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.2018.129

CẢI TIẾN TIÊU CHUẨN KHOẢNG CÁCH TRONG XÂY DỰNG CHÙM CÁC PHẦN TỬ RỜI RẠC

Võ Văn Tài[1*], Lê Thị Kim Ngọc[2] và Bành Văn Viên[2]
[1]*Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*
[2]*Học viên cao học, Trường Đại học Cần Thơ*
**Người chịu trách nhiệm về bài viết: Võ Văn Tài (email: vvtai@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 07/02/2018
Ngày nhận bài sửa: 06/04/2018
Ngày duyệt đăng: 29/10/2018

Title:
Improving distance criterion in building the cluster of discrete elements

Từ khóa:
Chỉ số tương tự, chùm, hình ảnh, khoảng cách, phương pháp thứ bậc

Keywords:
Cluster, distance, image, hierarchical, similar index

ABSTRACT

This research is to proposed a new measure to evaluate the similarity of cluster for discrete elements called the cluster similar index (CSI). CSI is used as criterion to build the algorithms to analyze fuzzy and non-fuzzy cluster and to determine the suitable number of clusters. CSI is also used to evaluate the quality of established clusters and compare them together. These established algorithms can be quickly performed by the Matlab procedures. The numerical examples illustrate the proposed algorithms and show their benefits compared to existing algorithms. Finally, 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 độ đo mới để đánh giá sự tương tự chùm của các phần tử rời rạc được gọi là chỉ số tương tự chùm (CSI). CSI được sử dụng làm tiêu chuẩn để xây dựng các thuật toán phân tích chùm mờ, không mờ và xác định số chùm thích hợp. CSI cũng được sử dụng để đánh giá chất lượng của các chùm được thiết lập cũng như so sánh chúng với nhau. Các thuật toán được thiết lập có thể thực hiện nhanh chóng bởi những chương trình được viết trên phần mềm Matlab. Những ví dụ số minh họa các thuật toán đề nghị và cho thấy thuận lợi của chúng so với các 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, Lê Thị Kim Ngọc và Bành Văn Viên, 2018. Cải tiến tiêu chuẩn khoảng cách trong xây dựng chùm các phần tử rời rạc. Tạp chí Khoa học Trường Đại học Cần Thơ. 54(7A): 101-108.

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. Chúng ta có thể xây dựng chùm cho các phần tử rời rạc (CDE) và chùm cho các hàm mật độ xác suất (CDF). Trong những năm gần đây nhiều tác giả như Goh and Vidal (2008), Tai và Pham-Gia (2010), Chen and Hung (2015), Tai và Thao (2017a, 2017b) đã quan tâm đến CDF. CDE đã được đề xuất trước và có những ưu điểm nhất định

so với CDF. Nó có tính trực quan hơn và tốc độ tính toán trong các thuật toán của nó thường nhanh hơn so với CDF. Trong nhiều trường hợp của áp dụng thực tế, CDE cũng có sai lầm nhỏ hơn CDF. Theo Tai and Thao (2017a), có 3 lý do chính cho vấn đề này: (a) 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 cho các hàm mật độ xác suất (PDF) thì ngược lại; (b) Dữ liệu thực tế thường là rời rạc, do đó để áp dụng CDF, bước đầu các PDF phải được ước lượng. Mặc dù có nhiều tiến bộ cho vấn đề này trong những năm gần đây, nhưng tính chính xác của việc thực hiện cho đến nay vẫn là bài toán chưa có lời giải cuối cùng; (c) Các độ đo cho những phần tử rời rạc thường được tính nhanh hơn nhiều so với các PDF, 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, (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. Đã có một số khoảng cách phổ biến giữa hai yếu tố rời rạc như khoảng cách Euclide, khoảng cách Chebyshev, khoảng cách City-Block, khoảng cách Minkowski…. Trong khi đó, khoảng cách giữa hai tập dữ liệu là khoảng cách Min, khoảng cách Max, khoảng cách Mean và khoảng cách Ward. Các loại khoảng cách và những vấn đề liên quan trong CDE được trình bày tóm tắt trong Webb (2003). Mặc dù có nhiều phương pháp được đề nghị và áp dụng trong thực tế, tuy nhiên chưa có phương pháp nào được xem là tối ưu. Với (ii), 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 đã được đề cập 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. Đối với (iii), chất lượng của các chùm đã được đo bằng nhiều phương pháp như chỉ số S, chỉ số F, chỉ số Dunn, chỉ số Xie - Beni (Dunn, 1973; Xie and Beni, 1991; Pal and Bezdek, 1995; Babuška, 2012). Mặc dù chúng được đánh giá tốt, nhưng các chỉ số trên chỉ được tính toán sau khi các chùm đã được thành lập. Vì vậy, để tìm chùm tốt nhất trong số các phương pháp, chúng ta cần thực hiện tất cả các phương pháp. Hơn nữa, các chỉ số trên chỉ đánh giá tính chất tốt của tất cả các chùm, mà không thể đánh giá tính chất tốt của mỗi chùm được thiết lập. Xuất phát từ những vấn đề trên, Tai and Thao (2017b) đã đề nghị một độ đo mới 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 PDF, không phải cho các phần tử rời rạc.

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, một độ đo mới gọi là chỉ số tương tự chùm (CSI) được đề nghị 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 và không mờ. Hơn nữa, CSI được xem như một tham số để đánh giá chất lượng của các chùm được xây dựng. Điều này có nghĩa là chúng ta có thể xây dựng chùm và đánh giá chất lượng của chùm cùng một lúc. Các thuật toán đề nghị đã được thực hiện nhanh chóng và hiệu quả bởi những thủ tục Matlab. 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 VÀ THUẬT TOÁN ĐỀ NGHỊ

2.1 Một số khái niệm

Định nghĩa 1: Chuẩn hóa dữ liệu

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

$$d_i = \max\left(x_i^j\right), z_{i*}^j = \frac{x_i^j}{d_i}, z_j^* = (z_1^{*j}, z_2^{*j}, ..., z*_n^j),$$

$$i = 1, 2, ..., n; j = 1, 2, ..., N.$$

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 $Z^* = \{z_1^*, z_2^*, ..., z_N^*\}$ mà mỗi phần tử của nó đều có tọa độ trên đoạn [0;1].

Việc chuẩn hóa dữ liệu nhằm đảm bảo tính hợp lý trong đánh giá mức độ gần nhau của các phần tử

trong không gian nhiều chiều với thang đo khác nhau.

Định nghĩa 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 CSI như sau:

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

trong đó $d(z_i^*, z_j^*)$ là khoảng cách giữa hai phần tử z_i^* và z_j^*. Có nhiều khoảng cách giữa hai phần tử như được tổng kết trong (Webb, 2003). Trong bài báo này, chúng tôi chọn là khoảng cách Euclide cho các ví dụ số. Trong trường hợp N = 2, công thức (1) trở thành:

$$d(z_i^*, z_j^*) = n - nd(z_i^*, z_j^*). \quad (2)$$

Chúng ta có thể thấy rằng $\frac{1}{C_N^2} \sum_{i<j} d(z_i^*, z_j^*)$ 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

$$0 \le \frac{1}{C_N^2} \sum_{i<j} d(z_i^*, z_j^*) \le n.$$

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

Khi đó ta cũng nhận được

$$0 \le c(Z^*) \le 1. \quad (3)$$

d_s là trung bình của các khoảng cách của các 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 $c(Z^*)$ thì ngược lại đối với d_s, do đó nếu $c(Z^*)$ càng lớn thì chùm được xây dựng sẽ càng tốt.

Định nghĩa 3: Chỉ số điều chỉnh ARI

Chỉ số ARI do (Hubert and Arabie, 1985) đề xuất là một cải tiến của chỉ số Rand (RI). Hiện nay, ARI đã trở thành một trong những chỉ số đánh giá chùm phổ biến. Nó được sử dụng để so sánh chất lượng của các chùm có số lượng các phần tử khác nhau.

Cho $U = \{u_1, u_2, ..., u_R\}$ và $V = \{v_1, v_2, ..., v_C\}$ là hai phân hoạch có cùng tập dữ liệu đại diện cho chùm R và C. ARI được tính theo công thức sau:

$$ARI = \frac{C_n^2 \sum_{r=1}^{R} \sum_{c=1}^{C} C_{t_{rc}}^2 - \left[c \sum_{c=1}^{C} C_{t_c}^2 \right]}{\frac{1}{2} C_n^2 \sum_{r=1}^{R} C_{t_r}^2 + \sum_{c=1}^{C} C_{t_c}^2 - \left[\sum_{r=1}^{R} C_{t_r}^2 \sum_{c=1}^{C} C_{t_c}^2 \right]}, \quad (4)$$

trong đó t_{rc} là số phần tử thuộc cả hai chùm u_r và v_c, t_r và t_c lần lượt là số phần tử thuộc chùm u_r và chùm v_c, và n là tổng số phần tử trong tập dữ liệu. Giá trị của ARI thuộc khoảng [-1; 1].

2.2 Thuật toán phân tích chùm không mờ dựa vào CSI

Bài toán: Cho một tập hợp gồm N phần tử $Z = \{z_1, z_2, ..., z_N\}$, ta cần phân chia chúng thành c chùm (c được chọn) sao cho hệ số CSI của chùm chứa một phần tử nào đó lớn hơn hệ số CSI của chùm khi ghép nó với nhóm khác.

Thuật toán: Thuật toán này được gọi là (NCA). Nó gồm 5 bước sau:

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

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

Bước 3: Tính hệ số CSI của chùm chứa mỗi phần tử. Nếu CSI này lớn hơn CSI của phần tử khi ghép với các chùm khác, ta giữ phần tử đó trong chùm. Ngược lại, ta gán nó vào chùm có CSI là lớn nhất.

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

2.3 Thuật toán phân tích chùm mờ dựa vào CSI

Bài toán: Cho một tập hợp gồm N phần tử $Z = \{z_1, z_2, ..., z_N\}$, ta cần phân chia chúng thành c chùm (c được chọn) sao cho xác suất của mỗi phần tử thuộc về đúng chùm chứa nó lớn hơn các xác suất khi ta gán phần tử đó vào chùm khác.

Thuật toán: Thuật toán này được gọi là FCA. Nó gồm 3 bước sau:

Bước 1: Khởi tạo ma trận phân vùng $U^{(0)}$ ngẫu nhiên. Tính phần tử đại diện của chùm v_i dựa vào công thức:

$$v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m z_k^*}{\sum_{k=1}^{N} (\mu_{ik})^m}.$$

Sau đó, tính CSI giữa mỗi phần tử và mỗi v_i.

Bước 2: Cập nhật ma trận phân vùng $U^{(1)}$ bằng công thức:

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

Bước 3: Lặp lại Bước 2 và Bước 3 cho tới khi $\left\|U^{(1)} - U^{(0)}\right\| < \varepsilon.$

Trong thuật toán trên, *m* là độ mờ. Khi $m = 1$ chùm mờ trở thành chùm không mờ. Khi $m \to \infty$, phân hoạch trở nên mờ với $\mu_{ik} = 1/c$. Cho đến hiện tại chưa có sự tối ưu trong xác định *m*. Trong bài viết này chúng tôi chọn *m* = 2 theo (Bora and Gupta, 2014) trong các ví dụ số. ε là số rất nhỏ được chọn tùy ý. Nếu ε càng nhỏ thì số bước lặp và thời gian tính toán sẽ càng nhiều. Trong các ví dụ số của bài viết này chúng tôi chọn $\varepsilon = 10^{-4}$.

3 MỘT SỐ VẤN ĐỀ LIÊN QUAN ĐẾN CÁC THUẬT TOÁN

3.1 Xác định số chùm thích hợp dựa vào CSI

Trong hai thuật toán trên, chúng ta cần phân tích bộ số liệu thành *k* chùm. Tuy nhiên, đối với bộ số liệu lớn, việc xác định *k* thích hợp và chọn số chùm khởi tạo như thế nào là một vấn đề khó khăn. Vì kết quả của hai thuật toán phụ thuộc nhiều vào việc chọn số chùm *k* và cách chọn các phần tử vào chùm khởi tạo nên vấn đề này được quan tâm rất nhiều. Nhiều phương pháp được áp dụng để tìm số chùm được đưa ra như dựa vào kiến thức tiên nghiệm về tập dữ liệu hay so sánh hệ số tương quan phân vùng, chỉ số phân vùng, chỉ số Dunn, chỉ số Xie-Beni (Dunn, 1973; Xie and Beni, 1991). Tuy nhiên, việc tính các chỉ số này được thực hiện sau khi phân tích bộ số liệu thành các trường hợp *k* chùm khác nhau, điều này làm cho việc tính toán trở nên cồng kềnh, kém hiệu quả. Trong bài báo này, dựa vào CSI, chúng tôi đề xuất thuật toán xác định số lượng chùm thích hợp cho thuật toán NCA và ma trận phân vùng cho phương pháp FCA. Thuật toán này được gọi là SUS.

Thuật toán: *(Thuật toán SUS)* Gọi $Z^* = \{z_1^*, z_2^*, ..., z_N^*\}$ là tập dữ liệu (đã được chuẩn hóa trên đoạn [0; 1]), $V^{(t)} = \left\{v_1^{(t)}, v_2^{(t)}, ..., v_N^{(t)}\right\}$ là dãy *N* trọng tâm ban đầu của chúng. Thuật toán SUS được trình bày như sau:

Bước 1: Khi *t* = 0, khởi tạo $V^{(0)} = \left\{v_1^{(0)}, v_2^{(0)}, ...v_N^{(0)}\right\} = Z^* = \{z_1^*, z_2^*, ..., z_N^*\}$, $\varepsilon > 0$ rất nhỏ.

Bước 2: Cập nhật dãy trọng tâm theo công thức:

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

trong đó

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

với $c_s = \frac{1}{nC_N^2} \sum_{i<j} c\left(z_i, z_j\right) \le 1$ là trung bình các hệ số tương tự chùm của các điểm dữ liệu và $\lambda = \frac{c_s}{r}.$

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

Trong thuật toán này, sau mỗi bước lặp thì mỗi $v_i^{(t)}$ sẽ hội tụ đến trọng tâm của chùm đang chứa nó (Chen and Hung, 2015). Quá trình này sẽ dừng lại khi các biến của tất cả $v_i^{(t)}$ thông qua hai bước liền kề lặp đi lặp lại nhỏ hơn ε. Khi ε lớn, thuật toán sẽ dừng nhanh hơn nhưng số lượng chùm có thể không thích hợp. Trong bài báo này, chúng tôi cũng chọn $\varepsilon = 10^{-4}$.

3.2 Vấn đề tính toán

Sử dụng phần mềm Matlab, những đoạn code để giải quyết vấn đề tính toán cho các thuật toán đề nghị đã được thiết lập. Như đã đề cập ở Phần giới thiệu, không có khoảng cách tối ưu giữa hai phần tử cũng như giữa các nhóm trong CDE. Các thuật toán truyền thống đã sử dụng khoảng cách Max, khoảng cách Min, khoảng cách Mean và khoảng

cách Ward để tính sự tương tự giữa hai cụm và khoảng cách Euclide để tính sự tương tự giữa hai phần tử. Do đó, để so sánh hiệu quả khi sử dụng CSI cho các phương pháp đề xuất với các phương pháp khác, chúng tôi cũng sử dụng các khoảng cách này để tính CSI. .

4 VÍ DỤ SỐ

Trong phần này bài viết trình bày 2 ví dụ số 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 các thuật toán đã tồn tại và thể hiện tính ứng dụng của vấn đề được nghiên cứu. Ví dụ 1 được thực hiện trên 150 phần tử thuộc 3 nhóm có phân phối chuẩn hai chiều. 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. Trong mỗi ví dụ, từ số liệu rời rạc ban đầu, chúng tôi chuẩn hóa dữ liệu, áp dụng các thuật toán đề nghị và so sánh hiệu quả với các phương pháp đã tồn tại. Hai 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. Xét 150 phần tử rời rạc thuộc phân phối chuẩn hai chiều với trung bình và ma trận hiệp phương sai được cho như sau:

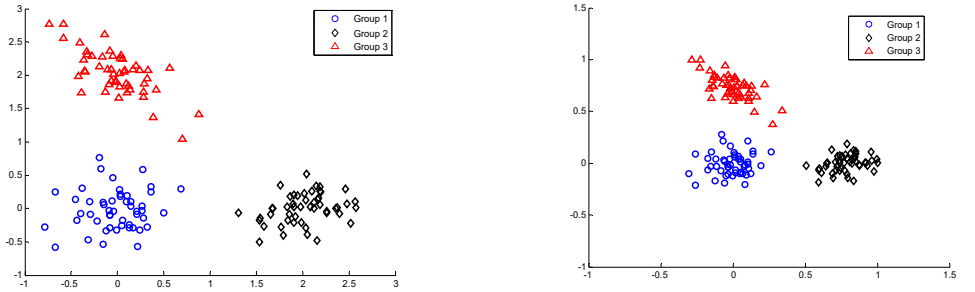
$$\text{Nhóm 1: } \mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \sum\nolimits_1 = \begin{pmatrix} 0,1 & 0 \\ 0 & 0,1 \end{pmatrix}$$

$$\text{Nhóm 2: } \mu_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}; \sum\nolimits_2 = \begin{pmatrix} 0,1 & 0,05 \\ 0,05 & 0,1 \end{pmatrix}$$

$$\text{Nhóm 3: } \mu_3 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}; \sum\nolimits_3 = \begin{pmatrix} 0,1 & -0,05 \\ -0,05 & 0,1 \end{pmatrix}$$

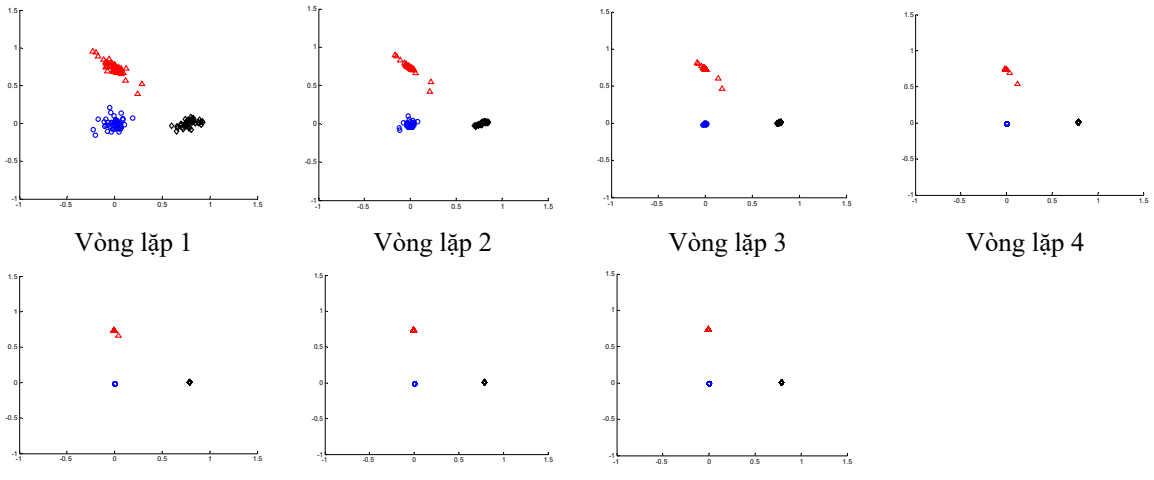
Biểu đồ phân tán của 150 phần tử với 50 phần tử trong mỗi nhóm và sự chuẩn hóa của nó được trình bày bởi Hình 1a và Hình 1b.

Áp dụng thuật toán SUS với dữ liệu đã chuẩn hóa, sau 7 vòng lặp thuật toán sẽ hội tụ. Các bước của thuật toán này được minh họa bởi Hình 2.



(a) (b)

Hình 1: Đồ thị phân tán của 3 nhóm (a) và đồ thị phân tán của 3 nhóm chuẩn hoá (b)



Vòng lặp 1 Vòng lặp 2 Vòng lặp 3 Vòng lặp 4

Vòng lặp 5 Vòng lặp 6 Vòng lặp 7

Hình 2: Các vòng lặp của thuật toán SUS

Từ Hình 2, ta được số lượng chùm thích hợp là $k = 3$. Kết quả của thuật toán SUS được lấy làm đầu vào của thuật toán FCA.

Phân tích chùm mờ FCA, 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 FCA với số chùm là 3 có tỉ lệ sai lầm là 0%. Thuật toán toán NCA cũng cho ta 3 chùm giống thuật toán FCA nghĩa là có tỉ lệ sai lầm của nó cũng là 0%. CSI của 3 chùm lần lượt là 0,8977; 0,8941 và 0,8961.

Cũng lấy số chùm thực hiện là 3 để thực hiện các phương pháp khác, ta có bảng tổng hợp các kết quả so sánh được cho bởi Bảng 1.

Bảng 1: So sánh thuật toán đề nghị và mộ số thuật toán tồn tại

Phương pháp	ARI	Sai số (%)	Thời gian tính (giây)	Độ lệch chuẩn
K-mean	0,73	76	6	0,18
K-medoids	0,61	33,33	40,5	0,21
Expectation-Maximization	0,71		27	0,18
NCA	1	0	25	0
FCA	1	0	34	0

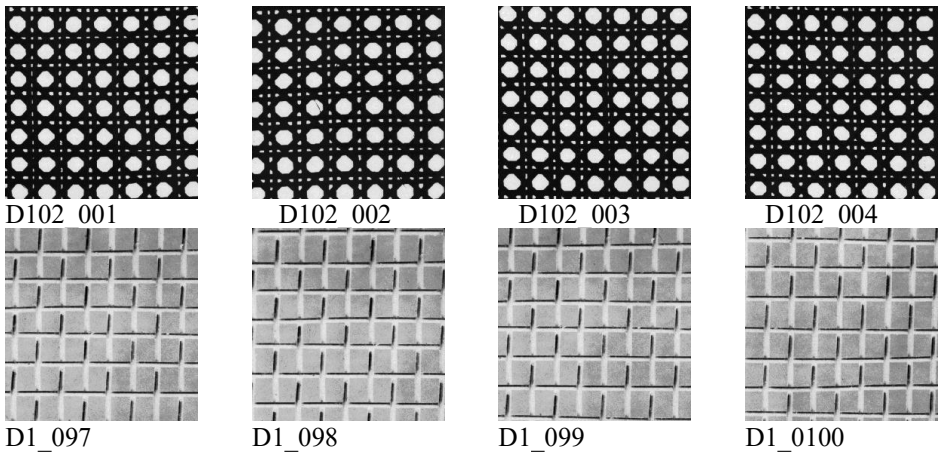
Từ Bảng 1, ta có thể thấy rằng sai lầm của hai thuật toán đề nghị tốt hơn các thuật toán còn lại. Hai thuật toán này này cũng cho kết quả tốt và ổn định hơn các thuật toán khác. Cụ thể thuật toán NCA và FCA đều cho chỉ số điều chỉnh là 1. Bởi vì thuật toán NCA và FCA cần thêm thời gian xác định số chùm nên nó không có ưu điểm hơn về thời gian tính toán so với thuật toán K-mean và Expectation-Maximization. Tuy nhiên với thời gian không quá lớn (từ 25-34 giây), chúng cũng không phải là trở ngại khi áp dụng. Hơn nữa, một vấn đề quan trọng khác là phương pháp đề nghị vừa thực hiện việc phân tích chùm vừa đánh giá được chất lượng của các chùm được xây dựng cùng lúc, trong khi các phương pháp được so sánh chỉ xây dựng chùm. Do đó sau khi thực hiện, để đánh giá chất lượng của chùm, chúng ta phải tốn thời gian tính các chỉ số. Nếu xét tổng thể thời gian để thực hiện cả hai giai đoạn: xây dựng chùm và đánh giá chất lượng chùm thì phương pháp đề nghị có thời gian tính toán chênh lệch không quá lớn với các phương pháp khác được so sánh.

Ví dụ 2. Ví dụ này áp dụng thuật toán đề xuất trong nhận dạng ảnh. Những ảnh này được lấy từ cơ sở dữ liệu kết cấu của Brodatz (1996) được thực hiện bởi nhiều nhà nghiên cứu về hình ảnh. Cụ thể, chúng tôi sử dụng 2 mẫu kết cấu D1, D102 (Hình 3), trong đó có 100 hình với kích thước (256x256) được lấy cho mỗi nhóm. Tính ma trận đồng hiện chất xám (GLCM) và trích xuất đặc trưng của ba kết cấu bao gồm độ tương phản, sự tương quan và tính đồng nhất (chi tiết về GLCM và các đặc điểm kết cấu, xem trong (Haralick, 1979; Celebi and Alpkocak, 2000).

Thực hiện trích xuất ba đặc trưng của 200 ảnh trên ta có kết quả Bảng 2.

Bảng 2: Đặc trưng kết cấu hai nhóm ảnh

STT	Độ tương phản	Sự tương quan	Tính đồng nhất	Lớp
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



Hình 3: Các ảnh mẫu của hai nhóm

Áp dụng thuật toán SUS, sau khi chuẩn hóa dữ liệu 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 FCA 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 NCA cũng cho ta hai chùm giống thuật toán FCA. CSI của hai chùm lần lượt là 0,9758 và 0,9727.

So sánh các thuật toán đề nghị với một số thuật toán đã tồn tại cho bộ ảnh này, ta có Bảng 3.

Bảng 3: So sánh kết quả phân tích chùm của các phương pháp

Phương pháp	ARI	Sai số	Thời gian tính	Độ lệch chuẩn
K-mean	0,41	82,51	0,07	0,28
K-mediods	0,43	74,50	91,67	0,23
Expectation-Maximization	0,50	83,13	0,47	0,25
NCA	0,98	0	1,99	0
FCA	1	0	0,9	0

Có thể thấy rằng cả hai thuật toán đề xuất cho kết quả chính xác hơn. Cụ thể là thuật toán NCA và FCA có chỉ số điều chỉnh lần lượt là 0,98 và 1. Mặc khác, cả hai thuật toán đề xuất đều ổn định hơn với độ lệch chuẩn bằng 0. Hơn nữa, nó chứng minh được 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 để thực hiện được cho hai mục đích quan trọng của bài toán phân tích chùm: Xây dựng các thuật toán phân tích chùm (thuật toán xác định số chùm, thuật toán phân tích chùm mờ và không mờ) và đánh giá được chất lượng của các chùm thiết lập. Các thuật toán này đã chứng minh được những ưu điểm khi so sánh trên các tập dữ liệu đối chứng và thực tế. 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 chúng tôi sẽ áp dụng thuật toán đề nghị cho việc nhận dạng các hình ảnh trong y học, môi trường, an ninh và nhiều lĩnh vực khác có yêu cầu. Tuy nhiên, trong các thuật toán đề nghị sự hội tụ của chúng vẫn chưa được xem xét. Đây sẽ hướng nghiên cứu mà chúng tôi sẽ tập trung thực hiện trong thời gian sắp tới.

TÀI LIỆU THAM KHẢO

Babuška, R., 2012. Fuzzy modeling for control. Science & Business Media. NewYork, 345 pages.

Bora, D. J. and Gupta, A. K., 2014. Impact of exponent parameter value for the partition matrix on the performance of fuzzy C means Algorithm. ArXiv. 109: 1-17.

Brodatz, P., 1996. Textures: A Photographic Album for Artists and Designers. Dover Publications. New York, 525 pages.

Celebi, E. and Alpkocak, A., 2000. Clustering of texture features for content-based image retrieval. Advances in Information Systems. 1901: 216-225.

Chen, J. H. and Hung, W. L., 2015. An automatic clustering algorithm for probability density functions. Journal of Statistical Computation and Simulation. 85(15): 3047-3063.

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.

Goh, A. and Vidal R., 2008. Unsupervised Riemannian clustering of probability density functions. Machine Learning and Knowledge Discovery in Databases. 11: 377-392.

Haralick, R. M., 1979. Statistical and structural approaches to texture. Proceedings of the IEEE. 67(5): 786-804.

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

Hung, W. L. and Yang, J.H., 2015. Automatic clustering algorithm for fuzzy data. Journal of Applied Statistics. 42(7): 1503-1518.

Li, J. and Wang, J. Z., 2008. Real-time computerized annotation of pictures. IEEE transactions on pattern analysis and machine intelligence. 30(6): 985-1002.

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

Tai, V. V. and Pham-Gia T., 2010. Clustering probability distributions. Journal of Applied Statistics. 37(11): 1891-1910.

Tai, V. V. and Thao N. T., 2017a. Fuzzy clustering of probability density function. Journal of Applied Statistics. 44(4): 583-601.

Tai, V. V. and Thao, N. T., 2017b. Similar Coefficient for Cluster of Probability Density Functions. Communications in Statistics - Theory and Methods. 47(8): 1792-1811.

Webb, A. R., 2003. Statistical pattern recognition. John Wiley & Sons. London, 725 pages.

Xie, X. L. and Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis & Machine Intelligence. 13(8): 841-847.