Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jvn.2020.027

CẢI TIẾN THUẬT TOÁN PHÂN TÍCH CHÙM CHO CÁC PHẦN TỬ RỜI RẠC

Võ Văn Tài[1*], Trần Thành Tiến[1], Châu Ngọc Thơ[1], Nguyễn Trang Thảo[2] và Huỳnh Văn Hiếu[3]

[1]*Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*

[2]*Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh*

[3]*Khoa Cơ bản, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh*

[*]*Người chịu trách nhiệm về bài viết: Võ Văn Tài (email: vvtai@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 14/11/2019
Ngày nhận bài sửa: dd/mm/yyyy
Ngày duyệt đăng: 29/04/2020

Title:
Improving the cluster analysis algorithm for discrete elements

Từ khóa:
Chỉ số tương tự, khoảng cách, phân tích chum, thuật toán

Keywords:
Algorithm, cluster analysis, distance, similar index

ABSTRACT

This study proposes a new concept to evaluate the similarity of cluster for discrete elements called the Cluster Similar Index (CSI). CSI is also used as a criterion to establish the algorithms in building the fuzzy and non-fuzzy clusters, and the algorithm used to determine the suitable number of clusters. The established algorithms can be quickly performed by the Matlab procedures. The numerical examples illustrate the proposed algorithms and show its benefits overcome the others. Analyzing the cluster of images from the proposed algorithm shows potential in the practical application of this research.

TÓM TẮT

Nghiên cứu này đề nghị một khái niệm mới để đánh giá mức độ gần nhau của các phần tử rời rạc gọi là chỉ số tương tự chùm (CSI). CSI được sử dụng như một tiêu chuẩn để xây dựng các thuật toán phân tích chùm mờ, không mờ và xác định số chùm thích hợp. Các thuật toán được thiết lập có thể thực hiện nhanh chóng bởi những chương trình được thiết lập trên phần mềm Matlab. Những ví dụ số minh họa các thuật toán đề nghị và cho thấy thuận lợi của nó so với một số thuật toán khác. Phân tích chùm các hình ảnh từ thuật toán đề nghị cho thấy tiềm năng trong áp dụng thực tế của vấn đề được nghiên cứu.

Trích dẫn: Võ Văn Tài, Trần Thành Tiến, Châu Ngọc Thơ, Nguyễn Trang Thảo và Huỳnh Văn Hiếu, 2020. Cải tiến thuật toán phân tích chùm cho các phần tử rời rạc. Tạp chí Khoa học Trường Đại học Cần Thơ. 56(2A): 30-36.

1 GIỚI THIỆU

Trong thời đại ngày nay, việc phân loại, lưu trữ và trích xuất dữ liệu đóng một vai trò rất quan trọng, ảnh hưởng đến sự phát triển của nhiều lĩnh vực, nhiều ngành khoa học khác nhau. Trong vấn đề này, bài toán phân tích chùm đóng vai trò nền tảng bởi vì kết quả của nó là việc chia dữ liệu thành những chùm sao cho những phần tử trong cùng một chùm có sự tương tự theo một tiêu chuẩn nào đó nhiều hơn so với những phần tử của chùm khác. Chính vì lý do này bài toán phân tích chùm đã được quan tâm bởi nhiều nhà nghiên cứu (Ester *et al*., 1973; Li and Wang 2008; Tai and Pha-Gia 2010; Wen and Yenn 2015). Chùm có thể được xây dựng cho các phần tử rời rạc (CDE) và cho các hàm mật độ xác suất (CDF). CDE đã được đề xuất trước và có những ưu điểm nhất định so với CDF. CDE có tính trực quan hơn và tốc độ tính toán trong các thuật toán của nó

cũng thường nhanh hơn so với CDF. Trong nhiều trường hợp của áp dụng thực tế, CDE cũng có sai lầm nhỏ hơn CDF. Thông thường có 3 lý do chính cho vấn đề này. Đó là tiêu chuẩn để đánh giá mức độ gần và xa của các phân tử rời rạc thường được minh họa trực quan rõ ràng, trong khi các hàm mật độ thì ngược lại. Dữ liệu thực tế thường là rời rạc, do đó để áp dụng CDF, bước đầu chúng ta phải ước lượng các hàm mật độ này. Mặc dù có nhiều tiến bộ cho vấn đề này trong những năm gần đây, tuy nhiên tính chính xác của việc thực hiện cho đến nay vẫn là bài toán chưa có lời giải cuối cùng (Tai and Thao 2017). Lý do cuối cùng là các độ đo cho những phần tử rời rạc thường được tính nhanh hơn nhiều so với các hàm mật độ, đặc biệt trong các phần mềm hiện nay. Các tiêu chuẩn để thực hiện CDF thường cũng không được tính chính xác trong các áp dụng thực tế mà phải tính gần đúng.

Trong CDE, có ba vấn đề quan trọng mà các nhà nghiên cứu đã quan tâm và cải tiến: (i) tìm một tiêu chuẩn thích hợp để đánh giá sự tương tự của hai và nhiều hơn hai phần tử, (ii) xây dựng các thuật toán phân tích chùm hiệu quả với sai lầm nhỏ nhất, và (iii) đánh giá chất lượng của các chùm đã xây dựng. Với (i), hầu hết các nghiên cứu đã sử dụng cho đến hiện tại là khoảng cách. Khoảng cách của hai phần tử rời rạc được sử dụng phổ biến là khoảng cách Euclide, khoảng các city-block, khoảng cách L^p, trong khi khoảng cách giữa hai tập hợp là khoảng cách max, khoảng cách min và khoảng cách trung bình. Mặc dù có nhiều thảo luận về việc chọn khoảng cách trong bài toán phân tích chùm, nhưng tới nay vẫn chưa có kết luận cuối cùng về việc chọn khoảng cách tối ưu (Ganti *et al.*, 1999; Xie and Beni, 1991). Với (ii), có hai phương pháp chính được áp dụng phổ biến: thứ bậc và không thứ bậc (Tai and Pham-Gia, 2010). Những phương pháp này cũng sử dụng tiêu chuẩn khoảng cách như đã chú ý ở trên để thực hiện. Thực tế ứng dụng cho thấy những phương pháp này có hiệu quả khi dữ liệu có sự phân nhóm tương đối rõ ràng. Khi dữ liệu không có nhiều sự tách rời, các phương pháp này thường dẫn đến những sai lầm lớn. Với (iii), có nhiều chỉ số đề nghị để đánh giá chất lượng chùm CDE được xây dựng như chỉ số S, chỉ số Dunn, chỉ số Xie-Beni (Dunn, 1973; Xie and Beni, 1991; Pal and Bezdek, 1995). Tuy nhiên, các chỉ số này chỉ được tính sau khi các chùm đã được thiết lập, vì vậy phải tốn thêm thời gian để đánh giá. Hơn nữa, các thuật toán này chỉ đánh giá một cách tổng quát chất lượng quá trình xây dựng chùm mà không đánh giá chất lượng của mỗi chùm. Tai and Thao (2017) đã đề nghị một tiêu chuẩn gọi là hệ số tương tự chùm để đánh giá chất lượng các chùm được thiết lập và xây dựng chùm, tuy nhiên độ đo mới chỉ thực hiện cho các hàm mật độ xác suất, không phải cho các phần tử rời rạc. Tiêu chuẩn hệ số tương tự chùm cũng được phát triển cho các phần tử rời rạc bởi Tai and Thao (2018) dựa vào khoảng cách cực đại. Tuy nhiên, nó cũng bộc lộ những hạn chế khi thực hiện với dữ liệu phức tạp.

Để khắc phục những hạn chế của các phương pháp như đã đề cập ở trên, dựa trên sự chuẩn hóa các biến về [0;1] của dữ liệu, khoảng cách của hai phần tử và hai tập hợp, chúng tôi đề nghị một tiêu chuẩn gọi là chỉ số tương tự chùm (CSI) mà nó được sử dụng như một tiêu chuẩn để phân tích chùm. Dựa trên CSI, nghiên cứu này đề xuất các thuật toán xây dựng chùm mờ, không mờ và xác định số chùm thích hợp. Các thuật toán đề nghị đã được thực hiện nhanh chóng và hiệu quả bởi những thủ tục Matlab được thiết lập. Những ví dụ số không những minh họa cho các thuật toán đã đề nghị mà còn cho thấy tính hiệu quả khi so sánh với các thuật toán đã tồn tại. Ứng dụng các thuật toán đề nghị trong nhận dạng ảnh cho thấy tiềm năng trong thực tế của vấn đề được nghiên cứu.

2 CHỈ SỐ TƯƠNG TỰ CHÙM

2.1 Chuẩn hóa dữ liệu

Để có tính hợp lý trong đánh giá mức độ gần nhau của các phần tử trong không gian nhiều chiều với thang đo khác nhau, chúng tôi đầu tiên đưa mỗi biến về thang đo [0;1]. Cụ thể như sau:

Trong không gian n chiều với các biến $x_1, x_2, \ldots, x_n$, cho một chùm có N phần tử $Z = \{z_1, z_2, \ldots, z_N\}$. Gọi $\{x_i^1, x_i^2, \ldots, x_i^N\}$, $i = 1, 2, \ldots, n$ là tập các giá trị của biến $x_i > 0$ trong tập dữ liệu Z. Đặt

$$d_i = \max\{x_i^j\}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, N.$$

$$z_{i*}^j = \frac{x_i^j}{d_i}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, N.$$

$z_j^* = (z_1^{*j}, z_2^{*j}, \ldots, z_n^{*j})$, j = 1, 2, ..., N là dữ liệu thứ j.

Ta có $z_i^{*j} \in [0;1]$, do đó các tọa độ của z_j^* luôn nằm trong [0;1], khi đó từ tập dữ liệu Z ban đầu chúng ta có tập dữ liệu $Z^* = \{z_1^*, z_2^*, \ldots, z_N^*\}$ mà mỗi phần tử của nó đều có tọa độ trên đoạn [0;1].

2.2 Chỉ số tương tự chùm

Cho một chùm gồm *N* phần tử trong không gian n chiều $Z = \{z_1, z_2, ..., z_N\}$, thực hiện chuẩn hóa dữ liệu để có tập dữ liệu Z^* như ở trên. Từ tập dữ liệu Z^* chúng ta định nghĩa hệ số tương tự của chùm như sau:

$$c\left(Z^*\right) = 1 - \frac{1}{n.C_N^2} \Sigma_{i<j}\, d^2\left(z_i^*, z_j^*\right), \quad (1)$$

trong đó, $d\left(z_i^*, z_j^*\right)$ là khoảng cách Euclide giữa hai phần tử z_i^* và z_j^*. Trong trường hợp $N = 2$, công thức (1) trở thành

$$c\left(z_i^*, z_j^*\right) = 1 - \frac{1}{n} d^2\left(z_i^*, z_j^*\right). \quad (2)$$

Chúng ta có thể thấy rằng $\frac{1}{C_N^2} \Sigma_{i<j}\, d^2\left(z_i^*, z_j^*\right)$ là trung bình các khoảng cách của tất cả các phần tử của chùm *Z* khi dữ liệu đã được chuẩn hóa về Z^* và $0 \le \frac{1}{C_N^2} \Sigma_{i<j}\, d^2\left(z_i^*, z_j^*\right) \le n.$

Đặt $d_S = \frac{1}{nC_N^2} \Sigma_{i<j}\, d^2\left(z_i^*, z_j^*\right)$, ta có $0 \le d_S \le 1.$

Khi đó, ta cũng nhận được $0 \le c(Z^*) \le 1.$

d_S là trung bình của bình phương các khoảng cách của những phần tử được chuẩn hóa [0;1]. Khi d_S càng nhỏ thì sự tương tự của các phần tử trong chùm càng lớn và ngược lại. Giá trị của $c\left(Z^*\right)$ thì ngược lại đối với d_S, do đó nếu $c\left(Z^*\right)$ càng lớn thì chùm được xây dựng sẽ càng tốt.

3 CÁC THUẬT TOÁN ĐỀ NGHỊ DỰA VÀO CSI

3.1 Phân tích chùm không mờ

Bài toán 1: Cho một tập hợp gồm *N* phần tử $N^{(0)} = \{z_1^{(0)}, z_2^{(0)}, ..., z_N^{(0)}\}$, chúng ta cần chia những phần tử này *k* chùm sao SCI của mỗi chùm là tối ưu.

Thuật toán 1: Thuật toán giải quyết Bài toán 1 gồm các bước sau:

Bước 1. Chuẩn hóa dữ liệu đã cho ban đầu như mục 2.1 để đưa tập dữ liệu ban đầu $N^{(0)}$ về $N^{*(0)} = \{z_1^{*(0)}, z_2^{*(0)}, ..., z_N^{*(0)}\}$.

Bước 2. Chia *N* phần tử vào *k* chùm một cách ngẫu nhiên.

Bước 3. Tính CSI của mỗi phần tử với chùm chứa nó. Nếu CSI này lớn hơn CSI của chùm khi kết hợp phần tử này với chùm khác thì ta giữ phần tử này trong chùm ban đầu. Ngược lại, ta di chuyển phần tử tới chùm có CSI lớn hơn.

Bước 4. Lập lại Bước 3 cho đến khi CSI của mỗi phần tử với chùm chứa nó là lớn nhất.

3.2 Phân tích chùm mờ

Bài toán 2: Cho một tập hợp gồm *N* phần tử $Z = \{z_1, z_2, ..., z_N\}$. Chúng ta chia *N* phần tử này thành *c* chùm sao cho mỗi phần tử đều xác định được xác suất gán vào mỗi chùm.

Thuật toán 2: Bài toán này được giải quyết với các thuật toán gồm các bước sau:

Bước 1. Chuẩn hóa dữ liệu đã cho ban đầu như mục 2.1 để đưa tập dữ liệu ban đầu $N^{(0)}$ về $N^{*(0)} = \{z_1^{*(0)}, z_2^{*(0)}, ..., z_N^{*(0)}\}$.

Bước 2. Bắt đầu ma trận phân chia $U^{(0)}$ với xác suất gán vào các chùm một cách ngẫu nhiên. Tìm phần tử đại diện cho các chùm bởi công thức:

$$v_i = \frac{\sum_{k=1}^{N} \left(\mu_{ik}\right)^2 z_k^*}{\sum_{k=1}^{N} \left(\mu_{ik}\right)^2}, i = 1, 2, ..., c. \quad (3)$$

Bước 3. Tính CSI giữa mỗi phần tử và v_i. Cập nhật ma trận phân chia mới $U^{(1)}$ công thức:

$$\mu_{ik} = \frac{c\left(v_i, z_k^*\right)^2}{\sum_{j=1}^{c} c\left(v_i, z_k^*\right)^2}, \; 1 \le i, j \le c, \; 1 \le k \le N. \quad (4)$$

Bước 4. Lập lại Bước 2 và Bước 3 đến khi $\left\|U^{(1)} - U^{(0)}\right\| < \varepsilon$.

Trong thuật toán trên ε là một số rất nhỏ nào đó. Khi ε càng nhỏ thì số vòng lặp càng lớn và thời gian tính toán càng nhiều. Trong bài báo này chúng tôi chọn ε = 0.0001.

3.3 Xác định số lượng thích hợp của chùm

Trong Thuật toán 1, chúng ta cần tách dữ liệu thành *k* chùm. Tuy nhiên, với dữ liệu lớn việc xác định *k* thích hợp là điều không dễ dàng. Việc xác định *k* phù hợp rất quan trọng trong bài toán phân tích chùm nên được rất nhiều nhà thống kê quan tâm. Có nhiều phương pháp được đề xuất để xác định giá trị *k* như dựa vào thông tin tiên nghiệm của mẫu, các chỉ số S, F, Dunn và Xie-Beni (Dunn, 1973; Xie and Beni, 1991). Tuy nhiên các phương pháp này vẫn còn hạn chế khi dữ liệu có sự chồng chéo phức tạp. Trong bài báo này, dựa vào CSI chúng tôi đề nghị thuật toán xác định số chùm thích hợp (Thuật toán 3) cho một tập hợp gồm *N* phần tử $Z = \left\{z_1, z_2, ..., z_N\right\}$. Thuật toán này gồm các bước sau:

Bước 1. Chuẩn hóa dữ liệu đã cho như mục 2.1 để đưa tập dữ liệu ban đầu $N^{(0)}$ về $N^{*(0)} = \left\{z_1^{*(0)}, z_2^{*(0)}, ..., z_N^{*(0)}\right\}$.

Bước 2. Khi $t = 0$, bắt đầu với $V^{(0)} = \left\{v_1^{(0)}, v_2^{(0)}, ...v_N^{(0)}\right\} = \left\{z_1^{*(0)}, z_2^{*(0)}, ..., z_N^{*(0)}\right\}$, và cho $\varepsilon > 0$.

Bước 3. Nâng cấp dãy trọng tâm bởi công thức sau:

$$v_i^{(t+1)} = \frac{\sum_{j=1}^{N} K_\lambda\left(v_i^{(t)}, v_j^{(t)}\right).z_i^{(t)}}{\sum_{j=1}^{N} K_\lambda\left(v_i^{(t)}, v_j^{(t)}\right)},\ i = 1, 2, ..., N, \quad (5)$$

trong đó

$$K_\lambda = \begin{cases} e^{\left(-\frac{n-nc}{\lambda}\right)} & \text{if } c = c(v_i, v_j) \geq c_S, \\ 0 & \text{if } c < c_S, \end{cases}$$

với $c_S = \frac{1}{nC_N^2} \sum i<j\, c\left(z_i, z_j\right) \leq 1$ trung bình của CSI cho từng đôi phần tử trong dữ liệu và $\lambda = \frac{c_S}{r}$.

Giá trị của λ ảnh hưởng số lượng chùm được thiết lập của dữ liệu. Khi $\lambda \to 0$, mỗi phần tử sẽ là một chùm và khi $\lambda \to \infty$ dữ liệu chỉ có một chùm. Giá trị λ phụ thuộc vào hằng số r và c_s. Mặc dù có nhiều thảo luận về việc chọn *r* (Wen and Yenn, 2015), nhưng sự tối ưu vẫn chưa được xác định. Trong bài báo này, thử nghiệm trên nhiều bộ dữ liệu, chúng tôi chọn *r* = 5 cho các ví dụ số và áp dụng.

Bước 4. Lặp lại Bước 3 cho đến khi $\max_i c\left(v_i^{(t)}, v_i^{(t+1)}\right) < \varepsilon$.

Trong thuật toán này, sau mỗi vòng lặp mỗi $v_i^{(t)}$ sẽ hội tụ đến trọng tâm của chùm chứa nó. Thuật toán dừng lại khi sự biến đổi các giá trị $v_i^{(t)}$ giữa hai vòng lặp liên tiếp không ít hơn ε. Khi ε lớn, thuật toán sẽ dừng nhanh hơn nhưng số lượng chùm có thể không phù hợp. Trong bài báo này chúng tôi chọn $\varepsilon = 10^{-4}$ cho tất cả ví dụ số và áp dụng.

Chúng tôi đã thiết lập các chương trình trên phần mềm Matlab để thực hiện nhanh chóng và hiệu quả các Thuật toán 1, Thuật toán 2 và Thuật toán 3. Các thuật toán này được áp dụng trong các tính toán cho những ví dụ của Phần 4.

4 VÍ DỤ SỐ

Trong phần này chúng tôi lấy 2 ví dụ để minh họa các bước của những phương pháp đề nghị, kiểm tra các chương trình đã thiết lập. Những ví dụ này cũng so sánh thuật toán đề nghị với một số thuật toán phổ biến hiện tại. Trong so sánh với các thuật toán chùm với nhau, chúng tôi sử dụng tham số ARI (Rand, 1971) và sai lầm khi thực hiện. Ví dụ 1 được thực hiện trên 150 phần tử thuộc 3 nhóm có phân phối chuẩn hai chiều mà nó được cho bởi chúng tôi. Ví dụ 2 áp dụng cho một vấn đề lý thú: nhận dạng hình ảnh. Đây là hướng áp dụng tiềm năng mà nhiều lĩnh vực thực tế đang đòi hỏi. Áp dụng này cũng thể hiện tính ứng dụng của vấn đề được nghiên cứu. Trong mỗi ví dụ, số liệu rời rạc ban đầu sẽ được chuẩn hóa, áp dụng các thuật toán đề nghị và so sánh hiệu quả với các phương pháp khác. Những ví dụ số với số phần tử khác nhau, đặc tính dữ liệu khác nhau, số chiều khác nhau cho thấy những ưu điểm của các thuật toán đề nghị so với các thuật toán được so sánh.

Ví dụ 1. Ví dụ này xem xét 150 phần tử rời rạc thuộc 3 nhóm (mỗi nhóm 50 phần tử) được tạo ra từ

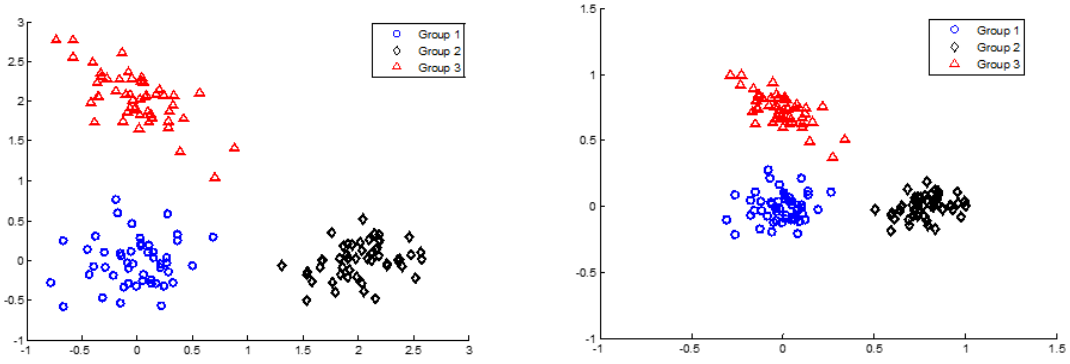
3 phân phối chuẩn hai chiều với vectơ trung bình và ma trận hiệp phương sai cụ thể như sau:

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

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

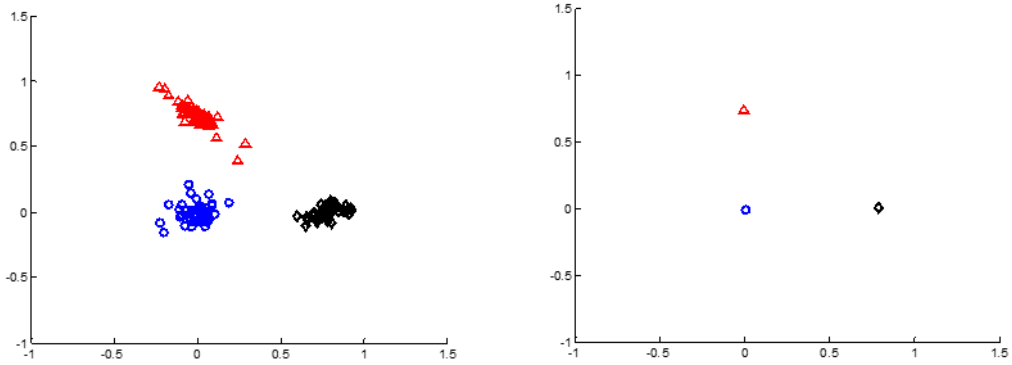
Nhóm 3: $\mu_3 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$; $\Sigma_3 = \begin{pmatrix} 0,1 & -0,05 \\ -0,05 & 0,1 \end{pmatrix}$.

Biểu đồ phân tán của 150 phần tử với 50 phần tử trong mỗi nhóm và sự chuẩn hóa của nó được trình bày bởi Hình 1a và Hình 1b.



(a) **(b)**

Hình 1: Đồ thị phân tán của 150 phần tử (a) và chuẩn hóa của nó (b)



(a) Vòng lặp 1 **(b) Vòng lặp 7**

Hình 2: Đồ thị phân tán của vòng lặp 1 (a) và vòng lặp cuối cùng (b) cho 150 phần tử

Áp dụng Thuật toán 1 với dữ liệu đã chuẩn hóa, sau 7 vòng lặp thuật toán sẽ hội tụ. Các bước của vòng lặp đầu tiên và cuối cùng được cho bởi Hình 2.

Từ Hình 2, ta được số lượng chùm thích hợp là $c = 3$. Kết quả của Thuật toán 1 được lấy làm đầu vào của Thuật toán 2. Áp dụng Thuật 2 với số chùm $k = 3$, sau 9 vòng lặp thuật toán dừng lại khi đó ta có 3 chùm cụ thể:

$C_1 = \{z_1, z_2, \ldots, z_{50}\}$, $C_2 = \{z_{51}, z_{52}, \ldots, z_{100}\}$, $C_3 = \{z_{101}, z_{102}, \ldots, z_{150}\}$.

Thuật toán này cho kết quả phù hợp hoàn toàn với dữ liệu đã cho ban đầu, nghĩa là sai lầm của nó là 0%.

Phân tích chùm mờ với Thuật toán 2, ta nhận được vòng lặp cuối cùng là ma trận U có 3 dòng và 150 cột. Một số cột của ma trận này được cụ thể như sau:

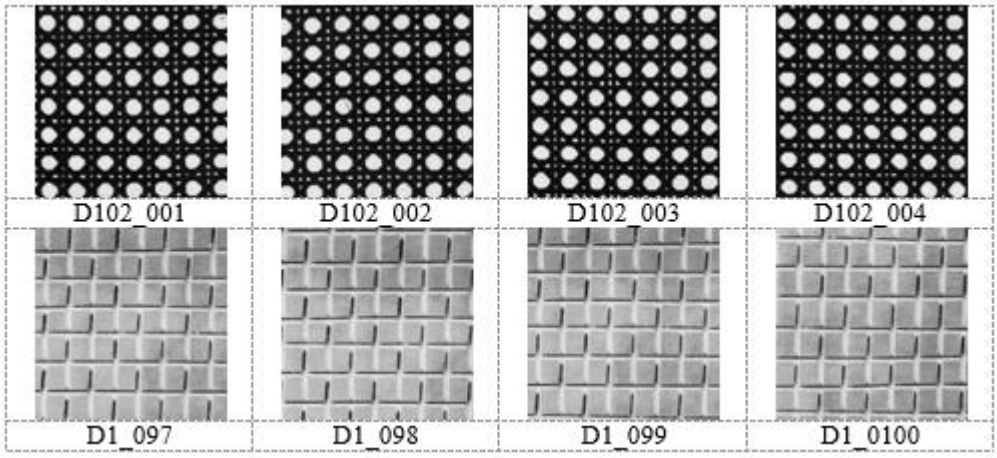
$$U = \begin{pmatrix} 0,4574 & 0,4628 & 0,3957 & ... & 0,2984 & 0,3109 & 0,2918 \\ 0,2804 & 0,2464 & 0,2251 & ... & 0,1750 & 0,1665 & 0,1889 \\ 0,2623 & 0,2909 & 0,3792 & ... & 0,5226 & 0,5226 & 0,5193 \end{pmatrix}.$$

Trong ma trận này, 50 cột đầu của hàng thứ nhất có xác suất lớn nhất, 50 cột kế tiếp có xác suất hàng thứ hai lớn nhất và 50 cột cuối có hàng thứ ba lớn nhất. Nó cũng có nghĩa rằng thuật toán này với số chùm là 3 có tỉ lệ sai lầm là 0%. CSI của 3 chùm lần lượt là 0,8977, 0,8941 và 0,8961,

So sánh chỉ số ARI và tỉ lệ sai lầm thực tế với các thuật toán phổ biến đang được sử dụng hiện tại như thuật toán k-trung bình, thuật toán k -trọng tâm, thuật toán EM (Lauritzen 1995), thuật toán của Tai và Thao (Tai and Thao 2018), ta có Bảng 1.

Bảng 1: So sánh thuật toán đề nghị và một số thuật toán phổ biến cho 150 phần tử

Phương pháp	ARI	Sai lầm (%)
k-trung bình	0,73	76,0
k-trọng tâm	0,61	33,3
EM	0,71	26,0
Tai and Thao	0,93	12,0
Thuật toán 1	1,00	0,0
Thuật toán 2	1,00	0,0



Hình 3: Các mẫu ảnh của hai nhóm hình ảnh

Từ Bảng 1, ta có thể thấy rằng sai lầm của hai thuật toán tốt hơn các thuật toán còn lại. Hai thuật toán đề xuất cũng cho kết quả tốt hơn các thuật toán khác với chỉ số ARI. Cụ thể Thuật toán 1 và Thuật toán 2 đều cho hiệu suất tốt nhất với chỉ số ARI = 1.

Ví dụ 2. Ví dụ này xem xét thuật toán đề nghị trong lĩnh vực nhận dạng ảnh. Những hình ảnh được xem xét được lấy từ tập dữ liệu kết cấu của Brodatz (1966). Đây là tập dữ liệu phổ biến được sử dụng trong nhiều nghiên cứu để so sánh hiệu quả cuả các thuật toán xây dựng chùm. Chúng tôi sử dụng hai tập dữ liệu mẫu kết cấu D1, D102 mà một số hình ảnh của nó được cho bởi Hình 3. Mỗi nhóm gồm có 100 hình ảnh với kích thước 256x256. Chúng tôi trích xuất 3 đặc tính kết cấu contrast, correlation và energy để làm đặc trưng cho mỗi hình ảnh (cho nhiều hơn chi tiết về việc trích xuất kết cấu ảnh xem Haralick 1979).

Thực hiện trích xuất ba đặc trưng của 200 ảnh trên ta có kết quả Bảng 3.

Bảng 3: Trích xuất 3 đặc trưng cho 200 ảnh

No	Contrast	Correlation	Energy	Nhóm
1	1,28	0,93	0,34	D102
2	1,15	0,93	0,33	D102
3	1,31	0,92	0,35	D102
4	1,26	0,93	0,36	D102
5	1,29	0,92	0,35	D102
...	...	...	...	…
196	0,47	0,86	0,16	D1
197	0,45	0,86	0,18	D1
198	0,43	0,88	0,17	D1
199	0,45	0,87	0,17	D1
200	0,47	0,86	0,16	D1

Áp dụng Thuật toán 3, sau 3 vòng lặp ta cũng được số chùm là 2. Sử dụng kết quả của thuật toán này làm đầu vào cho Thuật toán 2, sau 1 vòng lặp ta có ma trận xác suất sau:

$$U = \begin{pmatrix} 0,5709 & 0,5685 & 0,5716 & ... & 0,4282 & 0,4285 & 0,4289 \\ 0,4291 & 0,4315 & 0,4284 & ... & 0,5718 & 0,5716 & 0,5711 \end{pmatrix}$$

Ma trận xác suất này cũng cho ta hai chùm với các hình ảnh hoàn toàn được xếp đúng vào chùm của nó. Thuật toán 1 sau một vòng lặp cũng cho ta hai chùm giống thuật toán 2. CSI của hai chùm lần lượt là 0,9758 và 0,9727 chứng tỏ hai chùm được thiết lập rất tốt.

So sánh các thuật toán đề nghị với một số thuật toán phổ biến như Ví dụ 1, ta có Bảng 4.

Bảng 4: So sánh kết quả phân tích chùm của các phương pháp cho các hình ảnh

Phương pháp	ARI	Sai lầm (%)
k-mean	0,408	30,50
k-mediods	0,446	45,50
EM	0,502	0,25
Tai and Thao	0.940	0,12
Thuật toán 1	0,980	0,08
Thuật toán 2	1,000	0,00

Có thể thấy rằng cả hai thuật toán đề xuất cho kết quả chính xác hơn các thuật toán khác. Cụ thể là Thuật toán 1 và Thuật toán 2 có chỉ số điều chỉnh lần lượt là 0,98 và 1, sai lầm lần lượt là 0,08% và 0%. Hơn nữa, kết quả này cũng cho thấy tính khả thi của hai phương pháp khi áp dụng vào vấn đề thực tế, đặc biệt cho nhận dạng ảnh.

5 KẾT LUẬN

Bài báo đã đề nghị một tiêu chuẩn mới thực hiện được cho hai mục đích quan trọng của bài toán phân tích chùm: xây dựng được các thuật toán phân tích chùm hiệu quả (thuật toán xác định số chùm, thuật toán phân tích chùm mờ và không mờ) và đánh giá được chất lượng của các chùm đã được thiết lập. Các thuật toán này đã thể hiện được những ưu điểm khi so sánh trên một số tập dữ liệu. Với các chương trình được thiết lập trên phần mềm Matlab, các thuật toán đề nghị có thể áp dụng hiệu quả, nhanh chóng cho các tập dữ liệu lớn. Trong tương lai thuật toán đề nghị sẽ được áp dụng trong nhận dạng các hình ảnh trong y học, môi trường, an ninh và nhiều lĩnh vực khác có yêu cầu.

TÀI LIỆU THAM KHẢO

Brodatz, P., 1966. Textures: A photographic Album for Artists and Designers. New York: Dover Publications, 432 pages.

Dunn, J. C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3(3): 32 - 57

Ester, M., Kriegel, H.P., Sander, J. and Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD (Vol. 96, No. 34, pp. 226-231

Ganti, V., Gehrke, J. and Ramakrishnan, R., 1999. Clustering categorical data using summaries. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM 2: 73–83

Haralick, R. M., 1979. Statistical and structural approaches to texture. Proceedings of the IEEE 67: 786–804

Lauritzen, S.L., 1995. The EM algorithm for graphical association models with missing data. Computational Statistics & Data Analysis 19:191–201.

Li, J. and Wang, J. Z., 2008. Real-time computerized annotation of pictures. IEEE transactions on pattern analysis and machine intelligence 30: 985–1002.

Pal, N. R. and Bezdek, J. C., 1995. On cluster validity for the fuzzy c-means model. Fuzzy Systems, IEEE Transactions on 3: 370–379.

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association 66: 846–850.

Tai, V. V., Pham-Gia, T., 2010. Clustering probability distributions. Journal of Applied Statistics 37: 1891–1910.

Tai, V.V. and Thao, N.T., 2017. Communication in Statistics - Theory and Methods 47: 1792 - 1811.

Tai, V. V. and Thao, N.T., 2018. Similar coefficient of cluster for disdrete elements. Sankhya B. The Indian Journal of Statistics, 80(1): 19-36.

Wen, L. H. and Jenn, H. Y., 2015. Automatic clustering algorithm for fuzzy data. Journal of Applied Statistics 42(7):1503–1518

Xie, X. L. and Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis & Machine Intelligence 841–847.