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.2026.158

THUẬT TOÁN PHÂN TÍCH CHÙM MỜ TỰ ĐỘNG CHO DỮ LIỆU ĐIỂM DỰA TRÊN SỰ KẾT HỢP CỦA PHƯƠNG PHÁP KHUỶU TAY VÀ MÔ HÌNH HỖN HỢP GAUSS

Nguyễn Thanh Ngọc[1], Thái Văn Thành[1*], Nguyễn Ngọc Nhi[2], Nguyễn Tố Phương[2], Bùi Thị Kim Huệ[1], Lưu Thị Kiều Hương[1] và Thạch Hồng Sơn[3]

[1]*Khoa Công nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật Vĩnh Long, Việt Nam*
[2]*Trường Khoa học Tự nhiên, Đại học Cần Thơ, Việt Nam*
[3]*Trường Đại học FPT Cần Thơ, Việt Nam*
**Tác giả liên hệ (Corresponding author): thanhtv92@vlute.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 04/12/2025
Sửa bài (Revised): 23/12/2025
Duyệt đăng (Accepted): 29/06/2026

Title: *An automatic fuzzy clustering algorithm based on the elbow method and Gaussian mixture model*

Author(s): *Nguyen Thanh Ngoc[1], Thai Van Thanh[1], Nguyen Ngọc Nhi[2], Nguyen To Phuong[2], Bui Thi Kim Hue[1], Luu Thi Kieu Huong[1] and Thach Hong Son[3]*

Affiliation(s): *[1]Faculty of Information Technology, Vinh Long University of Technology Education, Viet Nam; [2]College of Natural Sciences, Can Tho University, Viet Nam; [3]FPT University – Can Tho Campus, Viet Nam*

TÓM TẮT

Trong nghiên cứu này, thuật toán phân tích chùm mờ tự động cho dữ liệu điểm đã được đề xuất dựa trên sự kết hợp của phương pháp khuỷu tay và mô hình hỗn hợp Gauss. Trong thuật toán này, tổng bình phương sai số trong chùm như hàm mục tiêu đã được sử dụng để xác định số chùm thích hợp dựa trên phương pháp khuỷu tay. Sau đó, số chùm nàyđược sử dụng làm đầu vào để xác định xác suất thuộc vào chùm của các phần tử qua mô hình hỗn hợp Gauss. Thông qua việc kết hợp các phương pháp này, một thuật toán phân tích chùm hiệu quả được thiết lập. Từ việc thực nghiệm trên một số tập dữ liệu đối chứng, mô hình đề nghị cho kết quả tốt, vượt qua nhiều mô hình nổi tiếng, được sử dụng phổ biến hiện nay. Mô hình đề nghị có thể được áp dụng thuận lợi cho dữ liệu thực bởi một chương trình Matlab được thiết lập.

Từ khóa: *Dữ liệu điểm, hỗn hợp Gauss, phân tích chùm mờ, phương pháp khuỷu tay*

ABSTRACT

This study proposes an automatic fuzzy clustering algorithm for point data based on the combination of the elbow method and the Gaussian Mixture Model. In this algorithm, the within-cluster sum of squared errors is employed as the objective function to determine the appropriate number of clusters using the elbow method. This number of clusters is then used as input to estimate the cluster membership probabilities of the elements through the Gaussian Mixture Model. By combining these methods, an effective clustering algorithm is established. Experiments on several benchmark datasets show that the proposed model achieves superior performance compared to many well-known and widely used models. Furthermore, the proposed model can be conveniently applied to real-world data through an implemented Matlab program.

Keywords: *Elbow method, fuzzy cluster analysis, Gaussian mixture model, point data*

1. GIỚI THIỆU

Phân tích chùm là việc chia dữ liệu thành các chùm sao cho những phần tử trong cùng một chùm có sự tương tự nhau theo những tiêu chí nào đó nhiều hơn những phần tử bên ngoài của nó (Hung et al., 2015). Với dữ liệu đồ sộ và tốc độ phát triển mạnh mẽ như hiện nay, phân tích chùm được xem là nền tảng không thể thiếu cho các ứng dụng liên quan đến khoa học dữ liệu. Do đó, phân tích chùm đã và đang nhận được sự quan tâm rất lớn từ các nhà thống kê và khoa học dữ liệu (Chen & Shiu, 2007; Shiu & Chen, 2016).

Phân tích chùm có thể thực hiện cho dữ liệu điểm (Vovan et al., 2023), dữ liệu khoảng (Hung et al., 2016) và dữ liệu hàm mật độ xác suất (Chen & Hung, 2015); trong đó, phân tích chùm cho dữ liệu điểm được xem là nền tảng (Vovan et al., 2023) với một tập dữ liệu cho trước, một phân tích chùm được thực hiện, ta phải giải quyết 3 vấn đề quan trọng. Đó là việc xác định số chùm thích hợp, những phần tử cụ thể trong mỗi chùm và xác suất thuộc vào chùm của các phần tử. Một thuật toán chỉ giải quyết được hai nhiệm vụ đầu tiên được gọi là phân tích chùm không mờ. Nếu một thuật toán giải quyết được cả 3 nhiệm vụ thì được gọi là thuật toán phân tích chùm mờ. Trong phân tích chùm không mờ, mỗi phần tử chỉ được gán vào duy nhất một chùm, trong khi phân tích chùm mờ một phần tử có thể được gán vào nhiều chùm với những xác suất cụ thể. Do đó, phân tích chùm mờ được xem là sự mở rộng của phân tích chùm không mờ (Hung & Yang, 2015).

Trong phân tích chùm, việc xác định số lượng chùm là một bước quan trọng và có ảnh hưởng quyết định đến chất lượng của kết quả phân chùm. Nếu số chùm không được lựa chọn hợp lý, mô hình có thể dẫn đến hiện tượng phân chùm quá mức hoặc không đủ chùm, gây ra sai lệch trong quá trình phân tích dữ liệu (Jain, 2010). Trong xác định số chùm, phương pháp tự cập nhật chùm của Chen and Shiu (2007) được sử dụng phổ biến. Trong phương pháp này, sự hội tụ cũng như kết quả phân tích chùm phụ thuộc vào một hằng số được cố định cho toàn bộ dữ liệu. Từ năm 2015, phương pháp tự cập nhật chùm được phát triển mạnh mẽ với nhiều cải tiến liên quan đến bán kính lân cận. Chen and Hung (2015) đã đề xuất cách xác định bán kính lân cận dựa trên giá trị trung bình của các khoảng cách và phương pháp này được áp dụng cho dữ liệu hàm mờ cũng như dữ liệu dạng cầu (Hung & Yang, 2015). Cách tiếp cận này cho phép thích ứng tốt hơn với các tập dữ liệu có đặc tính đa dạng trong thực tế; tuy nhiên, trong nhiều trường hợp vẫn chưa đảm bảo kết quả phân tích chùm tối ưu. Nhằm khắc phục hạn chế này, Shiu and Chen (2016) đã đề xuất mô hình tĩnh và động áp dụng ở mỗi lần lặp, giúp thuật toán thích ứng linh hoạt hơn với dữ liệu và thu được kết quả tốt trên nhiều tập dữ liệu. Tuy vậy, các tham số trong thuật toán của Shiu and Chen (2016) về bản chất vẫn là các hằng số; do đó, người dùng khi áp dụng phải thiết lập trước ba tham số, gồm bán kính lân cận, tham số hội tụ và hàm hạt nhân.

Về việc xác định xác suất thuộc vào chùm, có 3 phương pháp được xem là quan trọng, được sử dụng phổ biến hiện nay. Đó là phương pháp Fuzzy C-means (FCM) (Devi & Devi, 2013), chùm thứ bậc (Kuchaki et al., 2012) và mô hình Gauss hỗn hợp (Xuetong et al., 2024). FCM hoạt động bằng cách lặp lại quá trình tính toán tâm chùm dựa trên trung bình có trọng số và cập nhật ma trận mức độ thành viên cho đến khi hội tụ. FCM linh hoạt hơn so với K-means (Kodinariya & Makwana, 2013) trong xử lý dữ liệu giao thoa, nhưng lại nhạy cảm với nhiễu và cần xác định trước số chùm. Phương pháp FCM ban đầu chỉ thực hiện cho dữ liệu điểm, sau đó được cải tiến cho dữ liệu khoảng và hàm mật độ xác suất. FCM là một hướng mở rộng của phân chùm phân cấp truyền thống, nhằm cho phép mỗi đối tượng có thể tham gia vào nhiều chùm với các mức độ thành khác nhau. Thay vì xây dựng cây phân cấp với ranh giới "cứng", phương pháp này gán xác suất hoặc mức độ mờ cho từng đối tượng tại mỗi mức phân cấp. Phương pháp hỗn hợp Gauss trong phân tích chùm là một kỹ thuật dựa trên mô hình xác suất, trong đó dữ liệu được giả định sinh ra từ sự kết hợp của nhiều phân phối chuẩn khác nhau. Mỗi chùm được mô tả bởi một phân phối chuẩn với các tham số trung bình và ma trận hiệp phương sai. Việc gán một điểm dữ liệu vào chùm nào được xác định bởi xác suất hậu nghiệm thay vì ràng buộc cứng, nghĩa là mỗi điểm có thể thuộc về nhiều chùm với các mức độ khác nhau. Thuật toán thường sử dụng là cực đại hoá kỳ vọng nhằm ước lượng tham số và cập nhật phân bố chùm cho đến khi hội tụ. Tuy nhiên, phương pháp này có nhược điểm lớn vì cần biết trước số chùm, nhạy cảm với giá trị khởi tạo và có thể hội tụ vào cực tiểu cục bộ.

Trong nghiên cứu này, chúng tôi sử dụng phương pháp khuỷu tay để xác định số chùm (Syakur et al., 2018). Phương pháp này dựa trên việc quan sát tổng bình phương sai số trong chùm (Within-Cluster Sum of Squares – WCSS) khi số chùm k tăng dần. Khi biểu diễn WCSS theo k, điểm tại đó WCSS bắt đầu giảm chậm lại và tạo thành một "góc khuỷu tay" được xem là số chùm tối ưu. Mặc dù đơn giản, phương pháp WCSS vẫn là một cách

tiếp cận hiệu quả để khám phá cấu trúc chùm trong dữ liệu (Kodinariya & Makwana, 2013). Kết quả từ phương pháp khuỷu tay được sử dụng làm dữ liệu đầu vào cho mô hình hỗn hợp Gauss (Gaussian Mixture Model - GMM). Cách làm này đã khắc phục được yếu điểm về thông tin số chùm và điểm khởi tạo vốn có của mô hình GMM.

2. ĐÁNH GIÁ KẾT QUẢ PHÂN TÍCH CHÙM

2.1. Trong phân tích chùm không mờ

Trong phân tích chùm không mờ, để so sánh và đánh giá hiệu quả của các thuật toán, ta có thể dựa vào nhiều tham số thống kê. Trong nghiên cứu này, tham số được sử dụng là ARI (Adjusted Rand Index) (Hubert & Arabie, 1985) được sử dụng. Tham số này được xác định như sau:

$$\mathrm{AR}I = \frac{\sum_{jk}\binom{n_{jk}}{2} - \frac{\sum_{j=1}^{n}\binom{a_j}{2}\cdot\sum_{k=1}^{n}\binom{b_k}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_{j=1}^{n}\binom{a_j}{2} + \sum_{k=1}^{n}\binom{b_k}{2}\right] - \frac{\sum_{j=1}^{n}\binom{a_j}{2}\cdot\sum_{k=1}^{n}\binom{b_k}{2}}{\binom{n}{2}}} \quad (1)$$

Trong đó:

n: tổng số phần tử trong tập dữ liệu,

k=1,...,c: số chùm do thuật toán dự đoán,

n_{jk}: số điểm dữ liệu nằm đồng thời trong chùm thật j và chùm dự đoán k,

$a_j = \sum_k n_{jk}$: tổng số phần tử trong chùm thực tế thứ j,

$b_k = \sum_j n_{jk}$: tổng số phần tử trong chùm thứ k của thuật toán,

$\binom{n}{2}$: số tổ hợp chập hai của n, biểu thị tổng số cặp mẫu.

Tham số *ARI* nhận giá trị trong khoảng [−1,1]. Khi *ARI* càng lớn, chất lượng phân chùm càng tốt. Giá trị *ARI* bằng 1 nếu kết quả phân chùm hoàn toàn trùng khớp với thực tế và bằng -1 nếu hoàn toàn trái ngược. Giá trị *ARI* bằng 0 nếu việc phân chùm tương đương với phân bổ ngẫu nhiên các đối tượng vào chùm.

2.2. Trong phân tích chùm mờ

Cho n phần tử được chia thành k chùm. Gọi μ_{ij} là xác suất để gán phần tử thứ j vào chùm thứ i, với j = 1, 2, …, n và i = 1, 2, …, k. Trong phân tích chùm không mờ, μ_{ij}=1 khi phần tử thứ i thuộc vào chùm thứ j và $\mu_{ij} = 0$ đối với trường hợp ngược lại. Trong phân tích chùm mờ thì $0 < \mu_{ij} < 1$. Khi đó, $U = [\mu_{ij}]_{kxn}$ được gọi là ma trận phân vùng của n phần tử.

Trong nghiên cứu này, để đánh giá hiệu quả của các thuật toán phân tích chùm mờ, hai tham số sau được xem xét:

- Hệ số phân chia (Partition Coefficient):

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}^2$$

- Năng lượng phân chia (Partition Entropy):

$$PC = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}\log\left(\mu_{ij}\right)$$

Tham số *PC* và *PE* đo lường mức độ hỗn loạn hoặc tính không chắc chắn trong thuật toán phân tích chum mờ. Tham số *PC* có giá trị trong khoảng $[\frac{1}{k}, 1]$, trong khi *PE* có giá trị [0;1]. Trong phân tích chùm mờ, tham số *PC* càng lớn thì thuật toán càng tốt, trong khi tham số *PE* thì càng nhỏ càng tốt.

3. THUẬT TOÁN ĐỀ NGHỊ

Cho tập dữ liệu $X = \{x_1, x_2, \ldots, x_n\}$, thuật toán phân tích chùm mờ tự động cho X đề nghị gồm hai giai đoạn sau:

- **Giai đoạn 1**. ***Xác định số chùm thích hợp***

***Bước* 1**. Chọn số chùm k, $(2 \le k \le n-1)$ một cách ngẫu nhiên. Áp dụng thuật toán k-means để nhận được k chùm $\{C_1, C_2, \ldots, C_k\}$ và tâm chùm m_k.

***Bước* 2.** Tính tổng bình phương khoảng cách nội bộ các chùm theo công thức sau:

$$SSE = \sum_{i=1}^{k}\sum_{x_i \in C_k}\left[d\left(x_i, m_k\right)\right]^2,$$

trong đó $d\left(x_i, m_k\right)$ là khoảng cách Euclide giữa điểm dữ liệu x_i và trọng tâm chùm m_k.

Khi số chùm tăng lên, SSE giảm, nhưng đến một điểm nhất định, việc gia tăng số chùm không mang lại sự cải thiện đáng kể trong sự phân tán của các điểm trong chùm. Điểm khuỷu tay là vị trí SSE giảm chậm lại, cho thấy k đã đạt tối ưu

Bước 3. Cho k thay đổi, đồ thị *SSE* được vẽ theo k (trục hoành là k và trục tung là SSE), tìm điểm khuỷu tay cho SSE. Số chùm thích hợp chính là giá trị k ứng với điểm khuỷu tay của SSE. Gọi m là số chùm thích hợp được tìm sau khi kết thúcbước 3.

- **Giai đoạn 2: *Phân tích chùm mờ***

Bước 4. Khởi tạo ngẫu nhiên trọng tâm μ_k, ma trận hiệp phương sai S_k và trọng số π_k cho các chùm.

Bước 5. Tính xác suất điểm dữ liệu x_i thuộc về chùm theo công thức sau:

$$\gamma_{ik} = \frac{\pi_k.\mathrm{N}\left(x_i|\mu_k,\Sigma_k\right)}{\sum_{h=1}^{m}\pi_h.\mathrm{N}\left(x_i|\mu_h,\Sigma_h\right)} \quad (2),$$

trong đó, hàm mật độ chuẩn nhiều chiều có dạng:

$$\mathrm{N}\left(x_i|\mu_k,\Sigma_k\right)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_i-\mu_k)^T\sum\nolimits_k^{-1}(x_i-\mu_k)\right) \quad (3)$$

Bước 6. Cập nhật trọng tâm, ma trận hiệp phương sai và trọng số theo nguyên tắc sau:

- Trọng số: $\pi_k = \frac{1}{n}\sum_{i=1}^{n}\gamma_{ik}$. (4)
- Trọng tâm chùm: $\mu_k = \frac{\sum_{i=1}^{n}\gamma_{ik}x_i}{\sum_{i=1}^{n}\gamma_{ik}}$. (5)
- Ma trận hiệp phương sai:

$$\Sigma_k = \frac{\sum_{i=1}^{n}\gamma_{ik}(x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^{n}\gamma_{ik}}. \quad (6)$$

Bước 7. Tính log-likelihood tổng quát theo công thức:

$$\log L = \sum_{i=1}^{n}\log\left(\sum_{k=1}^{m}\pi_k.\mathrm{N}\left(x_i|\mu_k,\Sigma_k\right)\right) \quad (7)$$

Nếu log-likelihood thay đổi dưới ngưỡng ε thì dừng thuật toán. Khi giai đoạn 2 dừng, chúng ta nhận được các trọng tâm chùm μ_k và xác suất thuộc vào chùm của mỗi phần tử γ_{ik}, $i = 1, 2, \dots, n$; $k = 1, 2, \dots, m$.

4. ÁP DỤNG

Trong phần này, hai áp dụng được thực hiện để minh họa các bước của thuật toán đề nghị cũng như thể hiện hiệu quả của nó.

4.1. Áp dụng 1

Trong áp dụng này, tập dữ liệu mô phỏng đã được sử dụng gồm 200 phần tử, được sinh từ bốn phân phối chuẩn hai chiều với các tham số đã được thiết lập trước. Cụ thể 200 điểm dữ liệu hai chiều được tạo ra từ bốn phân phối Gauss với vector trung bình và ma trận hiệp phương sai được cho như sau:

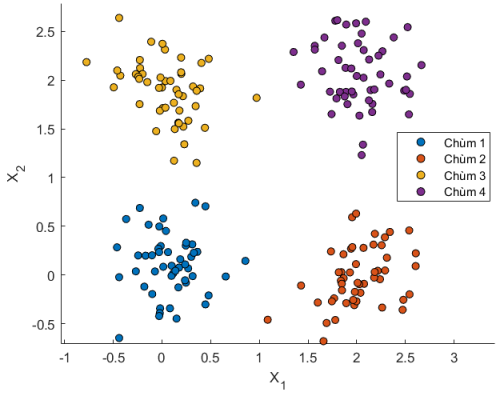
Chùm 1: $\mu_1 = \binom{0,1}{0,1}, \Sigma_1\begin{pmatrix}0,1 & 0\\ 0 & 0,1\end{pmatrix}$.

Chùm 2: $\mu_2 = \binom{2}{0}, \Sigma_2\begin{pmatrix}0,1 & 0,05\\ 0,05 & 0,1\end{pmatrix}$.

Chùm 3: $\mu_3 = \binom{0}{2}, \Sigma_3\begin{pmatrix}0,1 & -0,05\\ -0,05 & 0,1\end{pmatrix}$.

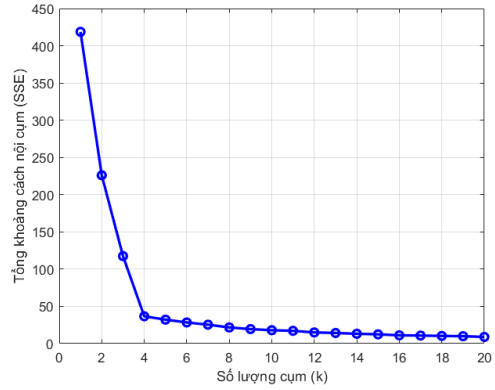
Chùm 4: $\mu_4 = \binom{0}{2}, \Sigma_4\begin{pmatrix}0,1 & -0,01\\ -0,01 & 0,1\end{pmatrix}$.

Mỗi chùm có 50 phần tử được minh họa trong Hình 1.



Hình 1. Đồ thị phân tán 200 phần tử của 4 chùm

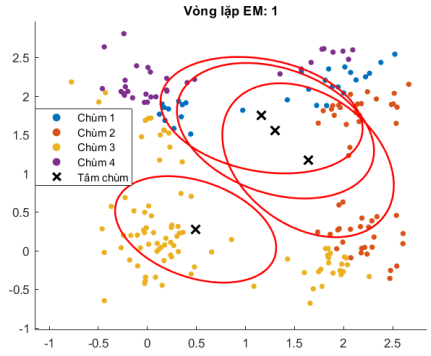
Kết quả minh họa từ việc áp dụng giai đoạn 1 được trình bày trong Hình 2.



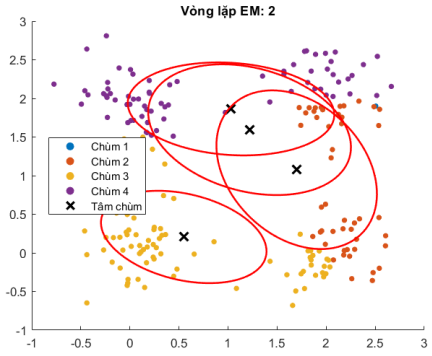
Hình 2. Giá trị SSE tương ứng với số chùm của 200 phần tử

Dựa trên Hình 2, có thể quan sát thấy một điểm gấp khúc rõ rệt tại $k = 4$, được xem là "khuỷu tay" trong biểu đồ. Điều này cho thấy việc phân chùm dữ liệu thành 4 nhóm là phù hợp.

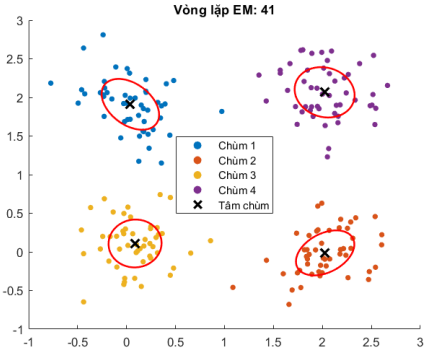
Sau khi thực hiện giai đoạn 2, thuật toán dừng sau 42 vòng lặp và được minh họa bởi Hình 3.



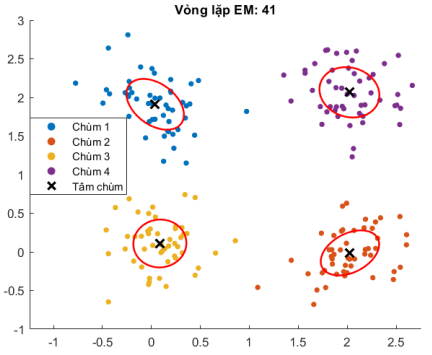
a) Vòng lặp 1



b) Vòng lặp 2



a) Vòng lặp 41



b) Vòng lặp 42

Hình 3. Minh họa một số vòng lặp của Giai đoạn 2

Sau khi giai đoạn 2 hội tụ, thuật toán cho kết quả các trọng tâm chùm như sau:

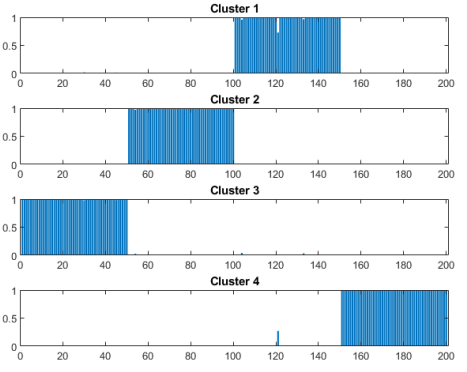
$$\mu_1 = (0{,}0386; 1{,}9103),$$

$$\mu_2 = (2{,}0275; -0{,}0130),$$

$$\mu_3 = (0{,}0865; 0{,}1088),$$

$$\mu_4 = (2{,}0223; 2{,}0658)$$

và ma trận phân chia được thể hiện trong Hình 4.



Hình 4. Xác suất thuộc vào 4 chùm của 200 phần tử

Để đánh giá hiệu quả của thuật toán đề nghị, nó được so sánh với các thuật toán được sử dụng phổ

biến ngày nay bao gồm thuật toán mờ và không mờ. Cụ thể:

- Thuật toán không mờ: Hierarchical Clustering (HC) (Johnson, 1967), k-Medoids (Kaufman & Rousseeuw, 1990) và k-means. Đối với thuật toán không mờ, tham số được so sánh là ARI.
- Thuật toán mờ: FCM (Bezdek, 1981) và phân chùm mờ tự động của Chen (2016).

Kết quả của các thuật toán được trình bày trong Bảng 1.

Bảng 1. So sánh các tham số của thuật toán được đề nghị với các thuật toán khác

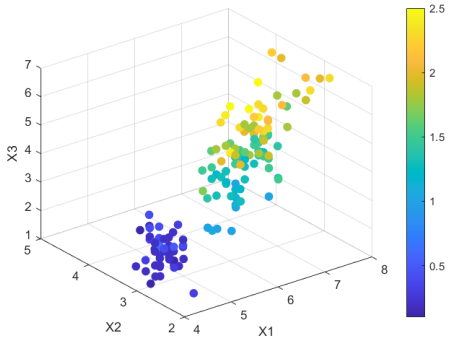
Phương pháp	*PC*	*PE*	*ARI*
k-means	-	-	1,0
HC	-	-	0,9732
k-medoids	-	-	1,0
Chen (2016)	0,9615	0,1007	1,0
FCM	0,8245	0,3846	0,9715
Đề nghị	**0,9966**	**0,0063**	**1,0**

Bảng 1 cho thấy thuật toán đề nghị tốt hơn hai thuật toán phân tích chùm mờ FCM và Chen (2016) vì giá trị *PC* lớn hơn và *PE* nhỏ hơn có ý nghĩa. Thuật toán đề nghị có giá trị tối ưu *ARI* = 1, trong khi thuật toán HC có giá trị là 0,9732. Tuy nhiên hai thuật toán phân tích chùm không mờ k-means và k-medoids cũng có giá trị bằng 1. Vì dữ liệu mô phỏng này có sự tách biệt tương đối rõ ràng của 4 chùm nên chúng ta chưa nhận thấy được ưu điểm của thuật toán đề nghị so với hai thuật toán k-means và k-medoids. Hiệu quả của thuật toán đề nghị được minh chứng trong những tập dữ liệu phức tạp hơn, cụ thể ở áp dụng 2.

4.2. Áp dụng 2

Áp dụng này được thực hiện cho tập dữ liệu Hoa Iris (https://archive.ics.uci.edu/dataset/53/iris), một tập dữ liệu đối chứng nổi tiếng trong lĩnh vực học máy để kiểm tra hiệu quả của bài toán phân tích chùm và phân loại.

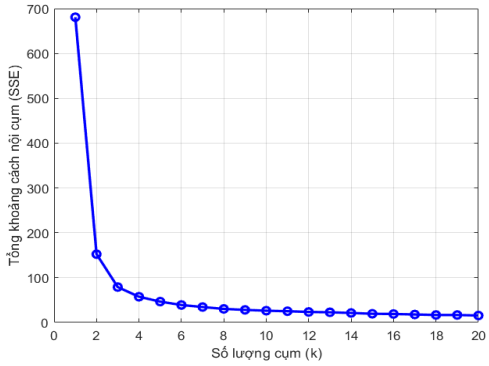
Hoa Iris gồm có 3 loại bao gồm *Setosa, Versicolor và Virginica* được phân biệt với nhau bởi 4 biến x_1= chiều dài hoa, x_2 = chiều rộng hoa (sepal width), x_3 = chiều dài cánh hoa (petal length) và x_4 = chiều rộng cánh hoa (petal width). Minh họa cho 3 loại hoa Iris được trình bày trong Hình 5.



Hình 5. Minh họa 3 loại hoa Iris

Tập dữ liệu về hoa Iris lần đầu tiên được trình bày bởi Fisher (1936) gồm 150 phần tử với 50 phần tử cho mỗi loại. Trong áp dụng này, kết quả giai đoạn 1 và giai đoạn 2 của thuật toán đề nghị cũng được trình bày. Bên cạnh đó, việc so sánh thuật toán đề nghị với các thuật toán khác như trong áp dụng 1 cũng đã được thực hiện.

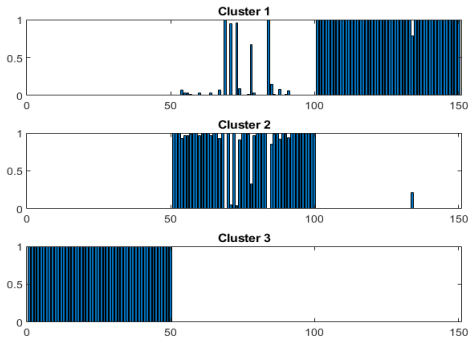
Kết quả giai đoạn 1 của thuật toán đề nghị được trình bày trong Hình 6.



Hình 6. Giá trị SSE tương ứng với số chùm của 150 phần tử

Dựa trên Hình 6, phương pháp khuỷu tay cho thấy điểm gãy rõ rệt tại $k = 3$, thể hiện sự giảm đáng kể của SSE so với các giá trị k khác. Do đó, giai đoạn 1 kết thúc với số chùm thích hợp là 3. Đây là một kết quả chia chùm đúng của tập dữ liệu.

Việc áp dụng giai đoạn 2 cho 150 phần tử, thuật toán dừng lại sau 38 vòng lặp. Khi đó, ta nhận được ma trận phân chia với xác suất thuộc vào 3 chùm của 150 phần tử được biểu diễn ở Hình 7.



Hình 7. Xác suất thuộc vào 3 chùm của 150 phần tử

Hình 7 cho thấy xác suất phân đúng vào chùm 1 (*Setosa*) và chùm 3 (*Virginica*) của các phần tử hầu như bằng 1. Có sự nhầm lẫn trong phân chùm cho một số ít phần tử của hoa *Versicolor* khi xếp vào chùm 1.

Từ việc so sánh thuật toán đề nghị với các thuật toán khác qua các tham số *PC*, *PE* và *ARI* ứng với các thuật toán phân tích chùm mờ và không mờ, ta có Bảng 2.

Bảng 2. Tham số đánh giá của các thuật toán trong áp dụng 2

Phương pháp	*PC*	*PE*	*ARI*
K-Means	-	-	0,7302
HC	-	-	0,7312
k-medoids	-	-	0,7302
C-Means	0,7832	0,3959	0,7294
Chen (2016)	0,9103	0.1237	0,8798
Đề nghị	**0,9833**	**0,0324**	**0,9039**

Bảng 2 cho thấy giá trị *ARI* của thuật toán đề nghị nhỏ nhất và nhỏ hơn rất nhiều so với các thuật toán được so sánh. Khi so sánh với 2 thuật toán mờ, thuật toán đề nghị cũng nhận được giá trị *PC* lớn nhất và giá trị *PE* nhỏ nhất. Sự khác biệt của chúng được xem là có ý nghĩa.

Tập dữ liệu trong áp dụng 2 được đánh giá phức tạp hơn trong áp dụng 1. Với tập dữ liệu này, thuật toán đề nghị đã thể hiện rõ ưu điểm so với 2 thuật toán phân tích chùm mờ cũng như 3 thuật toán phân tích chùm không mờ được so sánh.

5. KẾT LUẬN

Trong nghiên cứu này, một thuật toán phân tích chùm mờ tự động cho dữ liệu điểm đã được đề xuất, kết hợp giữa phương pháp khuỷu tay và GMM. Thuật toán đề nghị không chỉ giúp tự động xác định số chùm phù hợp, mà còn khai thác được đặc điểm phân phối xác suất của dữ liệu nhằm tăng độ chính xác và linh hoạt trong việc gán phần tử vào các chùm. Các kết quả thực nghiệm trên hai tập dữ liệu cho thấy phương pháp đề xuất có hiệu quả cao, thuận lợi hơn so với các thuật toán phân chùm nổi tiếng được sử dụng phổ biến. Thuật toán đề nghị có thể áp dụng thuận lợi cho dữ liệu thực bởi một chương trình được thiết lập trên phần mềm Matlab. Trong tương lai, thuật toán đề nghị có thể được mở rộng tích hợp với trí tuệ nhân tạo và các mô hình học sâu để áp dụng được cho dữ liệu ảnh một cách hiệu quả.

Mặc dù phương pháp khuỷu tay được sử dụng hiệu quả để xác định số chùm trong nghiên cứu này; cần lưu ý rằng, phương pháp vẫn tồn tại một số hạn chế, bao gồm tính chủ quan trong việc xác định điểm gãy và khả năng kém hiệu quả đối với các tập dữ liệu có cấu trúc phức tạp hoặc không thể hiện rõ đặc trưng khuỷu tay. Do đó, trong các nghiên cứu tiếp theo, có thể xem xét việc kết hợp phương pháp khuỷu tay với các tiêu chí đánh giá khác nhằm nâng cao tính khách quan và độ ổn định trong xác định số chùm.

TÀI LIỆU THAM KHẢO

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Springer. https://doi.org/10.1007/978-1-4757-0450-1

Chen, T. L., & Shiu, S. Y. (2007). An efficient automatic clustering algorithm based on a self-updating process. *Intelligent Data Analysis, 11*(5), 473–490

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/10.1080/00949655.2014.949715

Chen, T. L., & Shiu, S. Y. (2016). A new clustering algorithm based on self-updating process. In JSM Proceedings, Section on Statistical Computing (pp. 2034–2038). American Statistical Association.Devi, S. S., & Devi, M. R. (2013). Partition coefficient and partition entropy in fuzzy C-means clustering. *International Journal of Computer Trends and Technology*, *4*(10), 1554–1558.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density

estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. https://doi.org/10.1198/016214502760047131

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

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/FUZZ-IEEE.2016.7737867

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254. https://doi.org/10.1007/BF02289588

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York, NY: John Wiley & Sons. https://doi.org/10.1002/9780470316801

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of clusters in k-means clustering. *International Journal of Advanced Research in Computer Science and Management Studies*, *1*(6), 90–95.

Kuchaki, M. R., Asghari, Z. V., & Emami, C. (2012). A survey of hierarchical clustering algorithms. *The Journal of Mathematics and Computer Science*, *5*(3), 229–240. https://doi.org/10.22436/jmcs.05.03.11

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

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and Elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, 336, 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Vovan, T., Nguyenhoang, Y., & Danh, S. (2023). *An automatic fuzzy clustering algorithm for discrete elements*. *Journal of the Operations Research Society of China*, *11*(2), 309–325. https://doi.org/10.1007/s40305-021-00388-z

Xuetong, L., Jing, Z., & Hansheng, W. (2024). *Gaussian mixture models with rare events*. *Journal of Machine Learning Research*, 25, 1–40. https://doi.org/10.5555/3666122