Tạp chí Khoa học Trường Đại học Cần Thơ
Số chuyên đề: Khoa học tự nhiên
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jsi.2020.088

THUẬT TOÁN DI TRUYỀN TRONG PHÂN TÍCH CHÙM ẢNH DỰA TRÊN SỰ TRÍCH XUẤT NHỮNG KHOẢNG ĐẶC TRƯNG

Phạm Toàn Định[1,2*] và Võ Văn Tài[3]

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

[2]*Khoa Kỹ thuật, Trường Đại học Văn Lang*

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

***Người chịu trách nhiệm về bài viết: Phạm Toàn Định (email: phamtoandinh@vanlanguni.edu.vn)*

Thông tin chung:
Ngày nhận bài: 04/03/2020
Ngày nhận bài sửa: 18/03/2020
Ngày duyệt đăng: 29/06/2020

Title:
Genetic algorithm in cluster analysis for images based on extracting the feature intervals

Từ khóa:
Phân tích chùm, ảnh, thuật toán di truyền, độ đo chồng lấp

Keywords:
Cluster analysis, image, genetic algorithm, overlap divergence

ABSTRACT

Based on the extraction of interval data from gray level co-occurrence matrix, this study proposes the Genetic Algorithm in Cluster analysis for Images (GACI). This algorithm can determine the suitable number of clusters, and find the objects in each cluster. The GACI is quickly performed by the established Matlab procedure. The numerical examples illustrate step by step for the GACI, and compare it with the existing algorithms. The results have shown the advantage of the proposed algorithm and the potential in real application of this research.

TÓM TẮT

Dựa trên việc trích xuất khoảng dữ liệu từ ma trận đồng hiện mức xám, nghiên cứu này đề xuất thuật toán di truyền trong phân tích chùm cho các hình ảnh (GACI). Thuật toán có thể xác định số chùm thích hợp và tìm các phần tử trong mỗi chùm. GACI được thực hiện một cách nhanh chóng bởi một chương trình Matlab. Các ví dụ số minh họa từng bước cho GACI và so sánh nó với một số thuật toán đã công bố trước. Kết quả cho thấy ưu điểm của thuật toán đề nghị và tiềm năng trong áp dụng thực tế của nghiên cứu này.

Trích dẫn: Phạm Toàn Định và Võ Văn Tài, 2020. Thuật toán di truyền trong phân tích chùm ảnh dựa trên sự trích xuất những khoảng đặc trưng. Tạp chí Khoa học Trường Đại học Cần Thơ. 56(Số chuyên đề: Khoa học tự nhiên)(1): 8-16.

1 GIỚI THIỆU

Phân tích chùm là việc nhóm các phần tử đã cho 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 các phần tử của chùm khác. Nó là một hướng phát triển quan trọng của thống kê nhiều chiều, nền tảng của phân tích dữ liệu lớn và được ứng dụng trong rất nhiều lĩnh vực (Arivazhagan *et al.* 2010). Chính vì vậy nó đã và đang được rất nhiều các nhà thống kê và công nghệ thông tin quan tâm. Đối tượng phân tích chùm có thể là các phần tử rời rạc, các hàm mật độ xác suất và các khoảng. Phân tích chùm cho các phần tử rời rạc (CDE) đã được nghiên cứu sớm nhất với nhiều kết quả lý thuyết và ứng dụng được công bố (Cabanes *et al.*, 2013; Chen and Hung, 2016; Tai and Thao, 2018a, 2018b). Với dữ liệu lớn và phức tạp như các hình ảnh, mỗi đối tượng cần được biểu diễn thành một phân phối, từ đó phân tích chùm cho các hàm

mật độ xác suất (CDF) được đề xuất. Vì ý nghĩa thiết thực cho nhiều vấn đề phức tạp của thực tế nên CDF nhanh chóng được sự quan tâm của nhiều nhà thống kê. Các kết quả quan trọng trong những năm gần đây cho chủ đề này được nghiên cứu bởi Chen and Hung (2016). Với CDE và CDF, các nhà nghiên cứu đã sử dụng nhiều loại khoảng cách khác nhau làm tiêu chuẩn để xây dựng chùm theo phương pháp thứ bậc và không thứ bậc. Vấn đề xác định số chùm và các tính toán trong áp dụng thực tế đã được giải quyết.

Bên cạnh các phần tử rời rạc và các hàm mật độ xác suất, trong thực tế chúng ta còn lưu rất nhiều dữ liệu kiểu khoảng như nhiệt độ, lượng mưa, khoảng dự báo. Hơn nữa những tập dữ liệu như hình ảnh và nhiều vấn đề khác có thể được biểu diễn thành các khoảng dữ liệu để có thể áp dụng trong nhiều vấn đề thực tế (Kabi *et al.*, 2017). Từ các yêu cầu này, phân tích chùm cho các khoảng (CDI) được đề nghị. So với CDE và CDF, CDI vẫn chưa được nghiên cứu nhiều. De Souza *et al.* (2004) được xem là người đầu tiên nghiên cứu về vấn đề này. Thuật toán này sau đó được cải tiến bởi nhiều tác giả khác như Peng and Li (2006), De Carvalho *et al.* (2007), Chen and Hung (2016) và Kabi *et al.* (2017). Các thuật toán này đã sử dụng khoảng cách City-block (d_C), khoảng cách Euclide (d_E) và khoảng cách Hausdorff (d_H), tuy nhiên chưa tìm thấy các thuật toán sử dụng khoảng cách chồng lấp (d_O) trong xây dựng chùm cho dữ liệu khoảng. Kinh nghiệm cho thấy d_O có ưu điểm hơn d_C, d_E và d_H trong đánh giá sự tương tự của các khoảng. Một số ví dụ cụ thể cho thấy d_C, d_E và d_H không phân biệt được mức độ tương tự của nhiều khoảng trong khi d_O có thể thực hiện được điều này. Chính vì lý do này, các thuật toán đã tồn tại bộc lộ những hạn chế trong nhiều trường hợp. Trong bài viết này, d_O của hai phần tử trong không gian một chiều được sử dụng và cải tiến trong không gian nhiều chiều để đánh giá sự tương tự của hai khoảng. Dựa trên khoảng cách này và chỉ số DB (Davies and Bouldin 1979) của các phần tử rời rạc, nghiên cứu đề xuất chỉ số DB cải tiến (IDB) làm hàm mục tiêu trong thuật toán di truyền. Hơn nữa, ngoại trừ thuật toán của Chen and Hung (2016), các thuật toán khác không đề cập đến vấn đề xác định số chùm. Thuật toán đề nghị cũng giải quyết vấn đề này. Một vấn đề quan trọng của nghiên cứu này là việc áp dụng thuật toán đề nghị trong nhận dạng ảnh.

Trong nghiên cứu này, ma trận đồng hiện mức xám được sử dụng để biểu diễn thành các khoảng đại diện cho mỗi ảnh, sau đó xây dựng thuật toán di truyền phân tích chùm cho các hình ảnh. Thuật toán này có thể xác định số lượng chùm thích hợp cho các ảnh và những ảnh cụ thể cho mỗi chùm. Các tính toán phức tạp cho thuật toán đề nghị được thực hiện nhanh chóng và hiệu quả bởi một chương trình Mathlab được thiết lập. Những ví dụ số và áp dụng đã cho thấy ưu điểm của thuật toán đề nghị so với các thuật toán đang tồn tại.

2 CÁC ĐO ĐỘ VÀ KHOẢNG CÁCH TRONG XÂY DỰNG CHÙM CHO DỮ LIỆU KHOẢNG

2.1 Các khoảng cách phổ biến

Cho hai khoảng trong không gian p chiều: $a = ([a_1, \hat{a}_1], [a_2, \hat{a}_2], ..., [a_p, \hat{a}_p])$ và $b = ([b_1, \hat{b}_1], [b_2, \hat{b}_2], ..., [b_p, \hat{b}_p])$. Trong xây dựng chùm cho dữ liệu khoảng, các khoảng cách sau được sử dụng phổ biến:

Khoảng cách Hausdorff:

$$d_H(a,b) = \sum_{i=1}^{p} (\max\{| a_i - b_i |, | \hat{a}_i - \hat{b}_i |\}). \quad (1)$$

Khoảng cách City-block:

$$d_C(a,b) = \sum_{i=1}^{p} (| a_i - b_i | + | \hat{a}_i - \hat{b}_i |). \quad (2)$$

Khoảng cách Euclide:

$$d_E(a,b) = \sqrt{\sum_{i=1}^{p} [(a_i - b_i)^2 + (\hat{a}_i - \hat{b}_i)^2]}. \quad (3)$$

Khoảng cách Minskowski:

$$d_M(a,b) = \sqrt[p]{\sum_{i=1}^{p} [(a_i - b_i)^p + (\hat{a}_i - \hat{b}_i)^p]}. \quad (4)$$

Khoảng cách được định nghĩa bởi (1), (2), (3) và (4) đánh giá sự khác biệt giữa hai khoảng chỉ dựa vào đầu mút bên trái và bên phải của chúng. Những khoảng cách này không xem xét mức độ chồng lập giữa nên được xem là nguyên nhân chính dẫn đến những hạn chế trong xây dựng chùm.

2.2 Độ đo chồng lấp

Cho hai khoảng $a = [a_1, \hat{a}_1]$ và $b = [b_1, \hat{b}_1]$ trong không gian p chiều, khi đó độ đo chồng lấp của chúng được định nghĩa như sau:

$$d_O(a,b) = D(a,b).\left(1 - \frac{O(a,b)}{2r_a + 1}\right), \quad (5)$$

trong đó $r_a = \frac{1}{p}\sum_{i=1}^{p} | a_i - \hat{a}_i |$, $O(a,b)$ là vùng chồng lấp giữa a và b, và

$$D(a,b) = \max_{a' \in a}\{\min_{b' \in b}\{d_E(a',b')\}\}.$$

$$d_O(a,b) = \begin{cases} 0 & \text{khi } |c_a - c_b| \le r_b - r_a, \\ \left(|c_a - c_b| + r_a - r_b\right)\left(1 - \dfrac{2r_b}{2r_a + 1}\right) & \text{khi } |c_a - c_b| \le r_a - r_b, \\ |c_a - c_b| & \text{khi } r_a = r_b = 0, \\ \left(|c_a - c_b| + r_a - r_b\right)\left(1 - \dfrac{r_a + r_b - |c_a - c_b|}{2r_a + 1}\right) & \text{khi } |r_a - r_b| < |c_a - c_b| < r_a + r_b, \\ \left(|c_a - c_b| + r_a - r_b\right)\left(1 + \dfrac{|c_a - c_b| - (r_a + r_b)}{2r_a + 1}\right) & \text{khi } |c_a - c_b| \ge r_a + r_b, \end{cases} \quad (6)$$

với $c_a = \dfrac{a_1 + \hat{a}_1}{2}$, $r_a = \dfrac{1}{p}\sum_{i=1}^{p} |a_i - \hat{a}_i|$,

$c_b = \dfrac{b_1 + \hat{b}_1}{2}$, $r_a = \dfrac{1}{p}\sum_{i=1}^{p} |a_i - \hat{a}_i|$.

Trong trường hợp p chiều ($p > 1$), độ đo chồng lấp cũng được định nghĩa như (6), trong đó

Trong không gian một chiều, công thức (5) được cụ thể như sau:

$c_a = \dfrac{1}{p}\sum_{i=1}^{p}(a_i + \hat{a}_i)$, $r_a = \dfrac{1}{p}\sum_{i=1}^{p} |a_i - \hat{a}_i|$,

$c_b = \dfrac{1}{p}\sum_{i=1}^{p}(b_i + \hat{b}_i)$, , $r_b = \dfrac{1}{p}\sum_{i=1}^{p} |b_i - \hat{b}_i|$.

2.3 Tiêu chuẩn đánh giá chùm

Giả sử có N khoảng trong không gian p chiều được chia thành k chùm C_i, $i = 1,2, \ldots, k$, khi đó chỉ số *IDB* được cải tiến từ chỉ số *DB* nguồn được định nghĩa như sau:

$$IDB = \frac{1}{N}\sum_{i=1}^{k} \max_{i \ne j} \left\{ \frac{\frac{1}{|C_i|}\sum_{x_i \in C_i} d_O(x_i, \bar{x}_i) + \frac{1}{|C_j|}\sum_{x_j \in C_j} d_O(x_j, \bar{x}_j)}{d_E(\bar{x}_i, \bar{x}_j)} \right\}, \quad (7)$$

trong đó

$\bar{x}_i$ và $\bar{x}_j$ $(i, j = 1, 2, \ldots, k)$ lần lượt là trọng tâm của các khoảng trong chùm C_i và C_j,

$d_E(\bar{x}_i, \bar{x}_i)$ là khoảng cách Euclide của hai trọng tâm chùm C_i và C_j.

Trong bài viết này, khi so sánh hiệu quả các phương pháp xây dựng chùm cho các hình ảnh, các chỉ số như CR (Hubert and Arabie, 1985), chỉ số HI (Hubert, 1977), chỉ số MI (Mirkin and Chernyi, 1970), chỉ số RI (Rand 1971) được cùng lúc sử dụng. Trong xây dựng chùm, chỉ số MI càng nhỏ càng tốt, các chỉ số khác thì ngược lại.

3 THUẬT TOÁN ĐỀ NGHỊ

3.1 Phương pháp trích xuất dữ liệu ảnh

Ma trận đồng hiện mức xám cho một ảnh có kích thước $M \times N$ với G mức xám sẽ có kích thước $G \times G$. Mỗi phần tử $p_{d\theta}(i, j)$ của ma trận này thể hiện cường độ sáng i và j với một khoảng cách d và một góc định hướng θ xác định. Cụ thể nó được cho bởi công thức (8).

$$\begin{aligned} p_{d\theta}(i,j) = \{((r,c),(r',c')) \in M \times N \mid & d = ||(r,c),(r',c')||, \theta \\ & = \Theta((r,c),(r',c')), \\ I(r,c) = i, I(r',c') = j\}. & \end{aligned} \quad (8)$$

Sau khi tính toán ma trận đồng hiện mức xám cho mỗi ảnh, thực hiện trích xuất giá trị đặc trưng của nó thành khoảng theo công thức (9)

$$[\mu_x - r_1/2, \mu_x + r_1/2], [\mu_y - r_2/2, \mu_y + r_2/2], \quad (9)$$

trong đó

r_1 và r_2 là các giá trị ngẫu nhiên có luật phân phối điều trên [1;4].

$$\mu_x = \frac{1}{N_y}\sum_{j}^{N_y}\left(\frac{1}{N_x}\sum_{i}^{N_x}(\mathrm{i})p_{d\theta}(i,j)\right); \mu_y = \frac{1}{N_y}\sum_{i}^{N_x}\left(\frac{1}{N_x}\sum_{j}^{N_y}(\mathrm{j})p_{d\theta}(i,j)\right), \quad (10)$$

với N_x và N_y lần lượt là chiều thứ nhất và thứ hai của tập dữ liệu ảnh và $p_{d\theta}(i,j)$ được xác định bởi (8).

3.2 Mô hình đề nghị

Cho tập N ảnh $X = \{I_1, I_2, \ldots, I_N\}$. Chúng ta cần chia chúng thành các chùm với số lượng thích hợp tùy thuộc vào tập ảnh đã cho. Thuật toán đề nghị bao gồm những bước sau:

Bước 1. Trích xuất đặc trưng các ảnh đã cho thành N khoảng $X = \{a_1, a_2, \ldots, a_N\}$ theo (9) và (10).

Bước 2. Khởi tạo vectơ khoảng dữ liệu $\mathbf{V^{(0)}} = \left\{v_1^{(0)}, v_2^{(0)}, \ldots, v_N^{(0)}\right\} = X$ tại $t = 0$.

Bước 3. Cập nhật vectơ phân vùng bằng công thức (11)

$$v_i^{(t+1)} = \frac{\sum_{j=1}^{N} f\left(v_i^{(t)}, v_j^{(t)}\right).v_j^{(t)}}{\sum_{j=1}^{N} f\left(v_i^{(t)}, v_j^{(t)}\right)}, i = 1, \ldots, N, \quad (11)$$

trong đó

$$f\left(v_i^{(t)}, v_j^{(t)}\right) = \begin{cases} \exp\left(-\dfrac{d_O\left(v_i^{(t)}, v_j^{(t)}\right)}{\lambda}\right) & \text{khi } d_O\left(v_i^{(t)}, v_j^{(t)}\right) \le \mu\alpha_{ij}(t), \\ 0 & \text{khi } d_O\left(v_i^{(t)}, v_j^{(t)}\right) > \mu\alpha_{ij}(t), \end{cases}$$

với

$\alpha_{ij}(t) = \dfrac{\alpha_{ij}(t-1)}{1+\alpha_{ij}(t-1).f\left(v_i^{(t-1)}, v_j^{(t-1)}\right)}$ là hệ số cân bằng ($\alpha_{ij}(0) = 1$),

$\mu = \dfrac{1}{\binom{2}{N}}\sum_{i<j} d_O\left(v_i^{(0)}, v_j^{(0)}\right)$ là trung bình của các khoảng cách $d_O\left(v_i^{(0)}, v_j^{(0)}\right)$, $\lambda = \dfrac{\sigma}{r}$, với $\sigma = \sqrt{\dfrac{1}{\binom{2}{N}}\sum_{i<j}[d\left(v_i^{(0)}, v_j^{(0)}\right) - \mu]^2}$ là độ lệch chuẩn của khoảng cách và r là một hằng số.

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

Kết thúc bước này chúng ta có được số chùm là c.

Bước 5. Khởi tạo quần thể với các nhiễm sắc thể (NST) được mã hóa dạng số không nguyên được lấy ngẫu nhiên từ [min(V); max(V)] với kích thước cp.

Bước 6. Tính toán *IDB* bằng công thức (7) cho các NST đầu tiên.

Trong quá trình tính toán *IDB*, công thức (12) được sử dụng để phân chùm tạm thời:

$$U = \arg\max d_O(x, \bar{x}_i), i = 1, \ldots, c \quad (12)$$

Bước 7. Thực hiện các toán tử lai ghép, đột biến và chọn lọc, với xác suất lai ghép là 85% để có NST mới.

– **Toán tử lai ghép:** Với phương pháp lai ghép điểm, vị trí lai ghép được lựa chọn ngẫu nhiên, sao cho các giá trị của NST thuộc khoảng [min(V); max(V)]. Trong bài báo này, xác suất lai ghép là 85% được chọn. Khi đó các NST trong quần thể sẽ chịu ảnh hưởng trực tiếp của toán tử này.

Chẳng hạn, chúng ta có 100 NST được tạo ra ngẫu nhiên trong quần thể. Khi đó, sẽ có 100*0,85 = 85 NST thực hiện quá trình lai ghép. Trong trường hợp số thập phân, thuật toán sẽ làm tròn số NST.

– **Toán tử đột biến:** Các NST trong quần thể sẽ chịu tác động của toán tử lai ghép theo xác suất xác định, số lượng còn lại sẽ chịu ảnh hưởng của toán tử đột biến. Điểm đột biến được lựa chọn ngẫu nhiên và thay đổi giá trị của NST tại vị trí đó, các vị trí còn lại vẫn ổn định sau toán tử đột biến.

Ví dụ với 85 NST được lai ghép trong 100 NST. Khi đó, 15 NST còn lại sẽ chịu tác động của toán tử

đột biến. Trong trường hợp, xác suất lai ghép trong quần thể là 100%, khả năng đột biến bằng 0.

– **Toán tử lựa chọn:** Các NST được chọn trong vòng lặp tiếp theo với phương pháp vòng quay Roulette.

Bước 8. Tính toán lại chỉ số *IDB* cho NST mới.

Bước 9. Lặp lại Bước 5, Bước 6 và Bước 7 cho đến khi giá trị trung bình các hàm mục tiêu từ các NST trong vòng lặp thấp hơn hoặc bằng giá trị hàm mục tiêu tốt nhất trong quần thể. Cụ thể ở đây là hàm mục tiêu thấp nhất. Tuy nhiên, để thuật toán hoàn toàn hội tụ mạnh, sử dụng thêm điều kiện số vòng lặp của thuật toán sẽ đạt đến cực đại là 1000. Khi đó, thuật toán sẽ dừng và hội tụ toàn cục.

Thuật toán đề nghị có hai giai đoạn. Giai đoạn 1 gồm Bước 1, Bước 2, Bước 3 và Bước 4. Giai đoạn 2 gồm các bước còn lại. Giai đoạn 1 thực hiện việc trích đặc trưng cho các ảnh và tìm số chùm thích hợp cho các ảnh. Trong Bước 3, sau mỗi vòng lặp, các $v_i^{(t)}$ sẽ hội tụ đến trọng tâm của chùm chứa nó. Quá trình này sẽ ngừng khi sự biến đổi giữa hai vòng lặp cho tất cả $v_i^{(t)}$ nhỏ hơn ε. Khi Bước 4 kết thúc, nếu có c trọng tâm thì sẽ có số chùm là c. Trong thuật toán giá trị ε càng lớn, thuật toán sẽ ngừng càng nhanh, nhưng số lượng chùm có thể không thích hợp. Trong bài viết này, $\varepsilon = 10^{-4}$ được chọn cho các ví dụ số. Giai đoạn 2 xác định những ảnh cụ thể trong mỗi chùm. Một chương trình trên phần mềm Matlab được viết để thực hiện thuật toán đề nghị. Nó đã thực hiện một cách hiệu quả cho các ví dụ số của bài viết này.

4 VÍ DỤ SỐ

Trong ứng dụng này, 2 bộ dữ liệu được sử dụng để đánh giá tính hiệu quả của các phương pháp đề xuất. Mỗi bộ số liệu sẽ thực hiện trích xuất đặc trưng thành các khoảng như đã trình bày ở trên, thực hiện việc phân tích chùm theo phương pháp đề nghị và so sánh kết quả này với các mô hình khác để thấy được ưu điểm của các mô hình đề xuất. Nghiên cứu sử dụng các chỉ số CR, chỉ số HI, chỉ số MI và chỉ số RI để so sánh.

Ví dụ 1. Ví dụ này xem xét 30 ảnh của hai nhóm: 10 ảnh hoa mai và 20 ảnh hoa lan để thực hiện. Một số mẫu đại diện của tập dữ liệu được cho bởi Hình 1.

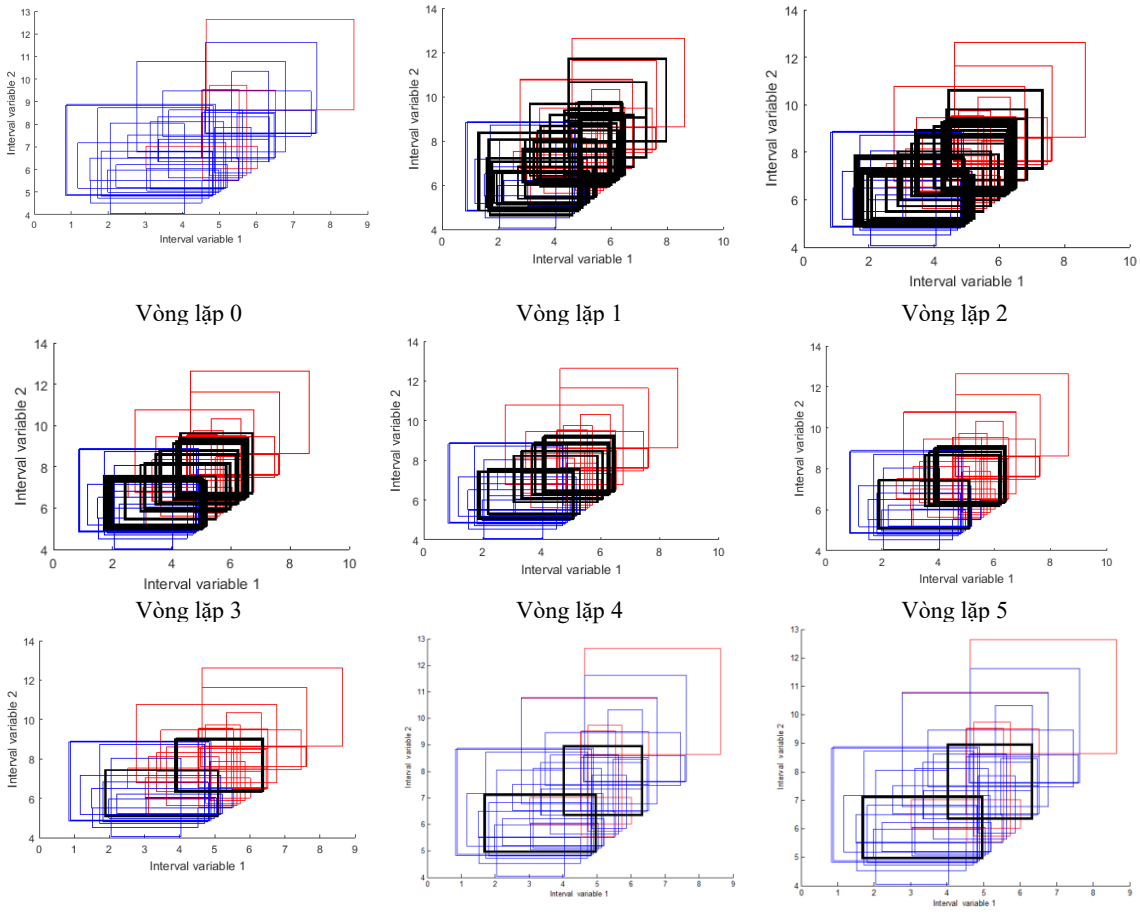
(a) Hoa mai **(b) Hoa lan**

Hình 1: Ảnh đại diện cho hoa mai và hoa lan của tập dữ liệu

Trích xuất đặc trưng các ảnh thành các khoảng (Bước 1) ta có Hình 2 (Vòng lặp 0). Thực hiện Bước 2, Bước 3 và Bước 4, sau 8 vòng lặp ta có Hình 2 và Bảng 1.

Bảng 1 Sự hội tụ của các khoảng trong Giai đoạn 1

Khoảng	V(0)				V(1)				….	V(8)			
1	4,6	8,34	4,5	8,44	4,58	8,06	4,42	8,23		3,98	6,36	3,85	6,50
2	4,65	5,86	3,42	7,09	4,57	6,04	3,64	6,99		3,98	6,36	3,85	6,50
..	…	..	…	..	…	..	…	..	…	..	…	..	…
6	4,52	5,78	3,28	7,01	4,51	6,01	3,56	6,96		3,98	6,36	3,85	6,50
7	1,46	5	2,17	4,29	1,61	4,94	2,2	4,35		1,73	4,98	2,19	4,52
..	…	..	…	..	…	..	…	..	…	..	…	..	…
16	2,09	5,18	3,06	4,22	1,97	5,14	2,87	4,24		1,73	4,98	2,19	4,52
17	2,81	5,25	3,41	4,66	2,63	5,41	3,3	4,74		1,73	4,98	2,19	4,52
18	5,07	6,31	3,9	7,49	4,81	6,29	3,95	7,15		3,98	6,36	3,85	6,50
..	…	..	…	..	…	..	…	..	…	..	…	..	…
28	5,18	6,46	4,68	6,98	4,87	6,44	4,35	6,96		3,98	6,36	3,85	6,50
29	3,72	7,03	4,29	6,47	3,92	6,66	4,1	6,49		3,98	6,36	3,85	6,50
30	4,57	7,45	5,34	6,7	4,54	7,28	5,01	6,81		3,98	6,36	3,85	6,50
	$\mu_1 = 3,2538; \sigma_1 = 2,2459$				$\mu_2 = 2,3578; \sigma_2 = 1,8071$					$\mu_8 = 1,4046; \sigma_8 = 1,4053$			



Vòng lặp 0 | Vòng lặp 1 | Vòng lặp 2

Vòng lặp 3 | Vòng lặp 4 | Vòng lặp 5

Vòng lặp 6 | Vòng lặp 7 | Vòng lặp 8

Hình 2; Các khoảng trích xuất cho hoa mai, hoa lan và sự hội tụ của Giai đoạn 1

Bảng 2 và Hình 2 cho thấy các ảnh này được chia thành 2 chùm. Thực hiện các bước còn lại của thuật toán, ta có

Bước 5: Khởi tạo quần thể gồm 100 NST có giá trị trong [Varmin; Varmax], ta có

– Varmin = [0,838 4,046 0,840 3,520 0,838 4,046 0,840 3,520].

– Varmax = [5,475 8,636 5,598 8,636 5,475 8,636 5,598 8,636].

– NST tốt đầu tiên:

$m^{(1)}$=[0,889 6,749 3,163 3,839 5,275 4,340 2,521 8,013].

– $IDB^{(1)}$ = 0,6566.

– U = [1 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 1 2 1 1 2 1 1 1].

Bước 6: Các toán tử của thuật toán

– Toán tử lai ghép: Từ 100 NST quần thể, toán tử lai ghép sử dụng 85% các NST để lai ghép với nhau.

– Toán tử đột biến: Sử dụng 15% số NST còn lại để thực hiện toán tử đột biến. Điểm đột biến được chọn ngẫu nhiên.

Bước 7: Tính toán chỉ số IDB cho 100 NST mới, ta có: $IDB^{(5)}$ = 0,5635 thấp nhất.

NST tốt trong vòng lặp 1:

$m^{(5)}$=[4,969 5,900 5,186 8,542 1,727 5,867 2,126 3,757].

– Kết quả phân chùm:

U=[2 2 1 2 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2 2 1 2 2 2 2 2 2].

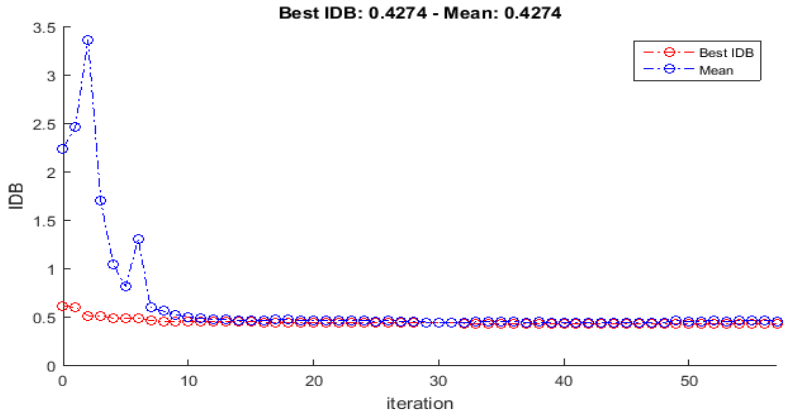
Bước 8: Lặp lại Bước 6 cho đến khi số vòng lặp đạt cực đại (1000 vòng) . Tuy nhiên, ở giai đoạn này, chúng ta có thể thấy rằng hàm mục tiêu IDB hội tụ

từ vòng lặp thứ 12. Sự hội tụ của thuật toán trong Giai đoạn 2, được mô tả trong Hình 3 và kết quả sau:

- NST tốt: $m^{(12)}$ =[1,347 5,327 5,519 3,521 5,152 6,796 4,018 8,018]
- Hàm mục tiêu tối ưu: IDB = 0,4274.
- Kết quả phân chùm tối ưu tối ưu:

U=[1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1]

Từ đây ta có kế quả 2 chùm $C_1 = \{I_1, I_2, ..., I_6, I_{18}, ..., I_{30}\}$; $C_2 = \{I_7, I_8, ..., I_{17}\}$.



Hình 3: Sự hội tụ của thuật toán trong Giai đoạn 2

So sánh với các phương pháp như De Souza *et al.* (2004), De Carvalho *et al.* (2007), Chen and Hung (2016), K-trung bình và thuật toán đề nghị với khoảng cách Euclide (AIGA-E), City-block (AIGA-C) và Hausdorff (AIGA-H) với số chùm là 2, ta có Bảng 2.

Bảng 2: Kết quả so sánh các phương pháp cho tập ảnh hoa mai và hoa lan

Phương pháp	CR	RI	MI	HI
Đề nghị	**1,0000**	**1,0000**	**0,0000**	**1,0000**
De Carvalho et al. (2007)	0,8333	0,9333	0,0667	0,8667
De Souza et al. (2004)	0,8667	0,9333	0,0667	0,8667
Chen & Hung (2016)	0,6280	0,8154	0,1846	0,6276
AIGA-E	0,8667	0,9333	0,0667	0,8667
AIGA-C	0,8667	0,9333	0,0667	0,8667
AIGA-H	0,8667	0,9333	0,0667	0,8667
k-means-C	0,8662	0,9333	0,0667	0,8667
k-means-E	0,8662	0,9333	0,0667	0,8667
k-means-H	0,8662	0,9333	0,0667	0,8667

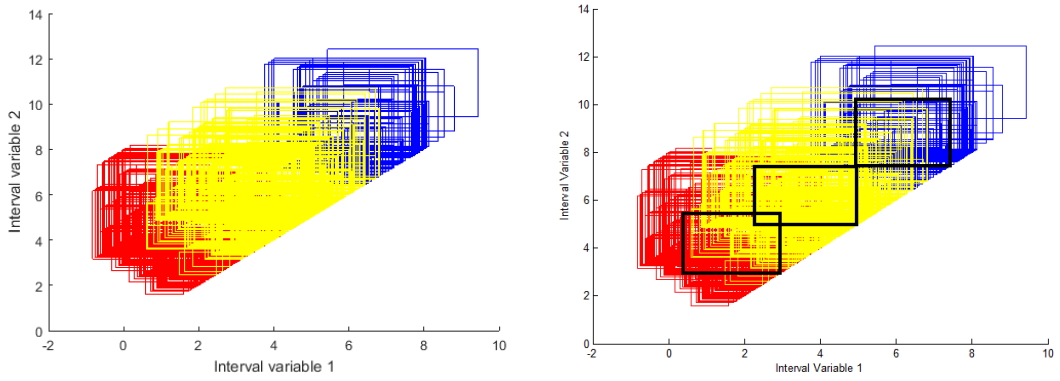
Bảng 2 cho thấy kết quả mô hình đề nghị cho kết quả tốt nhất trong tất cả các mô hình được so sánh, với tất cả các tham số CR, RI, MI và HI.

Ví dụ 2. Sử dụng bộ dữ liệu gồm 519 ảnh với 192 hoa Sen, 76 hoa Gazania và 251 hoa Passion. Tập dữ liệu ảnh được cung cấp tại http://www.robots.ox.ac.uk /~vgg/data /flowers /102/categories.html. Một số mẫu đại diện được cho bởi Hình 4.

Hình 4: Các mẫu ảnh đại diện cho tập dữ liệu 3 loại hoa

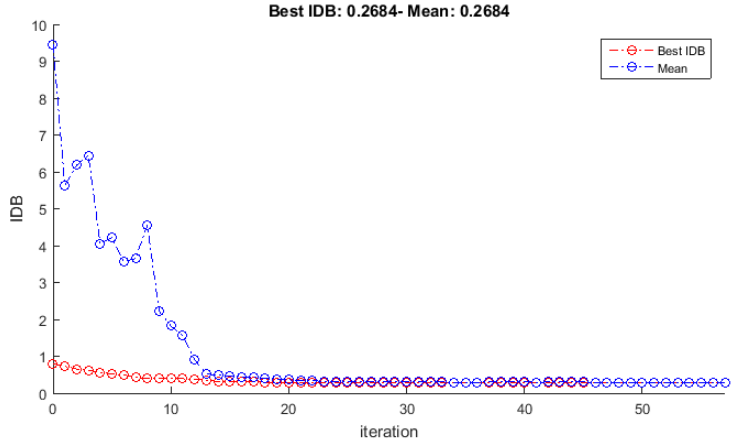
Trích xuất những ảnh thành các khoảng đại diện và thực hiện Giai đoạn 1 sau 18 vòng lặp, ta nhận được Hình 5.

Với 3 chùm, thực hiện tiếp Giai đoạn 2. Sau 45 vòng lặp thuật toán đã hội tụ (Hình 6).



(a) Khoảng dữ liệu trích xuất cho 519 ảnh **(b)Sự hội tụ của 519 ảnh thành 3 khoảng**

Hình 5 Các khoảng trích xuất cho 519 ảnh (a) và 3 khoảng hội tụ (b)



Hình 6: Sự quả hội tụ của Giai đoạn 2 cho 519 ảnh

Khi đó, ta có kết quả cụ thể sau:

Chùm tối ưu: $C_1 = \{I_1, I_2, ..., I_{192}\}$; $C_2 = \{I_{193}, I_{194}, ..., I_{268}\}$; . $C_3 = \{I_{269}, I_{194}, ..., I_{519}\}$

Chỉ số *IDB* = 0,2684.

So sánh với các mô hình khác ta có Bảng 3.

Bảng 3: Kết quả so sánh các phương pháp cho tập 519 ảnh

Thuật toán	CR	RI	MI	HI
Đề nghị	**0,9949**	**0,9976**	**0,0024**	**0,9951**
De Carvalho *et al.* (2007)	0,9326	0,9679	0,0321	0,9359
De Souza *et al*. (2004)	0,9326	0,9679	0,0321	0,9359
Chen and Hung (2016)	0,9693	0,9854	0,0146	0,9707
AIGA-E	0,9755	0,9884	0,0116	0,9767
AIGA-C	0,9755	0,9884	0,0116	0,9767
AIGA-H	0,9342	0,9689	0,0311	0,9377
k-means-C	0,8576	0,9326	0,0674	0,8651
k-means-E	0,8608	0,9334	0,0666	0,8608
k-means-H	0,8608	0,9334	0,0666	0,8608

Bảng 3 cho thấy, thuật toán đề nghị đã cho kết quả tốt nhất trong tất cả các phương pháp được xem xét.

5 KẾT LUẬN

Bài báo đã đề xuất phương pháp trích xuất đặc trưng của các hình ảnh thành các khoảng. Sau đó đề xuất một mô hình phân tích chùm dựa vào thuật toán di truyền. Thuật toán này cùng lúc xác định số chùm thích hợp cho mỗi tập ảnh và số ảnh cụ thể trong mỗi chùm. Thuật toán đề nghị được minh họa chi tiết bởi hai ví dụ số. Thực hiện trên hai tập ảnh này, mô hình đề nghị đã cho kết quả tốt. Chúng cũng cho kết quả tốt nhất khi so sánh với nhiều thuật toán khác. Tuy nhiên, trong mô hình đề nghị, vấn đề hội tụ của thuật toán vẫn chưa được xem xét. Đây sẽ là hướng nghiên cứu mở rộng trong thời gian tới.

TÀI LIỆU THAM KHẢO

Arivazhagan, S., Shebiah, R. N., Nidhyanandhan, S. S., and Ganesan, L. 2010. Fruit recognition using color and texture features. Journal of Emerging Trends in Computing and Information Sciences, 1(2): 90-94.

Cabanes, G., Bennani, Y., Destenay, R., and Hardy, A. 2013. A new topological clustering algorithm for interval data. Pattern Recognition, 46(11): 3030-3039.

Chen, J.H. and Hung, W.L., 2016. An automatic clustering algorithm for probability density functions. Journal of Statistical Computation and Simulation, 85(15): 3047-3063.

Davies, D.L. and Bouldin, D.W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2): 224-227.

De Carvalho, F.D.A., Pimentel, J.T., Bezerra, L.X. and de Souza, R.M., 2007. Clustering symbolic interval data based on a single adaptive Hausdorff distance. In 2007 IEEE International Conference on Systems, Man and Cybernetics: 451-455.

De Souza, R.M., de Carvalho, F.D.A. and Silva, F.C., 2004. Clustering of interval-valued data using adaptive squared Euclidean distances. In International Conference on Neural Information Processing: 775-780.

Hubert, L., 1977. Nominal scale response agreement as a generalized correlation. British Journal of Mathematical and Statistical Psychology, 30(1): 98-103.

Hubert, L. and Arabie, P., 1985. Comparing partitions. Journal of Classification, 2(1): 193-218.

Kabi, S., Wagner, C., Havens, T.C., Anderson, D.T. and Aickelin, U. 2017. Novel similarity measure for interval-valued data based on overlapping ratio. In: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1-6.

Mirkin, B.G. and Chernyi, L.B., 1970. Measurement of the distance between distinct partitions of a finite set of objects. Autom Tel, 5: 120-127.

Peng, W. and Li, T., 2006. Interval data clustering with applications. In 2006 18th IEEE International Conference on Tools with Artificial Intelligence: 355-362.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of The American Statistical Association, 66(336): 846-850.

Tai, V. V., and Trang, N, T., 2018a. Similar coefficient for cluster of probability density functions. Communications in Statistics-Theory and Methods, 47(8):1792-1811.

Tai, V. V., and Trang, N. T. 2018b. Similar coefficient of cluster for discrete elements. Sankhya B, 80(1): 19-36.