}\right\|_2.$$

***Bước** 5.4* Xây dựng ma trận phân vùng tối ưu:

Với mỗi PDF $p_j \in \mathcal{P}$, (j = 1, 2, …, *n*), xác định giá trị lớn nhất của ba ma trận phân vùng tại cột *j*:

$$m_j = \max\left\{\max_i u_{ij}, \max_i u_{ij}^{(upper)}, \max_i u_{ij}^{(lower)}\right\}.$$

Khi đó, ma trận phân vùng cuối cùng là

$$U^* = \left[u_{ij}^*\right]_{c\times n},$$

với các phần tử được xác định như sau:

$$u_{ij}^* = \begin{cases} u_{ij} & \text{nếu } m_j = \max_i\{u_{ij}\} \\ u_{ij}^{(upper)} & \text{nếu } m_j = \max_i\left\{u_{ij}^{(upper)}\right\}, \\ u_{ij}^{(lower)} & \text{nếu } m_j = \max_i\left\{u_{ij}^{(lower)}\right\} \end{cases} \qquad (13)$$

$$\forall i = 1,\dots,c, \forall j = 1,\dots,n.$$

4. VÍ DỤ MINH HỌA THUẬT TOÁN PHÂN TÍCH CHÙM ẢNH

Để minh họa cho cả hai thuật toán đề nghị, nghiên cứu sử dụng bộ dữ liệu trái cây gồm 1451 ảnh với kích thước 100 × 100 được định dạng JPG. Tập dữ liệu này gồm 3 loại trái cây: Apple (473 ảnh), Apple Rotten (488 ảnh) và Lychee (490 ảnh). Dữ liệu được lấy miễn phí từ website Kaggle tại link: https://www.kaggle.com/datasets/moltean/fruits. Mẫu ảnh của 3 loại trái cây được minh họa bởi Hình 2.

(a) Apple (b) Apple Rotten (c) Lychee

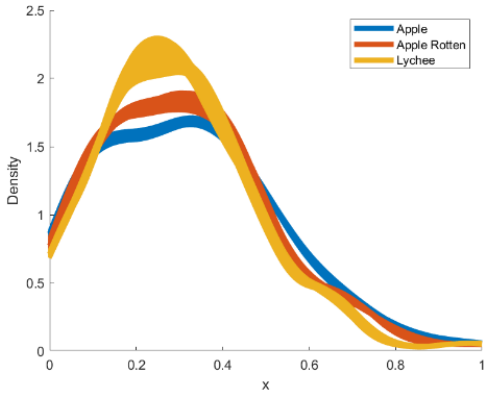
Hình 2. Mẫu ảnh của 3 loại trái cây

Các bước của thuật toán đề nghị trên tập dữ liệu này được cụ thể như sau:

***Bước* 1**. Áp dụng Thuật toán 1, trích xuất đặc trưng của 1451 ảnh ta có các PDF đại diện các ảnh của 3 nhóm

$$\mathcal{P} = \{p_1(x), p_2(x), \dots., p_{1451}(x)\},$$

mà chúng được minh họa bởi Hình 3.



Hình 3. Các PDF đại diện cho các ảnh trích xuất từ Thuật toán 1

Bước 2. Tại t = 0, xem mỗi PDF là một chùm, khi đó ta xây dựng tập đại diện PDFs ban đầu như sau:

$$\mathcal{C}^{(0)} = \{c_1{}^{(0)}, c_2{}^{(0)}, \dots, c_n{}^{(0)}\}$$

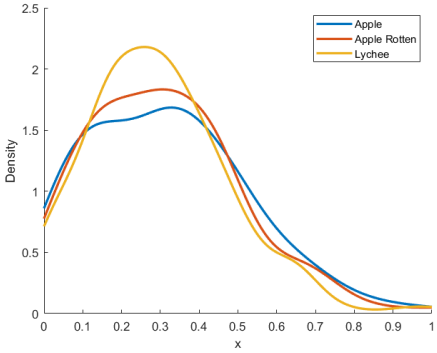
$$= \mathcal{P} = \{p_1(x), p_2(x), \dots, p_{1451}(x)\}.$$

Bước 3. Tại bước lặp thứ nhất $t = 1$, tiến hành cập nhật trọng tâm chùm để có tập $\mathcal{C}^{(1)}$. Ta có các kết quả sau

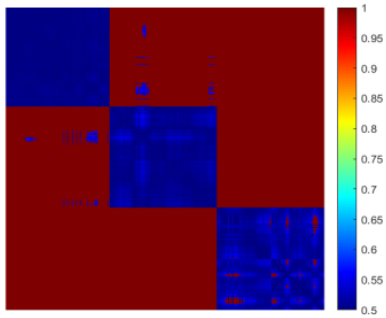
$$\mu = 0{,}7353 \text{ và } \sigma = 2{,}7154 .$$

Ma trận $\alpha(0) = [\alpha_{ij}(0)]_{1451\times1451}$ với $\alpha_{ij}(0) = 1$.

Khi đó, áp dụng công thức (2) ta tính được $\mathcal{C}^{(1)}$. Kết quả của quá trình hội tụ này, được minh họa bởi Hình 4.

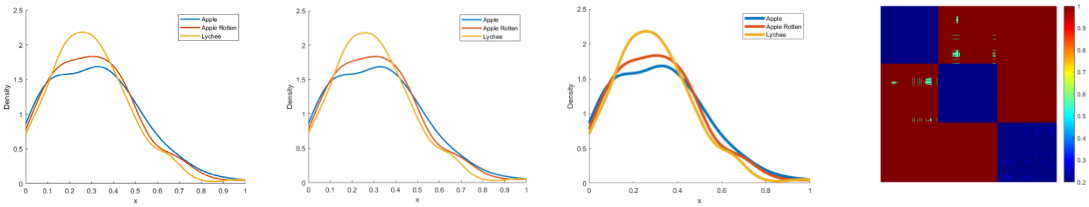


(a)



(b)

Hình 4. Minh họa kết quả sau lần cập nhật thứ nhất: (a) các PDF đại diện được cập nhật sau vòng lặp 1, (b) ma trận hệ số động α thu được tương ứng



(a) (b) (c) (d)

Hình 5. Quá trình cập nhật các PDFs đại diện qua các bước lặp: (a) $t = 2$, (b) $t = 3$, (c) $t = 4$, và (d) ma trận hệ số động α thu được tại $t = 4$

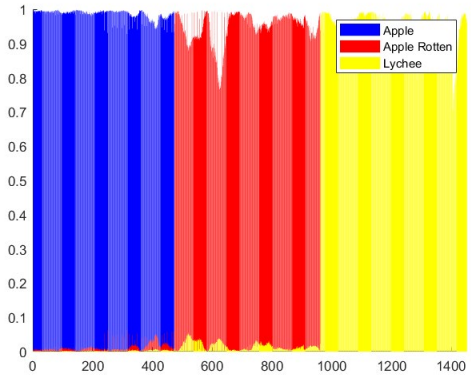
Bước 4. Trong lần lặp đầu tiên, ta có $\left\|\mathcal{C}^{(1)} - \mathcal{C}^{(0)}\right\| = 0{,}5247 > \epsilon$, nên thuật toán chưa dừng. Tiếp tục lặp lại **Bước** 3 cho đến lần thứ 4, ta thấy

$$\left\|\mathcal{C}^{(4)} - \mathcal{C}^{(3)}\right\| = 7{,}5165 \times 10^{-13} < \epsilon$$

thỏa mãn điều kiện dừng của thuật toán. Kết quả của quá trình hội tụ này được trình bày tại Hình 5.

Sau Bước 4, ta thu được $\mathcal{C}^*$gồm 3 PDF đại diện. Như vậy, sau **Bước** 4, 1451 ảnh ban đầu được chia thành 3 chùm. Đây là kết quả phân chùm đúng so với tập ảnh nguồn.

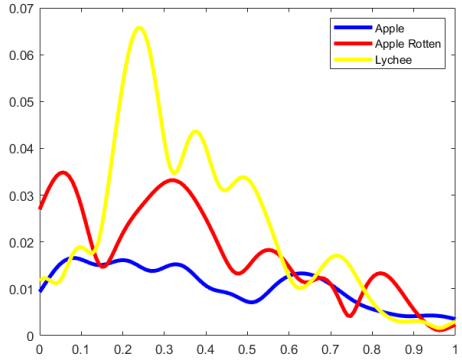
Bước 5. Với mỗi PDF $p_j \in \mathcal{P}$, tính khoảng cách từ mỗi phần tử đến từng chùm $c_i^*, i = 1{,}2{,}3$ và sử dụng công thức (8) để xây dựng ma trận phân vùng ban đầu $U = \left[u_{ij}\right]_{4\times 1941}$. Kết quả được minh họa bởi Hình 6



Hình 6. Ma trận phân vùng ban đầu

Từ ma trận phân vùng vừa tìm được, mỗi PDF p_j được gán nhãn theo công thức (9). Tiếp đó, công thức (10) được sử dụng để tính toán độ lệch chuẩn mật độ cho mỗi chùm, ta có 3 đường lệch chuẩn mẫu cho từng chùm, được trình bày trong Hình 7.

Dựa trên các giá trị độ lệch chuẩn đã thu được, tiến hành xác định biên trên $c^{*(upper)}$ và biên dưới $c^{*(lower)}$ cho mỗi PDF đại diện c_i^* theo công thức (11). Các phần tử đại diện sau khi được hiệu chỉnh bằng biên được thể hiện trong Hình 8.



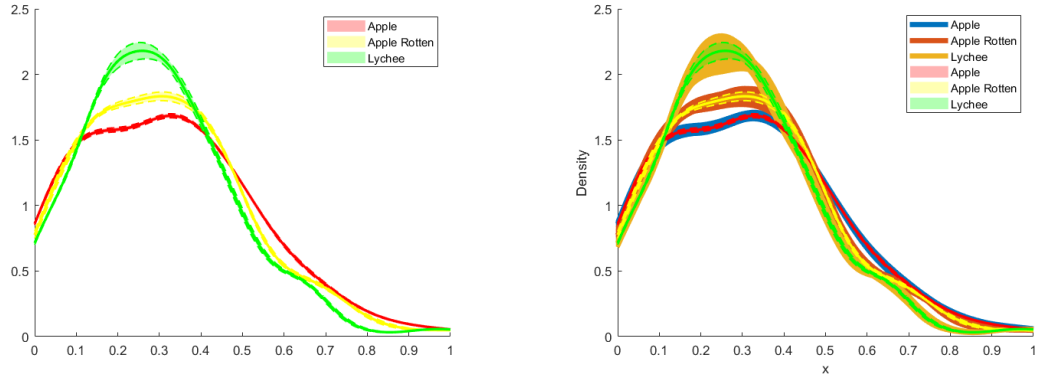
Hình 7. Độ lệch chuẩn mật độ trong mỗi chùm

Trong Hình (a), các biên cho thấy vùng biến thiên của phân bố quanh giá trị trung tâm, qua đó phản ánh mức độ phân tán dữ liệu trong từng chùm.

Trong Hình (b), khi chiếu các biên lên toàn bộ tập dữ liệu PDFs, ta quan sát được mức độ phù hợp của PDF đại diện với các quan sát thực tế. Các đường biên tạo thành một dải bao phủ, cung cấp cái nhìn trực quan về độ bao quát và khả năng khái quát hóa của phần tử đại diện đối với mỗi chùm.

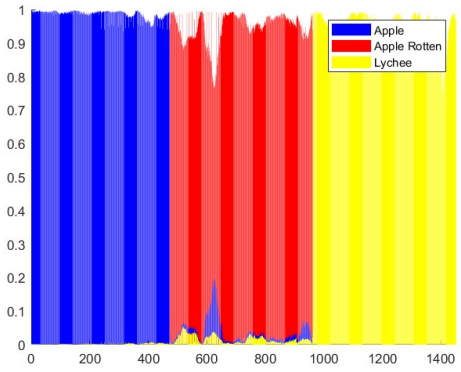
Điều này khẳng định rằng việc sử dụng biên trên và biên dưới không chỉ hỗ trợ định lượng sự phân tán, mà còn đóng vai trò quan trọng trong việc mô tả sự đa dạng của dữ liệu trong từng chùm, đồng thời giúp cập nhật lại ma trận phân vùng theo hướng tối ưu hơn.

Tiếp theo, bằng cách tính toán ma trận phân vùng tương ứng cho các biên trên và biên dưới của các PDF đại diện theo công thức (11), ta thu được các ma trận $U^{(upper)}$ và $U^{(lower)}$. Sau đó, áp dụng công thức (13) để xây dựng ma trận phân vùng tối ưu. Kết quả cuối cùng là ma trận phân vùng U^*, biểu thị xác suất thuộc 3 chùm của các ảnh được minh họa trong Hình 9.



(a) (b)

Hình 8. (a) Cận trên và dưới của 3 PDFs đại diện, (b) Tất cả các PDFs và PDFs đại diện



Hình 9. Xác suất thuộc 3 chùm của các ảnh

Kết quả thể hiện tại Hình 9 cho thấy xác suất phân vào chùm đúng của các ảnh rất cao. Các xác suất này hầu như đều bằng 1.

Với số chùm cố định là 3, các đặc trưng ảnh được trích xuất từ Resnet50 cho tập dữ liệu được xem xét, nghiên cứu áp dụng 2 thuật toán phân tích chùm nổi tiếng FCM (Kí hiệu Resnet50 + FCM) và k-means (Resnet50 + k-means). Nghiên cứu cũng thực hiện hai thuật toán phân tích chùm mờ gần đây cho tập dữ liệu ảnh được xem xét khi đặc trưng của nó được trích xuất từ phương pháp thống kê truyền thống bằng ma trận đồng hiện mức xám và biểu diễn thành vector biểu diễn thành khoảng hai chiều. So sánh kết quả thực hiện của các thuật toán thông qua 2 tham số PC và PE, ta có Bảng 1.

Kết quả từ Bảng 1 cho thấy, khi sử dụng đặc trưng của ảnh được trích xuất từ phương pháp học sâu Resnet50 (Thuật toán Resnet50 + FCM và Resnet50 + k-means) thì kết quả phân tích chùm sẽ tốt hơn các thuật toán sử dụng dữ liệu đầu vào trích xuất từ các phương pháp thống kê (Vo & Nguyen, 2018; Nguyen et al., 2024) vì giá trị PC lớn hơn và PE nhỏ hơn. Với cùng dữ liệu đầu vào đặc trưng được trích xuất từ Resnet50, thuật toán đề nghị đã cho giá trị PC lớn nhất và PE nhỏ nhất. Như vậy, thuật toán đề nghị đã cho kết quả phân tích chùm tốt nhất với tập dữ liệu này.

Bảng 1. Chỉ số đánh giá các thuật toán xây dựng chùm

Thuật toán	PC	PE
Đề nghị	0,9305	0,16250
Resnet50 + FCM	0,9298	0,16251
Resnet50 + k-means	0,9303	0,16223
Vovan and Nguyentrang (2018)	0,8971	0,2543
Nguyen et al. (2024)	0,9016	0,2108

5. KẾT LUẬN

Hai thuật toán để phân tích chùm được đề xuất nhằm hướng tới dữ liệu ảnh. Thuật toán 1 gồm các bước để trích xuất đặc trưng của mỗi ảnh thành 1 PDF đại diện dựa vào phương pháp Resnet50, thông qua quá trình GAP, GMP và phương pháp hàm hạt nhân trong ước lượng PDF. Kết quả của Thuật toán 1 được sử dụng làm dữ liệu đầu vào cho Thuật toán 2 để xây dựng chùm cho ảnh. Trong thuật toán 2, số chùm thích hợp, những phần tử trong mỗi chùm và xác suất thuộc vào chùm của mỗi phần tử được cùng lúc thực hiện. Kết hợp Thuật toán 1 và Thuật toán 2, có một phương pháp để phân tích chùm hiệu quả cho ảnh. Thuật toán đề nghị có thể thực hiện dễ dàng cho tập ảnh thực bởi một chương trình Matlab được thiết lập. Qua một tập ảnh cụ thể, thuật toán đề nghị đã cho kết quả hợp lý và vượt qua các thuật toán phân tích chùm phổ biến khác dựa vào các tham số đánh giá.

LỜI CẢM TẠ

Đề tài được tài trợ bởi Đại học Cần Thơ, mã số TSV2025-43.

TÀI LIỆU THAM KHẢO

Bezdek, C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*, 191–203. https://doi.org/10.1016/0098-3004(84)90020-7.

Chang, S. C., & Jeng, J. T. (2025). Interval generalized improved fuzzy partitions fuzzy c-means under Hausdorff distance clustering algorithm, *27,* 834 – 852. https://doi.org/10.1007/s40815-024-01809-w.

Chen, T. L., & Shiu, S. Y. (2007). A new clustering algorithm based on self-updating process. *Proceeding about Statistical Computing*, Salt Lake City, Utah, 2034, 1–5. https://api.semanticscholar.org/CorpusID:18819116

Chen, J. H., Chang, Y. C., & Hung, W. L. (2020). A self-organizing clustering algorithm for functional data. *Communications in Statistics - Simulation and Computation, 49*(5), 1237–1263. https://doi.org/10.1080/03610918.2018.1494280

Chen, J. H., & Hung, W. L. (2015). An automatic clustering algorithm for probability density functions. *Journal of Statistical Computation and Simulation, 85*(15), 3047–3063. https://doi.org/10.1111/stan.12315.

Conze, P.-H., Andrade-Miranda, G., Singh, V. K., Jaouen, V., & Visvikis, D. (2025). Current and emerging trends in medical image segmentation with deep learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, *7*(6), 545 - 569. https://doi: 10.1109/TRPMS.2023.3265863

Bussa, S.K., & Boppana, N. K. (2025). Enhanced ResNet50 deep learning algorithm for classification of crack images in RCC structures. *Asian Journal of Civil Engineering*, 26, 3773–3784. https://doi.org/10.1007/s42107-025-01396-7.

Hung, W. L., Chang-Chien, S. J., & Yang, M. S. (2015). An intuitive clustering algorithm for spherical data with application to extrasolar planets. *Journal of Applied Statistics, 42*(10), 2220–2232. https://doi.org/10.1080/02664763.2015.1023271

Hung, W. L., & Yang, J. H. (2015). Automatic clustering algorithm for fuzzy data. *Journal of Applied Statistics, 42*(7), 1503–1518. https://doi.org/10.1080/02664763.2014.1001326

Hung, W. L., Yang, J. H., & Shen, K. F. (2016). Self-updating clustering algorithm for interval-valued data. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1494–1500). https://doi.org/10.1109/FUZZIEEE.2016.7737867

Hung, W. L., Yang, J. H., Song, I. W., & Chang, Y. C. (2021). A modified self-updating clustering algorithm for application to dengue gene expression data. *Communications in Statistics - Simulation and Computation, 50*(2), 483–500. https://doi.org/10.1080/03610912018.1563149

Le, N. T. K., Le, T. H., & Vo, T. V (2023). Automatic clustering algorithm for interval data based on overlap distance. *Communications in Statistics - Simulation and Computation, 52*(5), 2194–2209. https://doi.org/10.1080/03610918.2021.1900248

Nguyen, T. T., Nguyen, T. T., Nguyen, K. N. T., & Vo, T. V. (2023). Balance-driven automatic clustering for probability density functions using metaheuristic optimization. *International Journal of Machine Learning and Cybernetics, 14,* 1063–1078. https://doi.org/10.1007/s13042-022-01683-8

Nguyen, T. D., & Vo, T. V. (2024). Classifying for interval and applying for image based on the extracted texture feature. *Granular Computing, 9*(29), 1-20. https://doi.org/10.1007/s41066-024-00450-0

Pham, D. T., & Vo, T. V. (2024). Improving fuzzy clustering model for probability density functions using the two-objective genetic algorithm. *Multimedia Tools and Applications, 83,* 45291-45314. https://doi.org/10.1007/s11042-023-17217-5

Vo, T. V., & Nguyen, T. T. (2018). Similar coefficient of cluster for discrete elements. *Sankhya B, The Indian Journal of Statistics*, 80(1), 19 - 36. https://doi.org/10.1007/s13571-018-0159-0

Vo, T. V., Nguyen, Y. H., & Dang, S. (2023). An automatic fuzzy clustering algorithm for discrete elements. *Journal of the Operations Research Society of China*, *11*, 309-325. https://doi.org/10.1007/s40305-021-00388-z

Vo, T. V. (2025). Classifying for images based on the extracted color scales and representative probability density functions. *Expert Systems with Applications, 291*, 128456. https://doi.org/10.1016/j.eswa.2025.128456.

Zhou, J., Liu, T., & Zhu, J. (2019). Weighted adjacent matrix for k-means clustering, *Multimedia Tools and Applications, 78*, 33415–33434. https://doi.org/10.1007/s11042-019-08009-x.