Một phương pháp phân cụm bán giám sát mờ đồng huấn luyện trên dữ liệu đa khung nhìn

Hoàng Thị Cành[1,2], Phùng Thế Huân[1], Vũ Thuỳ Trang[3], Phạm Huy Thông[4], Nguyễn Như Sơn[5], Lê Trường Giang[6]
[1] Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên, Thái Nguyên, Việt Nam
[2] Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Hà Nội, Việt Nam
[3] Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội, Hà Nội, Việt Nam
[4] Viện Công nghệ thông tin, Đại học Quốc gia Hà Nội, Hà Nội, Việt Nam
[5] Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Hà Nội, Việt Nam
[6] Trường Đại học Công nghiệp Hà Nội, Hà Nội, Việt Nam
Tác giả liên hệ: Phùng Thế Huân, pthuan@ictu.edu.vn
Ngày nhận bài: 30/06/2023, ngày sửa chữa: 10/09/2023, ngày duyệt đăng: 26/09/2023
Định danh DOI: 10.32913/mic-ict-research-vn.v2023.n2.1230

Tóm tắt: Trong thực tế hiện nay, dữ liệu đa khung nhìn ngày càng phổ biến. Dữ liệu đa khung nhìn (Multi-view data) đề cập đến loại dữ liệu bao gồm nhiều quan điểm hoặc góc nhìn về một đối tượng. Dữ liệu trong mỗi khung nhìn riêng lẻ có thuộc tính cụ thể thực hiện nhiệm vụ khám phá tri thức riêng và cung cấp các thông tin về cùng một vấn đề với độ chính xác và độ tin cậy khác nhau. Việc kết hợp nhiều loại thông tin từ các khung nhìn, có thể thu được biểu diễn đầy đủ và chính xác hơn về các đối tượng, dẫn đến việc phân tích dữ liệu và ra quyết định được cải thiện. Phân cụm đa khung nhìn là hướng nghiên cứu đã thu hút được sự quan tâm của các nhà khoa học trong nhiều năm gần đây. Tuy nhiên, chưa có nghiên cứu nào tập trung vào phân cụm bán giám sát mờ kết hợp thuật toán đồng huấn luyện để đánh giá độ chính xác và chất lượng phân cụm trên tập dữ liệu đa khung nhìn. Bài báo này đề xuất một phương pháp mới trong phân cụm bán giám sát mờ, sử dụng thuật toán đồng huấn luyện trên dữ liệu đa khung nhìn thu thập từ một nguồn dữ liệu. Đồng thời, bài báo cũng cung cấp các kết quả thực nghiệm để đánh giá tính hiệu quả và độ chính xác của thuật toán đề xuất.

Từ khóa: *Dữ liệu đa khung nhìn, phân cụm đa khung nhìn, phân cụm bán giám sát mờ, thuật toán đồng huấn luyện*

Title: **A Semi-Supervised Fuzzy Clustering Co-Training Approach on Multi-View Data**

Abstract: In today's practical reality, multi-view data is increasingly prevalent. Multi-view data refers to a type of data that encompasses multiple perspectives or viewpoints of an object. Data within each individual view possesses specific attributes dedicated to knowledge discovery and provides information on the same subject with varying degrees of accuracy and reliability. Combining various types of information from different views can yield a more comprehensive and accurate representation of objects, thereby improving data analysis and decision-making. Multi-view clustering has emerged as a research direction that has garnered the interest of scientists in recent years. However, there has been no research focusing on semi-supervised fuzzy clustering combined with co-training algorithms to assess the accuracy and quality of clustering on multi-view datasets. This paper proposes a novel method in semi-supervised clustering, utilizing co-training algorithms on multi-view data collected from a data source. Additionally, the paper provides experimental results to evaluate the effectiveness and accuracy of the proposed algorithm.

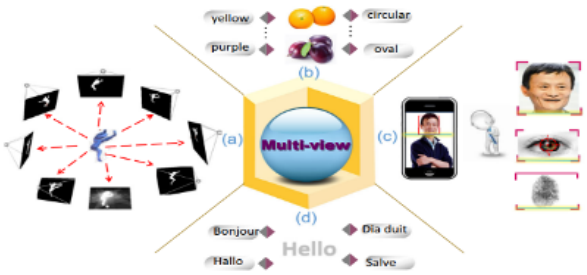
Keywords: *Multi-view data, multi-view clustering, semi-supervised fuzzy clustering, co-training algorithm.*

I. MỞ ĐẦU

Quá trình phân cụm dữ liệu đóng vai trò quan trọng trong khai phá dữ liệu nhằm tìm kiếm, phát hiện các nhóm dữ liệu theo đặc trưng để cung cấp tri thức cho quá trình ra quyết định. Phân cụm dữ liệu là quá trình phân chia các phần tử dữ liệu về các cụm dựa trên mức độ tương đồng giữa chúng, sao cho mức độ tương đồng giữa các phần tử trong cùng một cụm được tối đa hóa và mức độ tương đồng giữa các phần tử khác cụm được cực tiểu hóa [1]. Hầu hết các thuật toán phân cụm hiện tại được thiết kế cho dữ liệu một khung nhìn, trong khi các bài toán thực tế hiện nay nhu cầu sử dụng dữ liệu đa khung nhìn rất phổ biến.

Dữ liệu đa khung nhìn (Multi-view data) đề cập đến loại dữ liệu bao gồm nhiều quan điểm hoặc góc nhìn về một đối

tượng. Trong đó, dữ liệu được biểu diễn dưới nhiều khung nhìn khác nhau, mỗi khung nhìn cung cấp các thông tin và thuộc tính khác nhau về dữ liệu. Dữ liệu trong các khung nhìn được thu thập từ các phương thức, nguồn, các dạng khác nhau hoặc được quan sát từ các góc nhìn khác nhau [2]. Dữ liệu đa khung nhìn được áp dụng trong nhiều bài toán thực tế như: học máy, xử lý ảnh, kinh doanh, y tế, hệ thống tư vấn, khoa học dữ liệu [3], v.v. Điều này thúc đẩy sự cần thiết phát triển các phương pháp phân cụm tiên tiến, nhằm khám phá tri thức từ các bộ dữ liệu đa khung nhìn.

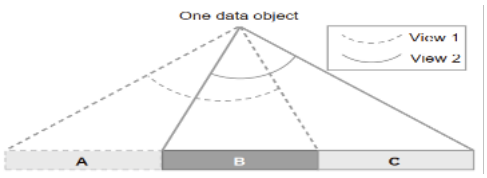


Hình 1. Ví dụ về dữ liệu đa khung nhìn [4]

Phân cụm đa khung nhìn (Multi-view Clustering - MvC) là một phương pháp phân cụm dữ liệu dựa trên việc sử dụng nhiều khung nhìn độc lập, nhằm tìm ra các nhóm dữ liệu tương đồng với nhau trong các khung nhìn. Rõ ràng, việc tích hợp thông tin từ nhiều khung nhìn và phát hiện ra tri thức tiềm ẩn chung được chia sẻ bởi nhiều khung nhìn mang lại lợi ích lớn cho việc phân cụm dữ liệu [5]. So sánh với các phương pháp phân cụm trên một khung nhìn thì phân cụm đa khung nhìn có một số ưu điểm vượt trội đó là nâng cao chất lượng phân cụm, giảm thiểu sự phụ thuộc vào khung nhìn và xử lý dữ liệu phức tạp [6]. Tuy nhiên, phân cụm đa khung nhìn đang đối diện với nhiều thách thức như đòi hỏi các khung nhìn độc lập, tốn nhiều chi phí cho việc thu thập, xử lý, lưu trữ dữ liệu từ nhiều khung nhìn và khó khăn trong việc kết hợp các phương pháp phân cụm khác nhau[6].

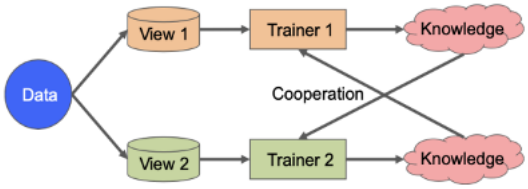
Trên thế giới, nhiều nhà nghiên cứu đang nỗ lực tìm cách giải quyết những khó khăn đã nêu ở trên. Vào năm 2004, Bickel và Scheffer [7] tiên phong trong nghiên cứu phương pháp phân cụm đa khung nhìn. Họ đã mở rộng phương pháp phân cụm K-means và Expectation Maximization (EM) để phù hợp với môi trường đa khung nhìn và xử lý dữ liệu văn bản với hai khung nhìn độc lập có điều kiện. Nhiều phương pháp phân cụm đa khung nhìn đã được đề xuất dựa trên nghiên cứu quan trọng này [8,9,10]. Phương pháp phân cụm bán giám sát sâu đa khung nhìn (DMSC - Deep Multi-view Semi-supervised Clustering) [11] do Rui Chen và cộng sự đề xuất, có thể tăng hiệu suất của phân cụm đa khung nhìn một cách hiệu quả bằng cách lấy thông tin được giám sát yếu có trong các ràng buộc theo cặp mẫu và bảo vệ các thuộc tính của dữ liệu đa khung nhìn. Trong một nghiên cứu khác, Li B và các cộng sự [12] đã đưa ra một phương pháp phân cụm đa khung nhìn mới dựa trên phân tích ma trận phi tuyến, với mục tiêu tối ưu hóa sự giống nhau giữa các khung nhìn. Phương pháp này sử dụng một mô hình ma trận phi tuyến hóa không âm (NMF - Non-negative matrix factorization) để tách các yếu tố ẩn chung từ các khung nhìn khác nhau của dữ liệu.

Hai nguyên tắc quan trọng đảm bảo sự thành công của thuật toán MvC chính là nguyên tắc bổ sung và nguyên tắc đồng thuận [6]. Nguyên tắc bổ sung khẳng định rằng nên sử dụng nhiều khung nhìn khác nhau để mô tả các đối tượng dữ liệu một cách toàn diện và chính xác hơn. Mặc dù mỗi khung nhìn riêng lẻ đã cung cấp đủ thông tin cho một nhiệm vụ khám phá tri thức cụ thể tuy nhiên, các khung nhìn khác nhau thường chứa các thông tin bổ sung cần được khai thác. Nguyên tắc đồng thuận nhằm mục đích tối đa hóa tính nhất quán giữa nhiều khung nhìn. Hình 2 minh hoạ nguyên tắc bổ sung và nguyên tắc đồng thuận [6]. Giả sử một đối tượng dữ liệu có hai khung nhìn, được ánh xạ vào một không gian dữ liệu ẩn: (i) thành phần A và thành phần C tồn tại trong khung nhìn riêng, phần A trong khung nhìn 1 và phần C trong khung nhìn 2, thể hiện tính bổ sung của hai khung nhìn; (ii) thành phần B của đối tượng được chia sẻ bởi cả hai khung nhìn, thể hiện sự đồng thuận giữa hai khung nhìn.



Hình 2. Minh hoạ nguyên tắc bổ sung và nguyên tắc đồng thuận

Thuật toán đồng huấn luyện, được giới thiệu bởi Blum và Mitchell vào năm 1998 [13], đã trở thành một công nghệ mang tính tiên phong và được coi là một trong những phương pháp phổ biến nhất trong lĩnh vực học đa khung nhìn. Mục tiêu của phương pháp này là tối đa hoá sự đồng thuận qua các khung nhìn để đạt được sự đồng thuận rộng nhất và cải thiện hiệu suất phân cụm.



Hình 3. Quy trình chung của thuật toán đồng huấn luyện [6]

Mặc dù đã có nhiều nghiên cứu trước đây tìm cách giải quyết các thách thức của phân cụm đa khung nhìn, nhưng các nghiên cứu lại chủ yếu tập trung sử dụng phương pháp phân cụm rõ. Mà đối với các loại dữ liệu phức tạp, khi các điểm dữ liệu có thể cùng lúc thuộc vào nhiều cụm với trọng số khác nhau thì việc áp dụng các thuật toán phân cụm rõ là hoàn toàn không hiệu quả. Trong khi đó, phân cụm bán giám sát mờ cho phép phân tích dữ liệu và phát hiện các cụm dữ liệu khác nhau. Điều này hỗ trợ trong việc đưa ra quyết định về phân tích dữ liệu, phát hiện dữ liệu nhiễu và tìm kiếm những đặc trưng quan trọng trong dữ liệu. Vì vậy, bài báo này nhằm lấp đầy khoảng trống nghiên cứu bằng cách đề xuất một phương pháp phân cụm bán giám sát mờ sử dụng thuật toán đồng huấn luyện trên dữ liệu đa khung nhìn có tên gọi SSFCC (Semi Supervised Fuzzy Multi-View Co-training Clustering).

Phần còn lại của bài báo được tổ chức như sau: Phần II trình bày một số phương pháp tiếp cận về phân cụm đa khung nhìn, phân cụm đồng huấn luyện. Phần III trình bày phương pháp đề xuất. Phần IV trình bày kết quả thực nghiệm và phân tích, đối sánh. Một số kết luận và định hướng nghiên cứu trong bài báo tiếp theo được đưa ra trong phần kết luận.

II. MỘT SỐ PHƯƠNG PHÁP TIẾP CẬN

1. Phân cụm K-means đa khung nhìn

Thuật toán K-means là một trong những thuật toán phân cụm dữ liệu phổ biến nhất và có khả năng xử lý tốt các tập dữ liệu quy mô lớn. K-means đã được áp dụng rộng rãi trong nhiều lĩnh vực như: phân tích mạng xã hội, kinh doanh, y tế v.v. Mặc dù nó đã được nghiên cứu kỹ trong nhiều thập kỷ qua, nhưng nhiều biến thể của K-means liên tục được đưa ra.

Đối với các bài toán phân cụm đa khung nhìn, thuật toán K-means được mở rộng như sau [14]:

Giả sử $\mathcal{X} = \{X^{(1)}, X^{(2)}, ..., X^{(V)}\} \in \mathbb{R}^V$ là tập dữ liệu tổng quát của tất cả V khung nhìn. Hàm mục tiêu của K-means đa khung nhìn có dạng như sau:

$$J = \sum_{v=1}^{V} \mu_v^{\gamma} J_v \tag{1}$$

Với các ràng buộc:

$$\begin{cases} \mu_v \geq 0 \\ \sum_{v=1}^{V} \mu_v = 1 \\ \gamma > 1 \end{cases}$$

Trong đó, μ_v là trọng số của view thứ v, γ là số mũ điều chỉnh μ_v , J_V tướng ứng với hàm mục tiêu K-means của view thứ v:

$$J_v = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik} ||x_i^{(v)} - v_k^{(v)}||^2 \tag{2}$$

Phương pháp phân cụm K-means đa khung nhìn sử dụng thông tin từ nhiều khung nhìn để cải thiện quá trình phân cụm. Bằng cách kết hợp các khung nhìn, thuật toán sẽ gán các điểm dữ liệu vào cùng một cụm, đảm bảo tính nhất quán trong quá trình phân cụm giữa các khung nhìn và tăng cường hiệu suất phân cụm.

2. Phân cụm quang phổ đa khung nhìn đồng huấn luyện

Phân cụm quang phổ đa khung nhìn cho phép khám phá các cấu trúc cụm tiềm ẩn bằng cách kết hợp thông tin từ nhiều đồ thị. Kumar và đồng nghiệp đã đưa ra một phương pháp phân cụm quang phổ đa khung nhìn sử dụng ý tưởng đồng huấn luyện [15]. Phương pháp này tuân theo tính nhất quán của việc học đa khung nhìn, cho phép kết hợp thông tin từ nhiều khung nhìn để cải thiện quá trình phân cụm. Trong đó, mỗi khung nhìn cung cấp các nhãn giống nhau cho tất cả các mẫu dữ liệu. Nhờ đó, phương pháp này có thể sử dụng vector riêng của một khung nhìn để "gán nhãn" cho một khung nhìn khác và ngược lại.

Thuật toán quang phổ đa khung nhìn đồng huấn luyện được triển khai thông qua các bước như sau [15]:

Bước 1: Sử dụng phân cụm quang phổ trên từng khung nhìn, thu được các vector riêng gọi là U_1 và U_2.

Bước 2: Phân cụm các điểm sử dụng U_1 và sử dụng phân cụm này để điều chỉnh cấu trúc đồ thị trong khung nhìn 2.

Bước 3: Phân cụm các điểm sử dụng U_2 và sử dụng phân cụm này để điều chỉnh cấu trúc đồ thị trong khung nhìn 1.

Bước 4: Quay lại **Bước 1** và lặp lại một số vòng lặp.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Bài báo đề xuất một giải pháp phân cụm bán giám sát mờ đa khung nhìn sử dụng thuật toán đồng huấn luyện. Nó được áp dụng cho dữ liệu đa khung nhìn được thu thập từ một nguồn có những đặc điểm sau: số lượng bản ghi trên mỗi khung nhìn là bằng nhau, số lượng thuộc tính trên mỗi khung nhìn có thể khác nhau và quan hệ giữa hai khung nhìn là ánh xạ một – một.

1. Ý tưởng thuật toán

Thuật toán SSFCC được đề xuất bao gồm 2 bước cụ thể như sau:

Bước 1: Phân cụm mờ cho dữ liệu chưa được gán nhãn trên mỗi khung nhìn. Trong bước này, thuật toán FCM (Fuzzy C-Means) được sử dụng độc lập cho từng khung nhìn (viewA và viewB) để chia các điểm dữ liệu vào các cụm. Kết quả

của bước này gồm các tâm cụm và độ thuộc tương ứng trên từng khung nhìn. Độ thuộc được ký hiệu là: $\bar{u}^A$ (độ thuộc của điểm dữ liệu trên viewA so với các tâm cụm) và $\bar{u}^B$ (độ thuộc của điểm dữ liệu trên viewB so với các tâm cụm).

Bước 2: Phân cụm bán giám sát mờ đồng huấn luyện đa khung nhìn. Trong bước này, kết quả của viewA được sử dụng làm dữ liệu huấn luyện cho viewB và ngược lại. Mục tiêu là tối đa hóa sự đồng thuận chéo qua tất cả các khung nhìn và đạt được sự đồng thuận rộng nhất. Đầu vào của bước này là các thành phần bán giám sát $\bar{u}^A$ và $\bar{u}^B$ thu được từ **Bước 1**. Đầu ra của bước này là các giá trị tâm cụm cuối cùng và giá trị độ thuộc tương ứng trên mỗi khung nhìn.

2. Chi tiết thuật toán SSFCC

Dựa trên ý tưởng đã trình bày ở phần trước, phần này sẽ trình bày mô hình hoá của phương pháp đề xuất. Hàm mục tiêu của phương pháp được biểu diễn bởi ba thành phần, như sau:

$$
\begin{aligned}
J(u^A, v^A, u^B, v^B) &= \sum_{k=1}^{N}\sum_{j=1}^{c} {u_{kj}^A}^2 ||x_k^A - v_j^A||^2 \\
&+ \sum_{k=1}^{N}\sum_{j=1}^{c} {u_{kj}^B}^2 ||x_k^B - v_j^B||^2 \\
&+ \sum_{k=1}^{N}\sum_{j=1}^{c} (u_{kj}^A - u_{kj}^B)^2 (||x_k^A - v_j^A||^2 + ||x_k^B - v_j^B||^2) \\
&+ \sum_{k=1}^{N}\sum_{j=1}^{c} (u_{kj}^A - \bar{u}_{kj}^A)^2 ||x_k^A - v_j^A||^2 \\
&+ \sum_{k=1}^{N}\sum_{j=1}^{c} (u_{kj}^B - \bar{u}_{kj}^B)^2 ||x_k^B - v_j^B||^2 \rightarrow Min
\end{aligned} \tag{3}
$$

Với các ràng buộc:

$$
\begin{cases}
\sum_{j=1}^{c} u_{kj}^A = \sum_{j=1}^{c} u_{kj}^B = 1 \\
u_{kj} \in [0,1]
\end{cases}
$$

Chú thích:

- x_k^A, x_k^B : đại diện cho điểm dữ liệu thứ k trên viewA và viewB tương ứng
- v_j^A, v_j^B : đại diện cho tâm cụm thứ j trên viewA và viewB tương ứng
- u_{kj}^A, u_{kj}^B : đại diện cho độ thuộc của điểm dữ liệu thứ k ở cụm thứ j trên viewA và viewB tương ứng
- $\bar{u}_{kj}^A, \bar{u}_{kj}^B$: đại diện cho độ thuộc của điểm dữ liệu thứ k ở cụm thứ j sau khi sử dụng thuật toán FCM trên viewA và viewB tương ứng
- $||x_k^A - v_j^A||^2, ||x_k^B - v_j^B||^2$: đại diện cho khoảng cách từ điểm dữ liệu thứ k đến tâm cụm thứ j trên viewA và viewB tương ứng

Hàm mục tiêu (3), thể hiện rõ thành phần của phân cụm mờ, phân cụm đồng huấn luyện và phân cụm bán giám sát, cụ thể:

- Thành phần thể hiện phân cụm mờ:

$$\sum_{k=1}^{N}\sum_{j=1}^{c} {u_{kj}^A}^2 ||x_k^A - v_j^A||^2$$

và

$$\sum_{k=1}^{N}\sum_{j=1}^{c} {u_{kj}^B}^2 ||x_k^B - v_j^B||^2$$

- Thành phần thể hiện phân cụm đồng huấn luyện:

$$\sum_{k=1}^{N}\sum_{j=1}^{c} (u_{kj}^A - u_{kj}^B)^2 (||x_k^A - v_j^A||^2 + ||x_k^B - v_j^B||^2)$$

- Thành phần thể hiện phân cụm bán giám sát:

$$\sum_{k=1}^{N}\sum_{j=1}^{c} (u_{kj}^A - \bar{u}_{kj}^A)^2 ||x_k^A - v_j^A||^2$$

và

$$\sum_{k=1}^{N}\sum_{j=1}^{c} (u_{kj}^B - \bar{u}_{kj}^B)^2 ||x_k^B - v_j^B||^2$$

Các tâm cụm và độ thuộc của bài toán tối ưu (3) được tính toán như sau:

Tâm cụm v_j^A và v_j^B có công thức như sau:

$$
v_j^A = \frac{\sum_{k=1}^{N} \left[{u_{kj}^A}^2 + (u_{kj}^A - u_{kj}^B)^2 + (u_{kj}^A - \bar{u}_{kj}^A)^2 \right] . x_k^A}{\sum_{k=1}^{N} \left[{u_{kj}^A}^2 + (u_{kj}^A - u_{kj}^B)^2 + (u_{kj}^A - \bar{u}_{kj}^A)^2 \right]} \tag{4}
$$

$$
v_j^B = \frac{\sum_{k=1}^{N} \left[{u_{kj}^B}^2 + (u_{kj}^A - u_{kj}^B)^2 + (u_{kj}^B - \bar{u}_{kj}^B)^2 \right] . x_k^B}{\sum_{k=1}^{N} \left[{u_{kj}^B}^2 + (u_{kj}^A - u_{kj}^B)^2 + (u_{kj}^B - \bar{u}_{kj}^B)^2 \right]} \tag{5}
$$

Phương pháp nhân tử Lagrange được sử dụng để xác định độ thuộc u_{ij}^A và u_{ij}^B , được xác định bằng công thức sau:

$$
u_{kj}^A = \frac{\frac{\lambda_A}{2} + \Delta_{kj}^A}{3{d_{kj}^A}^2 + {d_{kj}^B}^2} \tag{6}
$$

Trong đó:

$$\Delta_{kj}^A = u_{kj}^B ({d_{kj}^A}^2 + {d_{kj}^B}^2) + \bar{u}_{kj}^A {d_{kj}^A}^2$$

$$
\frac{\lambda_A}{2} = \frac{1 - \sum_{i=1}^{c} \frac{\Delta_{ki}^A}{3{d_{ki}^A}^2 + {d_{ki}^B}^2}}{\sum_{i=1}^{c} \frac{1}{3{d_{ki}^A}^2 + {d_{ki}^B}^2}}
$$

$$u_{kj}^{B} = \frac{\frac{\lambda_B}{2} + \Delta_{kj}^{B}}{3{d_{kj}^{B}}^2 + {d_{kj}^{A}}^2} \quad (7)$$

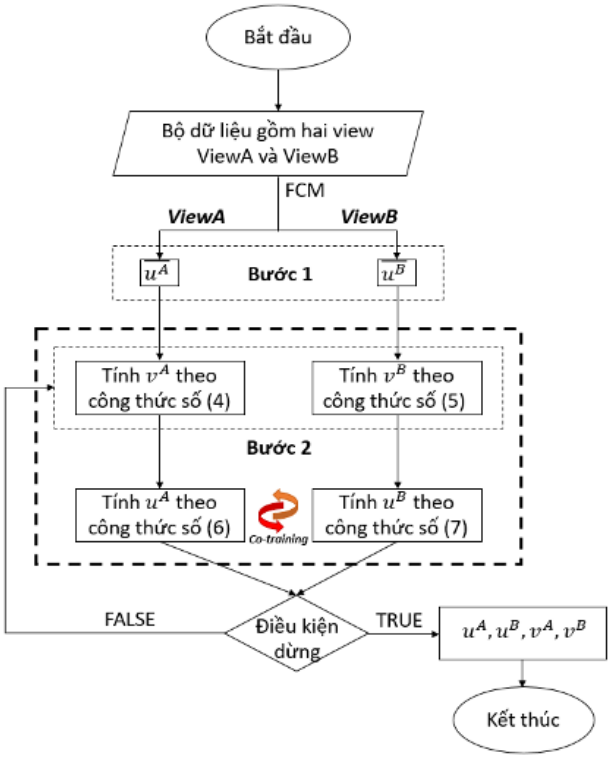
Trong đó:

$$\Delta_{kj}^{B} = u_{kj}^{A}({d_{kj}^{A}}^2 + {d_{kj}^{B}}^2) + \bar{u}_{kj}^{B}{d_{kj}^{B}}^2$$

$$\frac{\lambda_B}{2} = \frac{1 - \sum_{i=1}^{c} \frac{\Delta_{ki}^{B}}{3{d_{ki}^{B}}^2 + {d_{ki}^{A}}^2}}{\sum_{i=1}^{c} \frac{1}{3{d_{ki}^{B}}^2 + {d_{ki}^{A}}^2}}$$

3. Sơ đồ thuật toán SSFCC

Hình 4 dưới đây mô tả sơ đồ của thuật toán SSFCC.



Hình 4. Sơ đồ thuật toán

4. Thuật toán SSFCC

Thuật toán 1 SSFCC

Input:
- Tập dữ liệu X gồm hai view: viewA: $X^A = \{x_1^A, x_2^A, x_3^A, \ldots, x_n^A\}$, viewB: $X^B = \{x_1^B, x_2^B, x_3^B, \ldots, x_n^B\}$.
- Số lượng bản ghi trên viewA và viewB là bằng nhau.
- Số thuộc tính trên viewA và viewB có thể khác nhau.
- Số cụm c
- Số lần lặp Maxstep
- Sai số cho phép ϵ

Output: u^A, u^B, v^A, v^B

BEGIN

Bước 1: Phân cụm mờ cho dữ liệu chưa được gán nhãn trên mỗi khung nhìn

1.1. Áp dụng thuật toán FCM lên viewA thu được $\bar{u}^A$ và $\bar{v}^A$

1.2. Áp dụng thuật toán FCM lên viewB thu được $\bar{u}^B$ và $\bar{v}^B$

Bước 2: Phân cụm bán giám sát mờ đồng huấn luyện đa khung nhìn

2.1. Khởi tạo t = 0

Repeat

2.2. $t = t + 1$

2.3. Cập nhật v^A theo công thức số (4)

2.4. Cập nhật v^B theo công thức số (5)

2.5. Cập nhật u^A theo công thức số (6)

2.6. Cập nhật u^B theo công thức số (7)

Until

$Max\{||u^{A^{(t+1)}} - u^{A^{(t)}}||, ||u^{B^{(t+1)}} - u^{B^{(t)}}||\} \leq \epsilon$ hoặc $t \geq$ Maxstep

END.

IV. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

1. Các điều kiện thực nghiệm

Nhằm kiểm chứng hiệu suất của phương pháp đề xuất, nhóm nghiên cứu đã tiến hành cài đặt mô phỏng trên 9 bộ dữ liệu lấy từ kho dữ liệu học máy UCI [16]. Các bộ dữ liệu bao gồm Australian, Balance-scale, Heart, Iris, Spambase, Tae, Waweform, Wdbc, Wine.

Đối với mỗi bộ dữ liệu, chúng tôi đã chia thành hai view (viewA và viewB) và các thuộc tính trên mỗi view được lựa chọn ngẫu nhiên. Thông tin chi tiết về các bộ dữ liệu được trình bày trong Bảng I.

Việc cài đặt thực nghiệm được thực hiện trên máy tính thương hiệu Apple MacBook Air M1 2020 với cấu hình 8GB/256GB/7-core GPU, ngôn ngữ lập trình Python phiên bản 3.10.

Bảng I
THÔNG TIN CHI TIẾT CÁC BỘ DỮ LIỆU DÙNG CHO THỰC NGHIỆM

ST	Bộ dữ liệu	viewA		viewB		Số cụm
		Số lượng mẫu	Số thuộc tính	Số lượng mẫu	Số thuộc tính	
1	Australian	690	7	690	7	2
2	Balance-scale	625	2	625	2	3
3	Heart	270	6	270	7	2
4	Iris	150	2	150	2	3
5	Spambase	4601	29	4601	28	2
6	Tae	151	2	151	2	3
7	Waweform	5000	20	5000	20	3
8	Wdbc	569	15	569	15	2
9	Wine	178	6	178	7	3

Bảng II
BẢNG GIÁ TRỊ SO SÁNH HIỆU NĂNG VỀ ĐỘ CHÍNH XÁC PHÂN CỤM

Bộ dữ liệu	SSFCC		MKC		CMSC	
	Trung bình	Phương sai	Trung bình	Phương sai	Trung bình	Phương sai
Australian	**0.7109**	0.0037	0.6227	0.0162	0.4644	0.0009
Balance-scale	**0.5511**	0.013	0.549	0.0305	0.4032	0.0038
Heart	0.4517	0.2376	**0.5414**	0.0194	0.1533	0.0516
Iris	**0.8217**	0.0373	0.6411	0.0175	0.1573	0.0339
Spambase	0.3645	0.0563	**0.5144**	0.0143	0.3972	0.0001
Tae	**0.7785**	0.0009	0.458	0.0181	0.3278	0.0001
Waweform	0.3619	0.0214	0.5711	0.0253	**0.5885**	0.0087
Wdbc	**0.8493**	0.0625	0.5683	0.0153	0.2425	0.0097
Wine	0.3401	0.0477	**0.5889**	0.0189	0.2303	0.0126

Bảng III
BẢNG GIÁ TRỊ SO SÁNH HIỆU NĂNG VỀ CHẤT LƯỢNG PHÂN CỤM

Bộ dữ liệu	SSFCC		MKC		CMSC	
	Trung bình	Phương sai	Trung bình	Phương sai	Trung bình	Phương sai
Australian	**1.9306**	1.6895	2.2527	0.0377	3.5724	0.9613
Balance-scale	**1.0102**	0.0001	3.9196	0.1511	5.0247	0.6785
Heart	**2.3391**	0.0004	2.5519	0.9348	2.3409	0.3888
Iris	**0.7246**	2.8865	3.0464	0.0497	6.4878	0.9114
Spambase	**1.5221**	0.9296	2.7543	0.1186	2.2847	0.4525
Tae	**2.8793**	10.0895	2.8909	0.1649	6.3583	1.6738
Waweform	28.6892	0.0504	22.9106	1.937	**2.0502**	0.005
Wdbc	0.6525	0.0014	2.0328	0.0299	**0.5745**	0.0222
Wine	3.2525	2.4586	**2.8765**	0.0191	4.6377	1.8099

Để đánh giá hiệu suất, chúng tôi sử dụng các tiêu chí sau: i) Độ chính xác phân cụm: chúng tôi sử dụng độ chính xác phân cụm CA (Clustering Accuracy) [17]. Độ chính xác phân cụm đo lường mức độ chính xác của việc gán nhãn cho các điểm dữ liệu trong từng cụm. Giá trị CA càng cao cho thấy hiệu suất phân cụm càng tốt. ii). Chất lượng phân cụm: Chúng tôi sử dụng độ đo DB (Davies–Bouldin index) [18]. Chất lượng phân cụm đo lường độ tách biệt và đồng nhất giữa các cụm. Giá trị DB càng nhỏ cho thấy chất lượng phân cụm càng tốt.

2. Kết quả thực nghiệm

Phương pháp SSFCC được so sánh với hai phương pháp khác là Multi-view K-means Clustering (MKC) [14] và Co-trained Multi-view Spectral Clustering (CMSC) [15].

a) Kết quả thực nghiệm đánh giá theo độ chính xác phân cụm

Kết quả đánh giá độ chính xác phân cụm của phương pháp đề xuất (SSFCC) đã được so sánh với hai phương pháp MKC và CMSC trên 9 bộ dữ liệu được trình bày trong bảng II.

Trong Bảng II, phương pháp đề xuất SSFCC đã đạt giá trị tốt nhất trong 5/9 trường hợp (Australian, Balance-scale, Iris, Tae, Wdbc), trong khi phương pháp MKC chỉ có 3/9 trường hợp nhận giá trị tốt nhất (Heart, Spambase, Wine) và phương pháp CMSC chỉ có 1/9 trường hợp nhận giá trị tốt nhất (Waweform). Do vậy, phương pháp SSFCC cho kết quả tốt hơn phương pháp MKC và CMSC trong việc đạt được độ chính xác phân cụm.

b) Kết quả thực nghiệm đánh giá chất lượng phân cụm

Kết quả đánh giá chất lượng phân cụm của phương pháp đề xuất (SSFCC) đã được so sánh với hai phương pháp MKC và CMSC trên 9 bộ dữ liệu được trình bày trong bảng III.

Trong Bảng III, phương pháp SSFCC đã đạt giá trị tốt nhất trong 6/9 trường hợp (Autralian, Balance-scale, Heart, Iris, Spambase, Tae), trong khi phương pháp MKC chỉ có 1/9 trường hợp nhận giá trị tốt nhất (Wine) và phương pháp CMSC chỉ có 2/9 trường hợp nhận giá trị tốt nhất (Waveform, Wdbc). Do vậy, chất lượng phân cụm theo phương pháp SSFCC tốt hơn phương pháp MKC và phương pháp CMSC.

V. KẾT LUẬN

Bài báo này đã đề xuất một mô hình mới có tên gọi SSFCC, đó là một phương pháp phân cụm bán giám sát mờ đa khung nhìn đồng huấn luyện. Mô hình này áp dụng cho dữ liệu đa khung nhìn thu thập từ một nguồn dữ liệu và mang đặc trưng của dữ liệu trên cùng một nguồn. Kết quả thực nghiệm cho thấy phương pháp SSFCC vượt trội hơn so với các phương pháp MKC và CMSC khi đánh giá độ chính xác và chất lượng của việc phân cụm. Điều này khuyến khích và thúc đẩy nhiều nghiên cứu tiếp theo trong lĩnh vực phân cụm đa khung nhìn.

Mặc dù phương pháp SSFCC hiệu quả trong trường hợp dữ liệu trên hai khung nhìn được thu thập từ một nguồn dữ liệu, nhưng nó vẫn còn một số hạn chế. Mô hình có nhiều tham số, quá trình đồng huấn luyện lặp lại nhiều lần dẫn đến thời gian tính toán cao và chưa hiệu quả trong trường hợp dữ liệu được thu thập từ nhiều nguồn dữ liệu với các đặc điểm: số lượng bản ghi trên mỗi khung nhìn khác nhau, số lượng thuộc tính trên mỗi khung nhìn có thể khác nhau và quan hệ giữa hai khung nhìn là ánh xạ nhiều – nhiều.

Để giải quyết các vấn đề liên quan đến dữ liệu đa khung nhìn, đặc biệt là dữ liệu thu thập từ nhiều nguồn khác nhau, cần tiếp tục phát triển các thuật toán mới trong các nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO

[1] Al-Amri, Salem Saleh, and Namdeo V. Kalyankar. "Image segmentation by using threshold techniques." arXiv preprint arXiv:1005.4020 (2010).

[2] Li, Xin, et al. "A multi-view model for visual tracking via correlation filters." Knowledge-Based Systems 113 (2016): 88-99.

[3] Elkahky, Ali Mamdouh, Yang Song, and Xiaodong He. "A multi-view deep learning approach for cross domain user modeling in recommendation systems." Proceedings of the 24th international conference on world wide web. (2015).

[4] Xu, Chang, Dacheng Tao, and Chao Xu. "A survey on multi-view learning." arXiv preprint arXiv:1304.5634 (2013).

[5] Zhu, Zhenfeng, et al. "Shared Subspace Learning for Latent Representation of Multi-View Data." J. Inf. Hiding Multim. Signal Process. 5.3 (2014): 546-554.

[6] Yang, Yan, and Hao Wang. "Multi-view clustering: A survey." Big Data Mining and Analytics 1.2 (2018): 83-107.

[7] Bickel, Steffen, and Tobias Scheffer. "Multi-view clustering." ICDM. Vol. 4. No. 2004. (2004).

[8] Ye, Fanghua, et al. "New approaches in multi-view clustering." Recent applications in data clustering 195 (2018).

[9] Xu, Chang, Dacheng Tao, and Chao Xu. "A survey on multi-view learning." arXiv preprint arXiv:1304.5634 (2013).

[10] Sun, Shiliang. "A survey of multi-view machine learning." Neural computing and applications 23 (2013): 2031-2038.

[11] Chen, Rui, et al. "Deep multi-view semi-supervised clustering with sample pairwise constraints." Neurocomputing 500 (2022): 832-845.

[12] Li, B., Li, X., Zhang, L., Yu, Z., & Wu, X. (2021), "Multiview clustering via robust probabilistic non-negative matrix factorization", IEEE Transactions on Neural Networks and Learning Systems, 32(5), 1975-1986

[13] Blum, Avrim, and Tom Mitchell. "Combining labeled and unlabeled data with co-training." Proceedings of the eleventh annual conference on Computational learning theory. (1998).

[14] Ye, F., Chen, Z., Qian, H., Li, R., Chen, C., & Zheng, Z. New approaches in multi-view clustering. Recent applications in data clustering, 195.(2018).

[15] Kumar, Abhishek, and Hal Daumé. "A co-training approach for multi-view spectral clustering." Proceedings of the 28th international conference on machine learning (ICML-11). (2011).

[16] D. Dua and C. Graff, "UCI Machine Learning Repository," 2019.[Online]. Available. http://archive.ics.uci.edu/ml. [Accessed Jan. 10, 2022].

[17] Wang, Jiada, Yijun Liu, and Wujian Ye. "FMvC: Fast Multi-View Clustering." IEEE Access 11 (2023): 12808-12820.

[18] Bernard, Desgraupes. "Clustering Indices." University Paris Ouest Lab Modal'X November (2017).

Hoàng Thị Cành tốt nghiệp Đại học Thái Nguyên (ICTU). Nhận bằng Kỹ sư CNTT tại ICTU năm 2009. Nhận bằng Thạc sĩ KHMT tại ICTU năm 2012. Hiện đang NCS tại Viện Hàn lâm Khoa học và Công nghệ Việt Nam và là giảng viên Trường Đại học Công nghệ Thông tin và Truyền thông. Các lĩnh vực nghiên cứu bao gồm Thị giác máy tính, Nhận dạng mẫu, Khai phá dữ liệu, Tính toán mềm.
Email: htcanh@ictu.edu.vn

Phùng Thế Huân tốt nghiệp Đại học Thái Nguyên (ICTU). Nhận bằng Kỹ sư CNTT tại ICTU năm 2009. Nhận bằng Thạc sĩ KHMT tại ICTU năm 2012. Nhận bằng Tiến sĩ KHMT tại ICTU năm 2023. Hiện là giảng viên Trường Đại học Công nghệ Thông tin và Truyền thông.
Các lĩnh vực nghiên cứu bao gồm Thị giác máy tính, Nhận dạng mẫu, Khai phá dữ liệu, Tính toán mềm.
Email: pthuan@ictu.edu.vn

Vũ Thuỳ Trang hiện là sinh viên Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.
Các lĩnh vực nghiên cứu bao gồm Thị giác máy tính, Nhận dạng mẫu, Khai phá dữ liệu, Tính toán mềm, Logic mờ.
Email: vuthuytrangt_65@hus.edu.vn

Nguyễn Như Sơn nhận bằng Tiến sĩ Khoa học máy tính tại Đại học Queensland - Australia năm 2007. Hiện là giảng viên, nghiên cứu viên tại Viện Công nghệ Thông tin - Viện Hàn lâm Khoa học và Công nghệ Việt Nam.
Các lĩnh vực nghiên cứu bao gồm Thị giác máy tính, Học máy, Trí tuệ nhân tạo, Khai phá dữ liệu, Tính toán mềm, Logic mờ.
Email: nnson@ioit.ac.vn

Lê Trường Giang nhận bằng Kỹ sư Khoa học máy tính năm 2011 tại Trường Đại học Công nghiệp Hà Nội và Thạc sĩ Khoa học máy tính tại Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên (ICTU) năm 2014. Nhận bằng Tiến sĩ Hệ thống thông tin tại Học viện Khoa học Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam năm 2023. Hiện là cán bộ tại Trung tâm Đảm bảo Chất lượng, Trường Đại học Công nghiệp Hà Nội.
Các lĩnh vực nghiên cứu bao gồm Tối ưu hóa, Học máy, Khai phá dữ liệu, Trí tuệ nhân tạo, Tính toán mềm, Logic mờ.
Email: letruonggiang@haui.edu.vn

Phạm Huy Thông nhận bằng Tiến sĩ Toán Tin tại Đại học Quốc Gia Hà Nội. Hiện là giảng viên, nghiên cứu viên tại Viện Công nghệ Thông tin – Đại học Quốc Gia Hà Nội. Năm 2020.
Các lĩnh vực nghiên cứu bao gồm Tối ưu hóa, Khai phá dữ liệu, Tính toán mềm, Logic mờ.
Email: thongph@vnu.edu.vn