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

DOI:10.22144/ctu.jvn.2022.070

THUẬT TOÁN DI TRUYỀN TRONG PHÂN TÍCH CHÙM CHO DỮ LIỆU RỜI RẠC VÀ ỨNG DỤNG CHO NHẬN DẠNG ẢNH

Võ Văn Tài[1*], Nguyễn Hữu Thoại[1], Lê Thị Kim Cương[1], Phan Nguyễn Nhật Trang[1], Tăng Xuân Khánh[2] và Trần Đại Từ[3]

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

[2]*Trường Trung học Phổ thông Nguyễn Thông, Vĩnh Long*

[3]*Trường Bổ túc văn hóa Pali Trung cấp Nam bộ*

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

Thông tin chung:
Ngày nhận bài: 06/03/2022
Ngày nhận bài sửa: 08/04/2022
Ngày duyệt đăng: 15/04/2022

Title:
Genetic algorithm in building cluster for discrete data and applying for image

Từ khóa:
Dữ liệu rời rạc, khoảng cách, phân tích chùm, thuật toán di truyền

Keywords:
Cluster analysis, distance, discrete data, genetic algorithm

ABSTRACT

This study proposed a genetic algorithm in building cluster for discrete elements, in which the similarity coefficient of cluster was used to evaluate the similarity of the elements, and the improved Davis-Boudin index was used as the objective. Combined with the steps of a traditional cluster analysis algorithm and the operators such as crossover, mutation, and selection of the genetic algorithm, a new cluster analysis algorithm was proposed. The proposed algorithm is detailed with the implementation steps, and illustrated by numerical examples. It is also applied in image recognition, a problem that is still challenging today. The application also shows the potential of this research to many real-world problems related to image recognition.

TÓM TẮT

Thuật toán di truyền trong xây dựng chùm cho các phần tử rời rạc được đề xuất trong nghiên cứu, trong đó hệ số tương tự chùm được sử dụng để đánh giá sự tương tự của các phần tử và chỉ số Davis-Boudin cải tiến được sử dụng làm mục tiêu. Kết hợp với các bước của một thuật toán phân tích chùm truyền thống và các toán tử lai ghép, đột biến, chọn lọc của thuật toán di truyền, một thuật toán phân tích chùm mới được đề xuất. Thuật toán đề nghị được trình bày chi tiết các bước thực hiện và được minh hoạ bởi ví dụ số. Nó cũng được áp dụng trong nhận dạng ảnh, một vấn đề còn nhiều thách thức hiện nay. Áp dụng cũng cho thấy tiềm năng của nghiên cứu này cho nhiều vấn đề trong thực tế liên quan đến nhận dạng ảnh.

1. GIỚI THIỆU

Phân tích chùm là việc chia dữ liệu thành các nhóm sao cho những phần tử trong cùng nhóm có sự tương tự nhiều hơn so với những phần tử bên ngoài nhóm đó, dựa trên những biến quan sát của nó (Ester et al., 1986; Hung, 2016; Tai, 2017). Với ý nghĩa này, phân tích chùm trở thành nền tảng của việc lưu trữ và trích xuất dữ liệu lớn ngày nay. Trước khi thực hiện những áp dụng sâu của khoa học dữ liệu, bài toán phân tích chùm thường được áp dụng. Phân tích chùm có thể thực hiện cho những phần tử rời rạc (CDE) và các hàm mật độ xác suất (CDF) để từ đó áp dụng cho những vấn đề của thực tế. Khi xem

mỗi đối tượng là một phân phối, CDF sẽ được áp dụng. CDF đã được quan tâm về lý thuyết và áp dụng với những đóng góp tiêu biểu là Agusti et al. (2012), Chen & Hung (2015), Thao & Tai (2017).

CDE được ứng dụng phổ biến hơn rất nhiều so với CDF bởi các nhà thống kê (Bouguila & Elguebaly, 2009; Thao & Tai, 2018). Lý do quan trọng nhất cho vấn đề này là dữ liệu thực tế hầu như là các phần tử rời rạc, không phải là hàm mật độ xác suất. Một thuật toán hoàn chỉnh cho CDE bao gồm các bước chính gồm (i) Xây dựng độ đo đánh giá sự tương tự của hai phần tử cũng như giữa hai chùm chứa nhiều hơn hai phần tử; (ii) xác định số chùm thích hợp cần được chia cho tập dữ liệu và những phần tử cụ thể trong mỗi chùm; và (iii) tìm xác suất thuộc vào chùm của mỗi phần tử. Một thuật toán với 2 bước cơ bản (i) và (ii) được gọi là phân tích chùm không mờ và với 3 bước đầy đủ (i), (ii), (iii) được gọi là phân tích chùm mờ. Đối với (i), khoảng cách giữa hai phần tử thông thường được sử dụng là Euclide, city-block, Minskovky, trong khi các khoảng cách min, max, và trung bình được quan tâm cho hai chùm. Nhiều nghiên cứu đã khẳng định chưa có khoảng cách nào được xem là tối ưu trong bài toán xây dựng chùm (Sheng & Liu, 2006; Thao & Tai, 2018). Với nỗ lực khắc phục hạn chế của các khoảng cách trong xây dựng chùm, Thao & Tai (2018) đã đề xuất một độ đo gọi là hệ số tương tự chùm. Hệ số này có ưu điểm khi các chùm có số lượng phần tử khác nhau và có thể sử dụng nó để đo chất lượng chùm khi chúng được thiết lập. Hệ số tương tự chùm cũng được sử dụng để xây dựng chùm mà chúng có ưu điểm hơn phương pháp nổi tiếng k-trung bình trong nhiều trường hợp. Tuy nhiên, với dữ liệu có nhiều chồng lấp, độ đo này có nhiều hạn chế. Đối với (ii) một số nghiên cứu tiêu biểu đã được đề xuất như (Bouguila & Elguebaly, 2005; Thao & Tai, 2018; Dinh et al., 2020). Các nghiên cứu này sử dụng độ đo khoảng cách hoặc hệ số tương tự chùm để tìm số chùm thích hợp, sau đó áp dụng phương pháp k-trung bình để xác định các phần tử trong chùm. Chúng đã thể hiện sự hợp lý trong xây dựng chùm cho nhiều trường hợp cụ thể. Tuy nhiên, kết quả phụ thuộc rất lớn vào mức độ chồng lấp các phần tử trong nhóm.

Thuật toán di truyền (GA) là giải thuật nhằm tìm kiếm, chọn lựa các tính toán tối ưu cho một vấn đề nào đó dựa trên các nguyên tắc tiến hóa của chọn lọc tự nhiên và di truyền học. Các nguyên tắc tiến hóa được sử dụng trong GA thông thường là di truyền, đột biến, chọn lọc tự nhiên và trao đổi chéo (Agusti et al., 2012; Dinh et al., 2021). GA là thuật toán giải quyết bài toán bằng cách mô hình hóa chúng. Từ lời giải đơn giản ban đầu, qua nhiều bước tiến hóa ta được nhiều lời giải tốt, sau đó thông qua quá trình chọn lọc và tìm kiếm, hình thành nên lời giải tối ưu cho bài toán. Trong GA, để đơn giản hóa bài toán, ta mã hóa các đối tượng sang một cấu trúc hay một chuỗi phù hợp, tương tự như một nhiễm sắc thể. Mỗi cấu trúc hay chuỗi được xem như một lời giải có thể của bài toán. Qua các điều kiện chọn lọc, ta được các toán tử sinh ra trong quần thể là các cấu trúc hay chuỗi. Sau đó, ta tiến hành mã hóa các tham số trên mỗi thế hệ, định hướng, tìm kiếm cấu trúc hợp lý, thích nghi điều kiện chọn lọc để được cấu trúc tối ưu. Đối với bài toán phân tích chùm, GA cũng được quan tâm (Agusti et al., 2012; Dinh et al., 2021). Tuy nhiên, khoảng cách Euclide được sử dụng trong nghiên cứu này để xây dựng nên cũng không nhận được kết quả như mong đợi. Sự kết hợp giữa bài toán phân tích chùm truyền thống và GA là một hướng hấp dẫn, có thể cải thiện tốt kết quả thực hiện.

Trong nghiên cứu này, một độ đo mới gọi là chỉ số tương tự chùm được đề xuất để làm tiêu chuẩn xây dựng chùm cho các phần tử rời rạc. Kết hợp phương pháp xây dựng chùm truyền thống và di truyền, một thuật toán phân tích chùm cho các phần tử rời rạc được đề xuất. Thuật toán này có thể xác định số chùm thích hợp, những phần tử trong mỗi chùm và xác suất phụ thuộc vào chùm của mỗi phần tử cùng lúc. Thuật toán đề nghị được trình bày chi tiết bởi các bước và được minh hoạ bởi ví dụ số. Nó cũng được áp dụng trong nhận dạng ảnh và thể hiện được ưu điểm so với các phương pháp quan trọng khác.

2. CÁC VẤN ĐỀ LIÊN QUAN

2.1. Độ đo đánh giá sự tương tự của chùm

Cho 2 phần tử rời rạc X và Y trong không gian n chiều: $X = (x_1, x_2, ..., x_n)$, $Y = (y_1, y_2, ..., y_n)$. Khi đó, khoảng cách Euclide giữa X và Y được định nghĩa như sau:

$$d_E(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Giả sử ta có tập N các phần tử rời rạc $Z = \{z_1, z_2, ..., z_N\}$. Chuẩn hóa dữ liệu Z trong khoảng [0,1] theo nguyên tắc:

$$d^* = \max(z_i),\ i = 1, 2, ..., N;$$

$$z_i^* = \frac{z_i}{d^*},\ i = 1, 2, ..., N;\ Z^* = (z_1^*, z_2^*, ..., z_N^*)$$

Khi đó, hệ số tương tự chùm (SCI) của Z^* được định nghĩa như sau:

$$c_s = 1 - \frac{1}{nC_N^2}\sum_{i<j} d\left(z_i^*, z_j^*\right), \quad (1)$$

trong đó $d\left(z_j, z_j\right)$ là khoảng cách Euclide của z_i và z_j.

2.2. Hàm mục tiêu

Chỉ số *DB* được định nghĩa bởi Davies và Boudin (1979). Chỉ số này được thiết lập dựa vào khoảng cách nhỏ nhất của các phần tử với các chùm trung tâm và khoảng cách lớn nhất giữa các chùm khác nhau. Cụ thể chỉ số *DB* được định nghĩa như sau:

$$DB = \frac{1}{N}\sum_{i=1}^{k}\max_{i\neq j}\left\{\frac{\sum_{x\in C_i} d_E^2\left(x, \overline{x_i}\right) + \sum_{y\in C_j} d_E^2\left(y, \overline{x_j}\right)}{d_E^2\left(\overline{x_i}, \overline{x_j}\right)}\right\}, \quad (2)$$

trong đó

x và y là giá trị của các phần từ,

$\overline{x_i}$ và $\overline{x_j}$ lần lượt là trung bình của các phần tử trong chùm C_i và chùm C_j,

$d_E(x,y)$ là khoảng cách Euclide của x và y.

Chỉ số *DB* thường được sử dụng làm hàm mục tiêu trong xây dựng chùm, nhưng nó sẽ cho kết quả không tốt khi có sự chênh lệch lớn về số phần tử giữa các chùm. Khắc phục điều này, *DB* được điều chỉnh và được gọi là *FB*. Chỉ số *FB* được sử dụng làm mục tiêu trong xây dựng thuật toán đề nghị và được cho bởi công thức sau:

$$FB = \frac{1}{N}\sum_{i=1}^{k}\max_{i\neq j}\left\{\frac{\frac{1}{|C_i|}\sum_{x\in C_i}\mu_i d_E^2\left(x, \overline{x_i}\right) + \frac{1}{|C_j|}\sum_{y\in C_j}\mu_j d_E^2\left(y, \overline{x_j}\right)}{d_E^2\left(\overline{x_i}, \overline{x_j}\right)}\right\}, \quad (3)$$

trong đó

$$\mu_i = \frac{d_E^2\left(x_i, \overline{x_i}\right)}{\sum_{c=1}^{|C_i|} d_E^2\left(x_c, \overline{x_i}\right)}, \ 1 \le i \le N. \quad (4)$$

Cũng như chỉ số *DB*, chỉ số *FB* tính từng đôi khoảng cách giữa các chùm. Do đó, chỉ số *FB* càng nhỏ thì chùm xây dựng càng tốt.

2.3. Tham số đánh giá chùm mờ xây dựng

Hệ số phân hoạch và entropy được sử dụng để đánh giá chất lượng của thuật toán phân tích chùm mờ. Chúng được đưa ra như sau:

$$PC = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N}\mu_{ij}^2, \quad (5)$$

$$PE = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N}\mu_{ij}\log\left(\mu_{ij}\right), \quad (6)$$

Trong đó

μ_{ij} là xác suất để phần tử thứ j được xếp vào chùm thứ i.

k và N lần lượt là số chùm và số phần tử.

Đối với các thuật toán đã xây dựng, *PE* càng lớn thì thuật toán càng tốt và *PC* thì ngược lại.

2.4. Phần tử đại diện cho ảnh

Ma trận đồng hiện mức xám (GLCM) của ảnh $f(x, y)$ có kích thước $M\times N$ và có G mức độ xám là ma trận hai chiều P có kích thước $G\times G$. Mỗi phần tử $p(i, j)$ của ma trận thể hiện tần suất xảy ra cùng giá trị cường độ sáng của i và j tại khoảng cách d và một góc θ xác định. Công thức tính giá trị cụ thể cho phần tử $p(i, j)$ được thể hiện bởi

$$p_{d\theta}(i, j) = \left\{\begin{array}{l}\left\{(x, y), (x', y') \in M\times N \,\middle|\, d = \|(x, y), (x', y')\|, \right.\\ \left.\theta = \left((x, y), (x', y'),\ f(x, y) = i,\ f(x', y') = j\right)\right\}\end{array}\right., \quad (7)$$

Haralick (1979) đã đưa ra 14 đặc trưng kết cấu có thể tính được từ GLCM của kết cấu ảnh. Tuy nhiên, phần lớn các nghiên cứu sau đó chỉ sử dụng đến 3 đặc trưng quan trọng đại diện cho kết cấu (Panjwani & Healey, 1995; Zhang et al., 2018). Trong nghiên cứu này, 4 đặc trưng ảnh được sử dụng gồm: Entropy, tính đồng nhất, độ tương phản và hệ số tương quan để đặt trưng cho mỗi ảnh. Các đặc trưng ảnh được trình bày trong Bảng 1.

Bảng 1. Bốn đặc trưng kết cấu quan trọng của một ảnh

Đặc trưng	Biến	Công thức
Entropy	X_1	$\sum_{i,j} p(i, j)^2$
Độ tương phản	X_2	$\sum_{i,j} \lvert i-j\rvert^k\, p^l(i, j)$
Tính đồng nhất	X_3	$\sum_{i,j} \frac{p(i, j)}{1+\lvert i-j\rvert}$
Hệ số tương quan	X_4	$\sum_{i,j} \frac{(i-\mu_i)(j-\mu_i)\,p(i, j)}{\delta_i\delta_j}$

trong đó μ_i, μ_j là trung bình và độ lệch chuẩn của tổng hàng và cột trong ma trận GLCM tương ứng.

3. THUẬT TOÁN ĐỀ NGHỊ

Cho $Z = \{z_1, z_2, ..., z_N\}$ là dãy gồm N dữ liệu ban đầu với p chiều và $V^{(t)} = \{v_1^{(t)}, v_2^{(t)}, ..., v_N^{(t)}\}$ là tập gồm N chùm trọng tâm tại vòng lặp t. Để phân tích chùm cho Z, ta thực hiện các bước sau:

Bước 1: Khi $t = 0$, khởi tạo vector $V^0 = \{v_1^{(0)}, v_2^{(0)}, ..., v_N^{(0)}\} = Z$.

Bước 2: Cập nhật chùm trọng tâm theo công thức

$$v_i^{(t)} = \frac{\sum_{j=1}^{N} f\left(v_i^{(t-1)}, v_j^{(t-1)}\right).v_j^{(t-1)}}{\sum_{j=1}^{N} f\left(v_i^{(t-1)}, v_j^{(t-1)}\right)}, \; i = 1, 2, ..., N,$$

trong đó

$$f\left(v_i^{(t-1)}, v_j^{(t-1)}\right) = \begin{cases} e^{\left(-\frac{1-c\left(v_i^{(t-1)}, v_j^{(t-1)}\right)}{\lambda}\right)} & khi\, c\left(v_i^{(t-1)}, v_j^{(t-1)}\right) \le c_s, \\ 0 & khi\, c\left(v_i^{(t-1)}, v_j^{(t-1)}\right) > c_s, \end{cases}$$

với c_s là hệ số tương tự chùm và được tính theo công thức (1), $\lambda = \frac{\sigma}{5}$ là giá trị tham số và $c\left(v_i^{(t-1)}, v_j^{(t-1)}\right)$ là khoảng cách Euclide của phần tử $v_i^{(t-1)}$ và $v_j^{(t-1)}$.

Bước 3: Lặp lại Bước 2 cho đến khi $\max_i \left|v_i^{(t-1)} - v_j^{(t-1)}\right| < \varepsilon$.

Sau khi Bước 3 kết thúc, nếu $V^{(t)}$ có bao nhiêu phần tử thì ta sẽ được bấy nhiêu chùm. Giả sử kết thúc bước này ta có k chùm.

Bước 4: Bắt đầu với k chùm. Mã hoá các nhiễm sắc thể (NST) theo kích thước kp phần tử đại diện cho phân vùng có p chiều. Giá trị của mỗi NST nằm trong khoảng nhỏ nhất và lớn nhất của dữ liệu.

Bước 5: Khởi tạo p chuỗi NST với độ dài kp và tính toán giá trị hàm mục tiêu FB sử dụng công thức (3).

Bước 6: Sử dụng các toán tử: chọn lọc, lai ghép và đột biến.

– *Toán tử lai ghép*: Cho L_1 và L_2 là hai NST bố mẹ ban đầu. Khi đó, NST con được tạo ra theo công thức sau:

$$Child = L_1 + rand * (L_2 - L_1).$$

trong đó $rand$ là vector ngẫu nhiên có cùng độ dài với NST bố mẹ và có các giá trị nằm trong khoảng [0, 1].

– *Toán tử đột biến*: Cho x là giá trị tại vị trí được lựa chọn cho quá trình đột biến. Sau quá trình đột biến, giá trị x trở thành x' theo công thức sau:

$$x' = x + N(0, \sigma^2).$$

– *Toán tử chọn lọc*: Mục đích chính của toán tử này là lựa chọn NST bố mẹ cho quá trình tạo ra quần thể mới cho vòng lặp tiếp theo. Trong nghiên cứu này, phương pháp vòng quay Roulette được sử dụng. Xác suất lựa chọn mỗi NST phụ thuộc vào mỗi chùm C_i, được xác định bởi công thức:

$$p_i = \frac{FB_i}{\sum_{j=1}^{N} FB_j},$$

trong đó FB_i là giá trị hàm mục tiêu của NST thứ i trong quần thể đang xem xét và N là số NST có trong quần thể.

Bước 7: Tính giá trị FB của mỗi NST thu được từ Bước 4.

Bước 8: Lặp lại Bước 5, Bước 6 và Bước 7 cho đến khi số vòng lặp hiện tại đạt cực đại hoặc

$$\left|FB^{(t)} - \overline{FB}^{(t)}\right| < \varepsilon,$$

trong đó $FB^{(t)}$ là giá trị của hàm mục tiêu tại vòng lặp thứ t và $\overline{FB}^{(t)}$ là giá trị hàm mục tiêu trung bình của tất cả NST trong quần thể hiện tại.

Trong thuật toán đề nghị, ε là một số dương nhỏ. Nếu ε càng nhỏ thì các vòng lặp của thuật toán có thể càng nhiều và ngược lại. Trong nghiên cứu này, $\varepsilon = 0{,}001$ được chọn.

4. VÍ DỤ MINH HỌA VÀ ÁP DỤNG

4.1. Ví dụ minh họa

Trong ví dụ này, 200 phần tử được trích xuất từ bốn phân phối chuẩn hai chiều với trung bình và ma trận hiệp phương sai được cho như sau:

Nhóm 1: $\mu_1 = \begin{pmatrix} 0,1 \\ 0,1 \end{pmatrix}$, $\sum_1 \begin{pmatrix} 0,1 & 0 \\ 0 & 0,1 \end{pmatrix}$.

Nhóm 2: $\mu_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$, $\sum_2 \begin{pmatrix} 0,1 & 0,05 \\ 0,05 & 0,1 \end{pmatrix}$.

Nhóm 3: $\mu_3 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$, $\sum_3 \begin{pmatrix} 0,1 & -0,05 \\ -0,05 & 0,1 \end{pmatrix}$.

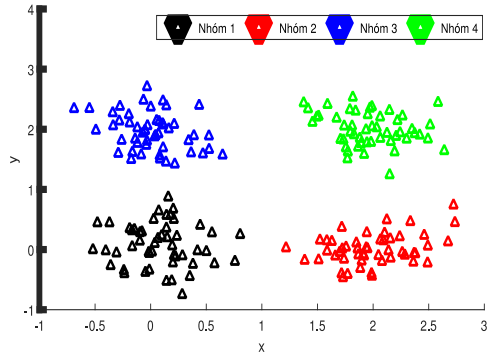
Nhóm 4: $\mu_4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, $\sum_4 \begin{pmatrix} 0,1 & -0,01 \\ -0,01 & 0,1 \end{pmatrix}$.

Trích xuất 50 phần tử hai chiều từ mỗi nhóm, chúng ta có được 200 phần tử rời rạc được của 4 nhóm được ký hiệu như sau:

$$C_1 = \{z_1, z_2, ..., z_{50}\};\ C_2 = \{z_{51}, z_{52}, ..., z_{100}\};$$

$$C_3 = \{z_{101}, z_{102}, ..., z_{150}\};\ C_4 = \{z_{151}, z_{152}, ..., z_{200}\}.$$

Các phần tử này được cho bởi Hình 1.



Hình 1. Đồ thị phân tán của 200 phần tử rời rạc

Bước 1: Khởi tạo vector $V^{(0)}$, trong đó

$v_1^{(0)} = (0,52;\ 0,12)$,

$v_2^{(0)} = (-0.19;\ 0,47), ..., v_{200}^{(0)} = (1,80;\ 2,38)$.

Bước 2: Cập nhật giá trị của mỗi phần tử trong tập dữ liệu, chúng ta thu được

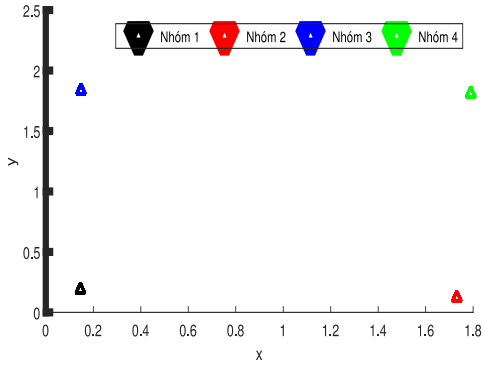
$v_1^{(1)} = (0,54;\ 0,16)$,

$v_2^{(0)} = (-0,03;\ 0,49), ..., v_{200}^{(0)} = (1,73;\ 2,05)$.

Bước 3: Kiểm tra điều kiện dừng, ta có

$$\max_i \left| v_i^{(1)} - v_i^{(0)} \right| = 0,8216 > \varepsilon \ (i = 1, ..., 200).$$

Ta thấy điều kiện thuật toán chưa thoả mãn nên lặp lại Bước 2 cho đến khi dữ liệu hội tụ về các chùm trọng tâm. Sau 5 vòng lặp của Bước 2 và Bước 3, chúng ta nhận được 4 phần tử được minh hoạ bởi Hình 2.



Hình 2. Sự hội tụ của 200 phần tử ban đầu về 4 phần tử sau 5 vòng lặp

Vì 200 phần hội tụ về 4 chùm trọng tâm nên chúng được chia thành 4 chùm. Kết quả này phù hợp với thực tế của dữ liệu.

Bước 4: Mã hóa NST đầu vào với các giá trị nằm trong khoảng $[V_{\min}; V_{\max}]$, trong đó

$V_{\min}$ = [-0,898; -0,771; -0,989; -0,771; -0,989; -0,771; -0,989; -0,771].

$V_{\max}$ = [2,872; 2,799; 2,872; 2,799; 2,872; 2,799; 2,872; 2,799].

NST_1 = [-0,275; 2,188; -0,287; - 0,186; 1,703; 1,996; 2,406; 0,289].

Bước 5: Thực hiện các toán tử của thuật toán.

Thực hiện toán tử lai ghép với 80% NST trong 100 NST từ quần thể thực hiện quá trình này.

Kết quả NST từ quá trình lai ghép, thực hiện tiếp tục quá trình đột biến có xác suất xảy ra đột biến là 1%.

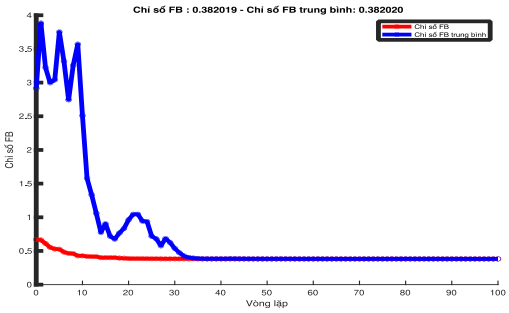
Bước 6: Tính toán giá trị hàm mục tiêu *FB* cho từng chùm trọng tâm sau Bước 5. Với 100 NST, thuật toán xác định NST (chùm trọng tâm) thứ 22 có giá trị *FB* = 0,44 thấp nhất.

NST_{22} = [-0,39; 2,10; -0,16; - 0,06; 2,11; -0,01; 2,31; 2,11].

Bước 7: Toán tử lựa chọn: chùm trọng tâm thứ 22 và 81 được lựa chọn cho vòng lặp tiếp theo.

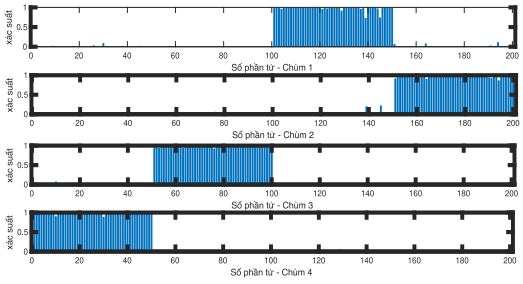
NST_{81} = [2,01; -0,36; 2,78; 2,51; 0,37; 1,55; 0,33; 0,01].

Bước 8: Lặp lại Bước 5, sau 100 vòng lặp thuật toán sẽ dừng. Quá trình thực hiện này được minh hoạ bởi Hình 3.



Hình 3. Sự hội tụ của thuật toán sau 100 vòng lặp

Khi đó ta có mối quan hệ giữa mỗi phần tử với 4 chùm được cho bởi Hình 4.



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

Từ Hình 4, chúng ta có thể thấy rằng thuật toán đề xuất cho kết quả xác suất mờ khá cao và gần với kết quả của phương pháp phân tích chùm thông thường. Ngoài ra, để thể hiện được ưu điểm của thuật toán đề xuất, tác giả đã so sánh với 2 thuật toán cùng dạng đã được công bố gồm Thuật toán c – Trung bình mờ (Fuzzy c-mean) và thuật toán của nhóm tác giả Tài và Thảo (2018) thông qua 2 chỉ số *PE* và *PC*. Kết quả so sánh được cho bởi Bảng 2.

Bảng 2. Chỉ số *PE* và *PC* của các thuật toán

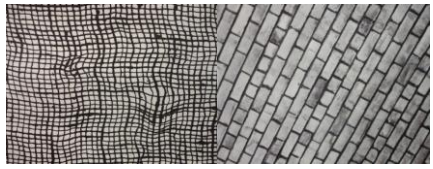
Thuật toán	*PE*	*PC*
Fuzzy C-mean	0,306	0,867
Tài và Thảo (2018)	0,128	0,898
Dinh et al. (2021)	0,098	0,965
Đề xuất	0,068	0,973

Theo Bảng 2, chúng ta có thể khẳng định rằng thuật toán đề xuất có kết quả phân tích chùm mờ hiệu quả hơn, với chỉ số *PC* = 0,973 là lớn nhất và thấp nhất là *PE* = 0,068.

4.2. Ứng dụng trong dữ liệu ảnh

Thuật toán đề nghị được ứng dụng cho dữ liệu ảnh. Các hình ảnh được sử dụng để nhận dạng là dữ liệu của Brodatz. Tập dữ liệu này được chia thành hai nhóm với 21 hình ảnh trong mỗi nhóm. Tập dữ liệu được lấy từ nguồn mở http://imagem.sel.eesc.usp.br/base/Brodatzrotated/index.html.

Một số ảnh mẫu của hai nhóm được cho bởi Hình 5.



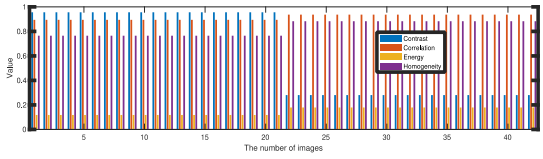
Hình 5. Hai mẫu ảnh của 2 nhóm

Đối với mỗi ảnh, việc trích xuất thành ma trận đồng hiện mức xám với 4 biến cụ thể được tóm tắt bởi Bảng 3.

Bảng 3. Kết quả trích xuất 4 đặc trưng của 42 hình ảnh

Ảnh	Độ tương phản	Hệ số tương quan	Entropy	Tính đồng nhất
I_1	0,808	0,908	0,133	0,807
I_2	1,036	0,887	0,129	0,761
…	…	…	…	…
I_{42}	0,331	0,965	0,151	0,881

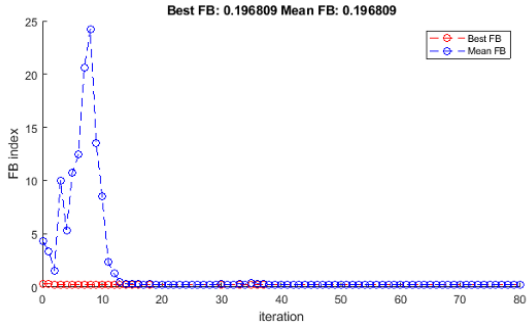
Sau khi trích xuất các đặc trưng của ảnh, chúng tôi xác định số lượng chùm thích hợp cho tập dữ liệu này. Kết quả của giai đoạn 1 được thể hiện trong Hình 6.



Hình 6. Giá trị 4 đặc trưng của hình ảnh trong lần lặp cuối cùng

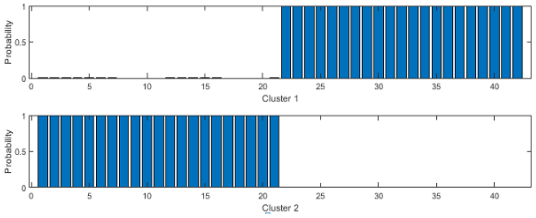
Từ Hình 6, chúng ta có thể thấy rằng mỗi đặc trưng đều hội tụ đến hai giá trị một cách rõ ràng. Do đó kết thúc giai đoạn này, chúng ta có hai chùm.

Tiếp tục thực hiện giai đoạn 2 với 80 lần lặp, ta có kết quả như Hình 7.



Hình 7. Sự hội tụ của Giai đoạn 2 qua các giai đoạn

Khi đó, xác suất để gán cho các chùm được cho bởi Hình 8.



Hình 8. Xác suất thuộc vào hai chùm của 42 phần tử

So sánh với thuật toán được đề xuất và các thuật toán khác, chúng ta có Bảng 4.

Bảng 4. Chỉ số *PE* và *PC* của các thuật toán

Thuật toán	*PE*	*PC*
Fuzzy c-mean	0,306	0,867
Tài và Thảo (2008)	0,265	0,907
Dinh et al. (2021)	0,109	0,945
Đề xuất	0,068	0,973

Trong Bảng 4, thuật toán đề xuất đã thu được kết quả vượt trội khi so sánh với các thuật toán hiện có cho cả chỉ số *PE* và *PC* (*PE* nhỏ nhất và *PC* lớn nhất).

5. KẾT LUẬN

Thuật toán phân tích chùm cho các phần tử rời rạc được đề xuất dựa vào kỹ thuật di truyền. Sự tối ưu trong chọn hàm mục tiêu cải tiến của kỹ thuật di truyền làm cho thuật toán đề nghị có ưu điểm hơn một số thuật toán phổ biến. Bên cạnh ví dụ và ứng dụng đã trình bày, nhiều tập dữ liệu khác đã được thực hiện và cho kết quả tốt khi so sánh với những phương pháp phổ biến. Ứng dụng nhận dạng ảnh có thể là nền tảng cho nhiều ứng dụng thực tế khác trong khoa học dữ liệu và trí tuệ nhân tạo. Đây cũng là hướng nghiên cứu tiếp theo trong thời gian tới.

TÀI LIỆU THAM KHẢO

Agusti, L., Salcedo, S. S., Jiménez, F. S., Carro, C. L., Del, S. J., & Portilla, F. (2012). A new grouping genetic algorithm for clustering problems. *Expert Systems with Applications, 39*(10), 9695–9703. https://doi.org/10.1016/j.eswa.2012.02.149

Bouguila, N., & Elguebaly, W. (2009). Discrete data clustering using finite mixture models. *Pattern Recognition, 42*(1), 33–42. https://doi.org/10.1016/j.patcog.2008.06.022

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

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1986). Adensity-based algorithm for discovering clusters in large spatial databases with noise. *In: KDD Proceeding*, pp. 226–231.

Haralick, R. M. (1979). Statistical and structur al approaches to texture. In *Proceedings of the IEEE, 67*(5), 786 – 804. https://doi.org/10.1109/PROC.1979.11328

Hung, W. L., & Yang, J. H. (2016). Automatic clustering algorithm for fuzzy data. *Journal of Applied Statistics,* 42, 1503–1518. https://doi.org/10.1080/02664763.2014.1001326

Panjwani, D. K., & Healey, G. (1995). Markov random field models for unsupervised segmentation of textured color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17*(10), 939–954. https://doi.org/10.1109/34.464559

Sheng, W., & Liu, X. (2006). A genetic k-medoids clustering algorithm. *Journal of Heuristics, 12*(6), 447–466. https://doi.org/10.1007/s10732-006-7284-z

Tai, V.V., & Thao, N. T. (2018). Similar coefficient of cluster for discrete elements. *Sankhya B, 80*(1), 19–36. https://doi.org/10.1007/s13571-018-0159-0

Tai, V. V., Trung, N. T., Trung V. D., Vinh, H. H., & Thao, N. T. (2017). Modified genetic algorithm based clustering for probability density

functions. *Journal of Statistical Computation and Simulation, 87*(10), 1964–1979. https://doi.org/10.1080/00949655.2017.1300663

Tai V.V., Dinh, P.T., & Dung T.T. (2021). Automatic genetic algorithm in clustering for discrete elements. *Communications in Statistics - Simulation and Computation.* https://doi.org/10.1080/03610918.2019.1588305

Thao, N. T., & Tai, N. T. (2017). Fuzzy clustering of probability density functions. *Journal of Applied Statistics, 44*(4), 583–601. https://doi.org/10.1080/02664763.2016.1177502

Zhang, N., Ruan, S., Lebonvallet, S., Liao, Q., & Zhu, Y. (2018). Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation. *Computer Vision and Image Understanding*, 155, 256–269. https://doi.org/10.1016/j.cviu.2010.09.007