TẠP CHÍ KHOA HỌC ĐẠI HỌC ĐÀ LẠT Tập 8, Số 2, 2018 13–23

CẢI TIẾN THUẬT TOÁN PHÂN CỤM MỜ DỰA TRÊN ĐỘ ĐO TRỌNG SỐ ENTROPY VÀ CHỈ SỐ CALINSKI-HARABASZ

Nguyễn Như Đồng[a*], Phan Thành Huấn[b]

[a]Phòng Đào tạo, Trường Cao đẳng Kỹ nghệ II, TP. Hồ Chí Minh, Việt Nam
[b]Bộ môn Tin học, Trường Đại học Khoa học Xã hội và Nhân văn, Đại học Quốc gia TP. Hồ Chí Minh, TP. Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ: Email: dongnhunguyen@gmail.com

Lịch sử bài báo
Nhận ngày 19 tháng 01 năm 2018
Chỉnh sửa ngày 22 tháng 03 năm 2018 | Chấp nhận đăng ngày 14 tháng 04 năm 2018

Tóm tắt

Phân cụm là kỹ thuật quan trọng trong khai thác dữ liệu và được ứng dụng rộng rãi trong các lĩnh vực như nhận dạng mẫu, thị giác máy tính và điều khiển mờ. Trong bài viết này, chúng tôi trình bày thuật toán cải tiến phân cụm mờ dựa vào sự kết hợp thuật toán phân cụm mờ dựa trên độ đo trọng số Entropy và chỉ số Calinski-Harabasz. Ưu điểm của phương pháp này là không những phân chia cụm hiệu quả, có độ chính xác cao mà còn có khả năng đo lường cụm, đánh giá cụm nhằm tìm ra được số cụm tối ưu đủ đáp ứng cho các nhu cầu thực tiễn. Sau cùng, chúng tôi trình bày kết quả thực nghiệm trên dữ liệu thực, cho thấy thuật toán cải tiến phân cụm hiệu quả và chính xác hơn.

Từ khóa: Chỉ số Calinski-Harabasz; Phân cụm mờ; Trọng số Entropy.

Mã số định danh bài báo: http://tckh.dlu.edu.vn/index.php/tckhdhdl/article/view/408

Loại bài báo: Bài báo nghiên cứu gốc có bình duyệt

AN IMPROVED FUZZY K-MEANS CLUSTERING ALGORITHM BASED ON WEIGHT ENTROPY MEASUREMENT AND CALINSKI-HARABASZ INDEX

Nguyen Nhu Dong[a*], Phan Thanh Huan[b]

[a]Training Department, Hochiminh Vocational College of Technology, Hochiminh City, Vietnam
[b]The Division of Information Technology, University of Social Sciences and Humanities, VNU Hochiminh, Hochiminh City, Vietnam
[*]Corresponding author: Email: dongnhunguyen71@gmail.com

Article history
Received: January 19th, 2018
Received in revised form: March 22nd, 2018 | Accepted: April 14th, 2018

Abstract

Clustering plays an important role in data mining and is applied widely in fields of pattern recognition, computer vision, and fuzzy control. In this paper, we proposed an improved clustering algorithm combined of both fuzzy k-means using weight Entropy and Calinski-harabasz index. The advantage of this method is that it does not only create efficient clustering but also has the ability to measure clusters and rate clusters to find the optimal number of clusters for practical needs. Finally, we presented experimental results on real-life datasets, which showed that the improved algorithm has the accuracy and efficiency of the existing algorithms.

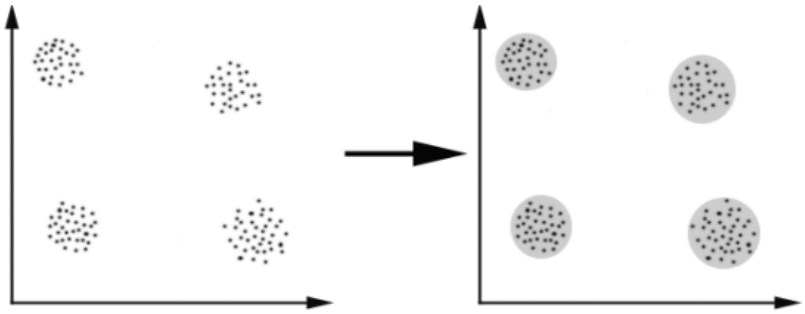
Keywords: *Calinski-Harabasz* Index; Fuzzy K-means; Weight entropy.

Article identifier: http://tckh.dlu.edu.vn/index.php/tckhdhdl/article/view/408

Article type: (peer-reviewed) Full-length research article

1. GIỚI THIỆU

Mục đích chính của phân cụm dữ liệu nhằm khám phá cấu trúc của mẫu dữ liệu để phân chia thành các nhóm dữ liệu từ tập dữ liệu lớn. Từ đó, người dùng có thể phân tích và nghiên cứu theo từng cụm dữ liệu, nhằm khám phá và tìm kiếm các thông tin tiềm ẩn, hữu ích hỗ trợ cho việc ra quyết định. Phân cụm dữ liệu là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "*tương tự*" với nhau và các phần tử trong các cụm khác nhau sẽ "*không tương tự*" với nhau. Độ tương tự được tính dựa trên giá trị các thuộc tính mô tả đối tượng.



Hình 1. Mô phỏng sự phân cụm dữ liệu

Hiện nay, kỹ thuật phân cụm có rất nhiều hướng tiếp cận. Trong bài viết này, nhóm tác giả tập trung cải tiến kỹ thuật phân cụm theo hướng tiếp cận phân hoạch. Ý tưởng chính của kỹ thuật này là phân một tập dữ liệu có *n* phần tử cho trước thành *k* nhóm dữ liệu sao cho mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Các thuật toán phân hoạch có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề phân cụm dữ liệu, vì nó phải tìm kiếm tất cả các cách phân hoạch có thể được. Chính vì vậy, trên thực tế người ta thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của các cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Với chiến lược này, thông thường người ta bắt đầu khởi tạo một phân hoạch ban đầu cho tập dữ liệu theo phép ngẫu nhiên hoặc theo *heuristic* và liên tục tinh chỉnh nó cho đến khi thu được một phân hoạch mong muốn, thoả mãn các điều kiện ràng buộc cho trước. Các thuật toán phân cụm phân hoạch cố gắng cải tiến tiêu chuẩn phân cụm bằng cách tính các giá trị đo độ tương tự giữa các đối tượng dữ liệu và sắp xếp các giá trị này, sau đó thuật toán lựa chọn một giá trị trong dãy sắp xếp sao cho hàm tiêu chuẩn đạt giá trị tối thiểu. Như vậy, ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược ăn tham để tìm kiếm nghiệm.

Trong phạm vi bài báo, nhóm tác giả trình bày cải tiến thuật toán phân cụm mờ kết hợp giữa phương pháp phân cụm mờ sử dụng trọng số *Entropy* do Jing, Ng, và Huang (2007) cũng như Li và Chen (2008, 2010) và kỹ thuật đánh giá cụm theo chỉ số *Calinski-Harabasz*. Phần 2 trình bày các khái niệm cơ bản về phân cụm rõ và phân cụm mờ. Phần 3 đề xuất mô hình phân cụm mờ dựa trên kết hợp giữa phân cụm mờ sử dụng trọng số *Entropy* và đánh giá cụm theo chỉ số *Calinski-Harabasz*. Kết quả thực nghiệm được trình bày trong Phần 4 và kết luận ở Phần 5.

2. CÁC VẤN ĐỀ LIÊN QUAN

2.1. Phân cụm rõ: Thuật toán K-Means

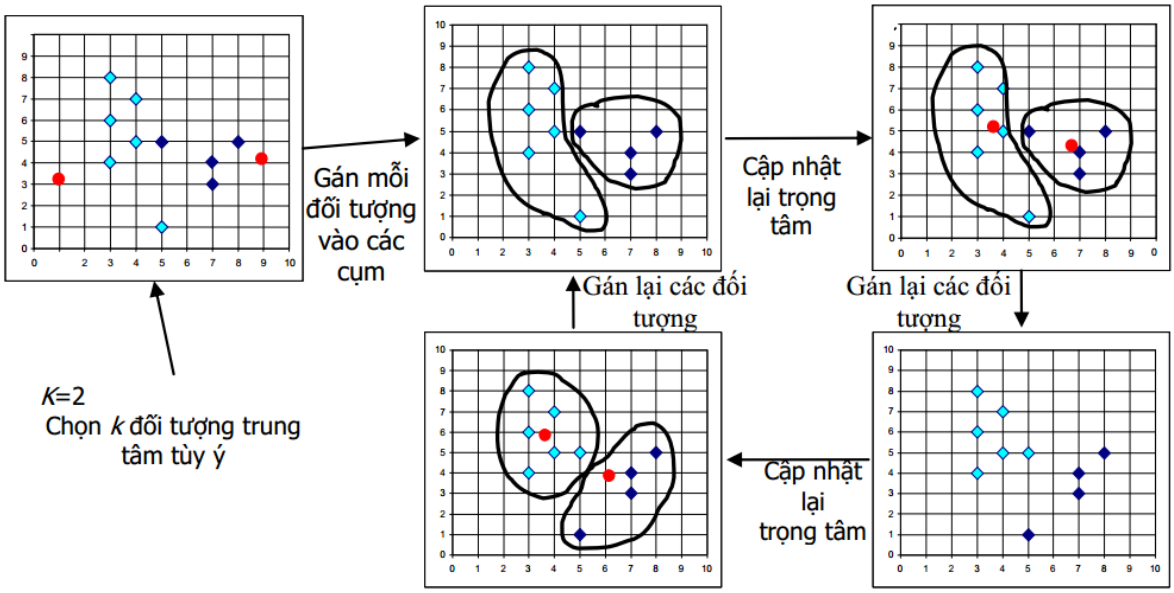
Cho một tập dữ liệu $X=\{x_1,x_2,\ldots,x_n\}$, với $x_i \in R^d$, gồm n đối tượng dữ liệu d chiều. Phân tách tập dữ liệu thành k cụm: $C_1,C_2,\ldots,C_k$ rời nhau thỏa mãn điều kiện sau: $C_i \neq \phi, \forall i = 1..k$;

$C_i \cap C_j \neq \phi, \forall i \neq j$; và $\bigcup_{i=1}^{k} C_i = X$. Trong đó: *k* là số cụm sẽ phân thành, cho trước, và nguyên dương; C_i là véc-tơ tâm cụm, dùng đề chỉ cụm thứ *i*.

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Ý tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng đã cho vào *k* cụm (*k* là số các cụm được xác đinh trước, *k* nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm là nhỏ nhất. Thuật toán có các bước như sau:

- *Đầu vào*: Cơ sở dữ liệu gồm *n* đối tượng *d* chiều và hằng số *k*;
- *Đầu ra*: Các cụm C_i (*i=1..k*) sao cho hàm tiêu chuẩn *F* đạt giá trị tối thiểu;
- *Bước 1*: Chọn *k* đối tượng m_j (*j=1...k*) là trọng tâm ban đầu của *k* cụm (*ngẫu nhiên hoặc theo kinh nghiệm*);
- *Bước 2*: Đối với mỗi đối tượng X_i ($1 \le i \le n$), tính toán khoảng cách từ nó tới mỗi trọng tâm m_j với *j=1...k*. Sau đó tìm trọng tâm gần nhất đối với mỗi đối tượng;
- *Bước 3*: Đối với mỗi *j=1...k*, cập nhật trọng tâm cụm m_j bằng cách tính trung bình cộng của các vector đối tượng dữ liệu;
- *Bước 4*: Lặp các Bước 2 và Bước 3 cho đến khi các trọng tâm của cụm không thay đổi.

Thuật toán K-Means được chứng minh là hội tụ và có độ phức tạp tính toán là: $O((n \times k \times d) \times \tau \times T^{flop})$. Trong đó: *n* là số đối tượng dữ liệu; *k* là số cụm dữ liệu; *d* là số chiều; τ là số vòng lặp; và T^{flop} là thời gian để thực hiện một phép tính cơ sở như phép tính nhân, chia… Như vậy, do K-Means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn. Tuy nhiên, nhược điểm của K-Means là chỉ áp dụng với dữ liệu có thuộc tính số và khám phá ra các cụm có dạng hình cầu, K-Means còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu. Hình 2 mô phỏng về một số hình dạng cụm dữ liệu khám phá được bởi thuật toán K-Means.



Hình 2. Mô phỏng kết quả phân cụm bằng thuật toán K-Means

Hơn nữa, chất lượng phân cụm của thuật toán K-Means phụ thuộc nhiều vào các tham số đầu vào như số cụm *k* và *k* trọng tâm khởi tạo ban đầu. Trong trường hợp các trọng tâm khởi tạo ban đầu quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm của K-Means là rất thấp, nghĩa là các cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế. Trên thực tế, người ta chưa có một giải pháp tối ưu nào để chọn các tham số đầu vào. Giải pháp thường được sử dụng nhất là thử nghiệm với các giá trị đầu vào *k* khác nhau rồi sau đó chọn giải pháp tốt nhất.

2.2. Phân cụm mờ: Thuật toán K-Means mờ

Các thực thể trong thế giới thực hay các khái niệm trừu tượng thường là các đối tượng phức tạp. Các đối tượng này chứa một tập nhất định các thông tin về đối tượng và các hành vi của chính đối tượng đó. Thông tin về đối tượng được gọi là thuộc tính đối tượng và được xác định bởi giá trị cụ thể. Chúng ta có thể thấy rằng, tùy thuộc vào mục tiêu phân cụm mà tính chất quan trọng của mỗi thuộc tính là khác nhau. Do đó, chúng ta cần đánh giá tính quan trọng của từng thuộc tính trong đối tượng để thu được kết quả phân cụm tốt hơn. Cụ thể là cung cấp một giá trị trọng số ω trong độ đo *F* để thể hiện mức độ quan trọng của thuộc tính. Phương pháp này được gọi là phân cụm mờ, do Friguiand và Nasraoui (2004) cũng như Chan, Ching, Ng, và Huang (2004) đề xuất. Độ đo *F* được tính như công thức (1).

$$F(T,W,C)=\sum_{l=1}^{k}\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}\omega_{li}^{\beta}(x_{ji}-c_{li})^2 \quad (1)$$

Trong đó: $\sum_{l=1}^{k}\tau_{lj}=1,(1\le j\le n),\tau_{lj}\in\{0,1\};\sum_{i=1}^{m}\omega_{li}=1,0\le\omega_{li}\le 1,(1\le l\le k)$; *n* là số phần tử trong cụm; *m* là số thuộc tính của phần tử; *k* là số cụm; và c_{li} là phần tử trung tâm của cụm ($1 \le i \le n$).

Thuật toán K-Mean mờ được mô tả như sau:

- *Đầu vào*: Cơ sở dữ liệu gồm *n* đối tượng và hằng số cụm *k*;
- *Đầu ra*: Các cụm C_i (*i=1…k*) sao cho hàm tiêu chuẩn *F* đạt giá trị tối thiểu;
- Bước 1: Chọn *k* đối tượng m_j (*j=1…k*) là trọng tâm ban đầu của *k* cụm (*ngẫu nhiên hoặc theo kinh nghiệm*); Khởi tạo trọng số $\omega_{li} = 1/m$ ($1 \le i \le n$; $1 \le l \le m$);
- *Bước 2*: Tính toán τ theo công thức (2);

$$\tau_{lj}=\begin{cases}1,\sum_{i=1}^{m}\omega_{li}(c_{li}-x_{ji})^2\le\sum_{i=1}^{m}\omega_{zi}(c_{zi}-x_{ji})^2\\0,\sum_{i=1}^{m}\omega_{li}(c_{li}-x_{ji})^2>\sum_{i=1}^{m}\omega_{zi}(c_{zi}-x_{ji})^2\end{cases} \quad (2)$$

- *Bước 3*: Tính hàm *F* theo công thức (3);

$$F(T,W,C)=\sum_{l=1}^{k}\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}\omega_{li}^{\beta}(x_{ji}-c_{li})^2 \quad (3)$$

- *Bước 4*: Cập nhật lại trọng tâm *C* theo công thức (4);

$$c_{li} = \frac{\sum_{j=1}^{n} \tau_{lj} x_{ji}}{\sum_{j=1}^{n} \tau_{lj}} \tag{4}$$

- *Bước 5*: Cập nhật lại ω theo công thức (5);

$$\omega_{li} = \sum_{i=1}^{m} \left[\frac{\sum_{j=1}^{n} \tau_{lj} (c_{li} - x_{ji})^2}{\sum_{j=1}^{n} \tau_{lj} (c_{lt} - x_{jt})^2} \right]^{1/(\beta-1)} \tag{5}$$

- *Bước 6*: Lặp lại các bước từ Bước 2 đến Bước 5 cho đến khi hàm *F* là nhỏ nhất.

Dựa vào công thức tính ω như trên, ta nhận thấy trong một số trường hợp giá trị ω có thể không tính toán được khi mẫu số bằng 0. Để giải quyết vấn đề này nhóm tác giả Huang, Ng, Rong, và Li (2005) đề xuất thuật toán "An Entropy weighting K-Means algorithm (EWKM)" nhằm khắc phục hạn chế khi tính toán ω bằng cách xây dựng hàm *F* cải tiến như công thức (6).

$$F(T,W,C) = \sum_{l=1}^{k} \left[\sum_{j=1}^{n} \sum_{i=1}^{m} \tau_{lj} \omega_{li}^{\beta} (x_{ji} - c_{li})^2 + \gamma \sum_{i=1}^{m} \omega_{li} \log \omega_{li} \right] \tag{6}$$

Mặc dù phương pháp EWKM đã giải quyết được vấn đề mẫu số bằng 0 khi tính toán giá trị ω nhưng nó vẫn còn hạn chế đó là chưa tách được các lớp một cách rõ rệt. Điều này dẫn đến kết quả là phần tử giữa các cụm có thể nằm gần nhau gây nên sự chồng chéo và không chính xác.

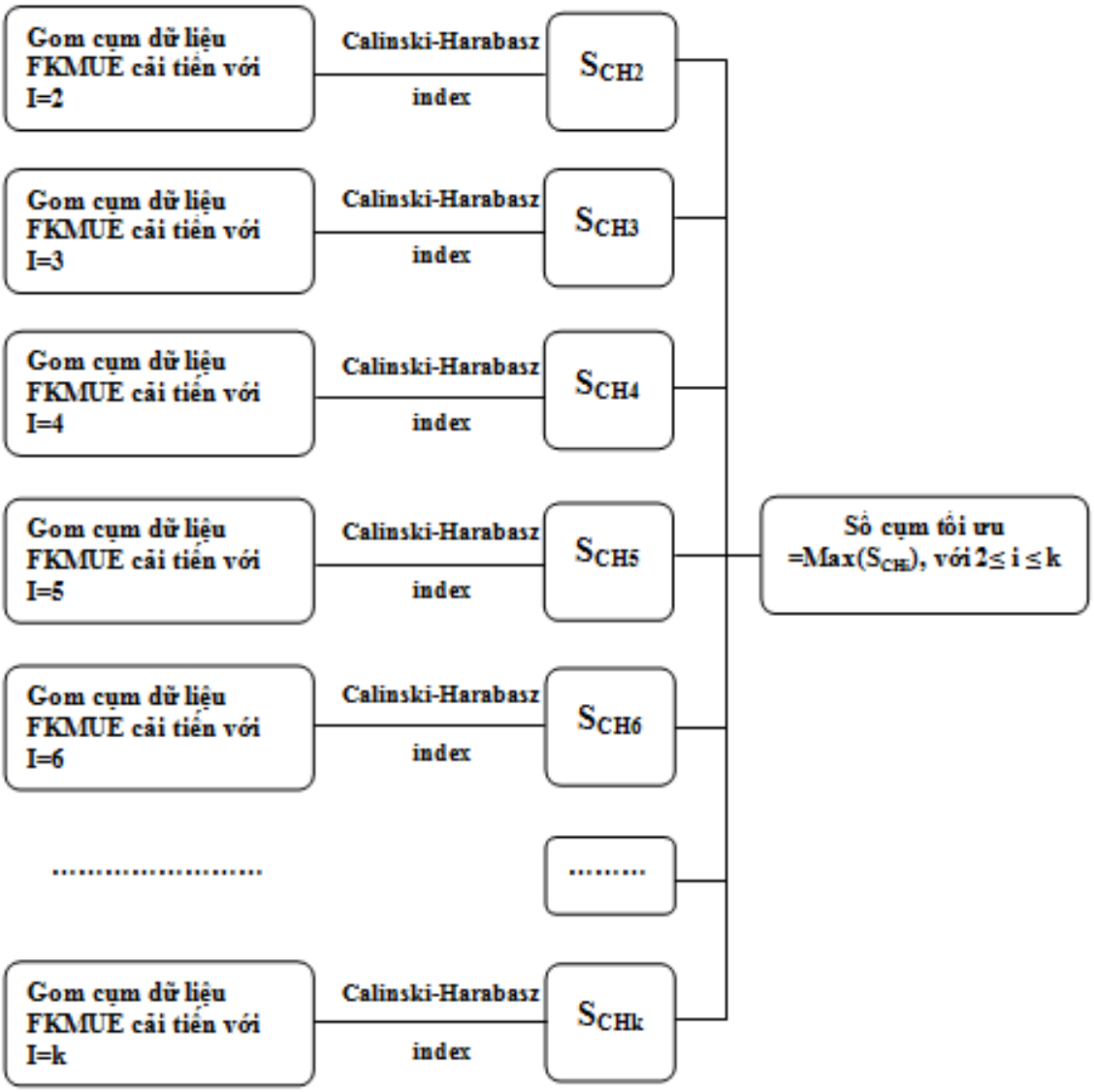
3. THUẬT TOÁN ĐỀ XUẤT

3.1. Mô hình kết hợp

Các thuật toán K-Means và K-Means mờ được khảo sát cho đến thời điểm này là một bước đột phá trong lĩnh vực phân cụm dữ liệu nhưng vẫn còn tồn tại một số khuyết điểm. Để khắc phục hạn chế của K-Means mờ, nhiều thuật toán cải tiến đã lần lượt ra đời và một trong những thuật toán cải tiến gần đây nhất được biết đến đó thuật toán phân cụm K-Means mờ sử dụng độ đo trọng số *Entropy* (*Fuzzy K-Means algorithm for clustering using Entropy - FKMUE*) của Li và Chen (2008, 2010). Mục tiêu của thuật toán này là điều chỉnh hàm chí phí *F* trong công thức (6) bằng cách thêm vào một biến liên quan đến khoảng cách trung bình giữa các phần tử và khoảng cách trung bình giữa các cụm với nhau. Điều này không những giúp cho các mẫu trong cùng một cụm gom lại gần nhau mà còn hướng đến việc phân chia các cụm cách xa nhau nhất.

Trong phần này sẽ trình bày cải tiến thuật toán phân cụm dựa vào sự kết hợp thuật toán K-Means mờ sử dụng độ đo trọng số *Entropy* và phương pháp đánh giá cụm theo chỉ số *Calinski-Harabasz*. Ưu điểm của phương pháp này không những phân chia cụm hiệu quả, có độ chính xác cao mà còn có khả năng đo lường cụm, đánh giá cụm nhằm tìm ra được số cụm tối ưu đủ đáp ứng cho các nhu cầu thực tiễn. Thuật toán cải tiến được chia thành hai giai đoạn như sau (Hình 3):

- *Giai đoạn 1*: Sử dụng thuật toán FKMUE để phân cụm tập dữ liệu đầu vào;
- *Giai đoạn 2*: Sau khi phân hoạch dữ liệu thành từng cụm bằng thuật toán FKMUE, sử dụng độ đo *Calinski-Harabasz* để đánh giá tính hiệu quả ứng với số cụm vừa tìm được. Quá trình trên sẽ được lặp đi lặp lại cho các số cụm khác nhau và số cụm có giá trị S_{CH} lớn nhất được chọn làm kết quả sau cùng.



Hình 3. Mô hình kết hợp giữa phương pháp FKMUE và *Calinski-Harabasz*

3.2. Thuật toán đề xuất

Hầu hết các thuật toán phân cụm mờ cho đến thời điểm này thường mang tính chất ngẫu nhiên khi chọn tâm ban đầu cho các cụm. Vì vậy trong một số trường hợp có thể xảy ra hiện tượng tâm của các cụm được chọn nằm cạnh nhau hoặc có tính chất tương tự nhau. Điều này dẫn đến kết quả là chi phí lặp lại cho số lần phân chia cụm là rất lớn làm giảm hiệu năng của hệ thống. Để giải quyết vấn đề này, nhóm tác giả đã làm một số điều chỉnh cho việc chọn tâm cho *k* cụm ban đầu bằng cách tính toán khoảng cách sao cho tâm của cụm tiếp theo có khoảng cách xa nhất so với tâm của các cụm đã chọn trước đó. Ngoài ra, do dữ liệu trong thực tế thường không đồng nhất và hay bị

dư thừa nên thường dẫn đến việc phân cụm có độ chính xác không cao. Để loại bỏ những thuộc tính dư thừa, không tham gia vào quá trình phân cụm cũng như đảm bảo dữ liệu luôn được nhất quán, chúng tôi tiến hành chuẩn hóa dữ liệu về gốc tọa độ theo công thức (7) như sau:

$$x_{ji} = \left| \frac{x^{orginal} - \min\limits_{t} x_{ti}}{\max\limits_{t} x_{ti} - \min\limits_{t} x_{ti}} \right| \ (1 \le i \le m) \tag{7}$$

Thuật toán đề xuất có các bước như sau:

- *Đầu vào*: Cơ sở dữ liệu gồm *n* đối tượng và các giá trị đầu vào *m, l, maxValue = 0, k, γ,* và số lần lặp tối đa *s* và trọng số ban đầu của các thuộc tính $\omega_{li} = 1/m$;
- *Đầu ra*: Số cụm có giá trị S_{CH} lớn nhất;
- *Bước 1*: Chuẩn hóa thuộc tính của các mẫu về gốc tọa độ theo công thức (7);
- *Bước 2*: Đưa tất cả các mẫu sau khi được chuẩn hóa vào tập *H* và gán tập *C* (chứa tâm của mỗi cụm);
- *Bước 3*: Chọn ngẫu nhiên một mẫu bất kỳ trong *H* và đưa vào *C*, đồng thời loại bỏ mẫu này ra khỏi *H*. Mẫu này được xem là tâm của cụm đầu tiên;
- *Bước 4*: Tìm một mẫu tiếp theo trong *H* đặt vào *C* sao cho khoảng cách từ mẫu này đến tất cả các mẫu trong *C* là xa nhất;
- *Bước 5*: Nếu số tâm trong *C* bằng với *k* thì chuyển sang Bước 7, ngược lại quay sang Bước 4;
- *Bước 6*: Tính giá trị *T* theo công thức (8) và giá trị *F(T,W,C)* theo công thức (9);

$$\rho = \frac{1}{n} \sum_{j=1}^{n} \tau_{lj} d(x_j, c_l) \times \sum_{i=1}^{m} \sqrt{\omega_i} \tag{8}$$

$$F(T,W,C) = \sum_{l=1}^{k} \left[\frac{\sum\limits_{j=1}^{n} \sum\limits_{i=1}^{m} \tau_{lj} \omega_{li}^{\beta} (x_{ji} - c_{li})^2}{\sum\limits_{i=1}^{m} (c_{li} - \overline{x_i})^2} \right] + \sum_{l=1}^{k} \left[\gamma \sum_{i=1}^{m} \omega_{li} \log \omega_{li} \right] \tag{9}$$

- *Bước 7*: Cập nhật *C* theo công thức (10) và *W* theo công thức (11);

$$\rho = \frac{\sqrt{m}}{n} \sum_{l=1}^{k} \sum_{j=1}^{n} \tau_{lj} d(x_j, c_l) \tag{10}$$

$$\omega_{lt} = \frac{1}{\sum\limits_{i=1}^{m} exp\left(\frac{-\psi_{lt}}{\gamma} \right)} exp\left(\frac{-\psi_{lt}}{\gamma} \right) \tag{11}$$

Trong đó $\psi_{lt} = \frac{1}{\sum_{i=1}^{m}(c_{li} - \overline{x_i})^2} \sum_{j=1}^{n} \tau_{lj}(c_{lt} - \overline{x_{jt}})^2$

- *Bước 8*: Lặp lại Bước 6 và Bước 7 cho đến khi F không thể nhỏ hơn được nữa hoặc số lần lặp bằng s, khi đó chuyển sang Bước 10;

- *Bước 9*: Đánh giá số cụm vừa tìm theo công thức (12) để được giá trị S_{CH};

$$S_{CH} = \frac{n-k}{k-1} \times \frac{\sum_{l=1}^{k} n_l(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T}{\sum_{l=1}^{k}\sum_{p=1}^{r}\sum_{x_l \in C_i}\left(x_{ip} - \bar{x}_{lp}\right)} \quad (12)$$

- *Bước 10*: So sánh giá trị S_{CH} với *maxValue*. Nếu S_{CH} lớn hơn *maxValue* thì gán *maxValue* = S_{CH};

- *Bước 11*: Tăng giá trị $l = l + 1$. Nếu $l \leq k$ thì chuyển sang Bước 3. Ngược lại số cụm được chọn l là số cụm có giá trị S_{CH} lớn nhất.

Trong phần này, nhóm tác giả đã trình bày một phương pháp mới cho việc phân cụm bằng cách kết hợp thuật toán FKMUE và độ đo *Calinski-Harabasz*. So với các thuật toán K-Means mờ truyền thống tính chất mờ thường đi kèm với cả đối tượng vì vậy không phản ánh rõ mức độ quan trọng của thuộc tính khi xem xét gom cụm đối tượng. Trái lại thuật toán K-Means mờ với trọng số *Entropy* chú trọng đến tính chất mờ trên từng thuộc tính, cho phép người dùng điều chỉnh để tăng hệ số mờ với các thuộc tính quan trọng và giảm giá trị với các thuộc tính không cần thiết. Thêm vào đó sự điều chỉnh của hàm F để làm gia tăng khoảng cách giữa các cụm và thu hẹp giữa các mẫu trong cụm đã góp phần rất lớn vào mức độ chính xác, giúp cho các cụm rõ ràng và có nhiều ý nghĩa hơn. Cuối cùng tính hiệu quả của phương pháp còn được đo lường qua tiêu chuẩn đánh giá cụm bằng phương pháp *Calinski-Harabasz*, giúp cho kết quả phân cụm được chính xác, có ý nghĩa và tin cậy hơn. Với những ưu điểm vừa được trình bày ở trên, có thể thấy rằng phương pháp đề xuất của chúng tôi có đủ tính khả thi cho bài toán phân cụm dữ liệu, đáp ứng nhu cầu ứng dụng thực tiễn trong nhiều lĩnh vực khác nhau.

4. THỰC NGHIỆM

Trong phần này, nhóm tác giả sử dụng bộ dữ liệu *Iris* từ kho dữ liệu học máy của Đại học California (Lichman, 2013). Bộ dữ liệu được rút gọn gồm năm thuộc tính, trong đó bốn thuộc tính kiểu số bao gồm: Chiều dài đài hoa; Chiều rộng đài hoa; Chiều dài cánh hoa; và Chiều rộng cánh hoa. Một thuộc tính còn lại là tên của loài hoa Iris (*Iris Setosa*; *Iris Versicolour*; *Iris Virginica*). Tỷ lệ phân chia cho mỗi loài trong ba loài Iris là 33.3%.

Bảng 1. Mô tả dữ liệu loài hoa Iris

Đặc trưng	Giá trị nhỏ nhất	Giá trị lớn nhất	Giá trị trung bình
Chiều dài đài hoa	4.3	7.9	5.84
Chiều rộng đài hoa	2.0	4.4	3.05
Chiều dài cánh hoa	1.0	6.9	3.76
Chiều rộng cánh hoa	0.1	2.5	1.20

Để đánh giá độ chính xác của thuật toán phân cụm, tỷ lệ lỗi được tính như công thức (13). Tỷ lệ lỗi càng nhỏ, thuật toán có độ chính xác càng cao, có tính phù hợp tốt.

Tỷ lệ lỗi = (số mẫu gom cụm sai/ Tổng số mẫu kiểm tra) x 100% (13)

Bảng 2. Tỷ lệ lỗi

Trường hợp	ω				Số mẫu	Số mẫu sai	Tỷ lệ lỗi (%)
	Thuộc tính 1	Thuộc tính 2	Thuộc tính 3	Thuộc tính 4			
1	0.40	0.40	0.10	0.10	150	82	55.0
2	0.10	0.10	0.40	0.40	150	5	3.0
3	0.10	0.40	0.40	0.10	150	16	10.6
4	0.40	0.10	0.10	0.40	150	12	8.0
5	0.25	0.25	0.25	0.25	150	7	4.6

Từ kết quả thống kê tỷ lệ lỗi trong Bảng 2 ta thấy quá trình phân cụm tốt nhất khi ω = {0.1, 0.1, 0.4, 0.4} với tỷ lệ chính xác là 97% và quá trình phân cụm xấu nhất khi ω = {0.4, 0.4, 0.1, 0.1} với tỷ lệ chính xác là 45%. Điều này cho thấy trọng số của các thuộc tính đóng góp rất lớn vào tính chính xác của quá trình phân cụm. Ngoài ra với tỷ lệ phân cụm chính xác lên đến 97% nếu chọn các thuộc tính phù hợp cho chúng ta kết luận rằng phương pháp phân cụm mờ với trọng số *Entropy* cải tiến là phương pháp phân cụm chính xác, hiệu quả, mang lại hiệu quả khi áp dụng cho các bài toán ứng dụng thực tiễn.

Một trong những khuyết điểm của các bài toán phân cụm đó là việc chọn số cụm đều dựa vào kinh nghiệm của người sử dụng. Nó không có cơ sở để đánh giá việc chọn số cụm nào là tối ưu. Có thể nói, phân cụm dữ liệu là một ví dụ của phương pháp học không giám sát, vì không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế có thể coi phân cụm dữ liệu là một cách học bằng quan sát (*learning by observation*). Trong phương pháp này ta không thể biết kết quả các cụm thu được sẽ thế nào khi bắt đầu quá trình. Vì vậy, thông thường cần có một chuyên gia để đánh giá các cụm thu được. Trong bài viết này, nhóm tác giả sử dụng phương pháp đánh giá theo độc đo *Calinski-Harabasz*. Kết quả đầu ra của quá trình phân cụm với trọng số *Entropy* cải tiến sẽ là đầu vào cho phương pháp đánh giá cụm *Calinski-Harabasz*. Quá trình trên sẽ được thực hiện với việc chọn nhiều số cụm khác nhau, sau đó kết quả đầu ra của mỗi lần phân cụm sẽ được đánh giá để chọn ra cụm tối ưu nhất như trong Bảng 3. Kết quả cho thấy với bộ dữ liệu *Iris*, số cụm k =3 là tối ưu nhất vì có tỷ lệ là 1.1438.

Bảng 3. Kết quả đánh giá cụm theo phương pháp *Calinski-Harabasz*

STT	Số cụm	Tỷ lệ đánh giá
1	3	1.1438
2	4	0.8229
3	5	0.7417
4	6	0.7221
5	7	0.6879

5. KẾT LUẬN

Trong bài viết này, nhóm tác giả đã trình bày cải tiến thuật toán phân cụm mờ bằng cách kết hợp phương pháp phân cụm K-Means mờ với độ đo trọng số *Entropy* và đánh giá cụm theo chỉ số *Calinski-Harabasz*. Các kết quả thực nghiệm cho thấy việc chọn hệ số phù hợp phương pháp phân cụm mới với trọng số *Entropy* là phương pháp hiệu quả có độ chính xác cao. Ngoài ra, để nâng cao độ tin cậy cho hệ thống phân cụm, chúng tôi đã sử dụng phương pháp đánh giá cụm theo chỉ số *Calinski-Harabasz* như là thước đo tính chính xác cho đầu ra khi chọn số cụm. Trong tương lai, nhóm tác giả cố gắng nghiên cứu thực nghiệm trên nhiều bộ dữ liệu khác nhau và đưa cải tiến trên vào ứng dụng thực tế.

TÀI LIỆU THAM KHẢO

Chan, Y., Ching, W., Ng, M. K., & Huang, J. Z. (2004). An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition, 37*(5), 943-952.

Friguiand, H., & Nasraoui, O. (2004). Unsupervised learning of prototypes and attribute weights. *Pattern Recognition, 37*(3), 567-581.

Hoàng, X. H., & Nguyễn, T. X. H. (2006). Mở rộng thuật toán gom cụm K-means cho dữ liệu hỗn hợp. *Tạp chí Tin học và Điều khiển học, 22*(3), 267-274.

Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in K-Means type clustering. *IEEE Transactions on Pattern Analysis, 27*(5), 657-668.

Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting K-Means algorithm for subspace clustering of high dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering, 19*(8), 1026-1041.

Li, T., & Chen, Y. (2008). *An improved K-means algorithm for clustering using Entropy weighting measures*. Paper presented at The 7th World Congress on Intelligent Control and Automation, China.

Li, T., & Chen, Y. (2010). Fuzzy K-Means incremental clustering based on K-Center and vector quantization. *Journal of Computer, 5*(11), 1670-1677.

Lichman, M. (2013). *UCI machine learning repository.* California, USA: University of California. Retrieved from http://archive.ics.uci.edu/ml.