AN IMPROVEMENT OF FUZZY CLUSTERING METHOD WITH FUZZY PARAMETER FOR EACH DATA CLUSTER

Nguyen Hong Tan[1*], Le Khanh Duong[1], Tran Thi Ngan[2]
[1]TNU - University of Information and Communication Technology, [2]Thuyloi University

ARTICLE INFO

Received: 09/9/2021
Revised: 29/11/2021
Published: 30/11/2021

KEYWORDS

Fuzzy clustering
Fuzzy parameters
Cluster data
Performance
Rating measure

ABSTRACT

Recently, fuzzy clustering is widely used to group data. Fuzzy clustering is studied and applicable in many technical applications like crime hot spot detection, tissue differentiation in medical images, software quality prediction etc. The researches on fuzzy clustering focuses mainly on the objective function to increase the performance of the clustering process. However, the fuzzy parameter is an important factor affecting the performace of the clustering process. The fuzzy parameter is used to reflect the degree of fuzzifier. In this study, the research team focuses on improving the fuzzy clustering algorithm with fuzzy parameters for each data cluster. Main contributions of the paper: i) building an improved algorithm from fuzzy clustering algorithm; ii) building a fuzzy parameter caculation function for each data cluster; iii) Execuction and evaluation the improved algorithm compared to other algorithms in the same field. The experimental results of study show that the improved algorithm is more efficient than the original algorithm.

MỘT CẢI TIẾN PHÂN CỤM MỜ VỚI THAM SỐ MỜ CHO TỪNG CỤM DỮ LIỆU

Nguyễn Hồng Tân[1*], Lê Khánh Dương[1], Trần Thị Ngân[2]
[1]Trường Đại học Công nghệ Thông tin và Truyền thông - ĐH Thái Nguyên
[2]Trường Đại học Thủy lợi

THÔNG TIN BÀI BÁO

Ngày nhận bài: 09/9/2021
Ngày hoàn thiện: 29/11/2021
Ngày đăng: 30/11/2021

TỪ KHÓA

Phân cụm mờ
Tham số mờ
Cụm dữ liệu
Hiệu năng
Độ đo đánh giá

TÓM TẮT

Phân cụm mờ được sử dụng nhiều trong thời gian gần đây để phân nhóm dữ liệu. Phân cụm mờ thường được nghiên cứu nhiều trong lĩnh vực phát hiện điểm nóng tội phạm, phân biệt mô trong ảnh y tế, dự đoán chất lượng phần mềm... Các nghiên cứu phân cụm mờ tập trung chủ yếu vào việc cải tiến hàm mục tiêu để tăng hiệu năng của quá trình phân cụm. Tuy nhiên để tăng hiệu năng của quá trình phân cụm, một yếu tố có ảnh hưởng lớn đó là tham số mờ. Khi đó, tham số mờ được sử dụng để phản ảnh mức độ mờ hóa. Do vậy, trong nghiên cứu này, nhóm nghiên cứu tập trung cải tiến từ thuật toán phân cụm mờ với tham số mờ cho từng cụm dữ liệu. Đóng góp chính của bài báo: i) Xây dựng một thuật toán cải tiến từ thuật toán phân cụm mờ; ii) Xây dựng hàm tính tham số mờ cho từng cụm dữ liệu; iii) Cài đặt, đánh giá thuật toán cái tiến so với các thuật toán cùng loại. Kết quả thực nghiệm của nghiên cứu cũng cho thấy thuật toán cải tiến cho hiệu năng tốt hơn so với thuật toán gốc ban đầu.

DOI: https://doi.org/10.34238/tnu-jst.4970

* Corresponding author. *Email: nhtan@ictu.edu.vn*

1. Giới thiệu

Phân cụm dữ liệu là việc phân chia các điểm dữ liệu về các cụm dữ liệu, sao cho 2 điểm dữ liệu có độ tương đồng cao thuộc về cùng một cụm, 2 điểm dữ liệu có độ tương đồng thấp thuộc về 2 cụm khác nhau [1]. Các thuật toán phân cụm chia thành 2 loại cơ bản: phân cụm cứng và phân cụm mờ. Trong phân cụm cứng, mỗi điểm dữ liệu thuộc về một cụm xác định. Với phân cụm mờ, mỗi điểm dữ liệu có thể thuộc về nhiều cụm dữ liệu khác nhau với một độ thuộc vào từng cụm là khác nhau. Các bài toán trong thế giới thực thường rất khó phân chia rõ ràng 1 điểm dữ liệu thuộc về cụm nào, do vậy thời gian gần đây các phương pháp phân cụm mờ được sử dụng nhiều. Các phương pháp phân cụm mờ đã ứng dụng trong phân loại tài liệu [2], phân đoạn ảnh [3], phân loại phương tiện tham gia giao thông [4], dự báo thời tiết [5].

Các phương pháp nghiên cứu mới phát triển từ thuật toán phân cụm mờ (Fuzzy C-Mean: FCM) [6] thường được giới thiệu để khắc phục và nâng cao khả năng phân cụm của thuật toán này. Một số nghiên cứu nhằm bổ sung thêm các thông tin để trợ giúp phân cụm mờ, khi đó người ta phát triển phân cụm bán giám sát mờ [7]-[9]. Một nhóm tác giả phát triển phân cụm mờ với các tập mờ nâng cao [10], [11]. Một số nhóm phát triển phân cụm mờ cho bài toán ứng dụng như phân đoạn ảnh thì bổ sung thêm thông tin không gian [12], bổ sung thông tin đặc trưng nha khoa để phân đoạn ảnh nha khoa [13]. Các nghiên cứu trên đều thực hiện với tham số mờ bằng2 (m=2), mà tập trung vào việc điều chỉnh các thành phần trong cụm để làm tăng hiệu suất, từ đó làm tăng chất lượng của phân cụm dữ liệu. Tuy nhiêu, một yếu tố có ảnh hưởng đến quá trình nâng cao chất lượng cụm là tham số mờ chưa được đề cập đến. Năm 2020, tác giả Trần Đình Khang và cộng sự [14] đã nghiên cứu đề cập đến việc lựa chọn một cách tính tham số mờ với từng điểm dữ liệu để làm tăng chất lượng của quá trình phân cụm dữ liệu.

Trong nghiên cứu này, nhóm nghiên cứu đưa ra một cải tiến thuật toán phân cụm mờ với tham số mờ cho từng cụm dữ liệu. Khi đó sẽ thấy được các mối quan hệ giữa trọng số mũ m trong thuật toán phân cụm và bán kính, kích thước mỗi cụm, cũng như khoảng cách tương đối giữa các phần tử đang xét vào tâm mỗi cụm. Nhóm nghiên cứu, cài đặt đánh giá thử nghiệm thuật toán cải tiến với thuật toán phân cụm mờ và thuật toán phân cụm mờ với tham số mờ của từng điểm dữ liệu.

Các phần tiếp theo của bài báo được cấu trúc như sau: mục 2 chúng tôi trình bày các nghiên cứu liên quan để phát triển trong nghiên cứu này. Mục 3, chúng tôi trình bày chi tiết cải tiến phân cụm mờ với tham số mờ cho từng cụm dữ liệu. Mục 4, chúng tôi trình bày các kết quả thực nghiệm, đánh giá so sánh của thuật toán cải tiến phân cụm mờ với tham số mờ cho từng cụm dữ liệu với một số thuật toán khác. Cuối cùng, là kết luận chỉ ra những đóng góp của bài báo và hướng phát triển của bài báo.

2. Nghiên cứu liên quan

2.1. Thuật toán Fuzzy C-Mean

Thuật toán phân cụm mờ được Bezdek [6] đề xuất dựa trên độ thuộc u_{kj} của phần tử dữ liệu X_k từ cụm j. Hàm mục tiêu được xác định như sau:

$$J = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\|X_i - V_j\right\|^2 \rightarrow min \quad (1)$$

Trong đó:

- m là tham số mờ hóa
- C là số cụm dữ liệu; N là số phần tử dữ liệu.
- u_{ij} là độ thuộc của phần tử dữ liệu X_i từ cụm j.
- $X_i \in R^r$ là phần tử thứ k của $X = \{X_1, X_2, ..., X_N\}$.
- V_j là tâm của cụm j.

Khi đó ràng buộc của (1) là:

$$\sum_{j=1}^{C} u_{ij} = 1; \qquad u_{ij} \in [0,1]; \qquad \forall i = \overline{1,N} \tag{2}$$

Sử dụng phương pháp Lagrange giải tối ưu hàm mục tiêu (1) với rằng buộc (2), xác định được tâm của cụm dựa vào (3) và độ thuộc dựa vào (4).

$$V_j = \frac{\sum_{i=1}^{N} u_{ij}^m X_i}{\sum_{i=1}^{N} u_{ij}^m} \qquad \forall j = \overline{1,C} \tag{3}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|X_i - V_k\|}{\|X_i - V_j\|} \right)^{\frac{1}{m-1}}} \qquad \forall i = \overline{1,N};\ \forall j = \overline{1,C} \tag{4}$$

Khi đó thuật toán Fuzzy C-means như sau (xem bảng 1).

Bảng 1. *Thuật toán Fuzzy C-means*

Input	Tập dữ liệu X gồm N phần tử trong không gian r chiều; số cụm C; mờ hóa m; ngưỡng ε; số lần lặp lớn nhất MaxStep>0.
Output	Ma trận U và tâm cụm V.
FCM	
1	Khởi tạo t=0
2	$u_{ij}^{(t)} \leftarrow random;\ \left(i = \overline{1,N}; j = \overline{1,C}\right)$ thỏa mãn điều kiện (2)
3	Repeat
3.1	t=t+1
3.2	Tính $V_j^{(t)}; \left(j = \overline{1,C}\right)$ bởi công thức (3)
3.3	Tính $u_{ij}^{(t)}; \left(i = \overline{1,N}; j = \overline{1,C}\right)$bởi công thức (4)
3.4	Until $\left\|U^{(t)} - U^{(t-1)}\right\| \le \varepsilon$ hoặc t > MaxStep

2.2. Thuật toán phân cụm mờ với tham số mờ cho từng điểm dữ liệu

Bảng 2. *Thuật toán MCFCM*

Input	Tập dữ liệu X gồm N phần tử , số cụm C, m_i, ngưỡng ε, số lần lặp tối đa maxStep > 0.
Output	Ma trận U và tâm cụm V.
MCFCM	
1	Khởi tạo t=0
2	Khởi tạo ngẫu nhiên V^t
3	Repeat
3.1	t=t+1
3.2	Tính ma trận U^t dựa trên công thức $u_{ij} = \frac{1}{\sum_{j=1}^{C} \left(\frac{\|X_i - V_k\|}{\|X_i - V_j\|} \right)^{\frac{2}{m_i - 1}}}$
3.3	Tính ma trận V^t dựa trên công thức $V_k = \frac{\sum_{i=1}^{N} u_{ik}{}^{m_i} X_i}{\sum_{k=1}^{N} u_{kj}{}^{m_i}}$
3.4	Until $\left\|V^{(t)} - V^{(t-1)}\right\| \ge \varepsilon$ or t > MaxStep

Trong thuật toán phân cụm mờ với nhiều tham số mờ được Trần Đình Khang và cộng sự [14] xây dựng dựa trên thuật toán phân cụm mờ với mỗi điểm dữ liệu xây dựng một tham số mờ riêng cho từng điểm dữ liệu. Khi đó, việc xác định tham số mờ được xác định bởi công thức (5).

$$m_i = m_1 + (m_2 - m_1)\left(\frac{S_i - S_{min}}{S_{max} - S_{min}}\right)^{\alpha};\ i = \overline{1,N} \tag{5}$$

Trong đó:

- m_1, m_2 là các giá trị cận trên và cận dưới của tham số m_i ($1 \le m_1 \le m_2$)

- α là tham số đầu vào.
- $S_i = \sum_{j=1}^{N/C} D_{ij}$. ; $D_{ij} = \| X_i - X_j \|$ $\left(\forall i, j = \overline{1,N}\right)$.
- $S_{max} = max_{i \in N}(S_i)$. $S_{min} = min_{i \in N}(S_i)$

Thuật toán phân cụm mờ với tham số mờ cho từng điểm dữ liệu (MCFCM) như sau (Bảng 2).

3. Cải tiến phân cụm mờ với tham số mờ theo từng cụm dữ liệu

Trong mục này, chúng tôi trình bày nội dung cải tiến phân cụm mờ với tham số mờ cho các cụm dữ liệu. Khi đó các mối quan hệ giữa tham số mờ trong thuật toán phân cụm và bán kính, kích thước mỗi cụm, cũng như khoảng cách tương đối giữa các điểm dữ liệu với tâm từng cụm. Khi xét độ thuộc của một phần tử x_i nào đó vào cụm j:

- Nếu bán kính cụm j lớn thì m nên nhỏ và ngược lại, khi bán kính cụm j nhỏ thì m nên lớn để có thể tối ưu hóa vùng mờ tối đa về phía cụm đó.
- Nếu khoảng cách tương đối giữa điểm x_i vào cụm j lớn so với khoảng cách tới các cụm khác thì m nên nhỏ và ngược lại, khi khoảng cách tương đối giữa điểm x_i vào cụm j nhỏ so với khoảng cách tới các cụm khác thì m nên lớn vì khả năng x_i thuộc vào cụm j là cao hơn.
- Nếu một điểm có xu hướng thuộc vào một cụm nào đó sẵn, ví dụ như điểm thuộc vùng tập trung đông các điểm khác thì m nên nhỏ vì khi đó, khả năng x_i được xét vào một cụm cụ thể nào đó cao hơn các điểm khác.
- Mô hình này đang thực nghiệm dựa trên kinh nghiệm.

Khi đó việc xác định mô hình được thực hiện như sau:

Hàm mục tiêu của phân cụm mờ với tham số mờ theo từng cụm được xác định bởi công thức (6).

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m_j} \left\| X_i - V_j \right\|^2 \rightarrow min \tag{6}$$

Với các ràng buộc xác định bởi (2).

Với đề xuất tính giá trị tham số m_j bởi công thức (7).

$$m_j = 1 + \frac{2}{log\, u_j + log|C_j|} \; j = \overline{1,C} \tag{7}$$

Trong đó:

$$u_j = \frac{1}{|C_j|} \sum_{X_i \in C_j} \left(\frac{1}{\sum_{k=1}^{C} \left(\frac{\|X_i - V_k\|}{\|X_i - V_j\|} \right)^{\frac{1}{m-1}}} \right) \qquad j = \overline{1,C} \tag{8}$$

- $|C_j|$: là lực lượng của các phân tử ở cụm j;
- C_j là tập các điểm dữ liệu có độ thuộc lớn nhất là cụm j.

Sử dụng phương pháp Lagrange giải tối ưu hàm mục tiêu (6) với ràng buộc (2).

$$\begin{cases} L = J - \sum_{i=1}^{N} \lambda_i \left(\sum_{j=1}^{C} u_{ij} - 1 \right) \\ \frac{\partial J}{\partial V_j} = 0 \\ \frac{\partial L}{\partial u_{ij}} = 0 \end{cases}$$

Xác định được tâm của cụm dựa vào (9) và độ thuộc dựa vào (10).

$$u_{ij} = \frac{1}{\sum_{j=1}^{C} \left(\frac{\|X_i - V_k\|}{\|X_i - V_j\|} \right)^{\frac{2}{m_j - 1}}} \tag{9}$$

$$V_k = \frac{\sum_{i=1}^{N} {u_{ik}}^{m_j} X_i}{\sum_{k=1}^{N} {u_{kj}}^{m_j}} \quad (10)$$

Thuật toán phân cụm mờ với tham số mờ của các cụm (MCFCM-C) được thực hiện như sau (Bảng 3).

Bảng 3. *Thuật toán MCFCM-C*

Input	Tập dữ liệu X gồm N phần tử , số cụm C, m_j, ngưỡng ε, số lần lặp tối đa maxStep > 0.
Output	Ma trận U và tâm cụm V.
MCFCM-C	
1	Khởi tạo t=0
2	Khởi tạo ngẫu nhiên V^t
3	Repeat
3.1	t=t+1
3.2	Tính ma trận U^t dựa trên công thức (9)
3.3	Tính ma trận V^t dựa trên công thức $V_k = \frac{\sum_{i=1}^{N} {u_{ik}}^{m_j} X_i}{\sum_{k=1}^{N} {u_{kj}}^{m_j}}$
3.4	Until $\left\|V^{(t)} - V^{(t-1)}\right\| \geq \varepsilon$ or t > MaxStep

4. Kết quả thực nghiệm

Dữ liệu thực nghiệm được là các bộ dữ liệu Liver, Diabetes, Arhythmia lấy trên kho dữ liệu chuẩn UCI Machine Learning Repository. Các độ đo dùng để đánh giá và so sánh hiệu năng của các thuật toán được cài đặt trong bài báo này gồm Davies-Bouldin (DB) [15], PBM [15], Partition Coefficient (PC) [16] and Classification Entropy (CE) [16], Rand index (RI) [14]. Thuật toán cải tiến phân cụm mờ với nhiều tham số mờ theo từng cụm (MCFCM-C) được cài đặt cùng với các thuật toán đã có bao gồm thuật toán phân cụm mờ với nhiều tham số (MCFCM [14]), phân cụm mờ (FCM [6]).

Kết quả thực nghiệm với các độ đo đánh giá hiệu năng giữa thuật toán phân cụm mờ với nhiều tham số mờ theo từng cụm (trình bày mục 3) với các thuật toán phân cụm cùng loại trên các bộ dữ liệu Liver, Diabetes, Arhythmia thể hiện ở bảng 4. Kết quả thực nghiệm cũng cho thấy: với độ đo DB thì phương pháp MCFCM-C tốt hơn 2 phương pháp FCM, MCFCM ở cả 3 bộ dữ liệu; với độ đo PBM thì phương pháp MCFCM-C tốt hơn 2 phương pháp FCM, MCFCM ở cả 3 bộ dữ liệu; với độ đo CE thì phương pháp MCFCM tốt ở 2 bộ dữ liệu Liver, Arrhythmia còn phương pháp MCFCM-C tốt ở bộ dữ liệu Diabetes, với độ đo RI thì phương pháp MCFCM-C tốt ở 2 bộ dữ liệu Diabetes, Arrhythmia còn phương pháp MCFCM-C tốt ở bộ dữ liệu Liver. Dựa trên 3 độ đo đánh giá hiệu năng của thuật toán thì hiệu năng của thuật toán MCFCM-C cải tiến cho giá trị tốt với 9/12 giá trị đánh giá và thuật toán MCFCM cho giá trị tốt với 3/12 giá trị đánh giá. Với kết quả này thì thuật toán MCFCM-C tốt hơn các thuật toán so sánh là FCM và MCFCM.

Bảng 4. *Kết quả thực nghiệm trên bộ dữ liệu Wine*

Data	Độ đo	FCM	MCFCM	MCFCM-C
Liver	DB-	4,78	3,89	**3,78**
	PBM+	193,27	273,47	**372,37**
	CE-	0,243	**0,223**	0,235
	RI+	0,637	**0,643**	0,641
Diabetes	DB-	3,27	3,19	**3,07**
	PBM+	283,63	344,76	**382,37**
	CE-	0,321	0,289	**0,273**
	RI+	0,837	0,874	**0,883**
Arrhythmia	DB-	4,92	4,67	**4,52**
	PBM+	482,73	492,38	**503,47**
	CE-	0,427	**0,352**	0,398
	RI+	0,746	0,782	**0,802**

5. Kết luận

Trong nghiên cứu này, chúng tôi tập trung vào việc cải tiến thuật toán Fuzzy C-Mean với tham số mờ theo từng cụm. Đóng góp chính của nhóm tác giả là cải tiến thuật toán Fuzzy C-Mean với tham số mờ theo từng cụm, xây dựng cách tính tham số mờ theo từng cụm. Đồng thời, chúng tôi đã cài đặt thực nghiệm để đánh giá so sánh giữa MCFCM-C với 2 thuật toán FCM và MCFCM. Các kết quả thử nghiệm cho thấy, thuật toán MCFCM-C cho hiệu năng chất lượng cụm tốt hơn so với thuật toán FCM, MCFCM. Trong nghiên cứu tiếp theo, chúng tôi sẽ phân tích với nhiều loại dữ liệu để đưa ra khuyến cáo phù hợp với dữ liệu loại gì, xây dựng cách tính tham số mờ phù hợp với từng loại dữ liệu.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] Bezdek and C. James, *Pattern recognition with fuzzy objective function algorithms*, Springer Science & Business Media, 2013.

[2] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," *Information Processing & Management*, vol. 57.2, 2020, Art. no. 102034.

[4] W. Ding, M. Abdel-Basset, and H. Hawash, "RCTE: A Reliable and Consistent Temporal-ensembling Framework for Semi-supervised Segmentation of COVID-19 Lesions," *Information sciences*, vol. 578, pp. 559-573, 2021.

[5] L. Cao, C. Wang, and J. Li, "Vehicle detection from highway satellite images via transfer learning," *Information sciences*, vol. 366, pp. 177-187, 2016.

[6] H. T. Pham and H. S. Le,"Some novel hybrid forecast methods based on picture fuzzy clustering for weather nowcasting from satellite image sequences", *Applied Intelligence*, vol 46.1, pp. 1-15, 2017.

[7] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-mean clustering algorithm," *Comput. Geosci*, vol. 10, pp. 191-203, 1984.

[8] E. Yasunori, H. Yukihiro, Y. Makito, and M. Sadaaki, "On semi-supervised fuzzy c-means clustering," *Fuzzy Systems, FUZZ-IEEE 2009. IEEE International Conference on*, IEEE, 2009, pp. 1119-1124.

[9] X. Yin, T. Shu, and Q. Huang, "Semi-supervised fuzzy clustering with metric learning and entropy regularization," *Knowledge-Based Systems*, vol. 35, pp. 304-311, 2012.

[10] H. Zhang and J. Lu, "Semi-supervised fuzzy clustering: A kernel-based approach," *Knowledge-Based Systems*, vol. 22, no. 6, pp. 477-481, 2009.

[11] H. S. Le, "Generalized picture distance measure and applications to picture fuzzy clustering," *Applied Soft Computing*, vol. 46(C), pp. 284-295, 2016.

[12] E. H. Ruspini, J. C. Bezdek, and J. M. Keller, "Fuzzy clustering: A historical perspective," *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 45-55, 2019.

[13] L. T. Ngo, D. S. Mai, and W. Pedrycz, "Semi-supervising Interval Type-2 Fuzzy C-Means clustering with spatial information for multi-spectral satellite image classification and changedetection," *Computers & geosciences*, vol. 83, pp. 1-16, 2015.

[14] M. T. Tran, T. N. Tran, and H. S. Le, "A novel semi-supervised fuzzy clustering method based on interactive fuzzy satisficing for dental X-ray image segmentation," *Applied Intelligence*, vol. 45, no. 2, pp. 402-428, 2016.

[15] T. D. Khang, N. D. Vuong, M. K. Tran, and M. Fowler, "Fuzzy C-Means Clustering Algorithm with Multiple Fuzzification Coefficients," *Algorithms*, vol. 13, no. 7, p. 158, 2020.

[16] L. Vendramin, R. J. Campello, and E. R. Hruschka, "Relative clustering validity criteria: A comparative overview," *Statistical analysis and data mining: the ASA data science Journal*, vol. 3-4, pp. 209-235, 2010.