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.2025.111

CẢI TIẾN THUẬT TOÁN TỰ CẬP NHẬT CHÙM CHO CÁC HÀM MẬT ĐỘ XÁC SUẤT

Trần Nam Hưng, Châu Ngọc Thơ và Võ Văn Tài*
Khoa 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): 06/01/2025
Sửa bài (Revised): 29/01/2025
Duyệt đăng (Accepted): 11/05/2025

Title: *Improving the cluster seft-update algorithm for probability density functions*

Author: *Tran Nam Hung, Chau Ngoc Tho and Vo Van Tai**

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

TÓM TẮT

Nghiên cứu được thực hiện nhằm cải tiến thuật toán tự cập nhật chùm cho các hàm mật độ xác suất (PDF) khi tham số trong thuật toán được chọn thích ứng với từng tập dữ liệu thay vì là một hằng số như các thuật toán trước đó. Thuật toán đề nghị được trình bày chi tiết các bước thực hiện và ví dụ số minh hoạ. Thuật toán đề nghị cũng được áp dụng cho dữ liệu ảnh khi các pixel của chúng được trích xuất và biểu diễn thành các PDF đại diện. Áp dụng cho dữ liệu số và dữ liệu ảnh, thuật toán đề nghị đã cho kết quả tốt và thuận lợi hơn một số thuật toán được công bố gần đây.

Từ khoá: *Dữ liệu ảnh, hàm mật độ xác suất, phân tích chùm, tự cập nhật chùm*

ABSTRACT

This research improves the cluster self-update algorithm for probability density functions (PDFs) when the parameter is selected based on each dataset rather than being constant, as in previous algorithms. The proposed algorithm is presented in detail with step-by-step procedures and numerical examples. It is also applied to image data, where the pixels are extracted and represented as representative PDFs. Applying numerical examples and images demonstrates that the proposed algorithm obtains good results and outperforms some recently published algorithms.

Keywords: *Image data, probability density function, cluster analysis, cluster self-update*

1. GIỚI THIỆU

Phân tích chùm là công cụ học không được giám sát, với mục tiêu khai phá và phân bổ dữ liệu phi cấu trúc thành các chùm, sao cho những đối tượng trong cùng chùm có được sự tương đồng theo những đặc tính chung nào đó. Trong xử lý dữ liệu lớn, phân tích chùm được xem là nền tảng không thể thiếu nên nó đã và đang nhận được rất nhiều sự quan tâm của các nhà nghiên cứu (Hung et al., 2015; Chen et al., 2020; Le et al., 2023).

Phân tích chùm có thể thực hiện cho dữ liệu điểm, dữ liệu khoảng và dữ liệu hàm mật độ xác suất (PDF). Phân tích chùm cho dữ liệu điểm được đề nghị từ rất sớm, được phát triển rất mạnh cả lý thuyết và ứng dụng (Vo & Nguyen, 2018; Hung et al., 2021; Vo et al., 2023). Trong thực tế, chúng ta cũng lưu trữ nhiều dữ liệu khoảng như giá thấp nhất và cao nhất của cổ phiếu, vàng, đô la. Các dữ liệu thời tiết như nhiệt độ, lượng mưa, độ ẩm,… cũng thường được ghi nhận dưới dạng khoảng. Do đó, việc phân tích chùm cho dữ liệu khoảng đã được đề nghị (Le et al., 2023). Khi dữ liệu lớn, mỗi đối tượng cần

được xem là một phân phối hay dữ liệu phức tạp như hình ảnh có thể được đại diện bởi một phân phối thì phân tích chùm cho các PDF đã được đề nghị (Chen & Hung, 2015; Chen et al., 2020). So với dữ liệu số và dữ liệu khoảng, phân tích chùm cho các PDF được đánh giá phức tạp hơn, đặc biệt là ứng dụng. Sự phức tạp thể hiện trong độ đo đánh giá sự tương tự của các PDF và trong các tính toán. Việc phân tích chùm cho các PDF được nghiên cứu này quan tâm.

Có nhiều thuật toán phân tích chùm khác nhau đã được đề nghị như: thuật toán thứ bậc, không thứ bậc, cực đại hoá kỳ vọng. Khi xây dựng thuật toán phân tích chùm, việc xác định số chùm thích hợp và những phần tử cụ thể trong mỗi chùm là hai vấn đề quan trọng cần phải được giải quyết. Khi hai vấn đề này được thực hiện trong cùng một thuật toán thì nó được gọi là thuật toán tự cập nhật chùm (cluster self-update algorithm, CSU). CSU lần đầu tiên được giới thiệu bởi Chen and Shiu (2007), với ý tưởng đơn giản là các đối tượng thuộc cùng một chùm qua thuật toán hội tụ đến cùng một phần tử. Trong CSU, sự hội tụ của thuật toán cũng như kết quả phân tích chùm phụ thuộc vào một hằng số mà nó được chọn cố định cho tất cả dữ liệu. Sau năm 2015, CSU được phát triển nhanh với nhiều cải tiến mới về vấn đề bán kính lân cận. Chen and Hung (2015) đã phát triển bán kính lân cận bởi giá trị trung bình các khoảng cách. Sự thay đổi này được áp dụng vào dữ liệu hàm mờ và dạng cầu (Hung & Yang, 2015). Việc sử dụng bán kính lân cận bởi trung bình khoảng cách thích ứng được với những tập dữ liệu có nhiều đặc tính khác nhau của thực tế, tuy nhiên nhiều trường hợp vẫn không cho kết quả phân tích chùm thích hợp. Để cải tiến nhược điểm này của chính mình, Shiu and Chen (2016) đã đề nghị mô hình tĩnh và động qua mỗi lần lặp. Sự cải tiến này đã tạo ra sự thích ứng hơn với dữ liệu của thuật toán nên nhận được những kết quả tốt cho nhiều tập dữ liệu. Tuy nhiên, các tham số trong thuật toán của Chen and Chiu (2016) thực chất vẫn là hằng số. Người sử dụng khi thực hiện vẫn phải cài đặt trước ba tham số như bán kính lân cận, tham số hội tụ và hàm hạt nhân.

Về tham số hội tụ trong CSU, một số công bố quan trọng sau đã được quan tâm. Trong Chen and Shiu (2007), tham số này được sử dụng cố định là 1. Với Chen and Shiu (2016), nó được thành lập dựa trên bán kính và số vòng lặp. Một đề nghị đáng chú ý của Chen and Hung (2015) là chọn tham số hội tụ bằng phương sai của từng đôi khoảng cách. Thuật toán cải tiến này cũng được áp dụng cho dữ liệu khoảng bởi Hung et al. (2016), nghiên cứu của Le et al. (2023) cho rằng, việc chọn tham số bằng trung bình và phương sai chưa giải quyết được vấn đề phát hiện nhiễu nên đã lần lượt thay bằng trung vị và khoảng tứ phân vị. Pham and Vo (2023) tập trung sử dụng tự cập nhật cho mờ hóa dữ liệu chuỗi thời gian với sự thay đổi tham số cho từng trường hợp cụ thể. Tất cả các nghiên cứu trên đều khẳng định rằng việc tham số trong thuật toán CSU quyết định số chùm thích hợp cũng như những phần tử trong mỗi chùm, nhưng việc xác định tối ưu cho nó vẫn là bài toán chưa có lời giải cuối cùng.

Thuật toán CSU được nghiên cứu với những đóng góp chính sau:

(i) Tổng quát hoá việc chọn tham số trong thuật toán CSU, phù hợp với từng tập dữ liệu để tối ưu kết quả xây dựng chùm.

(ii) Ứng dụng cho dữ liệu ảnh khi đặc trưng pixel của chúng được trích xuất để chứng minh sự hiệu quả của thuật toán đề nghị.

2. CÁC VẤN ĐỀ LIÊN QUAN

2.1. Hàm mật độ xác suất trọng tâm

Trong các thuật toán xây dựng chùm cho dữ liệu PDF, việc xác định hàm trọng tâm của các chùm là vấn đề quan trọng. Trong nghiên cứu này, hàm trọng tâm được đề nghị thông qua các định nghĩa sau.

Định nghĩa 1. *(Hàm hạt nhân)* Hàm ϕ được gọi là hàm hạt nhân khi và chỉ khi nó thoả các điều kiện sau:

(i) $0 \le \phi(u,v) \le 1$; $\phi(u,v) = 1$ khi $u = v$,

(ii) $\phi(u,v)$ chỉ phụ thuộc vào khoảng cách $d(u,v)$ giữa hai phần tử.

(iii) $\phi(u,v)$ là hàm nghịch biến theo khoảng cách của $d(u,v)$.

Định nghĩa 2. *(Hàm mật độ xác suất trọng tâm)* Cho m chùm $\{C_1, C_2, \dots, C_m\}$ của n-PDF $\{f_1, f_2, \dots, f_n\}$, $(m \le n)$. Khi đó, hàm

$$\theta_i = \sum_{i<j} \left(\frac{\phi(f_i, f_j)}{\sum_{i<j} \phi(f_i, f_j)} \times f_j \right), i = 1, 2, \dots m$$

được gọi là hàm trọng tâm của chùm C_i.

Ta có 2 nhận xét sau:

- Vì hàm $\phi(\cdot)$ không âm và có giá trị trong đoạn $[0;\ 1]$, nên hàm trọng tâm θ_i cũng không âm.
- Mặt khác ta có

$\sum_{i<j} \frac{\phi(f_i,f_j)}{\sum_{i<j}\phi(f_i,f_j)} = 1, \int f_j(x)dx = 1$ (vì f(x) là một PDF).

Do đó

$$\int \theta_i(x)dx = \int \sum_{i<j} \left(\frac{\phi(f_i, f_j)}{\sum_{i<j} \phi(f_i, f_j)} \times f_j(x) \right) dx$$

$$= \left(\sum_{i<j} \frac{\phi(f_i, f_j)}{\sum_{i<j} \phi(f_i, f_j)} \right) \int f_j(x)dx = 1.$$

Từ hai nhận xét trên, ta kết luận được hàm trọng tâm của một chùm cũng là một PDF.

2.2. Khoảng cách của hai hàm mật độ xác suất

Định nghĩa 3. *(Khoảng cách L_1)* Cho hai PDF $f_1(x)$ và $f_2(x)$, $x \in \mathbb{R}$. Khi đó, khoảng cách L_1 giữa $f_1(x)$ và $f_2(x)$ được định nghĩa là:

$$\|f_1, f_2\|_1 = \int |f_1(x) - f_2(x)| \, dx.$$

Từ định nghĩa trên, ta có $0 \leq \|f_1, f_2\|_1 \leq 2$. Khi hai PDF càng xa nhau thì khoảng cách này càng lớn và ngược lại. Nếu hai PDF trùng nhau thì khoảng cách bằng 1 và nếu hai PDF rời nhau hoàn toàn thì khoảng cách này bằng 2.

2.3. Trích xuất ảnh thành hàm mật độ xác suất đại diện

Để nhận dạng ảnh, trước hết chúng ta phải trích xuất đặc trưng của nó. Các đặc trưng của ảnh có thể được sử dụng là hình dạng, màu sắc và kết cấu. Trong nghiên cứu này, đặc trưng màu sắc được sử dụng thông qua các điểm pixel của ảnh.

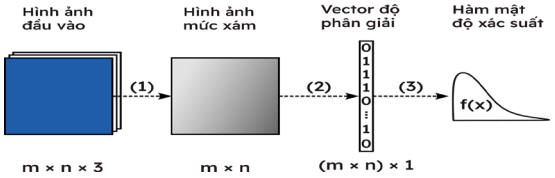
Sau khi các đặc trưng được trích xuất, chúng ta phải biểu diễn chúng thành một đối tượng nào đó để nhận dạng. Thông thường đối tượng đại diện cho ảnh là ma trận, các điểm và các khoảng. Đối tượng đại diện cho ảnh là PDF được sử dụng với các bước thực hiện cụ thể sau:

Bước 1. Chuyển đổi hình ảnh màu bất kỳ về dạng ảnh màu xám.

Bước 2. Chuyển ma trận điểm ảnh thành một véc tơ cột bằng cách duỗi thẳng tất cả giá trị pixel.

Bước 3. Ước lượng PDF cho mỗi ảnh dựa vào kết quả của Bước 2.

Như vậy, việc nhận dạng hay phân biệt các hình ảnh với nhau dựa vào sự khác biệt màu sắc thông qua các PDF đại diện. Kết quả thể hiện ở Hình 1 minh hoạ các bước trích xuất đặc trưng pixel của mỗi ảnh thành một PDF.



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

Có nhiều phương pháp tham số và phi tham số để ước lượng PDF. Trong nghiên cứu này, phương pháp hàm hạt nhân, một phương pháp phi tham số được sử dụng phổ biến hiện nay được sử dụng. Với dữ liệu là các điểm $\{x_1, x_2, \ldots, x_N\}$, PDF ước lượng theo phương pháp hàm hạt nhân có dạng:

$$\hat{f}(x) = \frac{1}{N}\frac{1}{h}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \qquad (1)$$

trong đó,

- x_i là dữ liệu thứ i với $i = 1, 2, \ldots, N$,
- h là tham số trơn,
- $K(\cdot)$ là hàm hạt nhân thoả mãn hai điều kiện $K(\cdot) \geq 0$ và $\int K(x)dx = 1$.

Khi có một tập dữ liệu cụ thể, kết quả ước lượng PDF từ tập dữ liệu này phụ thuộc vào việc chọn tham số trơn và hàm hạt nhân. Mặc dù có nhiều nghiên cứu quan tâm đến hai vấn đề này nhưng vẫn chưa có một phương pháp nào được xem hiệu quả cho tất cả số liệu. Trong nghiên cứu này, hàm hạt nhân được chọn có dạng chuẩn $K(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ và tham số trơn được chọn theo đề nghị của Bowman and Azzalini (1997). Cụ thể, $h = (4/3)^{1/5} \cdot \sigma \cdot N^{1/5}$, với N và σ lần lượt là số phần tử và độ lệch chuẩn của tập dữ liệu.

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

Khi chùm được xây dựng, để đánh giá hiệu quả của các thuật toán phân tích chùm với nhau, có thể căn cứ vào nhiều tham số thống kê. Trong nghiên cứu này, tham số ARI và G-mean được sử dụng.

(i) Chỉ số ARI (*Adjusted Rand Index*) đo mức độ tương đồng giữa kết quả phân chùm của thuật toán và thực tế. Chỉ số này được cho bởi công thức sau:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_i \binom{b_j}{2}\Big/\binom{n}{2}\right]}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_i \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_i \binom{b_j}{2}\Big/\binom{n}{2}\right]},$$

trong đó,

- n_{ij}: Số phần tử chung giữa chùm thực tế thứ i và chùm của thứ j của thuật toán,
- $a_i = \sum_j n_{ij}$: Tổng số phần tử trong chùm thực tế thứ i,
- $=$: Tổng số phần tử trong chùm thứ j của thuật toán,
- n: Tổng số phần tử trong tập dữ liệu.
- : Số tổ hợp chập hai của n, biểu thị tổng số cặp mẫu.

ARI có giá trị trong đoạn $[-1;\ 1]$, với *ARI* càng lớn thì kết quả phân tích chùm càng tốt và ngược lại. Khi kết quả phân tích chùm hoàn toàn khớp với thực tế thì giá trị của *ARI* là 1 và ngược lại, khi kết quả phân tích chùm hoàn toàn khác thực tế thì *ARI* nhận giá trị là -1. Giá trị *ARI* bằng 0 khi các đối tượng được phân bổ ngẫu nhiên vào các chùm.

(ii) Độ nhạy ($Sensitivity$) có công thức là $\frac{TP}{TP+FN}$ và độ chuyên ($Specificity$) có công thức là $\frac{TN}{TN+FP}$, trong đó TP là số cặp điểm được phân chùm đúng vào cùng một chùm, FN là số cặp điểm thuộc cùng chùm nhưng bị phân chùm sai, TN là số cặp điểm thuộc các chùm khác nhau và được phân chùm đúng, FP là số cặp điểm thuộc các chùm khác nhau nhưng bị phân chùm sai.

(iii) Chỉ số trung bình nhân (*G-mean*) là căn bậc hai của tích độ nhạy và độ chuyên. Cụ thể:

G-means $= \sqrt{Sensitivity \times Specificity}$.

Ta có *G-means* có giá trị $[0;\ 1]$. *G-means* có giá trị càng lớn thì kết quả phân tích chùm càng tốt và ngược lại. Khi G-means bằng 1 thì kết quả phân tích chùm là hoàn hảo.

3. THUẬT TOÁN PHÂN TÍCH CHÙM ĐỀ NGHỊ

Cho tập n PDF: $F = \{f_1, f_2, \dots, f_n\}$ trong không gian d-chiều. Gọi ϵ là một số dương nhỏ tuỳ ý và $T \in N^*$ là số lần lặp tối đa của thuật toán. Thuật toán CSU cải tiến cho F gồm bốn bước sau:

Bước 1. Tại vòng lặp $t = 0$, ta gán dãy trọng tâm khởi tạo chính là các PDF ban đầu. Tức là

$$\Theta^{(0)} = \left\{\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_n^{(0)}\right\} \equiv \left\{f_1^{(0)}, f_2^{(0)}, \dots, f_n^{(0)}\right\}.$$

Bước 2. Lấy ngẫu nhiên không lặp theo phân phối đều một tập hợp con S của F có lực lượng $\lfloor n/2 \rfloor$.

Bước 3. Tại vòng lặp $t+1$, nâng cấp trọng tâm $\Theta^{(t+1)}$ theo công thức sau:

$$\theta_i^{(t+1)} = \sum_{\theta_j^{(t)} \in S} \left(\frac{\phi\left(\theta_i^{(t)}, \theta_j^{(t)}\right)}{\sum_{\theta_j^{(t)} \in S} \phi\left(\theta_i^{(t)}, \theta_k^{(t)}\right)} \times \theta_j^{(t)} \right) \quad (2)$$

với $i, j = 1, 2, \dots, n,\ i \le j$ và $\phi\left(\theta_i^{(t)}, \theta_j^{(t)}\right)$ là hàm hạt nhân chặt cụt được xác định bởi

$$\phi\left(\theta_i^{(t)}, \theta_j^{(t)}\right) = \begin{cases} \exp\left(L_1\left(\theta_i^{(t)}, \theta_j^{(t)}\right)/\lambda\right) & (a), \\ 0 & (b) \end{cases}$$

với điều kiện của (a) là $L_1\left(\theta_i^{(t)}, \theta_j^{(t)}\right) \le \mu_d \alpha_{ij}(t)$, và điều kiện của (b) là $L_1\left(\theta_i^{(t)}, \theta_j^{(t)}\right) > \mu_d \alpha_{ij}(t)$.

Trong hàm hạt nhân chặt cụt, ta có

- $L_1\left(\theta_i^{(t)}, \theta_j^{(t)}\right)$ là khoảng cách từng đôi giữa hai PDF trọng tâm bất kỳ ở vòng lặp t,
- $\mu_d = \frac{1}{\binom{n}{2}} \sum_{i<j} L_1\left(\theta_i^{(t)}, \theta_j^{(t)}\right)$ là trung bình của từng đôi các khoảng cách L_1,
- $\alpha_{ij}(t+1) = \frac{\alpha_{ij}(t)}{1+\alpha_{ij}(t)\phi\left(\theta_i^{(t)}, \theta_j^{(t)}\right)}, \alpha_{ij}(0) = 1$ là tốc độ học,
- $\lambda = \frac{\sigma_d}{1+t}$ là tốc độ hội tụ, trong đó tham số $\sigma_d = \sqrt{\frac{1}{\binom{n}{2}} \sum_{i<j} \left(L_1\left(\theta_i^{(t)}, \theta_j^{(t)}\right) - \mu_d\right)^2}$ là phương sai của từng đôi các khoảng cách.

Bước 4. Lặp lại Bước 2 và Bước 3 t lần hoặc điều kiện sau được thoả:

$$\left\|\Theta^{(t)} - \Theta^{(t-1)}\right\| = \max\left\{L_1\left(\theta_i^{(t)}, \theta_j^{(t)}\right)\right\} < \epsilon, \quad (3)$$

trong đó ϵ là một số dương nhỏ tuỳ ý. Trong các ví dụ số và ứng dụng của nghiên cứu này, $\epsilon = 10^{-6}$ được chọn.

Khi thuật toán đề nghị kết thúc, nếu $\Theta^{(t)}$ có bao nhiêu phần tử thì chúng ta chia tập dữ liệu ban đầu F thành bấy nhiêu chùm. Những phần tử cùng hội tụ về chung một PDF đại diện thì được xếp chung một chùm.

Nhận xét. Ta có một số nhận xét sau về thuật toán đề nghị.

(i) Ở Bước 2, các tập hợp con được chọn chỉ chiếm 50% dữ liệu ban đầu và không ảnh hưởng đến kết quả của thuật toán nhưng lại làm giảm đáng kể

sự lặp lại của thuật toán và tăng khả năng tìm kiếm các đối tượng mới một cách ngẫu nhiên.

(ii) Điều kiện (a) và (b) là một sự cải tiến quan trọng trong thuật toán đề nghị so với các thuật toán trước đó. Bán kính của lân cận $r(t) = \mu_d \alpha_{ij}(t)$ thay đổi qua từng vòng lặp t bởi tốc độ học $\alpha_{ij}(t)$ và co giãn độ lớn với khoảng cách trung bình μ_d thay vì cố định như các thuật toán trước đó. Các lân cận cũng có sự tự cập nhật dựa trên ý tưởng của Self-Organizing Map về lượng tử hóa véc tơ mạng nơ-ron. Khi khoảng cách L_1 của hai PDF bất kỳ lớn hơn bán kính lân cận, trọng tâm đó không được xem xét và gán trọng số là 0. Ngược lại, khi khoảng này nằm trong bán kính cho phép, ta xem xét trọng số và di chuyển trọng tâm.

(iii) Việc chọn $\lambda = \frac{\sigma_d}{1+t}$ cũng là một đóng góp của nghiên cứu này. Thay vì λ là một hằng số cố định (tĩnh), λ trong thuật toán đề nghị được xem là động. Giá trị của nó phụ thuộc vào phương sai của từng đôi khoảng cách (σ_d) và giảm liên tục từ σ_d về 0. Khi vòng lặp t càng lớn, tốc độ hội tụ λ càng nhỏ, làm cho trọng số ϕ càng cao. Từ đó, hàm ϕ giúp các trọng tâm di chuyển chậm hơn. Điều này cho phép thuật toán đề nghị có kết quả hợp lý hơn khi các phần tử có nhiều sự chồng lấp.

4. VÍ DỤ SỐ

4.1. Ví dụ 1

Gọi F là tập gồm 20 PDF có phân phối chuẩn 2 chiều với các véc tơ trung bình được sinh ngẫu nhiên theo phân phối chuẩn và hiệp phương sai theo phân phối đều. Chi tiết dữ liệu mô phỏng xem xét gồm bốn chùm, mỗi chùm gồm 5 PDF lần lượt có véc tơ trung bình như sau:

Chùm 1: $\mu_1 \sim N((-1,-1)^T;\ 0{,}05I_2)$,

Chùm 2: $\mu_2 \sim N((1,1)^T;\ 0{,}05I_2)$,

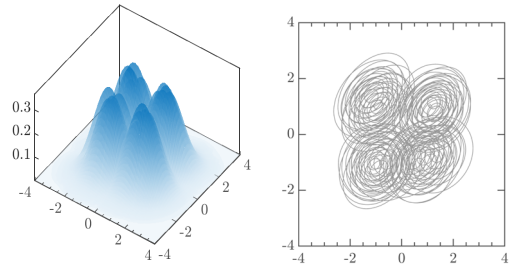
Chùm 3: $\mu_3 \sim N((1,-1)^T;\ 0{,}05I_2)$,

Chùm 4: $\mu_4 \sim N((-1,1)^T;\ 0{,}05I_2)$,

trong đó: I_2 là ma trận đơn vị cấp 2. Ma trận hiệp phương sai của các phân phối được xác định như sau:

$$\Sigma = \begin{pmatrix} 0{,}5 & u/2 \\ u/2 & 0{,}5 \end{pmatrix}, u \sim U[0,1].$$

Hình 2a thể hiện đồ thị bề mặt 3D của 20 PDF và Hình 2b là đường đồng mức của chúng.



(a) **(b)**

Hình 2. (a) Đồ thị bề mặt của 20 PDF có phân phối chuẩn hai chiều; và (b) Đường đồng mức của các PDF

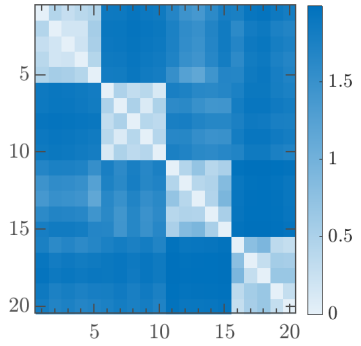
Hình 2 cho thấy các PDF ở ví dụ này có tỷ lệ chồng lấp khá cao giữa các nhóm nên có thể là một thách thức cho bài toán phân tích chùm.

Ta lần lượt thực hiện các bước trong thuật toán đề nghị với kết quả thực hiện như sau:

Bước 1. Xây dựng PDF trọng tâm ban đầu trùng với toàn bộ PDF đã cho. Cụ thể, ta có

$$\Theta^{(0)} = \left\{\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_{20}^{(0)}\right\} \equiv \{f_1, f_2, \dots, f_{20}\}.$$

Tính từng đôi các khoảng cách $L_1\left(\theta_i^{(0)}, \theta_j^{(0)}\right)$ giữa hai trọng tâm bất kỳ, ta nhận được kết quả được thể hiện ở Hình 3.



Hình 3. Ma trận khoảng cách các trọng tâm

Tính khoảng cách trung bình và phương sai của các khoảng cách, ta nhận được kết quả:

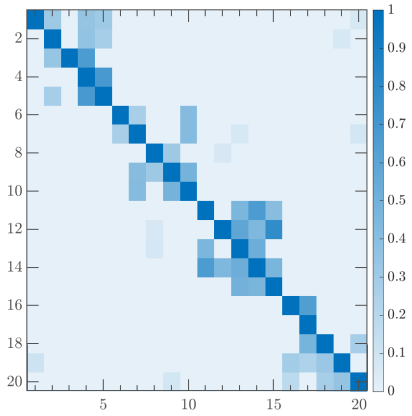
$$\mu_d = 1{,}50 \text{ và } \sigma_d = 0{,}54.$$

Bước 2. Chia $\lfloor n/2 \rfloor = 10$ PDF một cách ngẫu nhiên theo phân phối đều, tức là ta chọn các PDF trong tập S để khảo sát trọng số, trong đó,

$$S = \{1;\ 3;\ 4;\ 7;\ 8;\ 9;\ 14;\ 18;\ 19;\ 20\}.$$

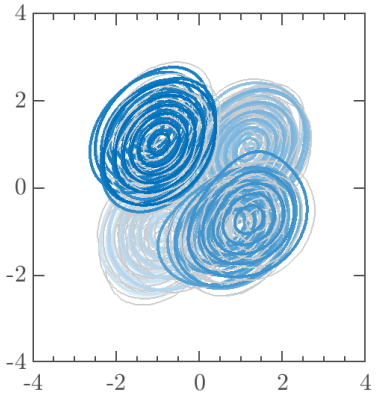
Bước 3. Vì tại vòng lặp đầu tiên $t = 0$, $\alpha_{ij}(0) = 1$, bán kính lân cận lúc này là $\mu_d = 1{,}50$ nên tốc độ hội tụ λ có giá trị bằng $\sigma_d = 0{,}54$. Khi đó, ma trận

ϕ giữa hai trọng tâm thuộc tập hợp S được tính với kết quả thể hiện trong Hình 4.



Hình 4. Ma trận hàm hạt nhân ϕ

Bước 3. Tiến hành nâng cấp các trọng tâm theo công thức (2) ta nhận được sự biến đổi của trọng tâm $\theta_i^{(1)}, i = 1;\ 2;\ 3;\ 4$, trong đó đồ thị đường đồng mức của chúng được thể hiện bởi Hình 5.

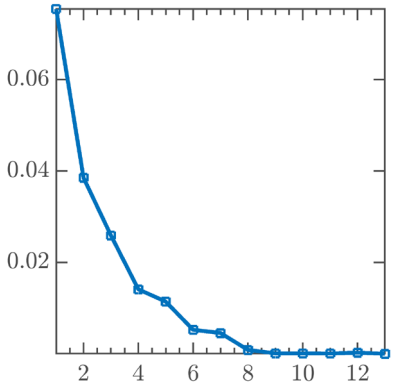


Hình 5. Đường đồng mức của các trọng tâm tại vòng lặp thứ nhất ($t = 1$)

Bước 4. Tính điều kiện dừng theo công thức (3), ta có

$$\epsilon = \left\|\Theta^{(1)} - \Theta^{(0)}\right\| = \max\left|{\theta_{ij}}^{(1)} - {\theta_{ij}}^{(0)}\right| = 0{,}08.$$

Vì điều kiện dừng chưa thỏa mãn ($\epsilon > 10^{-6}$), nên ta tiếp tục lặp lại Bước 2 và Bước 3 với sự cập nhật của các trọng tâm $\theta_i^{(1)}, i = 1;\ 2;\ 3;\ 4$. Sau 13 vòng lặp, thuật toán hội tụ. Quá trình dừng ứng với giá trị của ϵ qua 13 vòng lặp được mô tả trong Hình 6.



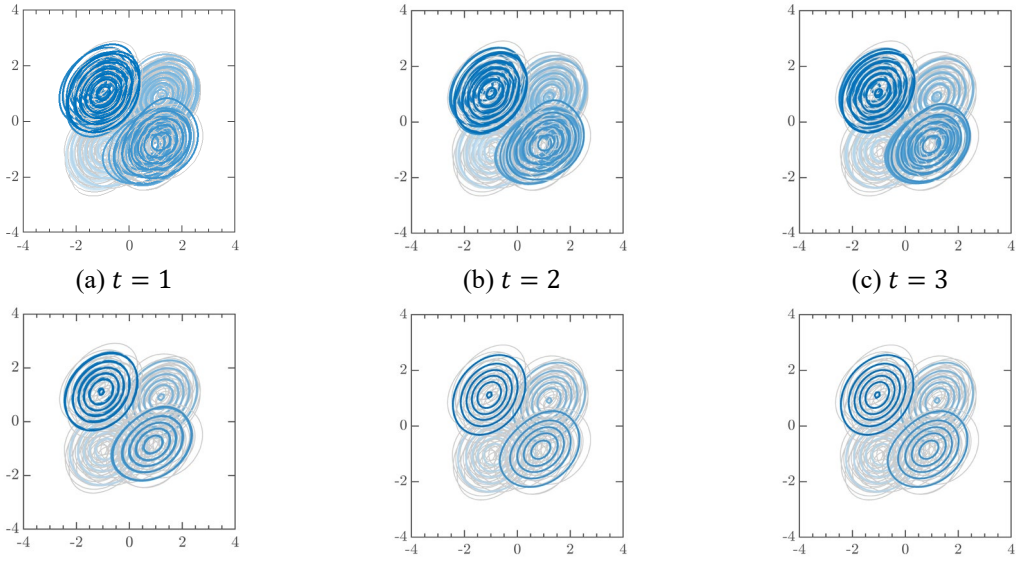
Hình 6. Quá trình dừng của thuật toán đề nghị qua 13 vòng lặp

Chi tiết PDF trọng tâm của một số vòng lặp có các đường đồng mức được trình bày trong Hình 7.

Kết quả thể hiện ở Hình 7 cho thấy rằng sau mỗi vòng lặp, các trọng tâm hội tụ về bốn chùm. Tại vòng lặp thứ 13, chúng ta có bốn trọng tâm một cách rõ ràng. Do đó, ta có bốn chùm. Vì các PDF chung một nhóm ban đầu đều hội tụ về cùng một PDF trọng tâm nên chúng ta cũng có các chùm với các PDF đúng như ban đầu.

Để đánh giá hiệu quả của thuật toán đề nghị so với các thuật toán khác, các chỉ số *ARI* và *G-Mean* được xem xét. Thuật toán đề nghị được so sánh với các thuật toán được công bố trong những năm gần đây như thuật toán của Chen and Hung (2015), Hung et al. (2021) và Le et al. (2023). Mỗi thuật toán được thực hiện 30 lần lặp lại và đánh giá kết quả trên trung vị của các hệ số đánh giá. Cụ thể sự khác biệt của các trung vị được kiểm tra dựa vào phương pháp Kruskal-Wallis với mức ý nghĩa 0,05. Kết quả so sánh của các thuật toán cho tập dữ liệu này được trình bày trong Bảng 1.

Bảng 1 cho thấy mức độ hiệu quả của các thuật toán phân tích chùm được xem xét. Về chỉ số *ARI*, các thuật toán được so sánh đều ở mức rất thấp (0,00 - 0,30), trong khi thuật toán đề nghị đã cho kết quả vượt trội, đạt mức tối ưu ($ARI = 1$). Kết quả kiểm định Kruskal-Wallis chứng minh sự khác biệt trung vị của thuật toán đề nghị với các thuật toán khác với mức ý nghĩa $p < 0{,}05$. Chúng ta cũng có những nhận xét tương tự về sự vượt trội của thuật toán đề nghị so với các thuật toán khác khi căn cứ vào tham số *G-means*.



(a) $t = 1$ (b) $t = 2$ (c) $t = 3$

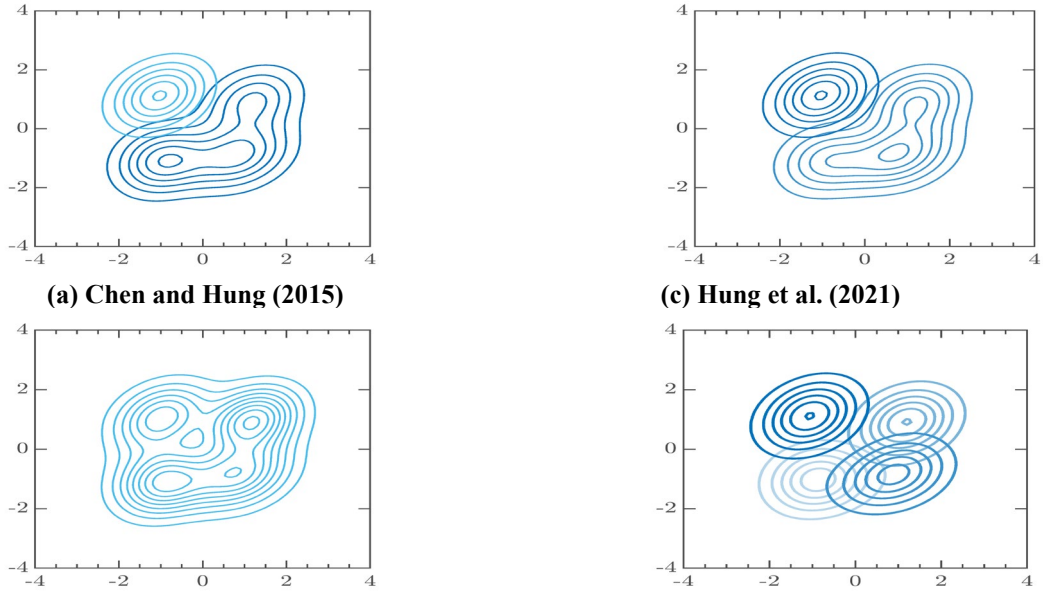
(d) $t = 5$ (e) $t = 10$ (f) $t = 13$

Hình 7. Đường đồng mức của các PDF trọng tâm tại một số vòng lặp

Bảng 1. Kết quả phân tích chùm của các thuật toán tự cập nhật cho 20 PDF

Thuật toán	*ARI*	*G-means*
Chen and Hung (2015)	0,30	0,71
Hung et al. (2021)	0,30	0,71
Le*et al.* (2023)	0,85	0,91
Thuật toán đề nghị	**1,00**	**1,00**
Kruskal-Wallis	$< 0{,}05$	$< 0{,}05$

Đường đồng mức các PDF trọng tâm ở vòng lặp cuối cùng cho các thuật toán được so sánh được vẽ, ta nhận được Hình 8.



(a) Chen and Hung (2015) **(c) Hung et al. (2021)**

(d) Le et al. (2023) **(e) Thuật toán đề nghị**

Hình 8. Trọng tâm ở vòng lặp cuối cùng của các thuật toán

Kết quả được thể hiện ở Hình 8 thể hiện sự khác biệt về mặt trọng tâm của các thuật toán được xem xét và thuật toán đề nghị. Rõ ràng, các thuật toán được so sánh hầu như có xu hướng co lại và cho số lượng chùm thấp. Cụ thể, thuật toán của Chen and Hung (2015) và Hung et al. (2021) cho kết quả hai chùm, trong khi thuật toán của Shiu and Chen (2016) và Le et al. (2023) chỉ cho một chùm duy nhất. Như vậy so với thực tế, các thuật toán được so sánh đã xác định sai số chùm, trong khi thuật toán đề nghị đã thể hiện sự hợp lý trong xác định số chùm với tập dữ liệu này. Hơn nữa, các tham số đánh giá đã cho thấy ưu điểm của nó so với các thuật toán mạnh đã được đề nghị trước đó. Sự khác biệt lớn nhất của thuật toán đề nghị với các thuật toán còn lại là sự thích ứng trong môi trường các phần tử gần nhau, ít có sự khác nhau về mặt khoảng cách và có sự chồng lấp cao.

4.2. Ứng dụng cho dữ liệu ảnh

Phần này áp dụng thuật toán đề nghị để phân tích chùm cho hình ảnh ba loại trái cây. Các ảnh này được trích từ tập dữ liệu có tên *Natural Color Dataset for Colorization Task* và được tải miễn phí từ website: https://www.kaggle.com/datasets/tyrionlannisterlzy/natural-color-dataset-for-colorization-task. Cụ thể, tập dữ liệu được xem xét gồm có 45 ảnh với 15 ảnh quả chuối, 25 ảnh quả lựu và 5 ảnh quả cà tím. Hình 9 minh họa một hình ảnh cho mỗi nhóm.

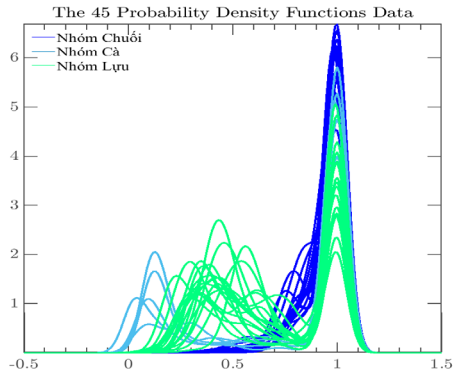
(a) Chuối **(b) Lựu** **(c) Cà tím**

Hình 9. Ba mẫu hình ảnh của tập dữ liệu ảnh được xem xét

Đầu tiên, chúng ta trích xuất đặc trưng của mỗi ảnh và biểu diễn chúng thành PDF như đã đề nghị. Các PDF nhận được đại diện cho 45 hình ảnh được trình bày ở Hình 9.

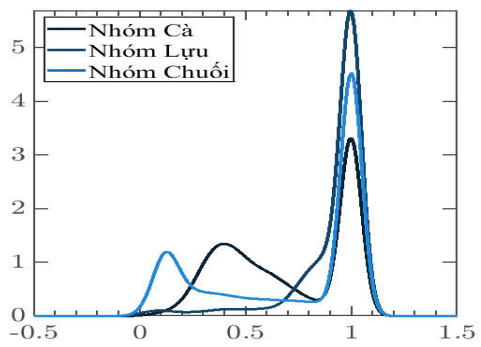
Tiếp tục thực hiện từng bước trong thuật toán đề nghị, sau 30 vòng lặp, thuật toán dừng, khi đó ta có được Hình 10.

Hình 10 cho thấy, 45 PDF ban đầu đã hội tụ về ba PDF trọng tâm. Như vậy, thuật toán đề nghị đã cho kết quả ba chùm. Đây là một kết quả phân tích chùm đúng so với thực tế.



Hình 10. Dữ liệu 45 hàm mật độ xác suất được trích xuất từ dữ liệu hình ảnh

Thuật toán đề nghị cũng được so sánh với các thuật toán khác qua các tham số *ARI* và *G-means* với 30 lần lặp lại. Kết quả so sánh của các thuật toán được trình bày trong Bảng 2.



Hình 11. Các PDF trọng tâm của vòng lặp cuối cùng

Bảng 2. Kết quả phân tích chùm của các thuật toán cho tập dữ liệu ảnh

Thuật toán	*ARI*	*G-means*
Chen & Hung (2015)	0,85	0,93
Hung et al. (2021)	-0,02	0,18
Lethikim et al. (2023)	0,82	0,89
Thuật toán đề nghị	**0,90**	**0,95**
Kruskal-Wallis	$< 0{,}05^{***}$	$< 0{,}05^{***}$

Bảng 2 một lần nữa cho thấy thuật toán đề nghị có hiệu quả cao và vượt trội một cách có ý nghĩa so với các thuật toán được so sánh với tập dữ liệu hình ảnh này.

5. KẾT LUẬN

Thuật toán tự cập nhật chùm cho các PDF được cải tiến trong nghiên cứu. Thuật toán đề nghị tập trung việc xác định tham số thích hợp mà các thuật toán trước đó chưa có sự giải quyết triệt để. Với các chọn tham số động, phụ thuộc vào sự biến động mức độ tương tự của các phần tử thông qua độ lệch chuẩn từng đôi khoảng cách giữa các trọng tâm, thuật toán đề nghị có thể thích ứng với nhiều tập dữ liệu khác nhau. Kết quả kiểm tra thuật toán đề nghị qua dữ liệu mô phỏng và dữ liệu ảnh đã cho thấy sự hợp lý của thuật toán. Kết quả so sánh trên các tập dữ liệu này thông qua các tham số đánh giá cũng cho thấy thuật toán đề nghị có ưu điểm vượt trội so với nhiều thuật toán mạnh được công bố trước đó. Trong thời gian tới, thuật toán đề nghị được kiểm tra trên những tập dữ liệu có tính đặc thù như chuỗi thời gian để áp dụng trong xây dựng mô hình dự báo.

LỜI CẢM TẠ

Công trình này được tài trợ bởi Chương trình học bổng đào tạo thạc sĩ, tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), mã số VINIF.2024.ThS.56.

TÀI LIỆU THAM KHẢO (REFERENCES)

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus Iilustrations*. https://global.oup.com/academic/"Oxford University Press

Chen, T. L., & Shiu, S. Y. (2007). A new clustering algorithm based on self-updating process. *Proceeding about Atatistical 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 & 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 & simulation, 85*(15), 3047–3063. https://doi.org/10.1111/stan.12315

Chen, T. L., & Shiu, S. Y. (2016). A new clustering algorithm based on self-updating process. *In proceedings Statistical Computing, Salt Lake City, Utah, 2034*, 2038. https://api.semanticscholar.org/CorpusID:18819116

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), 1494–1500. https://doi.org/10.1109/FUZZ-IEEE.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 & Computation, 50*(2), 483–500. https://doi.org/10.1080/03610912018.1563149

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

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

Shiu, S. Y., & Chen, T. L. (2016). On the strengths of the self-updating process clustering algorithm. *Journal of Statistical Computation & Simulation, 86*(5), 1010–1031. https://doi.org/10.1080/00949655.2015.1049605

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