BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Dương Tiến Dũng

Nghiên cứu phương pháp phân cụm bán giám sát mờ dựa trên phân tích biên và học chủ động với ràng buộc cặp

LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội – 2026

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Dương Tiến Dũng

Nghiên cứu phương pháp phân cụm bán giám sát mờ dựa trên phân tích biên và học chủ động với ràng buộc cặp

Chuyên ngành: Hệ thống thông tin
Mã số: 09 48 01 04

LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS. TS Hà Hải Nam
2. PGS. TS Nguyễn Long Giang

Hà Nội – 2026

Lời cam đoan

Tôi xin cam đoan luận án "Nghiên cứu phương pháp phân cụm bán giám sát mờ dựa trên phân tích biên và học chủ động với ràng buộc cặp" này là kết quả nghiên cứu của riêng tôi, được thực hiện dưới sự hướng dẫn của PGS.TS Hà Hải Nam và PGS.TS Nguyễn Long Giang tại Viện Công nghệ thông tin – Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Luận án sử dụng thông tin trích dẫn từ nhiều nguồn tham khảo khác nhau và các thông tin trích dẫn được ghi rõ nguồn gốc. Các kết quả nghiên cứu của tôi được công bố chung với các tác giả khác đã được sự nhất trí của đồng tác giả khi đưa vào luận án. Các số liệu, kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác ngoài các công trình công bố của tác giả. Luận án được hoàn thành trong thời gian tôi làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hà Nội, ngày tháng năm 2026

Dương Tiến Dũng

Lời cảm ơn

Để hoàn thành luận án này, tôi nhận được sự đồng hành và hỗ trợ của nhiều cá nhân, tổ chức. Thành quả này là nỗ lực của bản thân cùng sự giúp đỡ tận tâm từ thầy cô, đồng nghiệp, bạn bè và gia đình.

Trước hết, tôi xin bày tỏ lòng biết ơn sâu sắc tới hai thầy **PGS.TS Hà Hải Nam** và **PGS.TS Nguyễn Long Giang**; sự chỉ dẫn chuyên môn, lời khuyên hữu ích và nguồn động viên bền bỉ của các thầy là then chốt giúp tôi vượt qua mọi khó khăn, thách thức.

Tôi trân trọng cảm ơn Ban Giám đốc, các giảng viên Khoa Công nghệ thông tin và viễn thông và Phòng Đào tạo Nghiên cứu sinh của Học viện Khoa học và Công nghệ, Viện Hàn lâm KH&CN Việt Nam, vì sự hỗ trợ nhiệt tình và tạo điều kiện thuận lợi trong suốt quá trình thực hiện luận án.

Xin cảm ơn các thành viên Lab AI 4.0 (Viện CNTT – ĐHQGHN) đã cho tôi môi trường học hỏi, trao đổi và hỗ trợ quý báu; cảm ơn Ban Lãnh đạo và đồng nghiệp tại FPT Software vì sự thấu hiểu và linh hoạt thời gian để tôi tập trung nghiên cứu.

Cuối cùng, tôi tri ân Bố, Mẹ, Vợ và gia đình—điểm tựa vững chắc bằng tình yêu thương và sự sẻ chia. Luận án này không chỉ là một công trình khoa học mà còn là món quà tinh thần tôi gửi tới những người luôn tin tưởng và đồng hành cùng tôi.

Xin trân trọng cảm ơn!

Tác giả: **Dương Tiến Dũng**

Danh sách thuật ngữ và từ viết tắt

STT	Từ viết tắt	Từ tiếng Anh	Diễn giải
1	ALEXNET	AlexNet	Kiến trúc CNN AlexNet
2	AGNES	Agglomerative Nesting	Phân cụm phân cấp (kết tụ)
3	AFFC	Active Fuzzy Clustering	Phân cụm mờ chủ động
4	AL	Active Learning	Học chủ động
5	ASWC	Alternative Silhouette Width Criterion	Chỉ số Silhouette thay thế
6	ASSFBC	Active Semi-Supervised Fuzzy Clustering Based on Cluster Boundaries	Phân cụm bán giám sát mờ chủ động dựa trên vùng biên cụm
7	BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies	Thuật toán phân cụm phân cấp BIRCH
8	CA	Clustering Accuracy	Độ chính xác phân cụm
9	CLARA	Clustering LARge Applications	Phân cụm các ứng dụng lớn
10	CLARANS	Clustering Large Applications based on RANdomized Search	Phân cụm ứng dụng lớn dựa trên tìm kiếm ngẫu nhiên
11	CLIQUE	CLustering In QUEst	Phân cụm trong không gian tìm kiếm

Bảng tiếp tục ở trang sau

Tiếp tục từ trang trước

STT	**Từ viết tắt**	**Từ tiếng Anh**	**Diễn giải/tạm dịch**
12	CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
13	CRF	Conditional Random Field	Trường ngẫu nhiên có điều kiện (hậu xử lý phân đoạn)
14	CS3FCM	Confidence-Weighted Safe Semi-Supervised Fuzzy C-Means	Phân cụm mờ bán giám sát an toàn có trọng số tin cậy
15	CSDL	Database	Cơ sở dữ liệu
16	DB	Davies–Bouldin Index	Chỉ số Davies–Bouldin
17	DBSCAN	Density-Based Spatial Clustering of Applications with Noise	Phân cụm dựa trên mật độ (không giám sát)
18	DEEPLABV3P	DeepLabv3+	Mô hình DeepLabv3+ (phân đoạn ngữ nghĩa)
19	DENCLUE	DENsity-based CLUstEring	Phân cụm dựa trên mật độ
20	DIANA	Divisive Analysis	Phân tích phân chia (phân cụm phân cấp tách chia)
21	DPC	Density Peaks Clustering	Phân cụm đỉnh mật độ
22	DPFC	Density-Peak Fuzzy Clustering	Phân cụm mờ dựa trên đỉnh mật độ
23	EM	Expectation–Maximization	Thuật toán Kỳ vọng–Cực đại

Bảng tiếp tục ở trang sau

Tiếp tục từ trang trước

STT	**Từ viết tắt**	**Từ tiếng Anh**	**Diễn giải/tạm dịch**
24	F1	F1-score	Điểm F1 (theo cặp)
25	FC	Fractal Clustering	Phân cụm Fractal
26	FCM	Fuzzy C-Means	Phân cụm mờ Fuzzy C-Means
27	FCN	Fully Convolutional Network	Mạng tích chập hoàn toàn
28	FDC	Fuzzy Discriminant Clustering	Phân cụm mờ phân biệt
29	FPSO	Fuzzy Particle Swarm Optimization	Tối ưu hoá bầy đàn hạt mờ
30	GMM	Gaussian Mixture Model	Mô hình hỗn hợp Gaussian
31	HDBSCAN	Hierarchical DBSCAN	Phân cụm dựa trên mật độ phân cấp
32	HRNET	High-Resolution Network	Mạng độ phân giải cao HRNet
33	K-MEANS	K-means clustering	Phân cụm K-means
34	K-MEDOIDS	K-medoids clustering	Phân cụm K-medoids
35	K-PROTOTYPES	K-prototypes clustering	Phân cụm K-prototypes
36	K-SPANNING TREE	K-spanning Tree algorithm	Thuật toán cây bao trùm k
37	KNN	k-Nearest Neighbors	k láng giềng gần nhất

Bảng tiếp tục ở trang sau

Tiếp tục từ trang trước

STT	**Từ viết tắt**	**Từ tiếng Anh**	**Diễn giải/tạm dịch**
38	LHC-S3FCM	Localized Hierarchical Clustering Safe Semi-Supervised Fuzzy C-Means	Phân cụm mờ bán giám sát an toàn phân cấp cục bộ
39	MC-FCM	Multiple fuzzification Coefficients Fuzzy C-Means	Phân cụm mờ với nhiều hệ số mờ hoá
40	MCSSFC-P	Multiple fuzzification Coefficients Semi-Supervised Fuzzy Clustering (Point)	Phân cụm bán giám sát mờ nhiều hệ số (dạng điểm)
41	MK-SSFC	Multi-kernel Semi-Supervised Fuzzy Clustering	Phân cụm bán giám sát mờ đa nhân
42	MRF	Markov Random Field	Trường Markov ngẫu nhiên
43	NMI	Normalized Mutual Information	Thông tin tương hỗ chuẩn hoá
44	OPTICS	Ordering Points To Identify the Clustering Structure	Phân cụm theo thứ tự điểm
45	PCA	Principal Component Analysis	Phân tích thành phần chính
46	PBM	Pakhira Bandyopadhyay Maulik Index	Chỉ số PBM

Bảng tiếp tục ở trang sau

Tiếp tục từ trang trước

STT	Từ viết tắt	Từ tiếng Anh	Diễn giải/tạm dịch
47	PCCA	Pairwise-Constrained Competitive Agglomeration	Cặp ràng buộc có cạnh tranh
48	PFS	Picture Fuzzy Set	Tập mờ hình ảnh
49	QC	Quantum Clustering	Phân cụm lượng tử
50	RI	Rand Index	Chỉ số Rand
51	S3FCM	Safe Semi-Supervised Fuzzy C-Means	Phân cụm mờ bán giám sát an toàn
52	SEGFORMER	SegFormer	Mô hình Transformer cho phân đoạn
53	SFFD	Semi-Supervised Fuzzy Clustering with Feature Discrimination	Phân cụm mờ bán giám sát với phân biệt đặc trưng
54	SNN	Shared Nearest Neighbor	Láng giềng gần nhất chia sẻ
55	SSDBSCAN	Semi-Supervised DBSCAN	Phân cụm bán giám sát dựa trên mật độ
56	SSFCM	Semi-Supervised Fuzzy C-means	Thuật toán phân cụm bán giám sát mờ kmean
57	SSFC	Semi-Supervised Fuzzy Clustering	phân cụm bán giám sát mờ
58	STING	Statistical Information Grid	Lưới thông tin thống kê
59	SVM	Support Vector Machine	Máy véc-tơ hỗ trợ

Bảng tiếp tục ở trang sau

Tiếp tục từ trang trước

STT	**Từ viết tắt**	**Từ tiếng Anh**	**Diễn giải/tạm dịch**
60	SWIN-UNET	Swin-UNet	Kiến trúc Swin-UNet (Transformer cho y sinh)
61	TS3FCM	Trusted Safe Semi-Supervised Fuzzy Clustering Method	Thuật toán phân cụm mờ bán giám sát an toàn tin cậy
62	TV	Total Variation	Biến phân TV (giữ biên)
63	UCI	University of California, Irvine Machine Learning Repository	Kho dữ liệu học máy UCI
64	UNET	U-Net	Kiến trúc U-Net (phân đoạn ảnh)
65	UNET++	UNet++ (Nested U-Net)	Kiến trúc U-Net lồng nhau (UNet++)
66	V-NET	V-Net	Mạng 3D V-Net (phân đoạn thể tích)
67	WAVECLUSTER	Wavelet-Based Clustering Algorithm	Thuật toán phân cụm dựa trên sóng

Mục lục

Danh sách hình vẽ

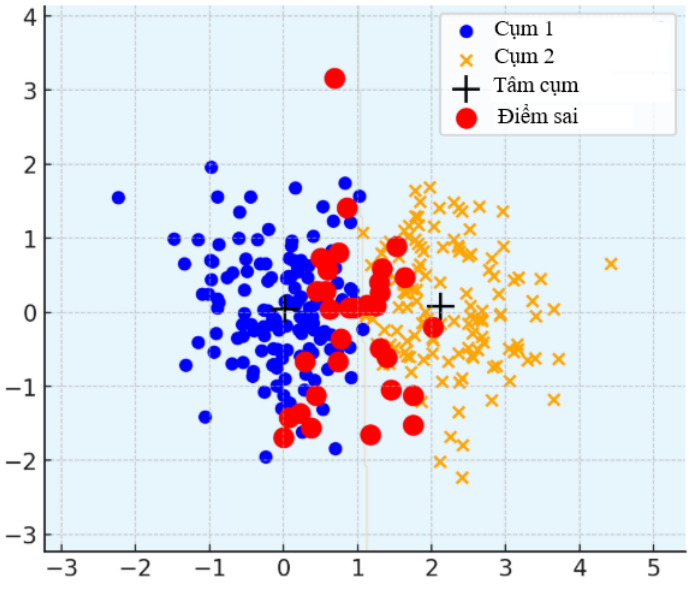
Danh sách bảng

Danh sách thuật toán

Mở đầu

Tính cấp thiết của đề tài luận án

Phân cụm dữ liệu [1–3] là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liệu vào các cụm, sao cho các đối tượng thuộc cùng một cụm thì tương đồng, còn các đối tượng thuộc các cụm khác nhau thì ít tương đồng hơn. Phân cụm dữ liệu là một kỹ thuật quan trọng trong khai phá dữ liệu trong bối cảnh dữ liệu lớn và đa dạng (web, văn bản, ảnh) [4–6], với nhu cầu ngày càng tăng. Các phương pháp phân cụm rõ gặp hạn chế khi dữ liệu chồng lấn; phương pháp phân cụm mờ Fuzzy C-Means (FCM) kế thừa khái niệm tập mờ của Zadeh [7] và trở thành phương pháp kinh điển nhờ gán độ phụ thuộc mềm, được phát triển mạnh mẽ trong các nghiên cứu gần đây [8–10] [11–15]. Dẫu vậy, FCM vẫn dễ sai ở ranh giới cụm, nhạy khởi tạo và kém ổn định với nhiễu.

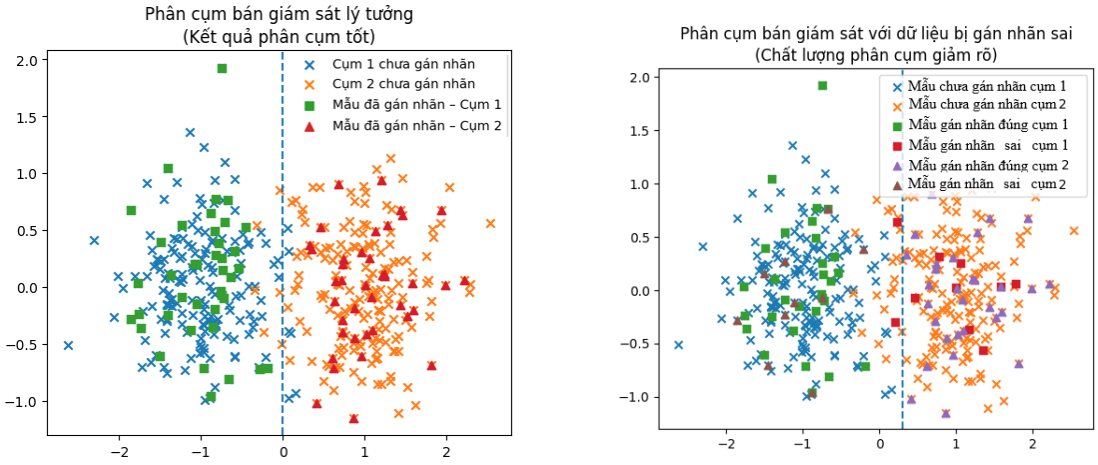


Hình 1: Vấn đề phân cụm ở biên giữa các cụm

Để khắc phục hạn chế cố hữu của FCM, các nghiên cứu phân cụm bán giám

sát mờ (SSFC) [16–21] đã kết hợp tính "mềm" của phân cụm mờ với tri thức bổ sung (nhãn ít, ràng buộc must-link/cannot-link(ML/CL), độ phụ thuộc cho trước) nhằm nâng cao chất lượng phân cụm [22–27]. Tuy nhiên, trong thực tế, các thông tin nhãn được cung cấp từ đầu thường mang tính chất cứng nhắc hoặc không chính xác hoàn toàn — đặc biệt khi được gán một cách thủ công hoặc thiếu kiểm chứng trong môi trường dữ liệu phức tạp. Sự hiện diện của các nhãn sai hoặc không đáng tin cậy dễ làm nhiễu toàn bộ quá trình lan truyền thông tin nhãn trong thuật toán SSFC. Vấn đề càng nghiêm trọng hơn khi các điểm dữ liệu được gán nhãn sai lại nằm ở vùng biên cụm, nơi vốn đã mơ hồ và khó phân định rạch ròi. Vấn đề này được minh họa ở Hình 2. Hiện tượng này càng rõ trong bài toán phân đoạn ảnh (y tế, viễn thám, giám sát), nơi ranh giới vùng thường mơ hồ và nhiễu cao. Gần đây, nhánh phân cụm bán giám sát mờ "an toàn" với các công trình nghiên cứu của Gan và các cộng sự nổi lên là một xu hướng hiệu quả khi giảm tác động nhãn sai bằng trọng số tin cậy và đồ thị cục bộ [17, 28, 29]. Đáng chú ý, nhóm nghiên cứu Thông, Huân đã phát triển phương pháp phân cụm bán giám sát mờ an toàn (TS3FCM) sàng lọc nhãn đáng tin, khởi tạo độ phụ thuộc rồi tối ưu toàn cục, cải thiện hiệu quả nhưng vẫn chịu ảnh hưởng khởi tạo ngẫu nhiên và vùng biên nhiễu [18].

Chính vì vậy, một yêu cầu cấp thiết đặt ra là: làm thế nào để nhận diện được các điểm dữ liệu có nhãn không đáng tin cậy, đặc biệt là tại biên cụm; đồng thời làm sao để chọn lọc hiệu quả những điểm thực sự cần thiết để gán nhãn hoặc điều chỉnh nhãn, thay vì lan truyền sai lệch từ những nhãn kém tin cậy. Trong bối cảnh đó, **học chủ động (Active Learning)** [30–32] đã nổi lên như một hướng tiếp cận đầy hứa hẹn để tăng cường hơn nữa hiệu quả của phân cụm bán giám sát mờ. Học chủ động cho phép mô hình chủ động chọn những điểm "giàu thông tin" thay vì gán nhãn dàn trải. Chiến lược học tập thông minh này



Hình 2: Phân cụm bán giám sát bị ảnh hưởng bởi nhãn sai

đặc biệt phù hợp với phân cụm bán giám sát mờ, bởi nó cho phép kết hợp hiệu quả giữa khả năng xử lý độ phụ thuộc mờ linh hoạt của FCM và khả năng chủ động thu thập tri thức từ chuyên gia của học chủ động, tạo ra một sức mạnh tổng hợp để vượt qua những thách thức của dữ liệu phức tạp, nhiễu và chồng lấn. Chính vì những ưu điểm vượt trội này, lĩnh vực nghiên cứu về phân cụm bán giám sát mờ chủ động (Active Semi-Supervised Fuzzy Clustering) [19, 33, 34] đang ngày càng thu hút sự quan tâm lớn từ cộng đồng khoa học trên toàn thế giới, đi đầu là Grira [31] với phương pháp phân cụm bán giám sát mờ chủ động dựa vào cặp ràng buộc và sau đó Novoselova đã tiếp tục hướng này với sự cải thiện về cách tạo cặp ràng buộc, đem lại những kết quả vô cùng ấn tượng cũng như vượt trội so với các phương pháp phân cụm mờ trước đó. **Tuy nhiên các tiếp cận này vẫn thiếu sự kết hợp toàn diện với việc nhận diện và xử lý nhãn sai, tinh chỉnh ranh giới cụm, tận dụng miền tri thức đã truy vấn để tăng độ chính xác kết quả phân cụm cuối cùng**.

Khoảng trống khoa học này tạo ra cơ hội cho luận án tập trung nghiên cứu, đề xuất và phát triển các phương pháp phân cụm bán giám sát mờ chủ động

mới, có khả năng vừa cải thiện độ chính xác phân cụm, vừa tối ưu hóa chi phí gán nhãn, đồng thời nâng cao tính ứng dụng trong bài toán phân đoạn ảnh, cụ thể với các lĩnh vực thực tiễn như xử lý ảnh viễn thám, ảnh y tế, và phân tích dữ liệu đa miền khi dữ liệu thường có nhiễu và có nhiều sai số khi gán nhãn. Trong phạm vi luận án này, nghiên cứu một số vấn đề như sau:

- Nghiên cứu đề xuất phương pháp phân cụm bán giám sát mờ chủ động dựa vào biên cụm (ASSFBC) với mục đích áp dụng học chủ động tập trung vào vùng biên cụm để cải thiện độ chính xác phân cụm.

- Nghiên cứu đề xuất phương pháp phân cụm bán giám sát mờ an toàn chủ động với cặp ràng buộc dựa vào biên cụm (AS3FCPC) với mục đích áp dụng học chủ động để cải thiện độ tin cậy của nhãn cũng như sử dụng cặp ràng buộc vùng biên cụm thông qua học chủ động để cải thiện độ chính xác phân cụm.

Mục tiêu nghiên cứu

Xuất phát từ những vấn đề cũng như hạn chế còn tồn tại của các phương pháp phân cụm bán giám sát cũ, luận án tập trung nghiên cứu phát triển một số phương pháp phân cụm bán giám sát mờ dựa vào học chủ động mới để giải quyết các vấn đề trên, cụ thể như sau:

1) Nghiên cứu đề xuất các phương pháp phân cụm bán giám sát mờ chủ động mới cụ thể là: phương pháp phân cụm bán giám sát mờ chủ động dựa vào biên cụm (ASSFBC), phương pháp phân cụm bán giám sát mờ an toàn chủ động với cặp ràng buộc dựa vào biên cụm (AS3FCPC) để cải thiện độ chính xác đối với các dữ liệu nhiễu, có vùng chồng lấn, gán nhãn sai cũng như tối ưu chi phí gán nhãn.

2) Áp dụng các phương pháp đề xuất trong bài toán phân đoạn ảnh cụ thể

là: phân đoạn ảnh thông qua phân cụm, phân đoạn ảnh để xác định vùng ngập lụt.

Đối tượng nghiên cứu

- Luận án tập trung nghiên cứu các thuật toán sử dụng phương pháp phân cụm bán giám sát, phân cụm bán giám sát mờ, phân cụm bán giám sát mờ an toàn, phân cụm bán giám sát chủ động, phân cụm bán giám sát mờ chủ động.

- Luận án tập trung nghiên cứu các vấn đề liên quan đến dữ liệu có nhiễu, dữ liệu nhạy cảm ở vùng biên và các dữ liệu người dùng cung cấp với độ tin cậy thấp, đồng thời khảo sát các tập dữ liệu phù hợp cho các thuật toán phân cụm (bao gồm cả tập dữ liệu tiêu chuẩn và tập dữ liệu ảnh) cùng các độ đo đánh giá độ chính xác và hiệu quả của các thuật toán.

- Luận án tập trung nghiên cứu đề xuất giải pháp phân cụm mờ bán giám sát chủ động để xử lý vấn đề biên cụm và dữ liệu thiếu tin cậy nhằm nâng cao chất lượng phân cụm.

- Luận án tập trung nghiên cứu ứng dụng các thuật toán phân cụm bán giám sát mờ chủ động trong các lĩnh vực đặc biệt là xử lý ảnh.

Phương pháp nghiên cứu:

Các kết quả nghiên cứu của luận án được đánh giá trên hai góc độ nghiên cứu gồm có:

- Góc độ nghiên cứu lý thuyết: Tổng hợp và nghiên cứu các tài liệu liên quan đến phân cụm mờ, phân cụm bán giám sát mờ. Tìm hiểu các hướng nghiên cứu mới liên quan đến phân cụm bán giám sát mờ và đề xuất phương pháp cải tiến.

- Góc độ nghiên cứu thực nghiệm: các thuật toán được cài đặt và thực nghiệm trên các bộ dữ liệu từ UCI và các bộ dữ liệu ảnh. Sử dụng các độ đo phù hợp để so sánh, đánh giá chất lượng cũng như độ hiệu quả của các thuật toán.

Đóng góp của luận án:

1. **Đề xuất ASSFBC** (Active Semi-Supervised Fuzzy Clustering Based on Cluster Boundaries): Xây dựng phương pháp nhận diện các điểm bất định ở *vùng biên cụm* và khai thác truy vấn chủ động để tinh chỉnh trực tiếp độ phụ thuộc để giảm sự không chắc chắn vùng biên, qua đó tận dụng vùng giàu thông tin này để xây dựng thuật toán tối ưu mới nâng cao hiệu quả phân cụm cũng như độ tin cậy của phân hoạch mờ.

2. **Phát triển AS3FCPC** (Active Safe Semi-Supervised Fuzzy Clustering with Pairwise Constraints based on Cluster Boundaries): tích hợp phân cụm mờ, học bán giám sát an toàn và học chủ động trong một khuôn khổ thống nhất; tự động truy vấn các điểm mơ hồ để sinh ràng buộc *must-link/cannot-link*, giúp ổn định ranh giới và tâm cụm, thích ứng tốt khi thiếu nhãn hoặc biên phức tạp.

3. **Chứng minh hiệu quả thực nghiệm và tính ứng dụng**: trên nhiều tập dữ liệu chuẩn, ASSFBC và AS3FCPC vượt trội so với các phương pháp mờ truyền thống và bán giám sát hiện có, cải thiện các chỉ số RI/F1/NMI, đạt PC cao và PE thấp, đồng thời cho chất lượng cụm ổn định theo DB và DB_fuzzy; phù hợp cho các bài toán dữ liệu phức tạp như nhận dạng mẫu và phân đoạn ảnh vệ tinh trong điều kiện nhiễu/thiếu nhãn/biên không rõ.

Cấu trúc của luận án:

Ngoài phần mở đầu và kết luận, luận án có 03 chương nội dung nghiên cứu như sau:

Chương 1: Luận án giới thiệu và định nghĩa bài toán sử dụng phương pháp phân cụm bán giám sát, phân cụm bán giám sát mờ, phân cụm bán giám sát mờ chủ động. Trình bày các khái niệm cơ bản về phân cụm, phân cụm bán giám sát, phân cụm bán giám sát mờ, học chủ động. Trình bày các chỉ số và phương

pháp đánh giá chất lượng mô hình phân cụm dữ liệu. Các đóng góp chính của luận án được trình bày trong các Chương 2, Chương 3

Chương 2: Luận án trình bày kết quả nghiên cứu về phương pháp phân cụm bán giám sát mờ chủ động dựa vào vùng biên. Các nội dung chính gồm: Phần mở đầu, đề xuất phương pháp phân cụm bán giám sát mờ chủ động dựa vào vùng biên, trình bày các kết quả thực nghiệm dựa trên bộ dữ liệu tiêu chuẩn UCI và dữ liệu ảnh.

Chương 3: Luận án trình bày phương pháp phân cụm bán giám sát mờ chủ động an toàn sử dụng cặp ràng buộc và vùng biên giữa các cụm, các nội dung chính gồm: Phần mở đầu, đề xuất phương pháp phân cụm bán giám sát mờ chủ động an toàn sử dụng cặp ràng buộc và vùng biên giữa các cụm, trình bày các kết quả thực nghiệm dựa trên bộ dữ liệu tiêu chuẩn UCI và dữ liệu ảnh xác định vùng ngập lụt.

Kết luận: nêu những kết quả đã đạt được của luận án, hướng phát triển trong tương lai và những vấn đề quan tâm của tác giả.

Chương 1

Tổng quan về phân cụm bán giám sát mờ, học chủ động và phân cụm bán giám sát mờ chủ động

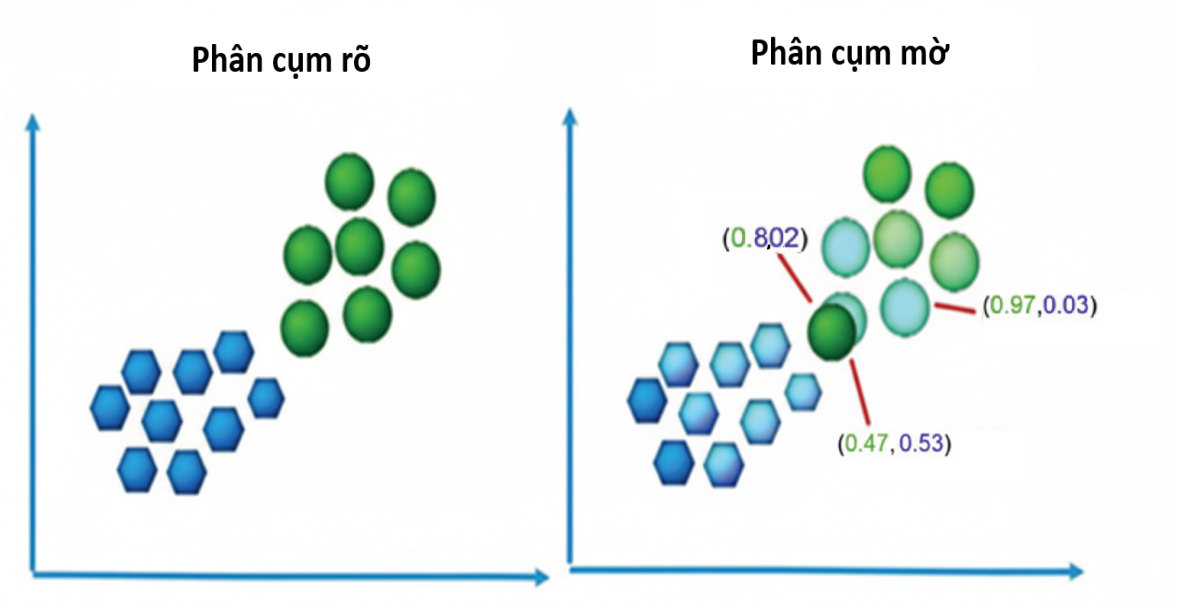
1.1 Tổng quan về phân cụm bán giám sát mờ

1.1.1 Phân cụm mờ

Trong lĩnh vực phân cụm dữ liệu, mục tiêu chính là nhóm các đối tượng có đặc điểm tương đồng vào cùng một cụm, đồng thời phân tách các đối tượng khác biệt vào các cụm riêng biệt. Trong các phương pháp phân cụm truyền thống (phân cụm cứng), mỗi điểm dữ liệu được gán một cách rõ ràng vào duy nhất một cụm. Tuy nhiên, trong thực tế của các hệ thống thông tin hiện đại, dữ liệu thường được thu thập từ nhiều nguồn và nhiều ngữ cảnh khác nhau (ví dụ: dữ liệu giao dịch, hành vi người dùng, nhật ký hệ thống, dữ liệu cảm biến/IoT, dữ liệu văn bản phản hồi), dẫn đến cấu trúc dữ liệu phức tạp, ranh giới cụm không rõ ràng và các nhóm đối tượng có thể chồng lấn. Trong những bối cảnh này, một đối tượng có thể đồng thời mang đặc trưng của nhiều nhóm (chẳng hạn một người dùng vừa thuộc nhóm "thích thể thao" vừa thuộc nhóm "quan tâm công nghệ"), do đó việc gán cụm theo kiểu cứng nhắc có thể làm mất đi thông tin quan trọng và hạn chế khả năng khai thác tri thức phục vụ ra quyết định.

Để giải quyết vấn đề trên, khái niệm phân cụm mờ (Fuzzy Clustering), còn

được gọi là phân cụm mềm (Soft Clustering), đã được giới thiệu như một hướng tiếp cận linh hoạt hơn cho phân tích dữ liệu. Thay vì đưa ra một quyết định cứng về việc một điểm thuộc về cụm nào, phân cụm mờ gán cho mỗi điểm dữ liệu một vector độ phụ thuộc, trong đó mỗi thành phần biểu diễn mức độ mà điểm dữ liệu đó thuộc về từng cụm. Giá trị độ phụ thuộc thường nằm trong khoảng $[0, 1]$, và tổng độ phụ thuộc của một điểm dữ liệu vào tất cả các cụm thường được chuẩn hóa về 1. Hình 1.1 là ví dụ minh họa về phân cụm rõ và phân cụm mờ.



Hình 1.1: Hình minh họa phân cụm mờ

Sự ra đời của phân cụm mờ đã mở ra một hướng tiếp cận phù hợp cho nhiều bài toán trong ngành hệ thống thông tin, nơi dữ liệu thường có tính không chắc chắn, nhiễu và biến đổi theo ngữ cảnh. Bằng cách cho phép các điểm dữ liệu có thể thuộc về nhiều cụm, phân cụm mờ cung cấp một mô tả chi tiết hơn về cấu trúc tiềm ẩn của dữ liệu, qua đó hỗ trợ tốt hơn cho các nhiệm vụ như phân khúc khách hàng, phát hiện bất thường/gian lận, phân tích hành vi người dùng, hệ gợi ý và hỗ trợ ra quyết định. Ngày nay có rất nhiều thuật toán phân cụm mờ

[35–38] đã được nghiên cứu cũng như phát triển, trong đó thuật toán phân cụm mờ Fuzzy C-Means (FCM) [39] là một thuật toán kinh điển được nghiên cứu và phát triển mạnh mẽ cho đến hiện nay.

Thuật toán Fuzzy C-Means (FCM) là một trong những thuật toán phân cụm mờ cơ bản và được sử dụng rộng rãi nhất. FCM là một thuật toán lặp dựa trên việc tối ưu hóa một hàm mục tiêu, nhằm tìm ra sự phân hoạch mờ tốt nhất của dữ liệu thành một số lượng cụm cho trước.

Hàm mục tiêu của FCM được định nghĩa như sau:

$$J_m(U, V) = \sum_{i=1}^{N} \sum_{j=1}^{C} (u_{ij})^m \|x_i - v_j\|^2 \tag{1.1}$$

Trong đó:

- N: số lượng điểm dữ liệu.
- C: số cụm mong muốn.
- x_i: vector biểu diễn điểm dữ liệu thứ i.
- v_j: vector biểu diễn tâm cụm thứ j.
- u_{ij}: độ phụ thuộc của x_i vào cụm j, với $0 \leq u_{ij} \leq 1$ và $\sum_{j=1}^{c} u_{ij} = 1$.
- m: hệ số mờ hóa ($m > 1$), điều khiển mức độ "mềm" của phân cụm. Thường chọn $m = 2$.
- $\|x_i - v_j\|^2$: bình phương khoảng cách Euclid giữa x_i và v_j.

Từ điều kiện cực tiểu hóa hàm mục tiêu bằng phương pháp nhân tử Lagrange, ta thu được các công thức cập nhật:

$$v_j^{(k)} = \frac{\sum_{i=1}^{N} \left(u_{ij}^{(k)}\right)^m x_i}{\sum_{i=1}^{N} \left(u_{ij}^{(k)}\right)^m} \tag{1.2}$$

$$u_{ij}^{(k+1)} = \frac{1}{\sum_{l=1}^{c} \left(\frac{\|x_i - v_j^{(k)}\|}{\|x_i - v_l^{(k)}\|} \right)^{\frac{2}{m-1}}} \tag{1.3}$$

Thuật toán 1.1 Thuật toán phân cụm mờ FCM

Đầu vào: Tập dữ liệu $X = \{x_1, x_2, \ldots, x_N\}$; số cụm c; ngưỡng hội tụ ε; số vòng lặp tối đa $MaxStep$.
Đầu ra: Ma trận độ phụ thuộc U và tập tâm cụm V.

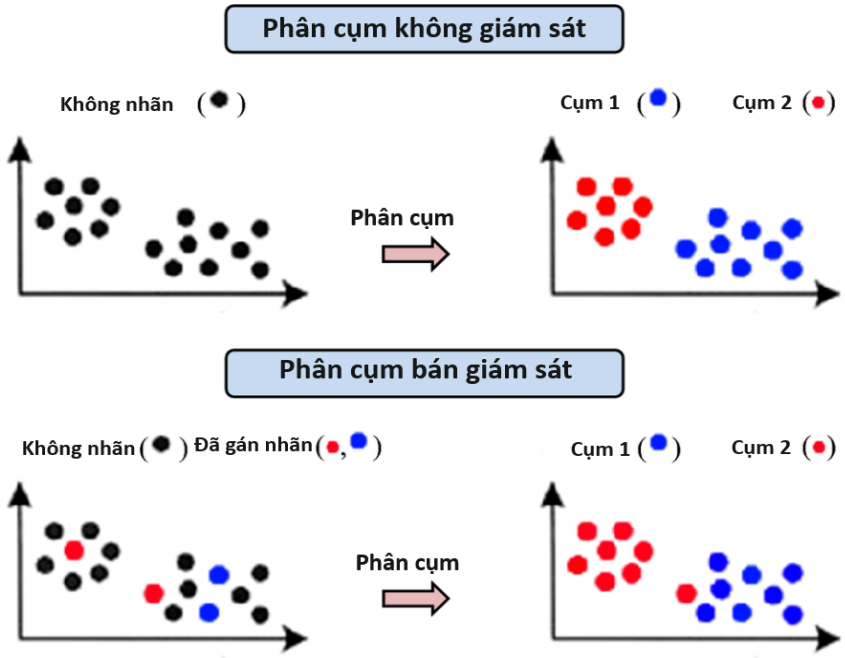
1 Khởi tạo $k = 0$.
2 Khởi tạo ngẫu nhiên các tâm cụm ban đầu $v_j^{(0)},\ j = 1, \ldots, c$.
3 **Thực hiện vòng lặp**:
4 $k \leftarrow k + 1$.
5 Cập nhật ma trận độ phụ thuộc $u_{ij}^{(k)}$ theo công thức 1.3.
6 Cập nhật tâm cụm $v_j^{(k)}$ theo công thức 1.2.
7 **Cho đến khi**: $\|v^{(k)} - v^{(k-1)}\| \leq \varepsilon$ **hoặc** $k \geq MaxStep$.
8 Trả về U, V.

Đây là một thuật toán dạng lặp và tại mỗi bước nó điều chỉnh tâm cụm và ma trận hàm thuộc sao cho thỏa mãn hàm mục tiêu đã định trước. Bezdek đã chứng minh thuật toán này hội tụ về điểm yên ngựa của hàm mục tiêu. Dù ra đời đã lâu nhưng cho đến nay thuật toán FCM vẫn là thuật toán phổ biến nhất và được ứng dụng rộng rãi trong các bài toán nhằm trích xuất luật và khai phá các mẫu tiềm ẩn có sự hiện diện của nhân tử mờ.

1.1.2 Phân cụm bán giám sát

Phân cụm bán giám sát kết hợp điểm mạnh của học không giám sát và học có giám sát nhằm nâng cao chất lượng phân nhóm khi chỉ có một lượng nhỏ thông tin giám sát nhưng rất nhiều dữ liệu chưa gán nhãn. Trong thực tế, gán nhãn diện rộng tốn kém và khó khả thi, trong khi dữ liệu thô dồi dào. Bằng cách đưa tín hiệu giám sát vào đúng chỗ, thuật toán vừa tận dụng cấu trúc nội

sinh của dữ liệu, vừa nhận được "định hướng" đủ để tránh những phân hoạch sai do biên mơ hồ, nhiễu hay chiều dữ liệu cao. Hình 1.2 là minh họa cho phân cụm bán giám sát.



Hình 1.2: Minh họa phân cụm bán giám sát

Về lợi ích, cách tiếp cận này (i) cải thiện độ chính xác và độ ổn định của cụm, (ii) giảm đáng kể chi phí gán nhãn nhờ dùng ít nhãn để dẫn dắt phần còn lại, và (iii) khắc phục hạn chế của phân cụm thuần không giám sát trong việc xác định số cụm hay tách cụm gần nhau. Về nguồn giám sát, ba hình thức phổ biến là: (a) mẫu đã gán nhãn—một tập con nhỏ đóng vai trò "định hướng" [40, 41]; (b) cặp ràng buộc—must-link/cannot-link giúp định hình ranh giới cụm [12, 42]; và (c) độ phụ thuộc mờ ưu tiên—gợi ý mức độ phụ thuộc mong đợi trong bối cảnh phân cụm mờ [18, 35].

Cách tích hợp thông tin giám sát thường đi theo bốn hướng chính nhưng có thể kết hợp linh hoạt: (1) dựa trên cặp ràng buộc—chèn ML/CL vào tiêu chí tối ưu hoặc bước gán cụm để tối đa thoả mãn ràng buộc [32]; (2) dựa trên hạt

giống—khởi tạo/tái khởi tạo tâm và lan truyền nhãn từ tập mẫu lõi [30]; (3) học độ đo khoảng cách—điều chỉnh metric sao cho điểm "cùng lớp" gần nhau, "khác lớp" xa nhau, từ đó làm rõ hình học cụm [43]; và (4) lai ghép—phối hợp nhiều tín hiệu/chiến lược (ví dụ ràng buộc + hạt giống + metric) để tận dụng ưu điểm từng nguồn [31]. Tựu trung, phân cụm bán giám sát là kết hợp giữa khai thác cấu trúc dữ liệu và tận dụng tri thức miền, giúp đạt phân hoạch có ý nghĩa hơn trong những kịch bản dữ liệu thực tế phức tạp.

1.1.3 Phân cụm bán giám sát mờ

Tuy đã ra đời khá lâu, FCM vẫn là một trong những thuật toán phân cụm phổ biến nhất và được ứng dụng rộng rãi trong nhiều bài toán khác nhau, đặc biệt là trong việc trích xuất luật và khám phá các mẫu tiềm ẩn có yếu tố mờ. FCM đã được ứng dụng thành công trong nhiều lĩnh vực, bao gồm phân đoạn ảnh [11, 44–46], nhận dạng khuôn mặt [47], hỗ trợ chẩn đoán bệnh [48], hỗ trợ nha khoa, nhận dạng cử chỉ và điệu bộ, phát hiện xâm nhập trái phép [49], ...

Mặc dù tính hiệu quả đã được chứng minh, thuật toán FCM vẫn tồn tại một số nhược điểm, đặc biệt là độ nhạy cảm với nhiễu và các điểm ngoại lệ, cũng như sự phụ thuộc vào quá trình khởi tạo ngẫu nhiên ban đầu. Để khắc phục những hạn chế này, nhiều nhà nghiên cứu đã đề xuất đưa một số thông tin bổ trợ vào như các ràng buộc, các nhãn, độ phụ thuộc của một số phần tử cho trước, v.v. để định hướng quá trình phân cụm. Phương pháp này được gọi là phân cụm bán giám sát mờ. Các phương pháp phân cụm bán giám sát mờ phổ biến được Thong và cộng sự trình bày trong bài tổng quan về phân cụm bán giám sát mờ [50], nhìn chung mang lại kết quả phân cụm tốt.

Semi-Supervised Fuzzy C-means (SSFCM) là phương pháp phân cụm bán giám sát mờ tiêu biểu cũng như là nền tảng của rất nhiều các phương pháp

phân cụm bán giám sát mờ sau này[51–54] [55–59], được đề xuất bởi Pedrycz and Waletzky [60]. Trong hàm mục tiêu của phương pháp SSFCM bao gồm hai thành phần: Thành phần thứ nhất là thành phần học không giám sát và thành phần thứ hai là thành phần học có giám sát. Với hàm mục tiêu được biểu diễn như sau:

$$J_{(u,d)} = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} d_{ij}^{2} + \alpha \sum_{i=1}^{N}\sum_{j=1}^{C} (u_{ij} - f_{ij} b_i)^{m} d_{ij}^{2} \rightarrow Min \tag{1.4}$$

Trong đó, tập dữ liệu $X = \{x_1, x_2, \ldots, x_N\}$, N là số điểm dữ liệu, C là số cụm, $V = \{v_1, v_2, \ldots, v_C\}$ là tập các tâm cụm, u_{ij} là độ phụ thuộc của điểm dữ liệu x_i đối với cụm j, d_{ij} là khoảng cách từ điểm x_i đến tâm cụm v_j, f_{ij} là độ phụ thuộc được gán nhãn của điểm x_i trong cụm j, và m là hệ số mờ.

Tham số α được sử dụng để cân bằng giữa thành phần giám sát và không giám sát, trong đó b_i phân biệt giữa phần tử được gán nhãn và chưa gán nhãn:

$$b_i = \begin{cases} 1, & \text{nếu } x_i \text{ được gán nhãn} \\ 0, & \text{ngược lại} \end{cases} \tag{1.5}$$

Các tâm cụm v_j và độ phụ thuộc u_{ij} được tính như sau:

$$v_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}\, x_i \;+\; \alpha \sum_{i=1}^{N} (u_{ij} - f_{ij} b_i)^{m}\, x_i}{\sum_{i=1}^{N} u_{ij}^{m} \;+\; \alpha \sum_{i=1}^{N} (u_{ij} - f_{ij} b_i)^{m}}, \quad j = 1, \ldots, C \tag{1.6}$$

$$u_{ij} = \frac{1}{1+\alpha}\left[\frac{1 + \alpha\left(1 - b_i \sum_{k=1}^{C} f_{kj}\right)}{\sum_{k=1}^{C} \frac{d_{ij}^2}{d_{kj}^2}} \;+\; \alpha f_{ij} b_i\right], \quad i = 1, \ldots, N;\; j = 1, \ldots, C \tag{1.7}$$

Thuật toán SSFCM lặp lại việc cập nhật v_j và u_{ij} cho đến khi thoả mãn điều kiện dừng $\|v^{(t)} - v^{(t-1)}\| \leq \varepsilon$ hoặc $t > MaxStep$:

Thuật toán 1.2 Thuật toán phân cụm bán giám sát mờ SSFCM

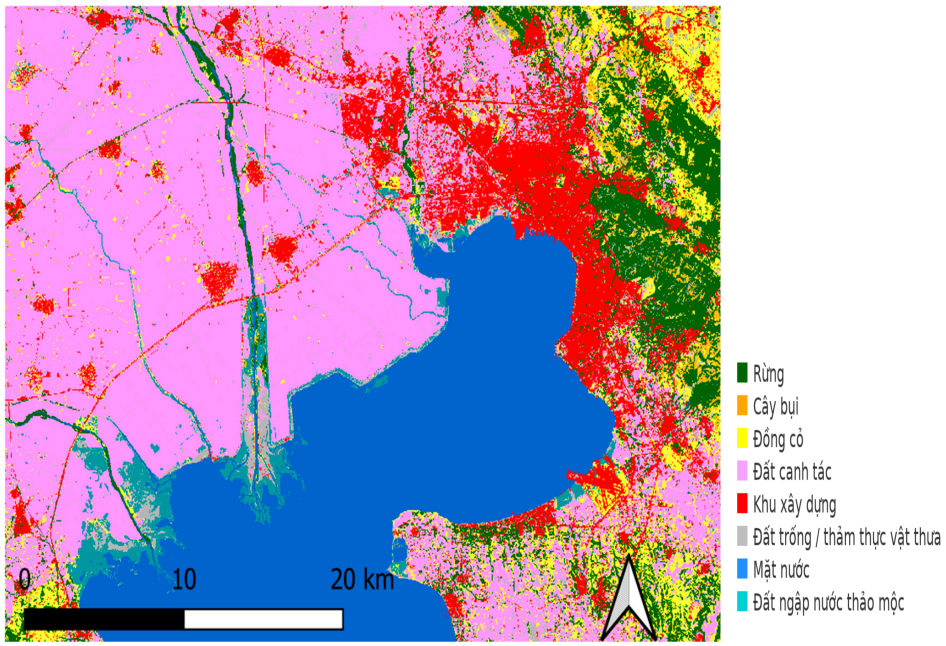
Đầu vào: Tập dữ liệu $X = \{x_1, \ldots, x_N\}$; số cụm C; tham số mờ m; ngưỡng hội tụ ε; số vòng lặp tối đa $MaxStep > 0$.
Đầu ra: Ma trận độ phụ thuộc U và tâm cụm V.

1 $t \leftarrow 0$
2 Khởi tạo ngẫu nhiên tâm cụm ban đầu $V^{(0)}$
3 **repeat**
4 $\quad t \leftarrow t + 1$
5 $\quad$ Cập nhật $u_{ij}^{(t)}$ theo công thức (1.7)
6 $\quad$ Cập nhật $v_j^{(t)}$ theo công thức (1.6)
7 **until** $\max_j \|v_j^{(t)} - v_j^{(t-1)}\| \leq \varepsilon$ **hoặc** $t > MaxStep$

1.1.4 Ứng dụng phân cụm bán giám sát mờ trong bài toán phân đoạn ảnh

Phân đoạn ảnh là gán mỗi điểm ảnh/superpixel vào một miền đồng nhất theo đặc trưng hình học–quang phổ–ngữ nghĩa. Trong thực tế, dữ liệu lớn, nhiễu, ít nhãn khiến gán nhãn đầy đủ tốn kém; do đó phân cụm bán giám sát mờ (SSFC) và đặc biệt SSFC chủ động (Active SSFC) được dùng để hỏi ít nhưng đúng chỗ (nhãn hạt giống, độ phụ thuộc cho trước, ràng buộc ML/CL) tại vùng ranh giới quan trọng rồi khuếch tán ràng buộc trên toàn ảnh/chuỗi thời gian. Ứng dụng trải rộng từ quan sát Trái Đất (lớp phủ, cháy rừng, lũ, băng tuyết), đô thị–giao thông (đường, mái nhà, vùng ngập), xe tự hành LiDAR/point cloud (mặt đường, vật thể), đến y–sinh học (cơ quan/khối u, tế bào) [61, 62] —những nơi nhãn khan hiếm, mất cân bằng lớp và yêu cầu gần thời gian thực. Hình 1.3, 1.4, 1.5 là các ví dụ minh họa về ứng dụng của phân đoạn ảnh trong thực tế.

Nhìn một cách hệ thống, có ba dòng phương pháp phân đoạn ảnh. (i) Xử lý ảnh cổ điển: ngưỡng hoá (Otsu), biên–điền vùng (Canny), phân vùng theo vùng, thuật toán tách nước (watershed), đường bao chủ động (active contour), và mô hình mức (level set), và điều chuẩn không gian (MRF/CRF) cho tốc độ–diễn



Hình 1.3: Phân đoạn ảnh cho phân loại lớp phủ bề mặt qua ảnh vệ tinh.

Hình 1.4: Phân đoạn hỗ trợ phát hiện cháy rừng.

Hình 1.5: Phân đoạn ảnh giúp xác định vùng ngập lụt

giải tốt, phù hợp khởi tạo/hậu xử lý nhưng nhạy tham số–tương phản–nhiễu [63–65]. (ii) Dựa trên phân cụm: coi điểm ảnh/superpixel là mẫu trong không gian đặc trưng; FCM cung cấp độ phụ thuộc mềm giúp mô hình hoá vùng chuyển tiếp, mở rộng với điều chuẩn không gian và SSFC để tận dụng nhãn và ràng buộc must–link/cannot–link [66]; Active SSFC truy vấn theo bất định (entropy-/margin), đại diện (mật độ/kNN), ảnh hưởng mô hình và đa dạng để đạt chất lượng tiệm cận gán nhãn dày đặc với rất ít câu hỏi [30]. Ngoài FCM, GMM/EM [67, 68] biểu đạt cụm elip nhưng nhạy khởi tạo/giả định Gaussian [69, 70]; DBSCAN/HDBSCAN và DPC phát hiện cụm phi lồi, loại nhiễu và nhận "đỉnh mật độ" có ảnh hưởng toàn cục [71–73] [74, 75], thường vận hành ở mức superpixel kèm CRF để xử lý biên [76, 77]. (iii) Học sâu: FCN, U-Net/UNet++, DeepLab, HRNet, SegFormer, Swin–UNet học đặc trưng đa tỉ lệ–ngữ cảnh cho độ chính xác cao; khi nhãn hạn chế, dùng pretrain/fine-tune, pseudo-label/consistency, tự giám sát, cộng thêm độ mất mát cân bằng (Dice/Focal) và ước lượng bất

định để duy trì hiệu năng [78, 79].

Tóm lại, các hệ thống phân đoạn hiệu quả thường (i) chọn biểu diễn phù hợp dữ liệu và ràng buộc hình học/ngữ nghĩa; (ii) tối ưu chuỗi tiền xử lý—mô hình—hậu xử lý cho mục tiêu vận hành (độ chính xác và độ trễ); và (iii) tận dụng SSFC để giảm chi phí nhãn nhưng vẫn kiểm soát ranh giới quan trọng. Vai trò của phân cụm đặc biệt là phân cụm bán giám sát mờ là cầu nối giữa công cụ cổ điển và học sâu: vừa tận dụng được dữ liệu không nhãn, vừa tổng hợp tri thức miền vào nơi mô hình "bối rối", từ đó đạt cân bằng giữa độ chính xác, độ tin cậy và chi phí trên nhiều miền ứng dụng.

1.1.5 Các phương pháp phân cụm bán giám sát mờ

Hiện nay có một số hướng nghiên cứu phân cụm bán giám sát mờ dựa trên phương pháp Fuzzy C-Means chuẩn [39] kết hợp với các kỹ thuật kinh điển như sử dụng hàm nhân kernel, hàm trọng số, hay hàm thích nghi, học sâu mang lại nhiều kết quả tích cực.

Phân cụm bán giám sát mờ dựa trên hàm nhân kernel

Hichem Frigui và cộng sự [80] đã tổng hợp các phương pháp phân cụm mờ dựa trên hàm nhân kernel. Nghiên cứu này cho thấy việc đưa tín hiệu giám sát từng phần trực tiếp vào hàm mục tiêu trong không gian đặc trưng giúp định hướng tối ưu hoá hiệu quả. Nhánh multiple kernel được khởi xướng bởi Zhao et al. [81], nơi tâm cụm có thể được loại khỏi mục tiêu; các mở rộng thực nghiệm cho thấy lợi ích rõ rệt trong nhiều tác vụ [82]. Sinh & Long [83] kết hợp nhiều kernel với trọng số theo thuộc tính để cải thiện phân loại đất, còn Kanzawa [84] sửa đổi ma trận kernel và điều khiển entropy nhằm dùng ràng buộc mềm. Tuy ưu điểm là mô hình hoá phi tuyến và tách cụm tốt hơn, hướng tiếp cận này đối mặt với ước lượng tham số kernel phức tạp và độ nhạy khởi tạo, dễ ảnh hưởng

đến độ ổn định kết quả.

Phân cụm Bán Giám sát Mờ Dựa trên Hàm Thích Nghi

Casalino và cộng sự [28, 85] đã phát triển thuật toán phân cụm bán giám sát mờ dựa trên FCM kết hợp hàm thích nghi để phân lớp dòng dữ liệu. Thuật toán này sử dụng thành phần động để ước lượng số cụm cần thiết dựa trên phân bố dữ liệu, sau đó tiến hành lặp bằng FCM để tìm kết quả tối ưu. Yu và cộng sự [29] đã đề xuất một cách tiếp cận khác bằng cách lan truyền ràng buộc dựa trên ràng buộc đóng bắc cầu, sử dụng toán tử bắc cầu để giải quyết vấn đề không sử dụng đầy đủ các ràng buộc must-link và cannot-link. Sau đó, một khung tổ hợp phân cụm bán giám sát dựa trên không gian con ngẫu nhiên với các hệ số tin cậy được thiết kế để xử lý dữ liệu nhiều chiều với nhiễu. Tiếp theo, một khung tập hợp phân cụm bán giám sát thích nghi được đề xuất để nâng cao tính thích ứng và hiệu suất của thuật toán [47].

Phân cụm bán giám sát mờ kết hợp học sâu

Một nhánh nghiên cứu quan trọng trong những năm gần đây là kết hợp học biểu diễn bằng mạng nơ-ron sâu với thuật toán FCM có ràng buộc. Ý tưởng chung là tận dụng khả năng trích chọn đặc trưng phi tuyến của mạng nơ-ron để tạo không gian thuận lợi cho phân cụm mờ, đồng thời tích hợp trực tiếp các ràng buộc bán giám sát. Neuro-Fuzzy FCM. Các nghiên cứu tiêu biểu như Arshad và cộng sự (2019) đã kết hợp mạng nơ-ron đa lớp (MLP) [86] hoặc CNN làm bộ mã hoá đặc trưng, tối ưu đồng thời hàm mất mát với hàm mục tiêu FCM có ràng buộc. Biến thể bán giám sát còn chèn thêm thành phần regularizer thành viên, buộc các mẫu có nhãn hội tụ gần tâm cụm tương ứng. Cách tiếp cận này giúp giảm nhiễu và cải thiện tách biệt cụm trên dữ liệu phức tạp như ảnh và tín hiệu. Hạn chế chính là việc lựa chọn kiến trúc mạng và tỷ lệ trọng số mất mát rất nhạy cảm, dễ gây mất cân bằng giữa hai mục tiêu. Deep Semi-Supervised

Fuzzy Clustering. Một hướng khác do Yang và cộng sự (2021) [87] phát triển là dùng autoencoder để học không gian tiềm ẩn "cụm-hoá tốt", sau đó thực hiện tối ưu FCM trong latent space. Các ràng buộc must-link/cannot-link được mã hoá thành một constraint loss và gộp với reconstruction loss thông qua một siêu tham số cân bằng.

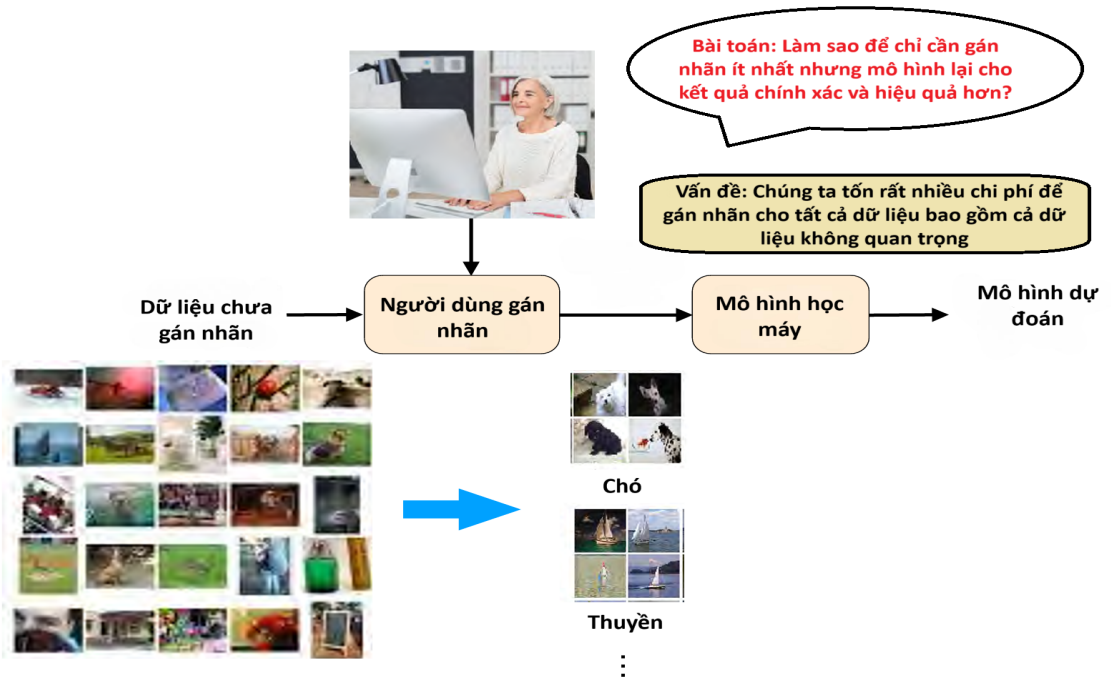
Các phương pháp neuro-fuzzy và deep SSFC có ưu điểm nổi bật là tận dụng được biểu diễn sâu để giảm chiều và khử nhiễu, đồng thời tạo ra không gian đặc trưng mà cụm tách biệt rõ hơn, đặc biệt hiệu quả với dữ liệu ảnh và văn bản. Tuy nhiên, chúng cũng bộc lộ hạn chế: (i) rất nhạy cảm với siêu tham số (tốc độ học, trọng số mất mát, kiến trúc mạng), (ii) dễ xảy ra hiện tượng overfitting khi số lượng nhãn ít, và (iii) chi phí huấn luyện cao, khó mở rộng cho tập dữ liệu cực lớn nếu không có phần cứng mạnh.

1.1.6 Sự cần thiết về việc chuyển hướng sang học chủ động

Mặc dù các phương pháp phân cụm bán giám sát mờ hiện nay đều có ưu điểm riêng, vẫn tồn tại các lỗ hổng: (i) phần lớn phương pháp đều nhạy cảm với nhiễu và chưa có cơ chế hiệu quả xử lý với vùng biên; (ii) chi phí cho dữ liệu bán giám sát còn cao và chưa có cơ chế lựa chọn dữ liệu tốt; (iii) độ nhạy tham số (ngưỡng biên, số cặp, bán kính lân cận, trọng số mất mát); (iv) các cặp ràng buộc được lựa chọn chưa tối ưu. Do đó, việc phát triển một phương pháp phân cụm bán giám sát mờ dựa trên học chủ động vừa tiết kiệm truy vấn, vừa giảm thiểu nhãn sai, và có hiệu quả cao hơn là cần thiết. Phần tiếp theo sẽ trình bày về học chủ động và các phương pháp học chủ động.

1.2 Học chủ động

1.2.1 Giới thiệu về học chủ động

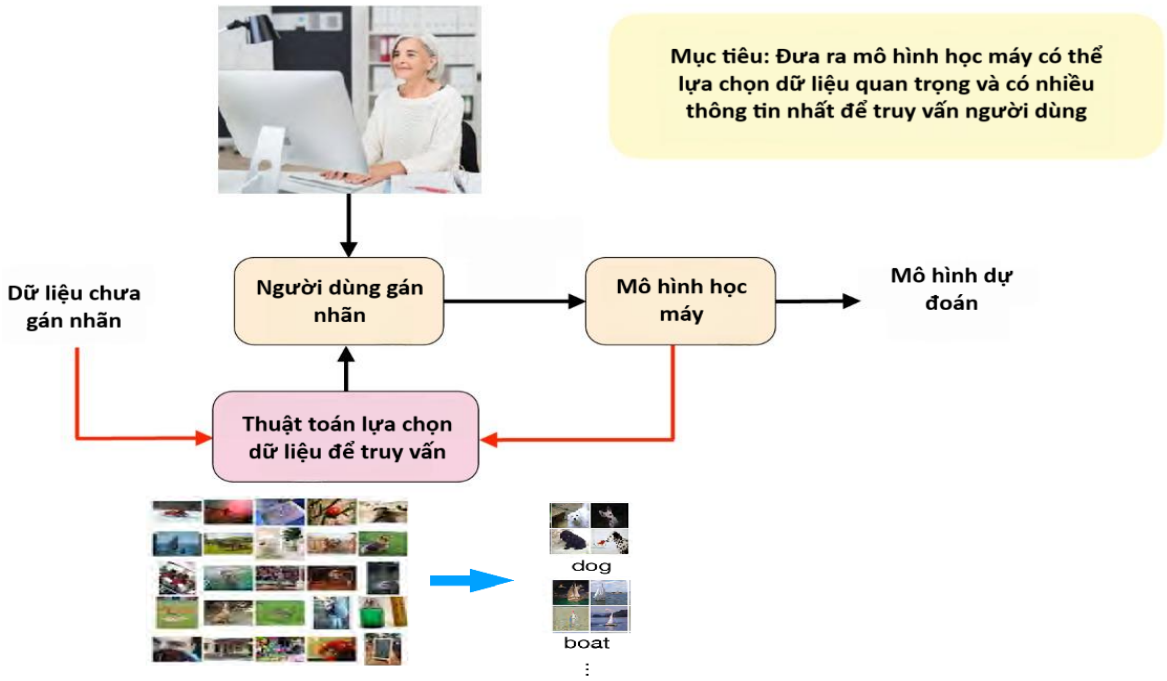


Hình 1.6: Học thụ động

Trong bối cảnh của học có giám sát và học không giám sát, các phương pháp truyền thống thường liên quan đến việc thu thập một lượng dữ liệu đủ lớn, sau đó dữ liệu này được lấy mẫu tự động từ phân phối mật độ cơ bản. Tập dữ liệu đó tiếp tục được rút gọn thành các lớp hoặc mô hình — phương pháp này thường được gọi là học thụ động (passive learning). Trong học thụ động, thuật toán học nhận dữ liệu một cách ngẫu nhiên từ môi trường (Hình 1.6) và xây dựng mô hình phân loại dựa trên tập dữ liệu đó.

Thông thường, trong nhiều ứng dụng, quá trình thu thập dữ liệu là bước tiêu tốn nhiều thời gian và chi phí nhất. Trong một số trường hợp, tài nguyên để thu

thập dữ liệu bị giới hạn, do đó việc sử dụng hiệu quả các tài nguyên này là rất quan trọng. Mặc dù phần lớn các bài toán giả định rằng việc thu thập dữ liệu diễn ra một cách ngẫu nhiên và tuân theo phân phối độc lập, đồng nhất, nhưng trong nhiều trường hợp, chúng ta có thể hướng dẫn quá trình lấy mẫu để tối ưu hóa hiệu quả học.



Hình 1.7: Học chủ động

Học chủ động (active learning) trong học máy là một chiến lược huấn luyện mô hình trong đó mô hình chủ động lựa chọn những dữ liệu "quan trọng nhất" để yêu cầu gán nhãn . Giả thiết cốt lõi là nếu thuật toán học được phép chọn lựa dữ liệu từ những gì nó học, thì nó sẽ có thể đạt hiệu suất tốt hơn ngay cả với lượng dữ liệu huấn luyện ít hơn. Hình 1.7 đã minh họa về mô hình học chủ động.

Các hệ thống học chủ động tìm cách vượt qua rào cản về chi phí gán nhãn bằng cách chủ động đưa ra các truy vấn yêu cầu người dùng hoặc chuyên gia gán nhãn cho một số điểm dữ liệu chưa có nhãn. Qua đó, hệ thống hướng tới

việc đạt độ chính xác cao với lượng dữ liệu có nhãn tối thiểu, từ đó tiết kiệm đáng kể chi phí gán nhãn. Học chủ động trở nên đặc biệt có giá trị trong các ứng dụng học máy hiện đại, nơi dữ liệu có thể rất dồi dào nhưng nhãn lại khan hiếm hoặc đắt đỏ.

Là một phương pháp học có giám sát, học chủ động tạo ra dữ liệu có nhãn với sự hỗ trợ từ con người thông qua các vòng phản hồi. Mục tiêu là huấn luyện mô hình bằng một tập dữ liệu nhỏ đã được gán nhãn và đồng thời giảm thiểu số lượng nhãn cần thiết bằng cách chiến lược chọn ra những điểm dữ liệu chứa nhiều thông tin nhất.

Phương pháp học chủ động đã được áp dụng thành công trong nhiều lĩnh vực khác nhau. Trong phân loại văn bản, nó được sử dụng cho các nhiệm vụ như phân tích cảm xúc, gán nhãn chủ đề và phân loại tin tức [88]. Trong trích xuất thông tin, học chủ động hỗ trợ hiệu quả cho các bài toán nhận dạng thực thể có tên và phân tích cú pháp [89]. Bên cạnh đó, trong phân loại hình ảnh, phương pháp này được ứng dụng trong các lĩnh vực như viễn thám và y sinh [61]. Một số nghiên cứu còn áp dụng học chủ động vào lựa chọn cảm biến, chẳng hạn như phát hiện xâm nhập hay nhận diện cử chỉ [90]. Ngoài ra, nó cũng được khai thác trong học cấu trúc mạng thông qua các phương pháp bán giám sát [29], cũng như trong hệ thống gợi ý với sự hỗ trợ của học chuyển giao.

1.2.2 Phương pháp học chủ động

Trong học chủ động, bước khởi đầu là xây dựng mô hình M và đánh giá mức độ chính xác của nó thông qua hàm tổn thất $Loss(M)$. Cách thiết kế mô hình và định nghĩa hàm tổn thất sẽ được điều chỉnh linh hoạt tùy thuộc vào yêu cầu của từng bài toán. Sau khi xác định hàm tổn thất, thuật toán sẽ chọn ra truy vấn tiếp theo nhằm đảm bảo mô hình sau cập nhật M' có giá trị tổn thất nhỏ

nhất. Với một truy vấn ứng viên q, cần ước lượng tổn thất kỳ vọng của mô hình M' khi nhận phản hồi x đi kèm. Do giá trị x chưa được biết trước, ta sẽ xét toàn bộ các trường hợp có thể xảy ra để tính tổn thất kỳ vọng theo công thức:

$$Loss(q) = \mathbb{E}_x[Loss(M')] \tag{1.8}$$

Từ công thức này, thuật toán học chủ động tiến hành chọn truy vấn q đem lại giá trị tổn thất kỳ vọng thấp nhất [91]. Tong đã đưa ra sơ đồ tổng quan của quá trình học chủ động như sau:

Thuật toán 1.3 Chiến lược chọn truy vấn tối ưu

for $i \leftarrow 1$ **to** $totalQueries$ **do**
  for all $q \in potentialQueries$ **do**
    Ước lượng $Loss(q)$
  end for
  Lựa chọn truy vấn q sao cho $Loss(q)$ đạt giá trị nhỏ nhất
  Cập nhật mô hình M bằng cách bổ sung truy vấn q và phản hồi x
end for
return mô hình M sau khi huấn luyện

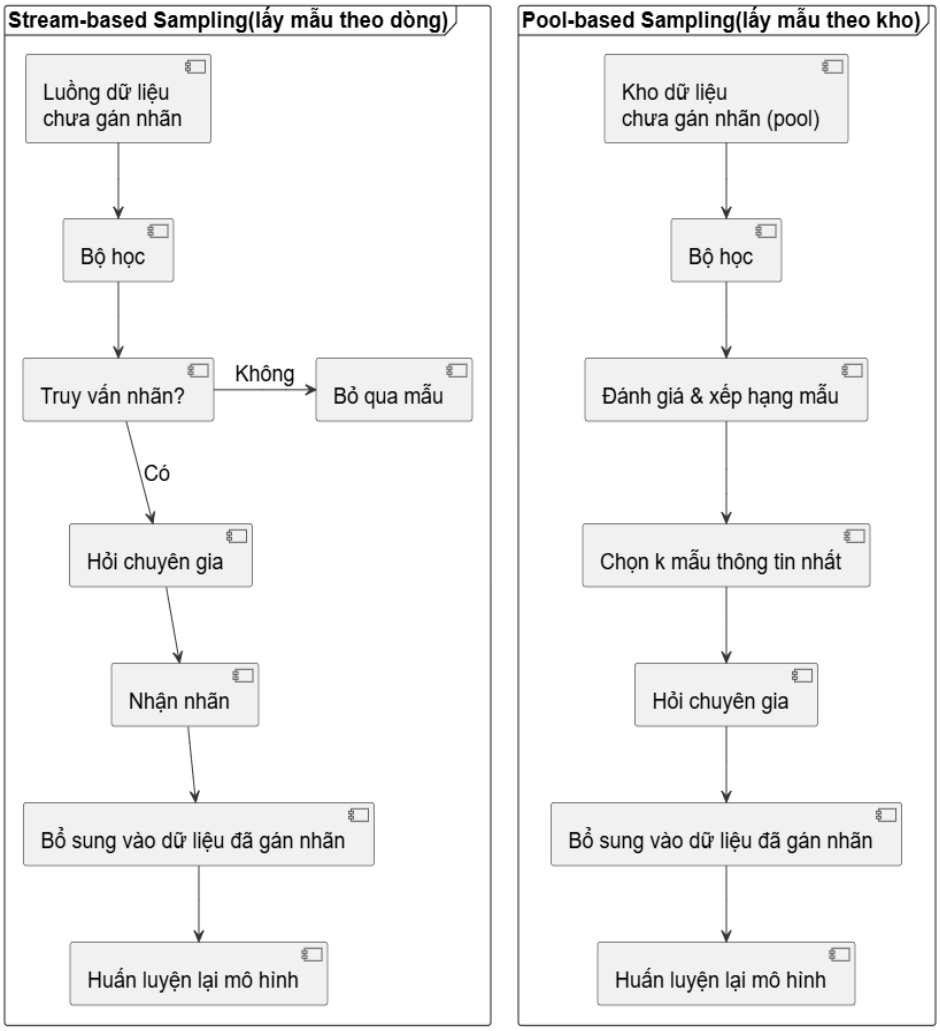
Trong lĩnh vực thống kê, thay vì tập trung vào việc tối thiểu hóa tổn thất kỳ vọng, người ta cũng có thể quan tâm đến phương án tối thiểu hóa tổn thất cực đại — tức giả định trong kịch bản bất lợi nhất:

$$Loss(q) = \max_x Loss(M') \tag{1.9}$$

Cách tiếp cận này trong nhiều tình huống lại mang đến nhiều thông tin bổ ích hơn, vì truy vấn gây ra tổn thất cực đại có thể giúp mô hình cải thiện năng lực học tốt hơn [92].

Trong các ứng dụng học máy, mô hình có thể được thiết kế để tự đặt câu hỏi về những mẫu dữ liệu chưa rõ ràng. Trong số nhiều kịch bản được nghiên

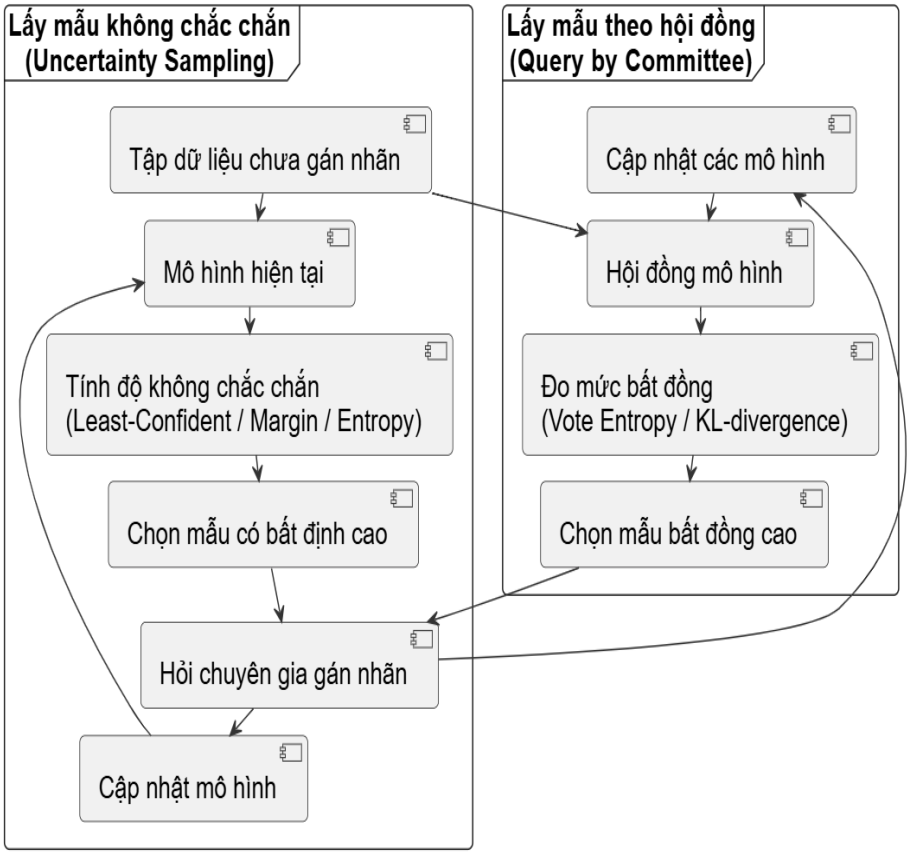
cứu, hai kịch bản điển hình và thường xuyên được áp dụng nhất trong học chủ động là lấy mẫu theo dòng (stream-based sampling) [93] và lấy mẫu theo kho (pool-based sampling) [94, 95] được minh họa trong hình 1.8.



Hình 1.8: Các kịch bản trong học chủ động

Ngoài hai kịch bản trên, hai phương pháp truy vấn quan trọng nhất trong học chủ động là Lấy mẫu không chắc chắn (uncertainty sampling) và Lấy mẫu theo hội đồng (query by committee) được minh họa như hình 1.9.

Bảng 1.1 thể hiện sự so sánh giữa 2 chiến lược truy vấn trong học chủ động:



Hình 1.9: Các phương pháp truy vấn trong học chủ động

Tiêu chí	**Lấy mẫu không chắc chắn**	**Lấy mẫu theo hội đồng**
Số mô hình cần dùng	Một mô hình duy nhất	Nhiều mô hình trong một committee
Nguyên tắc chính	Chọn mẫu mà mô hình bối rối nhất	Chọn mẫu gây bất đồng cao giữa các mô hình
Độ phức tạp tính toán	Đơn giản, chi phí thấp	Tốn kém hơn do huấn luyện nhiều mô hình
Trường hợp phù hợp	Khi chỉ có một bộ phân lớp	Khi muốn tăng độ tin cậy và giảm thiên lệch của một mô hình đơn

Bảng 1.1: So sánh hai chiến lược truy vấn trong học chủ động

1.3 Các nghiên cứu liên quan phân cụm bán giám sát mờ gần đây

1.3.1 Phân cụm bán giám sát mờ chủ động

A. Bối cảnh và động lực

Phân cụm bán giám sát mờ (SSFC) tận dụng thông tin bổ trợ bao gồm nhãn cho trước, các ràng buộc (ML/CL) và độ phụ thuộc cho trước để định hướng, cải thiện kết quả phân cụm so với phân cụm mờ không giám sát. Tuy nhiên, chi phí giám sát cao, dữ liệu nhiễu và các vấn đề ở các dữ liệu vùng biên khiến kết quả kém, nhạy tham số (băng thông, số cụm, ngưỡng peak) và dễ gặp ràng buộc mâu thuẫn, trong khi độ phức tạp tính toán lớn. Học chủ động khắc phục bằng cách chọn ít nhưng đúng truy vấn (điểm/cặp) theo các tín hiệu nội sinh như entropy hay fuzzy margin, từ đó cố định ranh giới, giảm sai số với rất ít phản hồi và tăng bền vững trước nhiễu so với hỏi ngẫu nhiên.

B. Các phương pháp áp dụng học chủ động với phân cụm bán giám sát mờ

Trong phân cụm bán giám sát mờ chủ động, các dạng truy vấn phổ biến gồm ba nhóm chính và thường được lựa chọn tùy theo chi phí phản hồi và mức độ gợi ý mà chúng mang lại. Thứ nhất, nhãn điểm (point-label) yêu cầu nhãn lớp đầy đủ cho một đối tượng x_i, phù hợp khi đã có ánh xạ cụm–lớp để điều chỉnh tâm μ_k hoặc liên thuộc μ_{ik}. Thứ hai, ràng buộc theo cặp (must-link/cannot-link) chỉ hỏi liệu hai điểm (x_i, x_j) có cùng cụm hay không mà không cần tên lớp; đây là kênh phản hồi thân thiện với người dùng và được khai thác rộng rãi, tiêu biểu là công trình kinh điển về chọn cặp giàu thông tin [30]. Cuối cùng, truy vấn cấu trúc/so sánh (triplet/graph/peak) như "x_i gần x_j hơn x_ℓ?" hoặc "điểm này có phải là đỉnh mật độ (peak) không?", tận dụng cấu trúc lân cận/đồ thị và đặc

biệt hữu ích trong họ density peak clustering (DPC) và các biến thể bán giám sát [96]; các hướng active learning qua cụm mật độ (ALEC) cũng cho thấy cách dùng cấu trúc mật độ để dẫn dắt truy vấn một cách hiệu quả [97].

Sau khi chọn loại truy vấn phù hợp, bước then chốt là xác định tiêu chí để quyết định "điểm/cặp/cấu trúc" tốt nhất cần hỏi. Trước hết, sự bất định (uncertainty) ưu tiên các đối tượng mà liên thuộc phân tán giữa các cụm, biên cách (margin) giữa hai cụm đứng đầu nhỏ, entropy cao hoặc có mức bất đồng lớn giữa các mô hình/khởi tạo (committee) — tinh thần cốt lõi của nhiều khung truy vấn ràng buộc chủ động [30, 98]. Bên cạnh đó, tính đại diện (representativeness) giúp tránh lãng phí truy vấn vào vùng nhiễu/thưa thớt, bằng cách ưu tiên các điểm/cặp ở vùng mật độ cao (ví dụ dựa trên kNN-density) hoặc các mẫu "điển hình" của lõi cụm [97]. Tiếp theo, tiêu chí ảnh hưởng/kỳ vọng đổi mô hình (influence/expected model change) trực tiếp ước lượng mức dịch chuyển của hàm mục tiêu, ma trận liên thuộc hoặc vị trí tâm cụm nếu nhận được phản hồi, từ đó ưu tiên các truy vấn có tác động lớn [98]. Cuối cùng, trong thiết lập truy vấn theo lô, đa dạng (diversity) đảm bảo thông tin thu về không trùng lặp bằng các cơ chế mô-đun con kiểu MMR để khử dư thừa — một yêu cầu thường xuyên được nhấn mạnh trong các khảo sát về phân cụm có ràng buộc [34]. Trên thực tế, các tiêu chí này thường được kết hợp để cân bằng giữa "nơi mô hình bối rối", "nơi đại diện cho cấu trúc", và "nơi có tác động lớn" mà vẫn duy trì sự phong phú thông tin trong mỗi đợt truy vấn.

C. Các phương pháp phân cụm bán giám sát mờ chủ động

Một trong những công trình quan trọng đặt nền móng cho hướng tiếp cận này là của Grira, Crucianu và Boujemaa (2008), với thuật toán phân cụm bán giám sát chủ động sử dụng cặp ràng buộc (Active fuzzy constrained clustering-AFCC)

[31, 99]. Ý tưởng chính là lựa chọn các cặp ràng buộc (must-link/cannot-link) một cách chủ động tại vùng biên không chắc chắn. Thuật toán sử dụng chỉ số độ chặt cụm (Fuzzy Hypervolume - FHV) để xác định cụm ít rõ ràng, trích các điểm biên (độ phụ thuộc thấp), rồi ghép với điểm gần nhất ở cụm lân cận để đưa vào tập truy vấn. Phương pháp này cho thấy có thể cải thiện đáng kể chất lượng phân cụm với số lượng nhỏ ràng buộc, tuy nhiên nhược điểm là phụ thuộc vào tham số (ngưỡng độ phụ thuộc, số cặp hỏi) và tốn chi phí tính toán trên dữ liệu lớn.

Hàm chi phí của AFCC kết hợp ba thành phần: độ chặt chẽ của cụm, sự tuân thủ ràng buộc, và thành phần điều chuẩn cạnh tranh để kiểm soát số cụm:

$$\begin{aligned}\mathcal{J}(V,U) = &\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2\, d^2(x_i,\mu_k) \\ &+ \alpha\left(\sum_{(x_i,x_j)\in\mathcal{M}}\sum_{k=1}^{C}\sum_{\substack{l=1\\ l\neq k}}^{C} u_{ik}u_{jl} + \sum_{(x_i,x_j)\in\mathscr{C}}\sum_{k=1}^{C} u_{ik}u_{jk}\right) \\ &- \beta\sum_{k=1}^{C}\left[\sum_{i=1}^{N} u_{ik}\right]^2 .\end{aligned} \tag{1.10}$$

Trong đó:

- C: số cụm,
- N: số điểm dữ liệu,
- u_{ik}: mức độ hội viên của điểm x_i với cụm k,
- μ_k: tâm cụm k,
- $d^2(x_i,\mu_k)$: khoảng cách bình phương giữa x_i và μ_k,
- $\mathcal{M}$ và $\mathscr{C}$: tập ràng buộc must-link và cannot-link,
- α: hệ số điều chỉnh ảnh hưởng của cặp ràng buộc,

- β: hệ số điều chỉnh ảnh hưởng của độ lớn cụm.

Phân rã mức độ phụ thuộc u_{rs} Mức độ phụ thuộc của điểm dữ liệu x_r vào cụm s được chia thành ba thành phần:

$$u_{rs} = u_{rs}^{\text{FCM}} + u_{rs}^{\text{constr}} + u_{rs}^{\text{bias}}, \tag{1.11}$$

trong đó:

$$u_{rs}^{FCM} = \frac{\frac{1}{d^2(x_r,\mu_s)}}{\sum_{k=1}^{C} \frac{1}{d^2(x_r,\mu_k)}}, \tag{1.12}$$

$$u_{rs}^{Constraints} = \frac{\alpha\Big(C_{v_r} - C_{v_{rs}}\Big)}{2\, d^2(x_r, \mu_s)}, \tag{1.13}$$

$$u_{rs}^{Bias} = \frac{\beta\,(N_s - \overline{N}_r)}{d^2(x_r, \mu_s)}. \tag{1.14}$$

Trong đó $C_{v_{rs}}$ là chi phí vi phạm ràng buộc khi gán x_r cho cụm s, C_{v_r} là chi phí trung bình theo tất cả các cụm, N_s là số phần tử trong cụm s, và $\overline{N}_r$ là giá trị trung bình có trọng số của kích thước cụm đối với điểm x_r.

Cập nhật tâm cụm và độ lớn cụm

$$\mu_k = \frac{\sum_{i=1}^{N} u_{ik}^2 x_i}{\sum_{i=1}^{N} u_{ik}^2}, \qquad N_k = \sum_{i=1}^{N} u_{ik}. \tag{1.15}$$

Tính động các hệ số α **và** β

$$\alpha = \frac{\frac{N}{M} \sum_{k=1}^{C} \sum_{i=1}^{N} u_{ik}^2 d^2(x_i, \mu_k)}{\sum_{k=1}^{C} \sum_{i=1}^{N} u_{ik}^2}, \tag{1.16}$$

$$\beta(t) = \frac{\eta_0 \exp\Big(-\frac{|t-t_0|}{\tau}\Big)}{\sum_{j=1}^{C} \Big(\sum_{i=1}^{N} u_{ij}\Big)^2} \times \left[\sum_{j=1}^{C} \sum_{i=1}^{N} u_{ij}^2 d^2(x_i, \mu_j) + \alpha\, \Phi(\mathcal{M}, \mathscr{C})\right], \tag{1.17}$$

với $\Phi(\mathcal{M}, \mathscr{C})$ ký hiệu gọn cho các tổng liên quan đến ràng buộc. β là yếu tố cân

bằng giữa các thành phần ở hàm mục tiêu và được cập nhật theo các vòng lặp.

Ngoài khoảng cách Euclid, AFCC sử dụng khoảng cách Mahalanobis để xử lý cụm elip:

$$d^2(x_i, \mu_k) = |C_k|^{1/p}(x_i - \mu_k)^\top C_k^{-1}(x_i - \mu_k), \tag{1.18}$$

với ma trận hiệp phương sai

$$C_k = \frac{\sum_{i=1}^{N} u_{ik}^2 (x_i - \mu_k)(x_i - \mu_k)^\top}{\sum_{i=1}^{N} u_{ik}^2}. \tag{1.19}$$

Trong AFCC, cặp hỏi được sinh có chủ đích tại nơi có giá trị phụ thuộc mơ hồ: mỗi vòng lặp chọn cụm kém rõ theo Fuzzy Hypervolume rồi lấy các điểm có cực đại u_{rs} thấp, ghép với điểm gần nhất ở cụm lân cận (từ giá trị hội viên lớn nhì) để hỏi ML/CL. Để giảm chi phí, thuật toán dùng hai xấp xỉ: Ambiguousness (chọn nhanh các điểm có $u < 0.3$ làm biên mở rộng) và Non-redundancy (chọn tuần tự theo tiêu chí max–min khoảng cách trong tập ứng viên). Nhờ vậy, chỉ cần rất ít truy vấn mà vẫn thu được các cặp giàu thông tin. Những cặp này được gọi là Most Valuable Pairs (MVP) vì đem lại lợi ích lớn cho việc làm rõ ranh giới cụm.

Thuật toán AFCC được mô tả thông qua thuật toán 1.4 như dưới đây:

Thuật toán 1.4 Thuật toán phân cụm mờ có ràng buộc chủ động (AFCC)

Đầu vào: X, $C_{\max}$, $T_{\max}$.
Đầu ra: U, V, $\mathcal{M}, \mathcal{C}$.

$C \leftarrow C_{\max}$, U, V khởi tạo; $\mathcal{M}, \mathcal{C} \leftarrow \emptyset$.
for $t = 1 : T_{\max}$ **do**
    Cập nhật U theo d_{ij} (Euclid/Mahalanobis) và ba thành phần mục tiêu.
    $v_k \leftarrow \frac{\sum_i u_{ik}^m x_i}{\sum_i u_{ik}^m}$, $N_k \leftarrow \sum_i u_{ik}^m$.
    Cập nhật $\alpha(t), \beta(t)$ theo quy tắc động.
    Thực hiện Competitive Agglomeration: $\{k : N_k < \tau_s\} \rightarrow$ loại, $\|v_k - v_\ell\| < \tau_m \rightarrow$ gộp.
    Xác định cụm mờ nhất: $k^\star = \arg\max_k H_k$ (Fuzzy Hypervolume).
    Chọn điểm biên: $\mathcal{B} = \{x_i : \max_j u_{ij} < \tau_{\text{bnd}}\}$, lọc không dư thừa.
    Sinh m MVP pairs: $(x_i, x_j) \in \mathcal{B} \times \text{NN}(k^\star)$, truy vấn nhãn $\Rightarrow$ thêm vào $\mathcal{M}$ hoặc $\mathcal{C}$.
    if hội tụ **then**
        break
    end if
end for
return $U, V, \mathcal{M}, \mathcal{C}$.

Kết quả thực nghiệm của AFCC chứng minh rằng chỉ cần một số lượng nhỏ ràng buộc must-link/cannot-link được lựa chọn có chủ đích, chất lượng phân cụm đã được cải thiện rõ rệt. Điều này khiến AFCC trở thành một cách tiếp cận bán giám sát hấp dẫn cho việc tổ chức cơ sở dữ liệu hình ảnh quy mô lớn.

Dựa trên phương pháp AFCC được đề xuất bởi Grira [31], Novoselova và Tom (2017) [32] mở rộng bằng cách thay đổi chiến lược lựa chọn các cặp ràng buộc. Cụ thể, trong khi mô hình gốc sử dụng cơ chế chọn cặp dựa trên fuzzy hypervolume và các tiêu chí biên mờ để phát hiện các vùng khó phân tách, thì nghiên cứu năm 2017 đề xuất một thuật toán active constraint selection dựa trên đồ thị k-nearest neighbors (k-NNG):

$$w(x_i, x_j) = \Big| NN(x_i) \cap NN(x_j) \Big|, \tag{1.20}$$

trong đó $NN(x)$ là tập k láng giềng gần nhất của x. Độ hữu ích (*utility*) cho một cặp ứng viên (x_i, x_j) được định nghĩa là

$$ASC(x_i, x_j) = \frac{k - w(x_i, x_j) + 1}{k + 1} \cdot \frac{1}{1 + \min\{\, LDS(x_i),\, LDS(x_j)\,\}}. \tag{1.21}$$

trong đó $LDS(x)$ đo mật độ cục bộ. Các cặp có ASC cao nhất được truy vấn (ML/CL) và có thể lan truyền ràng buộc để giảm dư thừa. Kết quả cho thấy đạt chất lượng tương đương với ít ràng buộc hơn so với chọn ngẫu nhiên, đồng thời hỗ trợ xác định đúng số cụm trong quá trình gộp cạnh tranh.

Gần đây, Gabriel Santos và cộng sự (2024) đã đề xuất K-GBS3FCM [100], một biến thể bán giám sát mờ "an toàn" dựa trên đồ thị KNN. Điểm mới của phương pháp này là sử dụng cấu trúc lân cận KNN để đánh giá "độ an toàn" của nhãn trước khi lan truyền ảnh hưởng đến các điểm chưa nhãn. Cơ chế này giúp giảm thiểu rủi ro do nhãn sai hoặc không chắc chắn. Tuy vậy, mô hình vẫn phụ thuộc mạnh vào tham số số lân cận K và chi phí xây dựng đồ thị KNN, vốn trở nên đắt đỏ khi kích thước dữ liệu tăng. Tiếp theo, Hong và cộng sự (2025) đã phát triển SFCM-PM [19], trong đó thành phần mới là khai thác thông tin độ phụ thuộc trước (độ phụ thuộc cho trước) bằng một điều chuẩn entropy trong hàm mục tiêu. Điều này giúp mô hình tận dụng tri thức miền để định hướng quá trình phân cụm, đặc biệt hiệu quả khi dữ liệu mất cân bằng. Điểm hạn chế là nếu thông tin trước bị nhiễu hoặc sai lệch, kết quả có thể bị ảnh hưởng tiêu cực; đồng thời việc lựa chọn trọng số giữa dữ liệu và độ phụ thuộc cho trước cũng khá nhạy cảm.

D. Nhận xét chung và vấn đề còn tồn tại:

Tóm lại, qua các nghiên cứu trên có thể thấy phân cụm bán giám sát mờ chủ động đã tiến triển từ việc chọn cặp truy vấn thông minh [31], sang hướng đảm bảo "an toàn" trong sử dụng nhãn [100], và gần đây là tích hợp thông tin độ phụ thuộc trước [19]. Điểm mạnh chung của các phương pháp này là cải thiện đáng kể chất lượng phân cụm với số lượng nhãn ít, đồng thời tận dụng tri thức miền hoặc cấu trúc dữ liệu để giảm nhiễu. Tuy nhiên, hạn chế chung của các phương pháp này là vẫn chưa có cơ chế xử lý triệt để các vấn đề phân cụm không chắc chắn vùng biên cũng như chưa có cơ chế tận dụng được nguồn thông tin quý giá khi truy vấn ở vùng biên khi phân cụm.

E. Các đề xuất của luận án:

Từ các hạn chế của các phương pháp phân cụm bán giám sát mờ chủ động gần đây, luận án định hướng xây dựng chiến lược để giải quyết các vấn đề này và đề xuất như sau:

- Đề xuất phương pháp xác định biên cụm, nơi các cụm có thể chồng lấn khiến cho các điểm dữ liệu ở vùng này bị phân cụm không chắc chắn, dẫn đến sai số tập trung quanh biên cụm.
- Đề xuất phương pháp chỉnh sửa vùng biên thông qua học chủ động nhằm cải thiện chất lượng phân cụm tại các vùng có độ mơ hồ cao, tận dụng phản hồi có chọn lọc từ chuyên gia để tinh chỉnh nhãn và ranh giới cụm.
- Đề xuất thuật toán phân cụm bán giám sát mờ chủ động dựa vào vùng biên nhằm nâng cao hiệu quả phân cụm tổng thể, bằng cách khai thác tri thức đã truy vấn và tinh chỉnh thu được từ vùng biên.

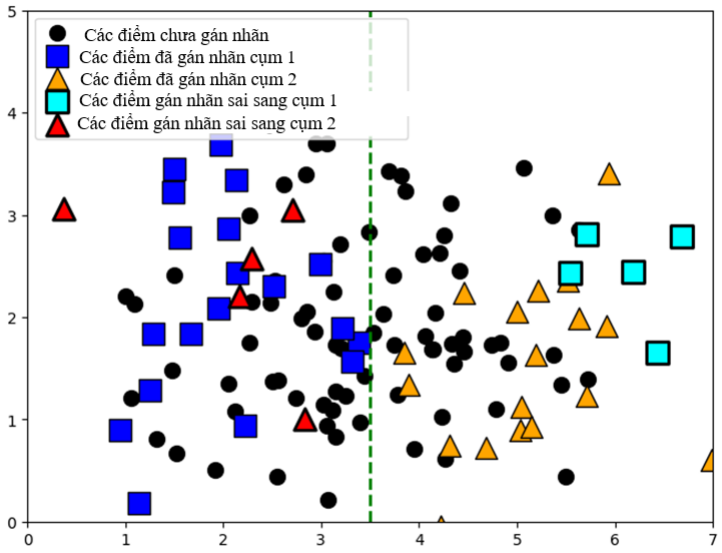
1.3.2 Phân cụm bán giám sát mờ an toàn

A. Bối cảnh và động lực:

Trong bối cảnh phân cụm bán giám sát mờ (SSFC), hiệu quả mô hình phụ thuộc chặt chẽ vào chất lượng dữ liệu và tín hiệu giám sát (nhãn hạt giống, ràng buộc must-link/cannot-link). Khi hai nguồn này thiếu tin cậy, quá trình tối ưu dễ lệch khỏi cấu trúc thật của dữ liệu: nhãn do con người gán có thể nhiễu hoặc không nhất quán giữa chuyên gia, kéo các giá trị liên thuộc μ_{ik} về những tâm không phù hợp; ràng buộc theo cặp có thể mâu thuẫn hoặc chỉ phản ánh quan hệ cục bộ, không đại diện cho hình học cụm toàn cục; điểm ngoại lai và điểm vùng biên mang tính ngẫu nhiên cao nên trở thành điểm kéo gây thiên lệch nếu bị neo quá cứng; dữ liệu khuyết thuộc tính làm tăng phương sai của khoảng cách, từ đó ảnh hưởng trực tiếp tới cập nhật tâm v_k và ma trận liên thuộc; cuối cùng, lệch mật độ và mất cân bằng kích thước cụm khiến các tiêu chí dựa trên khoảng cách/mật độ trở nên nhạy tham số, cụm nhỏ–đậm có thể lấn át cụm thưa–rộng. Đáng chú ý, độ tin cậy trong thực tế hiếm khi đồng nhất: nó biến thiên theo điểm (instance-level), cặp (pairwise-level), cụm (cluster-level) và thậm chí theo thuộc tính (feature-level).

Hình 1.10 minh họa một tập dữ liệu gồm hai cụm: các hình vuông thể hiện các điểm có nhãn thuộc lớp 1, các tam giác đại diện cho lớp 2, và các vòng tròn đen là dữ liệu chưa có nhãn. Một số điểm bị gán nhãn sai (hình vuông màu cyan và tam giác đỏ), ảnh hưởng đến việc xác định ranh giới quyết định (đường chấm xanh lá), dẫn đến việc phân cụm sai lệch.

Những điều kiện bất lợi này dẫn tới ba nguy cơ chủ đạo: (i) khuếch đại sai lệch khi các nhãn/ràng buộc nhiễu được lan truyền qua các vòng lặp cập nhật, làm méo ranh giới quyết định; (ii) tối ưu mục tiêu sai khi các quan sát có độ tin



Hình 1.10: Hình minh họa khi dữ liệu bị gán nhãn sai

cậy thấp vẫn được gán trọng số như nhau trong hàm mục tiêu; và (iii) hội tụ cục bộ do mô hình bị "kéo" bởi một số ràng buộc mạnh nhưng không đáng tin, đặc biệt trong miền biên nơi mật độ thay đổi và các cụm giao thoa. Do đó, đặt ra yêu cầu phương pháp luận trọng tâm của SSFC là nhận diện, điều tiết và xử lý độ tin cậy trong toàn bộ quy trình.

B. Các phương pháp phân cụm bán giám sát mờ an toàn gần đây:

Gần đây, trong các thuật toán bán giám sát mờ, một số cơ chế học an toàn đã được bổ sung và thường được tích hợp nhằm cải thiện quá trình phân cụm, chẳng hạn như hàm trọng số và các chiến lược thích ứng. Haitao Gan và cộng sự [17, 21, 101] chính là những nhà nghiên cứu đi đầu cho xu hướng học an toàn đã đề xuất một số phương pháp phân cụm bán giám sát mờ an toàn như S^3FCM, LHC-S3FCM và CS3FCM. Trong số đó, phương pháp CS3FCM [17] cho thấy hiệu suất ổn định qua các thực nghiệm và được trình bày chi tiết dưới đây. Ý tưởng chính của CS3FCM là mỗi điểm dữ liệu có ảnh hưởng khác nhau đến chất

lượng phân cụm.

Cụ thể, có hai tập dữ liệu: tập dữ liệu có nhãn $X = [x_1, x_2, \ldots, x_j]$ và tập dữ liệu chưa có nhãn $X_u = [x_{j+1}, x_{j+2}, \ldots, x_n]$. C là số lượng cụm. Phần tử x_k có nhãn $y_k \in \{1, \ldots, C\}$.

Trong CS3FCM, toàn bộ dữ liệu được phân chia thành C cụm bằng thuật toán FCM. Sau đó, ma trận phân cụm được tính toán: $\hat{U} = [\hat{u}]_{c\times n}$ và nhãn đầu ra được ước lượng $\hat{Y} = [\hat{y_1}, \ldots, \hat{y_n}]$. Sử dụng thuật toán Kuhn-Munkres, nhãn ước lượng $\hat{y_k}|_{k=1}^{l}$ và nhãn thực được so sánh để tạo ra ma trận N_c chứa các phần tử ρ_{ij}, trong đó ρ_{ij} đo lường xác suất phần tử thuộc lớp i được phân loại thành lớp j.

$$N_c = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1c} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{c1} & \rho_{c2} & \cdots & \rho_{cc} \end{bmatrix} \tag{1.22}$$

với điều kiện $\sum_{j=1}^{c} u_j = 1$ và $0 \leq \rho_j \leq 1$.

Với phần tử có nhãn x_k, nếu $y_k = \hat{y_k}$ và $p_{y_k,\hat{y_k}}$ lớn thì độ tin cậy (safe confidence) của nó cao. Trọng số s_k của x_k được tính như sau:

$$s_k = \begin{cases} p_{y_k,\hat{y_k}} \times \hat{u}_{y_k} & \text{nếu } y_k = \hat{y_k} \\ p_{y_k,\hat{y_k}} \times (1 - \hat{u}_{y_k}) & \text{ngược lại} \end{cases} \tag{1.23}$$

Tác giả cũng xây dựng đồ thị lân cận $W = [w_{lr}]_{n\times n}$ để xác định lân cận (với dữ liệu chưa có nhãn) và trọng số đồ thị:

$$w_{lr} = \begin{cases} \exp\left(\frac{-\|x_k - x_l\|^2}{\delta}\right) & \text{nếu } x_l \in N_p(x_k) \text{ và } \hat{y_k} = \hat{y_l} \\ 0 & \text{ngược lại} \end{cases} \tag{1.24}$$

trong đó $N_p(x_k)$ là tập các điểm lân cận gần nhất của x_k, và x_r là phần tử chưa có nhãn.

Hàm mục tiêu của CS3FCM là:

$$\begin{aligned} J_c = &\sum_{k=1}^{N}\sum_{i=1}^{C} u_{ik}^2 d_{ik}^2 + \gamma_1 \sum_{i=1}^{C}\sum_{i=1}^{N_l}(u_{ik} - f_i)^2 d_{ik}^2 \\ &+ \gamma_2 \sum_{r,s=1}^{N} \frac{1}{s_k} W_{rs} \sum_{k=1}^{C}(u_k - u_r)^2 \to \min \end{aligned} \tag{1.25}$$

với điều kiện $\sum_{i=1}^{C} u_{ik} = 1, \quad \forall k = 1, \ldots, N$

trong đó γ_1 và γ_2 là các tham số điều chỉnh.

Mức độ thành viên u_{ik} cho phần tử có nhãn x_k được tính như sau:

$$u_{ik} = \frac{p_{ik} + \frac{1 - \sum_{j=1}^{c} \frac{p_{ik}}{q_{ik}}}{\sum_{i=1}^{c} \frac{1}{q_{ik}}}}{q_{ik}} \tag{1.26}$$

Mức độ thành viên u_{ir} cho phần tử chưa có nhãn x_r được tính như sau:

$$u_{ir} = \frac{z_{ir} + \frac{1 - \sum_{i=1}^{c} \frac{z_{ir}}{t_{ir}}}{\sum_{i=1}^{c} \frac{1}{t_{ir}}}}{q_{ir}} \tag{1.27}$$

Tâm cụm v_i được tính như sau:

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^2 x_k + \lambda_1 \sum_{k=1}^{N} s_k (u_{ik} - f_{ik})^2 x_k}{\sum_{k=1}^{N} u_{ik}^2 + \lambda_1 \sum_{k=1}^{N} s_k (u_{ik} - f_{ik})^2} \tag{1.28}$$

tuy nhiên vẫn còn nhiều thách thức, đặc biệt là trong việc xử lý các cặp ràng buộc (pairwise constraints) và đánh giá rủi ro giữa các điểm dữ liệu có nhãn và chưa có nhãn.

Để khắc phục vấn đề này, nhóm nghiên cứu của Thong và Huan đã đề xuất phương pháp phân cụm bán giám sát mờ tin cậy gọi là TS3FCM (Trusted Safe Semi-Supervised Fuzzy Clustering Method). Phương pháp này triển khai theo hai giai đoạn: đầu tiên, sàng lọc các dữ liệu có nhãn để xác định mức độ tin cậy; sau đó thực hiện SSFCM sử dụng các nhãn đáng tin cậy cùng với thông tin thành viên ban đầu (độ phụ thuộc cho trước). Thuật toán TS3FCM [18] được phát triển nhằm giải quyết bài toán phân cụm dữ liệu với độ tin cậy trong bối cảnh dữ liệu có nhãn bị nhiễu hoặc không đáng tin cậy. Thiết kế của TS3FCM xuất phát từ các hạn chế trong các phương pháp phân cụm bán giám sát mờ an toàn (Safe SSFC) trước đó như hiệu suất tính toán thấp và nhạy cảm với nhiễu.

Thuật toán gồm ba giai đoạn:

- **(1) Phân cụm FCM cho dữ liệu có nhãn**: sử dụng hàm mục tiêu cải tiến với trọng số dựa trên phân tích láng giềng để lọc ra nhãn có độ tin cậy thấp.
- **(2) Chuyển đổi dữ liệu**: sử dụng các nhãn đáng tin cậy để tính giá trị hội viên ban đầu cho dữ liệu chưa có nhãn, tạo thành ma trận hội viên ban đầu $\bar{U}$.
- **(3) phân cụm bán giám sát mờ**: tối ưu hóa quá trình phân cụm cho toàn bộ tập dữ liệu.

Giai đoạn 1: Phân cụm dữ liệu có nhãn với FCM cải tiến

Hàm mục tiêu được định nghĩa như sau:

$$J = \sum_{k=1}^{L} \sum_{i=1}^{C} \frac{n_{1k} + n_{2k}}{n_{3k} + 1} u_{ki}^{m} d_{ki}^{2} \rightarrow \min \tag{1.29}$$

Với các ràng buộc:

$$u_i \in [0,1]; \quad k = 1, ..., L, \quad i = 1, ..., C \tag{1.30}$$

$$\sum_{i=1}^{C} u_{ki} = 1; \quad k = 1, ..., L \tag{1.31}$$

Trong đó:

- n_{1k}: số lượng láng giềng không có nhãn của điểm x_k
- n_{2k}: số lượng láng giềng có cùng nhãn với x_k
- n_{3k}: số lượng láng giềng có nhãn khác với x_k
- d_{ki}: khoảng cách từ điểm x_k đến tâm cụm thứ i

Cập nhật tâm cụm:

$$v_i = \frac{\sum_{k=1}^{L} \frac{n_{1k}+n_{2k}}{n_{3k}+1} u_{ik}^m x_k}{\sum_{k=1}^{L} \frac{n_{1k}+n_{2k}}{n_{3k}+1} u_{ik}^m}, \quad i = 1, ..., C \tag{1.32}$$

Cập nhật giá trị hội viên:

$$u_{ki} = \frac{1}{\sum_{j=1}^{C} \left(\frac{d_{ki}}{d_{kj}}\right)^{\frac{2}{m-1}}}, \quad k = 1, ..., L, \quad i = 1, ..., C \tag{1.33}$$

Sau mỗi vòng lặp, áp dụng hàm rút gọn (defuzzification):

$$u_{ik} = \begin{cases} \frac{u_{ik}}{2}, & \text{nếu } l(i) = l(x_k) \\ u_{ik} + \frac{u_{kj}}{2(C-1)}, & \text{nếu } i \neq j \text{ và } l(j) = l(x_k) \end{cases} \tag{1.34}$$

Điều kiện dừng: $||v^{(t)} - v^{(t-1)}|| \leq \varepsilon$ hoặc $t >$ Maxsteps.

Thuật toán 1.5 FCM cải tiến

Đầu vào: Tập dữ liệu X với N phần tử, số phần tử có nhãn: $L < N$, số cụm C, ngưỡng ε, tham số mờ $m = 2$, số vòng lặp tối đa $Maxsteps$.
Đầu ra: Ma trận hội viên U và tâm cụm V.

1 $t \leftarrow 0$
2 Khởi tạo tâm cụm $v_i^{(0)}$ ngẫu nhiên
3 **Thực hiện vòng lặp**:
4 $\quad t \leftarrow t + 1$
5 $\quad$ Tính $u_{ki}^{(t)}$ theo phương trình 1.33
6 $\quad$ Áp dụng rút gọn hội viên bằng phương trình 1.34
7 $\quad$ Cập nhật tâm cụm $v_i^{(t)}$ bằng phương trình 1.32
8 **Cho đến khi:** thỏa mãn điều kiện dừng

Giai đoạn 2 và 3: Phân cụm bán giám sát

Hàm mục tiêu:

$$J_{TS3FCM} = \sum_{k=1}^{N}\sum_{i=1}^{C} u_{ki} d_{ki}^2 + \lambda \sum_{k=1}^{N}\sum_{i=1}^{C} (u_{ki} - \bar{u}_{ki})^2 d_{ki}^2 \to \min \tag{1.35}$$

Ràng buộc:

$$u_{ki} \in [0,1]; \quad \sum_{i=1}^{C} u_{ki} = 1 \quad \text{với mọi } k \tag{1.36}$$

Tính toán tâm cụm:

$$v_i = \frac{\sum_{k=1}^{N} \left(u_{ki}^2 + \lambda (u_{ki} - \bar{u}_{ki})^2\right) x_k}{\sum_{k=1}^{N} \left(u_{ki}^2 + \lambda (u_{ki} - \bar{u}_{ki})^2\right)}, \quad i = 1, \ldots, C \tag{1.37}$$

Cập nhật hội viên:

$$u_{ki} = \frac{1 + \lambda - \lambda \cdot \sum_{j=1}^{C} \bar{u}_{kj}}{(1+\lambda) \sum_{j=1}^{C} \left(\frac{d_{ki}}{d_{kj}}\right)^2} - \frac{\lambda \bar{u}_{kj}}{1+\lambda}, \quad k = 1, \ldots, N, i = 1, \ldots, C \tag{1.38}$$

TS3FCM hướng đến mục tiêu "phân hoạch dữ liệu với độ tin cậy cao" và

Thuật toán 1.6 Thuật toán phân cụm bán giám sát mờ an toàn

Đầu vào: Tập dữ liệu X với N phần tử, số phần tử có nhãn: $L < N$, số cụm C, ngưỡng ε, tham số mờ $m = 2$, hệ số điều chỉnh λ, ma trận hội viên ban đầu U
Đầu ra: Ma trận hội viên cuối cùng U và tâm cụm V.

1 $t \leftarrow 0$
2 **Thực hiện vòng lặp**:
3 $t \leftarrow t + 1$
4 Tính $u_{ki}^{(t)}$ theo phương trình 1.38
5 Tính $u_i^{(t)}$ theo phương trình 1.37
6 **Cho đến khi**: $\left\| v_i^{(t)} - v_i^{(t-1)} \right\| \leq \varepsilon$ hoặc $t > Maxsteps$

cho thấy một số ưu điểm vượt trội so với các phương pháp trước đó như Safe SSFC của Gan, đặc biệt trong việc cải thiện hiệu quả tính toán và độ tin cậy mô hình. Tuy vậy, TS3FCM vẫn còn một số hạn chế đáng kể: phương pháp phụ thuộc vào việc khởi tạo ngẫu nhiên các tâm cụm, sử dụng công thức bán kính lân cận cố định, và gặp khó khăn khi xử lý các điểm nhiễu, ngoại lệ, đặc biệt là trong các vùng biên giao nhau giữa các cụm mờ.

D. Nhận xét chung và vấn đề còn tồn tại:

Các phương pháp phân cụm bán giám sát mờ hiện nay chủ yếu sử dụng cơ chế đánh trọng số và sàng lọc dữ liệu có nhãn thông qua hàng xóm gần, các cơ chế này đã tạo hiệu quả đáng kể trong việc giảm các thông tin giám sát sai lệch khiến cho hiệu quả phân cụm tăng nhất là đối với các dữ liệu có độ tin cậy thấp. Tuy nhiên các phương pháp trên vẫn còn hạn chế do còn nhạy cảm với việc khởi tạo ngẫu nhiên, việc xử lý nhiễu cũng chưa thực sự hiệu quả với những dữ liệu có độ nhiễu cao nhất là ở vùng biên cụm.

E. Các đề xuất của luận án:

Dựa vào các điểm hạn chế còn tồn tại của các phương pháp cũ, luận án đã nghiên cứu và đưa ra một số đề xuất nhằm cải thiện các hạn chế cũng như nâng cao hiệu quả của phân cụm như sau:

- **Áp dụng phương pháp khởi tạo có chọn lọc kmean++** để cải thiện chất lượng đầu vào của quá trình phân cụm, giúp tăng khả năng hội tụ và độ ổn định của thuật toán.
- **Áp dụng học chủ động trong giai đoạn tiền xử lý** khi xác định hàng xóm của các điểm đã gán nhãn, nhằm truy vấn các nhãn có khả năng sai cao và giảm chi phí gán nhãn.
- **Áp dụng cơ chế đồng thuận** để xử lý việc thay đổi nhãn dữ liệu ngay trong quá trình xác định hàng xóm, đảm bảo sự nhất quán và thích nghi của mô hình.
- **Đề xuất cơ chế xác định vùng biên** nhằm phát hiện các vùng có độ mơ hồ cao, nơi các nhãn có thể bị gán sai, giúp mô hình nhận biết và xử lý tốt hơn các điểm dữ liệu khó phân cụm.
- **Bổ sung cơ chế sử dụng cặp ràng buộc** được tạo ra dựa trên truy vấn biên để tăng độ chính xác của phân cụm và cải thiện khả năng học thích nghi trong các tình huống dữ liệu phức tạp.

1.4 Đánh giá hiệu năng thuật toán phân cụm

Trong luận án, tôi sử dụng bốn độ đo phổ biến để đánh giá chất lượng phân cụm so với nhãn tham chiếu (nếu có) và đánh giá nội tại cấu trúc cụm: RI, F1-score, NMI và DB. Ba độ đo đầu (RI, F1, NMI) được sử dụng để so sánh phân hoạch dự đoán với phân hoạch chuẩn (ground truth); chỉ số DB đánh giá

nội tại dựa trên độ chặt trong cụm và độ tách giữa các cụm. Với RI, F1, NMI, giá trị lớn hơn thể hiện hiệu năng tốt hơn; với DB, giá trị nhỏ hơn thể hiện cụm tốt hơn.

Rand Index (RI) [102]

Gọi $\mathcal{S} = \{(i, j) \mid 1 \le i < j \le n\}$ là tập tất cả các cặp điểm dữ liệu. Kí hiệu TP là số cặp cùng lớp (theo nhãn thật) và cùng cụm (theo mô hình), TN là số cặp khác lớp và khác cụm; FP là số cặp khác lớp nhưng lại bị gom cùng cụm; FN là số cặp cùng lớp nhưng bị tách ra hai cụm khác nhau. Khi đó:

$$\mathrm{RI} \;=\; \frac{\mathrm{TP} + \mathrm{TN}}{\binom{n}{2}} \;\in\; [0, 1]. \tag{1.39}$$

RI phản ánh tỉ lệ các quyết định cặp đúng (cùng–cùng hoặc khác–khác) trong toàn bộ $\binom{n}{2}$ cặp; càng gần 1 càng tốt.

F1-score [40]

Từ cùng các thống kê cặp ở trên, *Precision* và *Recall* lần lượt là:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{1.40}$$

F1-score là trung bình điều hoà của P và R:

$$\mathrm{F1} \;=\; \frac{2\,\mathrm{P}\,\mathrm{R}}{\mathrm{P} + \mathrm{R}} \;=\; \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \;\in\; [0, 1]. \tag{1.41}$$

F1 nhấn mạnh cân bằng giữa bắt đúng các cặp cùng lớp (recall) và tránh gom nhầm các cặp khác lớp (precision).

Normalized Mutual Information (NMI) [103]

Giả sử nhãn thật tạo thành phân hoạch $\mathcal{Y} = \{Y_1, \ldots, Y_L\}$ và mô hình tạo thành phân hoạch dự đoán $\widetilde{\mathcal{Y}} = \{\widetilde{Y}_1, \ldots, \widetilde{Y}_C\}$. Kí hiệu $n_{ij} = |Y_i \cap \widetilde{Y}_j|$, $n_{i\cdot} = \sum_j n_{ij}$, $n_{\cdot j} = \sum_i n_{ij}$ và $n = \sum_{i,j} n_{ij}$. *Thông tin tương hỗ* giữa hai phân hoạch là:

$$I(\mathcal{Y}; \widetilde{\mathcal{Y}}) = \sum_{i=1}^{L} \sum_{j=1}^{C} \frac{n_{ij}}{n} \log\!\left(\frac{n_{ij}\, n}{n_{i\cdot}\, n_{\cdot j}}\right), \tag{1.42}$$

với quy ước hạng tử bằng 0 khi $n_{ij} = 0$. Entropy tương ứng:

$$H(\mathcal{Y}) = -\sum_{i=1}^{L} \frac{n_{i\cdot}}{n} \log\!\left(\frac{n_{i\cdot}}{n}\right), \quad H(\widetilde{\mathcal{Y}}) = -\sum_{j=1}^{C} \frac{n_{\cdot j}}{n} \log\!\left(\frac{n_{\cdot j}}{n}\right). \tag{1.43}$$

Dạng chuẩn hoá đối xứng:

$$\mathrm{NMI} = \frac{I(\mathcal{Y}; \widetilde{\mathcal{Y}})}{\sqrt{H(\mathcal{Y})\, H(\widetilde{\mathcal{Y}})}} \in [0, 1]. \tag{1.44}$$

NMI tiến gần 1 khi phân hoạch dự đoán trùng khớp mạnh với nhãn thật.

Davies–Bouldin (DB) [104]

DB là chỉ số nội tại đánh giá độ chặt và độ tách của các cụm trong phân hoạch *cứng*. Gọi V_i là tâm cụm, T_i là số điểm thuộc cụm i. Độ phân tán trong cụm:

$$S_i = \sqrt{\frac{1}{T_i} \sum_{x \in i} \|x - V_i\|^2}, \tag{1.45}$$

và độ cách biệt $M_{ij} = \|V_i - V_j\|$. Chỉ số DB:

$$\mathrm{DB} = \frac{1}{C} \sum_{i=1}^{C} \max_{j \neq i} \frac{S_i + S_j}{M_{ij}}, \tag{1.46}$$

giảm khi cụm chặt (S_i nhỏ) và tách xa (M_{ij} lớn). DB được tính sau bước defuzzification (gán cụm theo độ phụ thuộc lớn nhất), bổ sung góc nhìn hình

học cho RI/F1/NMI.

Partition Coefficient (PC) và Partition Entropy (PE) [105]

PC và PE là các chỉ số nội tại đánh giá mức độ "sắc nét" hay "mờ" trong phân hoạch mờ. Với ma trận độ phụ thuộc $U = (u_{ik})$:

$$\mathrm{PC} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C} u_{ik}^2 \in \left[\frac{1}{C}, 1\right], \tag{1.47}$$

PC càng cao $\rightarrow$ độ phụ thuộc càng sắc, biên cụm rõ ràng.

Entropy phân hoạch:

$$\mathrm{PE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C} u_{ik}\log u_{ik}, \tag{1.48}$$

PE càng thấp $\rightarrow$ phân hoạch càng rõ (ít mờ), phù hợp để đánh giá biên cụm trong dữ liệu chồng lấn.

Fuzzy Davies–Bouldin ($\mathrm{DB}_{\mathrm{fuzzy}}$) [105]

$\mathrm{DB}_{\mathrm{fuzzy}}$ là biến thể mờ của Davies–Bouldin, phản ánh đúng bản chất phân cụm mờ:

$$S_k = \sqrt{\frac{\sum_i u_{ik}^m \|x_i - V_k\|^2}{\sum_i u_{ik}^m}}, \qquad M_{kl} = \|V_k - V_l\|. \tag{1.49}$$

Khi đó:

$$\mathrm{DB}_{\mathrm{fuzzy}} = \frac{1}{C}\sum_{k=1}^{C}\max_{l\neq k}\frac{S_k + S_l}{M_{kl}}. \tag{1.50}$$

Chỉ số này đặc biệt phù hợp cho phân cụm bán giám sát mờ vì điểm gần biên (độ phụ thuộc thấp) có ảnh hưởng ít hơn so với điểm lõi cụm.

Tổng quan: Bộ bảy độ đo **RI, F1, NMI, PC, PE, DB/$\mathrm{DB}_{\mathrm{fuzzy}}$** cung cấp đánh giá toàn diện: - RI/F1/NMI phản ánh mức độ phù hợp ngữ nghĩa so với nhãn thật (ngoại tại), - PC/PE và $\mathrm{DB}_{\mathrm{fuzzy}}$ đánh giá độ sắc nét và hình học

của cụm (nội tại), - DB (cứng) cho phép so sánh mức độ tách/chặt sau khi defuzzification.

Sự kết hợp này đảm bảo việc đánh giá mô hình phân cụm bán giám sát mờ vừa đúng lý thuyết độ phụ thuộc, vừa phản ánh cấu trúc thật của dữ liệu.

1.5 Kết luận Chương 1

Chương 1 đã hệ thống hoá các nền tảng lý thuyết phục vụ trực tiếp cho bài toán phân đoạn ảnh. Trước hết, luận án nhắc lại những khái niệm cốt lõi về tập mờ và các mở rộng, cơ chế gán độ phụ thuộc và hàm mục tiêu của FCM, từ đó dẫn dắt lên phân cụm bán giám sát mờ (SSFC) — nơi thông tin phụ (nhãn hạt giống, scribbles, ràng buộc must–link/cannot–link, cấu trúc lân cận) được tích hợp để điều chỉnh phân hoạch. Trên cơ sở đó, chương đã điểm lược các nhánh tiếp cận SSFC gần đây như SSFC "an toàn"/"tin cậy", SSFC trên các mô hình tập mờ nâng cao, SSFC với nhiều hệ số mờ (multi–fuzzifier), cùng vai trò của học chủ động trong việc thiết kế truy vấn (điểm/cặp/cấu trúc) nhằm tối đa lượng thông tin thu được dưới ràng buộc chi phí gán nhãn thấp. Cũng trong chương này, bài toán phân đoạn ảnh, các ứng dụng thực tế cũng như việc áp dụng phương pháp phân cụm trong bài toán phân đoạn ảnh đã được trình bày. Bên cạnh đó chương cũng xác lập bộ tiêu chí đánh giá hiệu năng xuyên suốt luận án.

Từ các nghiên cứu tổng quan trong chương này, luận án đã tìm ra khoảng trống khoa học trong các phương pháp phân cụm bán giám sát mờ hiện nay nằm ở sự gắn kết chặt chẽ giữa mềm hoá ranh giới, độ tin cậy và tính chủ động của truy vấn. Dựa trên nền các khoảng trống nêu trên, luận án định hướng nghiên cứu giải quyết vấn đề theo các hướng sau: (i) học chủ động tập trung vùng biên nhằm "cố định" ranh giới nơi sự mập mờ độ phụ thuộc cao và nguy

cơ lan truyền sai lớn; (ii) khai thác truy vấn cặp ràng buộc (ML/CL) như kênh phản hồi hiệu quả–thân thiện để hiệu chỉnh cấu trúc cụm theo từng bước, với tiêu chí lựa chọn cặp cân bằng bất định–đại diện–đa dạng–ảnh hưởng; và (iii) xử lý độ tin cậy dựa vào học chủ động.

Các nghiên cứu đề xuất về các phương pháp phân cụm bán giám sát mờ chủ động mới được trình bày trong các chương tiếp theo.

Chương 2

Đề xuất phương pháp phân cụm bán giám sát mờ chủ động dựa vào biên cụm

Chương này trình bày kết quả nghiên cứu về phương pháp phân cụm bán giám sát mờ chủ động dựa trên vùng biên cụm mới (ASSFBC). Nội dung chính của chương bao gồm: Phần mở đầu đưa ra định nghĩa về biên cụm, thách thức khó khăn và động lực của thuật toán, ý tưởng thuật toán đề xuất, sơ đồ mô tả cùng chi tiết từng bước trong thuật toán, phân tích độ phức tạp tính toán, và kết quả thực nghiệm trên tập dữ liệu UCI. Bên cạnh đó, chương cũng đề cập đến các thí nghiệm được thực hiện trên dữ liệu ảnh nhằm phục vụ cho bài toán phân đoạn ảnh để phân loại lớp phủ mặt đất. Kết quả thu được chứng minh hiệu quả của phương pháp đề xuất khi so sánh với các phương pháp khác trong cùng lĩnh vực.

2.1 Mở đầu

Một trong những thách thức then chốt của phân cụm bán giám sát mờ là việc thường gặp khó khăn khi xử lý ở vùng biên giữa các cụm. Biên cụm ám chỉ những vùng mà các điểm dữ liệu nằm gần các ngưỡng quyết định giữa các cụm, tạo ra sự mơ hồ và không chắc chắn trong việc gán cụm. Trong dữ liệu thực tế, các điểm dữ liệu thường biểu hiện các đặc trưng pha trộn, khiến việc phân loại chính xác trở nên khó khăn và làm suy giảm độ chính xác phân cụm. Ví

dụ, trong phân đoạn ảnh, các biên cụm không rõ ràng có thể dẫn đến biên đối tượng bị mờ hoặc sai lệch; tương tự, trong sinh tin học, các cụm gene có chức năng chồng lấn gây ra thách thức lớn trong phân loại. Tính mờ của SSFC, vốn cho phép mức độ phụ thuộc cụm khác nhau, càng làm gia tăng sự mơ hồ tại các vùng biên, nhấn mạnh nhu cầu về các cách tiếp cận tinh vi hơn để xử lý bất định liên quan đến biên một cách tường minh.

Học chủ động được xem là một giải pháp tiềm năng để nâng cao hiệu quả của phân cụm bán giám sát mờ bằng cách tập trung vào các điểm dữ liệu quan trọng nhất. Phương pháp này cho phép mô hình lựa chọn các điểm dữ liệu có giá trị thông tin cao để gán nhãn, thay vì gán nhãn một cách ngẫu nhiên, giúp giảm thiểu số lượng dữ liệu có nhãn cần thiết mà vẫn đảm bảo hiệu suất phân cụm. Một số nghiên cứu trước đây đã áp dụng học chủ động vào phân cụm, chẳng hạn như Seed-based K-Means hay Seed-based FCM [30], nhằm tối ưu hóa quá trình gán nhãn và cải thiện khả năng phát hiện cấu trúc cụm. Mặc dù có nhiều ưu điểm, học chủ động vẫn gặp phải những thách thức nhất định như chi phí tính toán cao, phụ thuộc vào cách chọn điểm dữ liệu ban đầu và khả năng xử lý các tập dữ liệu lớn chưa thực sự tối ưu.

Kết hợp học chủ động với phân cụm bán giám sát mờ mở ra hướng đi mới nhằm nâng cao độ chính xác và khả năng thích ứng của mô hình phân cụm. Thay vì áp dụng phương pháp học chủ động trên toàn bộ tập dữ liệu, việc chỉ tập trung vào các điểm nằm gần biên cụm có thể giúp cải thiện chất lượng phân cụm và tối ưu hóa việc sử dụng dữ liệu có nhãn. Một cách tiếp cận hiệu quả là sử dụng ràng buộc cặp để hướng dẫn quá trình học, đảm bảo rằng các điểm dữ liệu có tính chất tương đồng sẽ được nhóm vào cùng một cụm, trong khi các điểm khác biệt được tách biệt rõ ràng. Phương pháp này giúp cải thiện độ chính xác của mô hình mà không làm tăng đáng kể chi phí tính toán.

Mặc dù đã có nhiều nghiên cứu kết hợp học chủ động và phân cụm bán giám sát mờ, vẫn còn một số hạn chế cần được giải quyết. Một trong những thách thức lớn nhất là cách xác định vùng biên cụm một cách chính xác để lựa chọn điểm dữ liệu phù hợp cho quá trình học chủ động. Nếu vùng biên không được xác định đúng, việc chọn sai điểm gán nhãn có thể làm giảm độ chính xác của mô hình. Hơn nữa, tính toán hiệu quả trong môi trường dữ liệu lớn cũng là một vấn đề quan trọng, khi mà các phương pháp hiện tại vẫn đòi hỏi tài nguyên tính toán đáng kể.

Những thách thức và khoảng trống khoa học này chính là động lực để luận án phát triển các phương pháp phân cụm bán giám sát mờ hiệu quả hơn, đặc biệt là trong việc xử lý các điểm nằm gần biên cụm. Trong chương 2 này, luận án trình bày **phương pháp phân cụm bán giám sát mờ chủ động dựa vào biên cụm** tích hợp học chủ động một cách cải tiến để tối ưu hóa quá trình phân cụm giúp giải quyết các hạn chế của các thuật toán cũ với vùng biên, nhiễu cũng như cải thiện hiệu suất phân cụm.

2.2 Ý tưởng thuật toán

Phương pháp đề xuất bắt đầu bằng việc phân cụm dữ liệu sử dụng thuật toán Fuzzy C-Means (FCM). Bước này không chỉ nhóm các điểm dữ liệu vào các cụm mà còn xác định các vùng chồng lấn hay còn gọi là **biên** giữa các cụm. Những vùng này bao gồm các điểm dữ liệu không được gán một cách chắc chắn vào một cụm cụ thể, điều này cho thấy khả năng bị phân loại sai.

Bước tiếp theo là áp dụng **kỹ thuật học chủ động** vào các vùng biên này. Bằng cách chọn lọc và truy vấn chuyên gia hoặc một hệ thống hỗ trợ (oracle) để xác định nhãn thực sự của các điểm dữ liệu mơ hồ, thuật toán có thể điều chỉnh lại độ phụ thuộc của chúng vào cụm phù hợp, từ đó cải thiện chất lượng

phân cụm tổng thể.

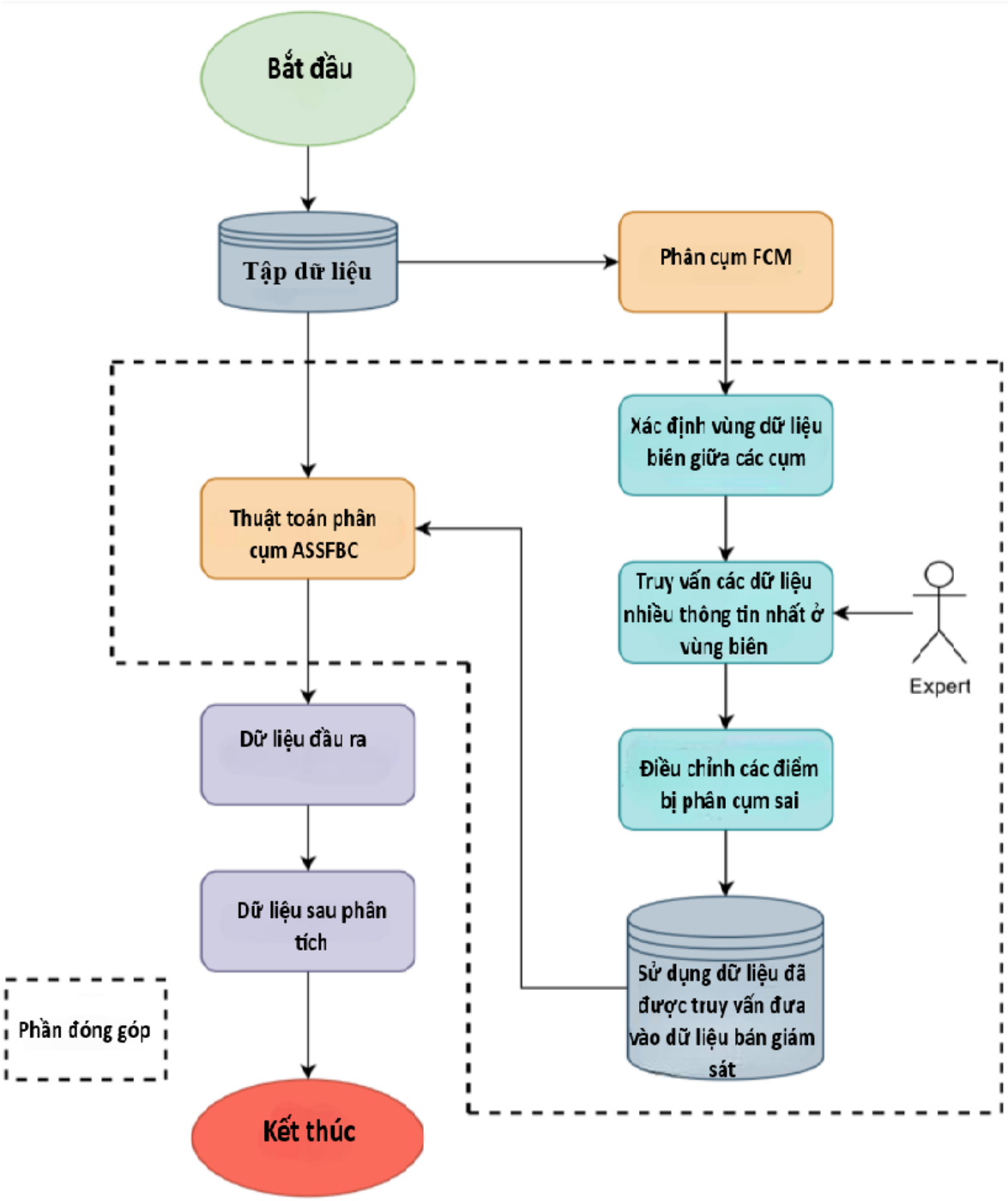
Trong giai đoạn tiếp theo, các tâm cụm của vùng biên và các điểm dữ liệu đã được điều chỉnh thông qua học chủ động sẽ được sử dụng để tinh chỉnh lại quá trình phân cụm. Một thuật toán phân cụm mới được phát triển với mục tiêu điều chỉnh độ phụ thuộc các điểm trong vùng chồng lấn sao cho vùng này nhỏ đi. Điều này giúp giảm số lượng phần tử vùng biên giữa các cụm, tăng độ rõ ràng của từng cụm và cải thiện khả năng phân tách giữa các cụm.

Dựa trên những ý tưởng chính này, chúng ta có thể xây dựng một sơ đồ minh họa phương pháp đề xuất, như thể hiện trong Hình 2.1.

2.3 Chi tiết thuật toán

Phương pháp ASSFBC được xây dựng nhằm giải quyết một trong những hạn chế quan trọng của các thuật toán phân cụm mờ truyền thống, đó là khả năng xử lý các phần tử dữ liệu nằm tại vùng biên giữa các cụm. Các điểm dữ liệu này thường có giá trị độ phụ thuộc gần như nhau đối với nhiều cụm, dẫn đến sự mơ hồ trong quá trình gán nhãn và làm giảm độ chính xác của kết quả phân cụm. Để khắc phục, ASSFBC đề xuất một hàm mục tiêu mới, tích hợp đồng thời ba thành phần: phân cụm mờ, phân cụm bán giám sát mờ, và hiệu chỉnh biên cụm bằng học chủ động. Toàn bộ quy trình vận hành của thuật toán được mô tả trong hình 2.1 thành một tiến trình liền mạch với các giai đoạn sau.

Đầu tiên là giai đoạn **phân cụm khởi tạo (Fuzzy Part)**. Tại bước này, thuật toán Fuzzy C-Means (FCM) được áp dụng trực tiếp lên tập dữ liệu để xác định các tâm cụm ban đầu cũng như ma trận độ phụ thuộc. Mục tiêu tối ưu là cực tiểu hóa tổng khoảng cách bình phương giữa các điểm dữ liệu và tâm cụm mà chúng thuộc về, qua đó hình thành nên cấu trúc phân cụm nền tảng. Đây là bước quan trọng, tạo tiền đề cho các hiệu chỉnh sau này, bởi nếu không có một



Hình 2.1: Mô hình phân cụm bán giám sát mờ chủ động dựa trên vùng biên cụm (ASSFBC)

khung phân cụm ban đầu đủ tốt thì việc xử lý biên cũng sẽ trở nên kém hiệu quả.

Tiếp theo, thuật toán chuyển sang giai đoạn **xác định biên cụm (Clusters Boundary Determination)**, được viết cụ thể ở thuật toán 7. Ở giai đoạn này, ASSFBC định nghĩa biên cụm là tập hợp các điểm dữ liệu mà sự chênh lệch giá trị hội viên đối với hai cụm gần nhất nhỏ hơn một ngưỡng θ định sẵn. Cụ thể, với một điểm dữ liệu x_k, nếu tồn tại hai cụm i và j sao cho $|u_{ki} - u_{kj}| < \theta$ trong đó u_{ki} là độ phụ thuộc lớn nhất của điểm, thì x_k được coi là một phần tử biên. Các phần tử này chính là những điểm dữ liệu gây khó khăn cho thuật toán phân cụm do mức độ mơ hồ cao. Như vậy, bước này cho phép hệ thống khoanh vùng và tập trung vào những điểm dễ gây nhầm lẫn, thay vì xử lý toàn bộ dữ liệu một cách đồng đều.

Sau khi đã xác định được các điểm biên, ASSFBC tiến hành bước **học chủ động và hiệu chỉnh độ biên cụm (Boundaries Adjustment)**, được viết cụ thể trong thuật toán 8. Đây là giai đoạn mà thông tin từ chuyên gia hoặc "oracle" được đưa vào quy trình. Các điểm biên được đưa ra để chuyên gia xác định nhãn chính xác. Trong khuôn khổ thực nghiệm của luận án thì việc truy vấn này sẽ được thực hiện bằng cách lấy nhãn trực tiếp từ bộ dữ liệu. Nếu kết quả chuyên gia cung cấp khác với phân cụm hiện tại của thuật toán, giá trị độ phụ thuộc của các điểm này sẽ được hiệu chỉnh lại theo hướng phù hợp với nhãn đúng. Việc điều chỉnh này đóng vai trò quan trọng, vì nó trực tiếp làm giảm sự mơ hồ tại vùng biên và khắc phục các sai sót mà bản thân FCM hay các biến thể bán giám sát khó xử lý. Qua đó, độ chính xác của kết quả phân cụm được nâng lên một cách đáng kể.

Tiếp nối giai đoạn hiệu chỉnh vùng biên các cụm là bước **tinh chỉnh biên cụm (Boundary Semi-Supervised Fuzzy Part)**. Ở bước này, các tâm cụm

và ma trận độ phụ thuộc được cập nhật lại bằng cách sử dụng thông tin hiệu chỉnh từ các phần tử biên. Về bản chất, quá trình này có mục tiêu kéo các tâm cụm dịch chuyển gần hơn một chút so với các điểm biên đã được xác định và hiệu chỉnh đúng. Nhờ đó, các vùng biên giữa các cụm trở nên rõ ràng hơn, độ chặt (compactness) được cải thiện, và sự tách biệt (separation) giữa các cụm được nâng cao. Đây chính là sự khác biệt then chốt của ASSFBC so với các thuật toán bán giám sát trước đây, bởi nó không chỉ sử dụng ràng buộc ở mức toàn cục mà còn tập trung trực tiếp vào vùng khó khăn nhất của quá trình phân cụm.

Cuối cùng, toàn bộ các bước trên được lặp đi lặp lại trong giai đoạn **tối ưu hoá (Optimization Process)**. Quá trình lặp này tiếp tục cho đến khi sự thay đổi giữa các tâm cụm ở hai vòng lặp liên tiếp nhỏ hơn một ngưỡng nhất định hoặc khi đạt đến số vòng lặp tối đa.

ASSFBC hình thành một hàm mục tiêu mới 2.1, trong đó sự kết hợp giữa phân cụm mờ, phân cụm bán giám sát mờ và hiệu chỉnh biên cụm bằng học chủ động được thể hiện một cách nhất quán. Hàm mục tiêu này không chỉ tối ưu phân cụm dựa trên khoảng cách dữ liệu và mức độ hội viên, mà còn chủ động xử lý các vùng biên, vốn là nơi dễ phát sinh sai sót :

$$\begin{aligned}\min_{u,v}\ J(u,v) = &\sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj}^2 \left\|x_k - v_j\right\|^2 + \alpha \sum_{k=1}^{L}\sum_{j=1}^{C} \left|u_{kj} - \bar{u}_{kj}\right|^2 \left\|x_k - v_j\right\|^2 \\ &+\beta \sum_{i=1}^{C-1}\sum_{d=i+1}^{C}\sum_{l\in N_{id}} (1-\mu_{id}^{l}) \left(\left\| v_i - \tfrac{1}{|N_i|}\sum_{h\in N_i} x_h \right\|^2 + \left\| v_d - \tfrac{1}{|N_d|}\sum_{h\in N_d} x_h \right\|^2 \right)\end{aligned} \tag{2.1}$$

Với các ràng buộc: $u_{kj}, \bar{u}_{kj} \in [0,1]$, $\forall k = 1,\ldots,N$ trong đó tập dữ liệu $X = \{x_1, x_2, \ldots, x_k, \ldots, x_N\}$ với số lượng điểm dữ liệu N, số lượng cụm C, và độ phụ thuộc của phần tử k trong cụm j: u_{kj}, bình phương khoảng cách

$\|x_k - v_j\|^2$ từ phần tử dữ liệu k đến tâm cụm v_j. L là lực lượng điểm dữ liệu ở tất cả các vùng biên bao gồm cả các điểm giám sát đã được truy vấn. N_{id} là số lượng điểm dữ liệu thuộc vùng biên giữa hai cụm i và d, với $N_{id} \ll N$. Tập các điểm này được phân hoạch thành hai nhóm có lực lượng N_i và N_d sao cho $N_{id} = N_i + N_d$, trong đó một điểm biên x_l thuộc nhóm i nếu $u_{li} > u_{ld}$, và thuộc nhóm d trong trường hợp ngược lại; chỉ số l ký hiệu một phần tử bất kỳ của tập biên N_{id}. μ_{id}^l là hiệu của độ phụ thuộc của phần tử l (được tính ở vòng lặp trước để sử dụng cho v_j), $\frac{1}{|N_i|}\sum_{h\in N_i} x_h$ được tính cho trọng tâm của các điểm vùng biên hướng về cụm i và $\frac{1}{|N_d|}\sum_{h\in N_d} x_h$ được tính cho trọng tâm của các điểm vùng biên hướng về cụm d.

Ta có thể xem xét ba phần của hàm mục tiêu:

$\sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj}\,\|x_k - v_j\|^2$	**là phần mờ (fuzzy part).**
$\alpha \sum_{k=1}^{L}\sum_{j=1}^{C} \|u_{kj} - \bar{u}_{kj}\|^2\,\|x_k - v_j\|^2$	**là một phần mờ bán giám sát chủ động, và trong phần này, chúng ta lấy các phần tử vùng biên được gắn nhãn dựa trên học chủ động cho hàm thành viên làm thành viên được giám sát.**
$\beta \sum_{i=1}^{C-1}\sum_{d=i+1}^{C}\sum_{l\in N_{id}} (1-\mu_{id}^l)\left\|v_i - \frac{1}{\|N_i\|}\sum_{h\in N_{id}} x_h\right\|^2$ $+\beta \sum_{i=1}^{C-1}\sum_{d=i+1}^{C}\sum_{l\in N_{id}} (1-\mu_{id}^l)\left\|v_d - \frac{1}{\|N_d\|}\sum_{h\in N_{id}} x_h\right\|^2$	**là thành phần tối ưu dựa vào vùng biên**.

Bảng 2.1: Các thành phần trong hàm mục tiêu

Ghi chú: Thành phần thứ ba trong hàm mục tiêu được thiết kế nhằm điều chỉnh vị trí các tâm cụm dựa trên thông tin từ các điểm biên đã được xác định và hiệu chỉnh nhãn. Cụ thể, với mỗi cặp cụm (i, d), trọng tâm T_{id} của tập điểm biên N_{id} phản ánh vị trí trung tâm của vùng chồng lấn giữa hai cụm này. Thành

phần thứ 3 của hàm mục tiêu không nhằm kéo các tâm cụm dịch chuyển mạnh vào vùng chồng lấn, mà đóng vai trò như một cơ chế điều chỉnh mềm. Cụ thể, trọng số $(1-\mu_{id}^{l})$ đảm bảo rằng chỉ những điểm biên có mức độ mơ hồ cao (tức μ_{id}^{l} nhỏ) mới tạo ra ảnh hưởng đáng kể lên vị trí tâm cụm.

Nhờ đó, các tâm cụm được điều chỉnh nhẹ nhàng hướng về vị trí trung tâm của vùng biên chưa ổn định, giúp giảm hiện tượng dao động hoặc lệch tâm bất thường trong quá trình lặp. Đồng thời, hệ số β cho phép kiểm soát mức độ ảnh hưởng của thành phần này, đảm bảo rằng cấu trúc toàn cục của các cụm không bị phá vỡ và mức độ chồng lấn không bị gia tăng quá mức.

Đạo hàm theo v_j ta có :

$$\begin{aligned}\frac{\partial J}{\partial v_j} &= 2\sum_{k=1}^{N} u_{kj}^{2}(v_j - x_k) + 2\alpha \sum_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 (v_j - x_k) \\ &\quad + 2\beta \sum_{l\in N_{id}} \big(1-\mu_{id}^{l}\big)(v_i - T_i) + 2\beta \sum_{l\in N_{id}} \big(1-\mu_{id}^{l}\big)(v_d - T_d) = 0\end{aligned} \tag{2.2}$$

Trong đó $T_i = \frac{1}{|N_i|}\sum_{h\in N_i} x_h$ và $T_d = \frac{1}{|N_d|}\sum_{h\in N_d} x_h$

Vì đạo hàm riêng theo v_j chỉ xét đến các thành phần mà v_j thực sự xuất hiện, nên ta có hai trường hợp:

$$\frac{\partial}{\partial v_j}\|v_i - T_i\|^2 = \begin{cases} 2(v_j - T_i), & \text{nếu } v_j = v_i, \\ 0, & \text{nếu } v_j \neq v_i. \end{cases} \tag{2.3}$$

$$\frac{\partial}{\partial v_j}\|v_d - T_d\|^2 = \begin{cases} 2(v_j - T_d), & \text{nếu } v_j = v_d, \\ 0, & \text{nếu } v_j \neq v_d. \end{cases} \tag{2.4}$$

Tại mỗi lần đạo hàm, v_j chỉ có thể trùng với một trong hai tâm v_i hoặc v_d. Hai trường hợp này cho kết quả cùng dạng $2(v_j - T_{id})$, trong đó $T_{id} = T_i$ nếu $v_j = v_i$ và $T_{id} = T_d$ nếu $v_j = v_d$.

Từ đó ta có:

$$2\sum_{k=1}^{N} u_{kj}^2(v_j - x_k) + 2\alpha \sum_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2(v_j - x_k) + 2\beta \sum_{l \in N_{id}} (1 - \mu_{id}^l)(v_j - T_{id}) = 0. \quad (2.5)$$

Nhóm các hạng chứa v:

$$\begin{aligned} &\left[2\sum_{k=1}^{N} u_{kj}^2 + 2\alpha \sum_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 + 2\beta \sum_{l \in N_{id}} (1 - \mu_{id}^l)\right] v_j \\ &= 2\sum_{k=1}^{N} u_{kj}^2 x_k + 2\alpha \sum_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 x_k + 2\beta \sum_{l \in N_{id}} (1 - \mu_{id}^l) T_{id}. \end{aligned} \quad (2.6)$$

Rút gọn và chia hai vế cho hệ số chung:

$$v_j = \frac{2\sum\limits_{k=1}^{N} u_{kj}^2 x_k + 2\alpha \sum\limits_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 x_k + 2\beta \sum\limits_{l \in N_{id}} (1 - \mu_{id}^l) T_{id}}{2\sum\limits_{k=1}^{N} u_{kj}^2 + 2\alpha \sum\limits_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 + 2\beta \sum\limits_{l \in N_{id}} (1 - \mu_{id}^l)}. \quad (2.7)$$

Chúng ta có điều kiện ràng buộc:

$$\sum_{j=1}^{C} u_{kj} = 1, \quad \forall k = 1, \ldots, N.$$

Do đó, áp dụng hệ số nhân Lagrange cho mỗi $k = 1, \ldots, N$:

$$J(u, v, \lambda) = J(u, v) - \sum_{i=1}^{N} \lambda_k \left(\sum_{j=1}^{C} u_{kj} - 1 \right). \quad (2.8)$$

Bằng cách đặt

$$\frac{\partial J(u, v, \lambda)}{\partial u_{kj}} = 0$$

cho phần tử k và cụm j đã cho (với khoảng cách $d_{kj}^2 = \|x_k - v_j\|^2$), ta có:

$$2u_{kj}d_{kj}^2 + 2\alpha(\bar{u}_{kj} - u_{kj})(-d_{kj}^2) - \lambda_k = 0. \quad (2.9)$$

Điều này dẫn đến:

$$u_{kj} = \frac{2\alpha\bar{u}_{kj}d_{kj}^2 + \lambda_k}{2(1+\alpha)d_{kj}^2} = \frac{\alpha\bar{u}_{kj}}{1+\alpha} + \frac{\lambda_k}{2(1+\alpha)d_{kj}^2}. \tag{2.10}$$

Bằng cách đặt

$$\frac{\partial J(u,v,\lambda)}{\partial\lambda} = 0,$$

ta có:

$$\sum_{j=1}^{C} u_{kj} = 1. \tag{2.11}$$

Dẫn đến:

$$\sum_{j=1}^{C}\left(\frac{\alpha\bar{u}_{kj}}{1+\alpha} + \frac{\lambda_k}{2(1+\alpha)d_{kj}^2}\right) = 1. \tag{2.12}$$

Do đó, ta tìm được:

$$\lambda_k = \frac{1 - \frac{\alpha}{1+\alpha}\sum_{j=1}^{C}\bar{u}_{kj}}{\sum_{j=1}^{C}\frac{1}{2(1+\alpha)d_{kj}^2}}. \tag{2.13}$$

Thay (2.11) vào (2.8) ta có:

$$u_{kj} = \frac{\alpha\bar{u}_{kj}}{1+\alpha} + \frac{1 - \frac{\alpha}{1+\alpha}\sum\limits_{i=1}^{C}\bar{u}_{ki}}{d_{kj}^2\sum\limits_{i=1}^{C}\frac{1}{d_{ki}^2}}. \tag{2.14}$$

Đặt : $s_k = \sum_{j=1}^{C}\bar{u}_{kj}$ và $w_{kj} = \frac{d_{kj}^{-2}}{\sum\limits_{h=1}^{C} d_{kh}^{-2}}$

(2.12) sẽ thành:

$$u_{kj} = \frac{\alpha}{1+\alpha}\,\bar{u}_{kj} + \left(1 - \frac{\alpha}{1+\alpha}s_k\right)w_{kj}$$

Bên cạnh đó, với mỗi điểm dữ liệu x_j, ký hiệu

$$Z_k = \{\, k \mid d_{kj}^2 = 0 \,\}$$

là tập các cụm mà điểm x_j có khoảng cách bằng 0 đến tâm cụm và $1_{\{k \in Z_j\}}$ là hàm chỉ mục.

Ta xây dựng công thức cập nhật:

$$u_{kj} = \begin{cases} \frac{1}{|Z_j|} 1_{\{k \in Z_j\}}, & \text{với } |Z_j| \geq 1, \\ \frac{\alpha}{1+\alpha} \bar{u}_{kj} + \left(1 - \frac{\alpha}{1+\alpha} s_k\right) w_{kj}, \quad , & \text{với } |Z_j| = 0. \end{cases} \tag{2.15}$$

Chứng minh rằng $0 < u_{kj} < 1$**:**

Trường hợp tồn tại cụm có $d_{kj}^2 = 0$

Theo công thức (2.13) hiển nhiên ta có $u_{kj} \geq 0$ và $\sum_{j=1}^{C} u_{kj} = 1$

Trường hợp không tồn tại cụm có $d_{kj}^2 = 0$, khi đó:

(1) Trường hợp chuẩn hóa:

$$\sum_{j=1}^{C} \bar{u}_{kj} = 1, \quad 0 \leq \bar{u}_{kj} \leq 1.$$

Khi đó:

$$u_{kj} = \frac{\alpha}{1+\alpha} \bar{u}_{kj} + \frac{1}{1+\alpha} w_{kj}.$$

Vì $\bar{u}_{kj} \in [0, 1]$ và $w_{kj} \in (0, 1)$, ta có:

$$u_{kj} = \frac{\alpha}{1+\alpha} \bar{u}_{kj} + \frac{1}{1+\alpha} w_{kj}$$

là tổ hợp lồi của 2 số $\in [0, 1]$ nên $u_{kj} \in [0, 1]$

hơn nữa do $\sum_{j=1}^{C} \bar{u}_{kj} = 1$ và $\sum_{j=1}^{C} w_{kj} = 1$ nên dễ thấy :

$$\sum_{j=1}^{C} u_{kj} = \frac{\alpha}{1+\alpha} \cdot 1 + \frac{1}{1+\alpha} \cdot 1 = 1.$$

Vậy:

$$0 < u_{kj} < 1.$$

(2) Trường hợp chuẩn hóa dưới mức:

$$0 \leq \bar{u}_{kj} \leq 1, \qquad s_k = \sum_{j=1}^{C} \bar{u}_{kj} \leq 1.$$

Công thức cập nhật:

$$u_{kj} = \frac{\alpha}{1+\alpha}\,\bar{u}_{kj} + \left(1 - \frac{\alpha}{1+\alpha} s_k\right) w_{kj}.$$

Dễ thấy

$$\sum_{j=1}^{C} u_{kj} = \frac{\alpha}{1+\alpha} \cdot s_k + 1 - \frac{\alpha}{1+\alpha} s_k = 1.$$

Bên cạnh đó do $s_k \leq 1$ nên:

$$1 - \frac{\alpha}{1+\alpha} s_k \geq \frac{1}{1+\alpha} > 0$$

cùng với $\frac{\alpha}{1+\alpha}\,\bar{u}_{kj} \geq 0$ và $w_{kj} > 0$, ta có:

$$u_{kj} > 0.$$

Suy ra:

$$0 < u_{kj} < 1.$$

Kết luận

Trong mọi trường hợp:

$$0 < u_{kj} < 1$$

Chứng minh hội tụ: Hàm mục tiêu $J(u, v)$ bao gồm ba thành phần: (i) thành phần chuẩn FCM phản ánh độ chặt cụm, (ii) thành phần bán giám sát ép u tiệm cận $\bar{u}$ tại vùng biên được gán nhãn, và (iii) hạng tử điều chỉnh biên cụm làm co nhỏ vùng chồng lấn.

Khi cố định v, $J(u, v)$ là hàm bậc hai theo u, khả vi và lồi; do đó nghiệm cập nhật $u^{(t+1)}$ thu được từ công thức đóng là cực tiểu cục bộ:

$$u^{(t+1)} = \arg\min_{u} J(u, v^{(t)}).$$

Tương tự, khi cố định u, $J(u, v)$ là hàm bình phương khoảng cách có trọng số theo v, vì vậy việc cập nhật $v^{(t+1)}$ cũng làm giảm hoặc giữ nguyên giá trị của J:

$$v^{(t+1)} = \arg\min_{v} J(u^{(t+1)}, v).$$

Do mỗi bước lặp đảm bảo

$$J(u^{(t+1)}, v^{(t+1)}) \leq J(u^{(t)}, v^{(t)}),$$

và $J(u, v) \geq 0$, nên dãy $\{J^{(t)}\}$ giảm đơn điệu và bị chặn dưới, suy ra tồn tại giới hạn hữu hạn J^*:

$$\lim_{t \to \infty} J^{(t)} = J^*.$$

Như vậy, thuật toán ASSFBC hội tụ đơn điệu về một điểm dừng (u^*, v^*) thỏa các điều kiện cực trị $\frac{\partial J}{\partial u} = 0$ và $\frac{\partial J}{\partial v} = 0$, tương ứng với nghiệm cực tiểu cục bộ của hàm mục tiêu.

Thuật toán "Xác định biên cụm" bắt đầu bằng việc tính toán số lượng mẫu có thể truy vấn chuyên gia, xác định bởi tỷ lệ hạt giống (seed rate), thường là

một giá trị cố định cho một tập dữ liệu nhất định. Tỷ lệ seed xác định tỷ lệ của tập dữ liệu được chọn bởi chuyên gia, thay vì được điều chỉnh động. Sau đó, thuật toán đánh giá sự khác biệt giữa giá trị độ phụ thuộc lớn nhất và các cụm khác của mỗi điểm dữ liệu. Bằng cách xác định và sắp xếp các điểm có độ khác biệt nhỏ nhất, thuật toán chọn ra các phần tử vùng biên để truy vấn chuyên gia, đóng vai trò là các điểm chính để tinh chỉnh quá trình phân cụm.

Thuật toán 2.1 Xác định biên cụm

Đầu vào: $X = \{x_k\}_{k=1}^{N}$, số cụm C, ma trận độ phụ thuộc $U \in [0,1]^{N \times C}$, ngưỡng biên θ, tỉ lệ truy vấn $seed\_rate$. **Đầu ra**:Tập điểm biên L và tập truy vấn Q.

1 $N_q \leftarrow \lfloor N \cdot seed\_rate \rfloor$
2 **Lặp với** $k = 1$ đến L:
3 $\quad i^* \leftarrow \arg\max_j \, u_{kj}$
4 $\quad u_{k\_dif} \leftarrow \min_{j \neq i^*} |u_{k,i^*} - u_{kj}|$
5 **Kết thúc lặp**
6 $L \leftarrow \{x_k \mid u_{k\_dif} < \theta\}$
7 Sắp xếp các điểm dữ liệu x_k theo giá trị u_{k_dif} tăng dần
8 Chọn N_q điểm có u_{k_dif} nhỏ nhất làm các phần tử vùng biên được truy vấn.

Thuật toán "Điều chỉnh Biên Cụm" hoạt động bằng cách tinh chỉnh ma trận độ phụ thuộc dựa trên phản hồi của chuyên gia. Trong luận án, tôi thực hiện mô phỏng phản hồi chuyên gia bằng cách lấy nhãn cụm trực tiếp từ tập nhãn của tập dữ liệu. Trong mỗi lần lặp, thuật toán so sánh kết quả phân cụm với kết quả sau khi truy vấn cho các phần tử biên. Khi truy vấn chuyên gia, luận án sử dụng tiêu chí bất định trong học chủ động vì lựa chọn các điểm có hiệu độ phụ thuộc giữa 2 cụm là nhỏ nhất, điều này thể hiện điểm có độ "mập mờ" cao nhất nằm ở biên giữa 2 cụm. Các điều chỉnh được thực hiện bằng cách sửa đổi các giá trị độ phụ thuộc theo một ngưỡng δ được xác định trước, đảm bảo rằng việc phân cụm thích ứng với các chỉnh sửa của chuyên gia. Tại bước này $\bar{u}$

cũng được xác định chính bằng độ phụ thuộc của các phần tử được truy vấn.

Thuật toán 2.2 Điều chỉnh biên cụm

Đầu vào:

- U: Ma trận độ phụ thuộc của tập dữ liệu Y gồm N phần tử.
- R: Tập kết quả phân cụm từ FCM cho các điểm Y.
- δ: Một giá trị dương nhỏ để điều chỉnh biên.

Đầu ra: Ma trận độ phụ thuộc đã điều chỉnh $\bar{U}$.

1 $i \leftarrow 0$
2 **Thực hiện vòng lặp**:
3 $\quad i \leftarrow i + 1$
4 $\quad$ **Lặp qua từng** phần tử biên x_i được xác định:
5 $\quad\quad$ Truy vấn chuyên gia để lấy nhãn cụm thực tế ru_i
6 $\quad\quad$ **Nếu** $r_i \neq ru_i$ **thì**:
7 $\quad\quad\quad$ $(k, j) \leftarrow \arg\max_{k \neq j}\{u_{ik}, u_{ij}\}$
8 $\quad\quad\quad$ **Nếu** $|u_{ik} - u_{ij}| > \delta$ **thì**:
9 $\quad\quad\quad\quad$ $m \leftarrow (u_{ik} + u_{ij})/2$
10 $\quad\quad\quad\quad$ $u_{ik} \leftarrow m - \delta/2$
11 $\quad\quad\quad\quad$ $u_{ij} \leftarrow m + \delta/2$
12 $\quad\quad\quad$ **Ngược lại**:
13 $\quad\quad\quad\quad$ $(u_{ik}, u_{ij}) \leftarrow (u_{ij}, u_{ik})$
14 **Cho đến khi**: $i > N$
15 **Cập nhật các điểm đã truy vấn vào** $\bar{U}$

Thuật toán ASSFBC là sự kết hợp phân cụm FCM với điều chỉnh bán giám sát dựa trên các phần tử vùng biên.

Sau quá trình phân cụm ban đầu, thuật toán xác định các phần tử vùng biên và tinh chỉnh chúng thông qua phản hồi từ chuyên gia. Các bước cuối cùng bao gồm lặp lại quá trình cập nhật độ phụ thuộc và tâm cụm cho đến khi thuật toán hội tụ, đảm bảo kết quả phân cụm chính xác và phù hợp với dữ liệu đầu vào. Thuật toán được mô tả như sau:

Thuật toán 2.3 Phân cụm bán giám sát mờ chủ động dựa vào vùng biên cụm (ASSFBC)

Đầu vào: Tập dữ liệu $X = \{x_k\}_{k=1}^{N}$, ma trận độ phụ thuộc U, ngưỡng hội tụ ε, số vòng lặp tối đa $T_{\max}$.
Đầu ra: Ma trận độ phụ thuộc U, tâm cụm V.

1 $U \leftarrow \text{FCM}(X)$.
2 Xác định tập biên và tập truy vấn $L, N_q \leftarrow$ Xác định biên cụm(U).
3 Điều chỉnh: $U, \bar{U} \leftarrow$ Điều chỉnh vùng biên cụm(U, N_q)
4 $t \leftarrow 0$
5 Khởi tạo tâm cụm $v_j^{(t)}$ từ kết quả trước đó, với $j = 1, \ldots, C$
6 **Thực hiện vòng lặp**:
7 $\quad t \leftarrow t + 1$
8 $\quad$ Cập nhật $u^{(t)}$ theo (2.15)
9 $\quad$ Cập nhật $v^{(t)}$ theo (2.7)
10 **Cho đến khi**: $\left\|v^{(t)} - v^{(t-1)}\right\| \leq \varepsilon$ hoặc $t > T_{\max}$

Độ phức tạp thuật toán

Theo đúng quy trình của ASSFBC gồm bốn giai đoạn: (i) FCM trên toàn bộ dữ liệu: mỗi vòng lặp cập nhật ma trận độ phụ thuộc theo công thức chuẩn dạng tỉ lệ, trong đó việc tính toán mỗi giá trị u_{kj} yêu cầu tổng hợp trên toàn bộ C cụm còn lại, dẫn đến chi phí $O(NC^2)$ cho mỗi vòng lặp, do đó với I_0 vòng lặp chi phí là $O(I_0 NC^2)$; (ii) Xác định biên cụm: tính độ chênh lệch (hoặc chỉ số "gần biên") từ ma trận độ phụ thuộc và sắp xếp để ưu tiên các điểm mơ hồ, chi phí $O(NC + N \log N)$ (thường gộp viết $\approx O(N \log N)$); (iii) Tinh chỉnh biên cụm: Thuật toán chỉ truy vấn và hiệu chỉnh trên một tập con rất nhỏ các điểm biên $\mathcal{Q}$ với $|\mathcal{Q}| = N_q \ll N$ (các điểm có mức bất định cao nhất). Với mỗi điểm $x_i \in \mathcal{Q}$, thao tác cần thiết là (i) quét vector độ phụ thuộc $\{u_{i1}, \ldots, u_{iC}\}$ để xác định hai giá trị lớn nhất (hoặc các chỉ số liên quan) và (ii) cập nhật lại một số phần tử của hàng i trong U theo phản hồi chuyên gia/qui tắc ràng buộc; các bước này có chi phí $O(C)$ cho mỗi điểm. Do đó, tổng chi phí tính toán cho pha tinh chỉnh

biên là $O(N_q C)$ (không tính thời gian phản hồi của chuyên gia); (iv) Phân cụm theo hàm mục tiêu mới ASSFBC: lặp phép cập nhật độ phụ thuộc/tâm cụm dưới hàm mục tiêu đã hiệu chỉnh trong T vòng, mỗi vòng vẫn có chi phí trội $O(NC^2)$ (tương tự FCM chuẩn về bậc tính toán), tổng $O(TNC^2)$.

Kết hợp các thành phần, độ phức tạp thời gian tổng thể của ASSFBC là

$$\boxed{O\Big(NC^2(I_0+T) \;+\; NC \;+\; N\log N \;+\; N_q C\Big)}$$

với N là số điểm dữ liệu, C là số cụm, I_0 là số vòng FCM khởi tạo và T là số vòng lặp tối ưu hoá dưới hàm mục tiêu ASSFBC. Do $N_q \ll N$ và trong hầu hết các kịch bản thực tế NC bị át bởi NC^2 hoặc $N\log N$, chi phí chi phối của thuật toán đến từ các bước cập nhật độ phụ thuộc trên toàn bộ dữ liệu ở hai giai đoạn phân cụm chính, tức là $O\Big(NC^2(I_0+T)\Big)$, trong khi bước xác định biên có chi phí $O(N\log N)$ đóng vai trò là hạng phụ nhưng vẫn đáng kể khi C nhỏ.

2.4 Kết quả thực nghiệm

Phần này trình bày các kết quả thực nghiệm nhằm chứng minh hiệu quả của phương pháp đề xuất so với các phương pháp phân cụm bán giám sát mờ khác trong bài toán đánh giá chất lượng cụm.

2.4.1 Dữ liệu, độ đo và môi trường thực nghiệm

Với mong muốn chứng minh hiệu quả của phương pháp đề xuất trong trường hợp có nhiều điểm nhiễu tại vùng biên các cụm dữ liệu, thí nghiệm được thực hiện nhằm so sánh và đề xuất các kịch bản thử nghiệm trên ba loại dữ liệu sau: dữ liệu tiêu chuẩn UCI, tập dữ liệu được tạo thủ công, và dữ liệu ảnh. Những tập dữ liệu này được lựa chọn để đánh giá toàn diện hiệu suất của thuật toán trên nhiều loại dữ liệu và thách thức phân cụm khác nhau, khẳng định tính hiệu

quả của mô hình phân cụm trong nhiều tình huống thực tế khác nhau.

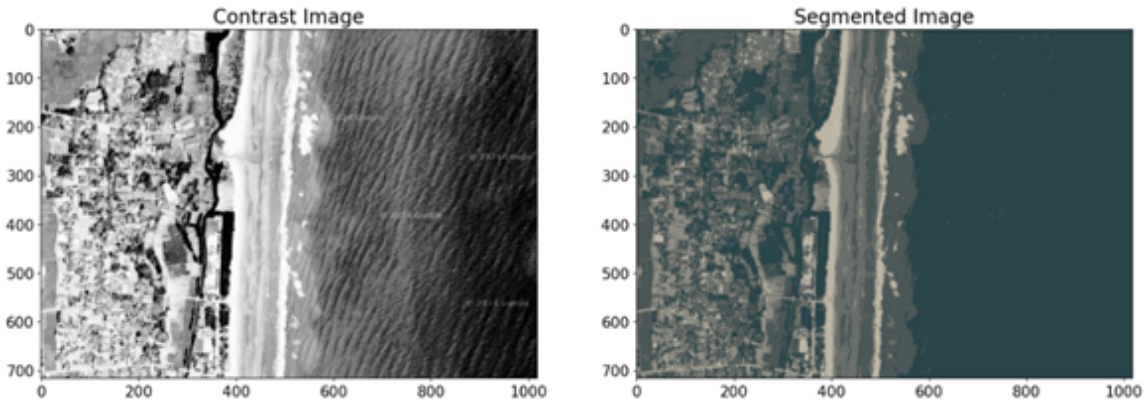
Đầu tiên, luận án sử dụng sáu bộ dữ liệu kinh điển của **UCI** [106], bao quát đa dạng về số chiều và số lớp: *Iris*, *Wine*, *Breast Cancer Wisconsin*, *Glass Identification*, *New Thyroid*, *Soybean (small)*. Thống kê chi tiết được tóm tắt ở Bảng 2.2.

STT	**Tập dữ liệu**	**Số mẫu**	**Số thuộc tính**	**Số nhãn**
1	Iris	150	4	3
2	Wine	178	13	3
3	Breast	569	30	2
4	Glass	214	9	6
5	Thyroid	215	5	3
6	Soybean	47	35	4

Bảng 2.2: Các tập dữ liệu UCI dùng trong thí nghiệm chương 2

Dữ liệu tự sinh 2D với vùng chồng lấn ngẫu nhiên: Để mô phỏng các tình huống cụm có ranh giới mờ và chồng lấn trong không gian 2D, luận án xây dựng thủ tục GEN2D tạo bộ dữ liệu tổng hợp gồm C cụm. Trước hết, các tâm cụm $V_1, \ldots, V_C$ được sinh ngẫu nhiên và các điểm dữ liệu được lấy mẫu quanh từng tâm (ví dụ theo phân phối chuẩn 2D), tạo nên các vùng "lõi" tương đối tách biệt. Tiếp theo, để hình thành vùng chồng lấn, một tỉ lệ điểm được chọn ngẫu nhiên từ các cụm và được tái bố trí về khu vực giao thoa giữa hai cụm: với mỗi điểm được chọn, thuật toán ngẫu nhiên chọn một cụm khác làm cụm đích và sinh điểm mới quanh vùng giữa hai tâm (chẳng hạn quanh trung điểm hoặc một vị trí ngẫu nhiên trên đoạn nối hai tâm). Nhờ việc lựa chọn ngẫu nhiên các cặp cụm và vị trí giao thoa, vùng chồng lấn xuất hiện đa dạng giữa nhiều cụm khác nhau thay vì cố định theo một cấu hình duy nhất. Bộ dữ liệu thu được cho phép kiểm tra tính ổn định của các phương pháp phân cụm trong điều kiện có nhiều điểm bất định tại rìa cụm, đồng thời phục vụ minh họa các tình huống

mà chiến lược truy vấn chủ động (AL) thường ưu tiên khai thác.



Hình 2.2: Kết quả phân đoạn thông qua phân cụm trên ảnh Landsat-8

Dữ liệu ảnh: Để đánh giá trong bối cảnh dữ liệu thực tế có nhiều nhiễu, luận án sử dụng ảnh Landsat-8 khu vực ven biển Thanh Hóa (Việt Nam) từ google map chế độ vệ tinh(toạ độ tâm: 19.518729°N, 105.807984°E) với kích thước (1358 pixels x 869 pixels) để phân đoạn (thông qua phân cụm). Sau khi thực hiện phân cụm, ảnh được tách thành các vùng khác nhau theo màu sắc tương ứng với các cụm. Dựa trên các vùng này, người dùng có thể tiến hành phân thành các lớp theo nhãn cho trước (nhãn của ảnh được đánh thủ công trên từng pixel theo các vùng tương ứng với các lớp được xác định trước của ảnh). Trong nội dung của luận án này, tôi chỉ tập trung vào việc phân cụm nên việc sử dụng ảnh phân đoạn để phân lớp sẽ không đề cập. Bộ dữ liệu này nhấn mạnh biên phức tạp giữa các cụm khác nhau. Hình 2.2 là kết quả việc phân đoạn ảnh sử dụng phân cụm.

Trong thực nghiệm, sau khi chuẩn hóa dữ liệu, α được chọn trong khoảng$\alpha \in [0.1, 1]$, và giá trị $\alpha = 0.5$ cho thấy sự cân bằng tốt giữa tính mềm của phân cụm và độ tin cậy của nhãn vùng biên. Ngược lại, hệ số β chỉ xuất hiện trong thành phần tối ưu dựa trên vùng biên, tác động trực tiếp lên công thức cập nhật tâm cụm v_j (xem (2.7)), và không tham gia vào cập nhật độ phụ thuộc u_{kj}. Do đó,

β điều khiển mức độ kéo các tâm cụm về trọng tâm của các vùng biên giữa các cụm lân cận, nhằm làm sắc nét ranh giới phân cụm.Trong thực nghiệm, β được chọn sao cho thành phần vùng biên đóng vai trò điều chỉnh mềm, tránh làm méo cấu trúc cụm tổng thể. Các thí nghiệm cho thấy khoảng giá trị $\beta \in [0.05, 0.2]$ là phù hợp, trong đó $\beta = 0.1$ mang lại kết quả ổn định nhất, đặc biệt trên các tập dữ liệu có vùng chồng lấn và nhiễu cao. Các chỉ số được sử dụng để đánh giá chất lượng phân cụm là: RI, F1-score, NMI , DB, PC, PE, DB_fuzzy. Bảng 2.3 là bảng các tham số được dùng trong thực nghiệm trong các thuật toán.

Thuật toán	m	ε	$T_{\max}$	Tham số riêng
FCM	2.0	10^{-6}	1000	–
SSFCM	2.0	10^{-6}	1000	Trọng số giám sát $\alpha = 0.5$
eFCM	2.0	10^{-6}	1000	Hệ số tin cậy $\gamma = 0.8$
AFFC (2008)	2.0	10^{-6}	1000	–
AFFC (2017)	2.0	10^{-6}	1000	–
ASSFBC	2.0	10^{-6}	1000	Ngưỡng biên $\theta = 0.05$, $\alpha = 0.5$, $\beta = 0.1$

Bảng 2.3: Bảng tham số thực nghiệm cho các thuật toán

Các thí nghiệm được thực hiện bằng MATLAB trên máy tính xách tay LG Gram, được trang bị bộ vi xử lý Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz đến 2.40GHz và 8GB RAM. Điều này đảm bảo rằng các điều kiện thí nghiệm được xác định rõ ràng để có thể tái lập. Các thí nghiệm được tiến hành để so sánh phương pháp đề xuất ASSFBC với các phương pháp liên quan như: FCM [39], SSFCM [27], eFCM [107], AFFC(2008) [31], và AFFC(2017) [32]. Bảng 2.4 trình bày so sánh các thuật toán sử dụng trong chương 2, gồm FCM, SSFCM, eFCM, AFFC (2008), AFFC (2017) và phương pháp đề xuất **ASSFBC**. Bảng này làm rõ kiểu học, thông tin bổ trợ, cơ chế học chủ động và ưu/nhược điểm của từng phương pháp, qua đó nhấn mạnh động cơ và lợi thế của **ASSFBC**

trong việc xử lý vùng biên và cải thiện độ chính xác phân cụm.

Thuật toán	**Kiểu học**	**Thông tin bổ trợ**	**Cơ chế áp dụng học chủ động cho thông tin bổ trợ**	**Đặc điểm nổi bật / Hạn chế**
FCM [39]	Không giám sát	Không có	Không có	Cơ bản; nhanh; xử lý nhiễu, biên kém.
SSFCM [27]	Bán giám sát	Có (ML/CL tĩnh)	Không có	Dùng ràng buộc cặp nhưng nhạy với nhiễu; hiệu quả phụ thuộc số lượng ràng buộc ban đầu.
eFCM [107]	Bán giám sát	Có (ML/CL tĩnh)	Không có	Ổn định hơn nhờ entropy; chưa tự giảm số cụm, chưa chọn cặp hỏi tối ưu.
AFFC(2008) [31]	Bán giám sát chủ động	Có (ML/CL động)	MVP (Most Valuable Pairs)	Sinh cặp hỏi hiệu quả, giảm số cụm linh hoạt; vẫn thiếu cơ chế đánh giá và xử lý biên cụm.
AFFC(2017) [32]	Bán giám sát chủ động	Có (ML/CL động)	MVP cải tiến	Ổn định hơn AFFC gốc; vẫn thiếu cơ chế đánh giá và xử lý biên cụm.
ASSFBC	Bán giám sát chủ động nâng cao	độ phụ thuộc đã điều chỉnh bằng học chủ động	truy vấn và tinh chỉnh tại vùng biên	Có cơ chế xác định biên; Cơ chế truy vấn cải tiến và tinh chỉnh biên ; tối ưu thuật toán dựa vào thông tin ở biên cụm.

Bảng 2.4: So sánh các thuật toán thực nghiệm chương 2

2.4.2 Đánh giá kết quả thực nghiệm

Đánh giá theo độ chính xác và chất lượng phân cụm

Sau khi tiến hành các thí nghiệm trên tập dữ liệu, kết quả được thể hiện chi tiết trong **các bảng 2.5, 2.7, 2.9, 2.11** tương ứng với **tập dữ liệu UCI, dữ liệu tổng hợp, và dữ liệu ảnh** .

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
IRIS						
RI	0.8741 ± 0.0735	0.9018 ± 0.0612	0.9094 ± 0.0841	0.9253 ± 0.0714	0.9492 ± 0.0589	**0.9939** ± 0.0321
F1	0.8364 ± 0.0962	0.8491 ± 0.0794	0.8614 ± 0.0869	0.9086 ± 0.0662	0.9304 ± 0.0548	**0.9898** ± 0.0347
NMI	0.7381 ± 0.1125	0.7704 ± 0.0981	0.7825 ± 0.0913	0.8241 ± 0.0814	0.8463 ± 0.0679	**0.9744** ± 0.0298
DB	**0.9414** ± 0.1598	1.3615 ± 0.1325	1.4173 ± 0.1421	1.5471 ± 0.1014	1.3932 ± 0.0893	1.2294 ± 0.0751
Breast						
RI	0.7594 ± 0.0841	0.7594 ± 0.0841	0.7641 ± 0.0793	0.7689 ± 0.0758	0.7731 ± 0.0702	**0.7848** ± 0.0651
F1	0.7894 ± 0.0791	0.7894 ± 0.0791	0.7961 ± 0.0724	0.8044 ± 0.0689	0.7971 ± 0.0738	**0.8667** ± 0.0619
NMI	0.4751 ± 0.1195	0.4751 ± 0.1195	0.4928 ± 0.1112	0.4974 ± 0.1074	0.5002 ± 0.1029	**0.5123** ± 0.0984
DB	**0.7485** ± 0.1184	0.8156 ± 0.1253	1.3481 ± 0.1344	2.3923 ± 0.1042	0.9658 ± 0.0964	0.9144 ± 0.1633
Glass						
RI	0.8214 ± 0.1014	0.8192 ± 0.1063	0.8258 ± 0.0982	0.8310 ± 0.0934	0.8322 ± 0.0909	**0.8494** ± 0.0851
F1	0.5983 ± 0.1271	0.5942 ± 0.1318	0.6029 ± 0.1214	0.6058 ± 0.1158	0.6091 ± 0.1125	**0.6154** ± 0.1084
NMI	0.6891 ± 0.1379	0.6804 ± 0.1441	0.6875 ± 0.1338	0.6971 ± 0.1273	0.6582 ± 0.1524	**0.7584** ± 0.1102
DB	**0.6383** ± 0.1161	2.3294 ± 0.1742	1.9192 ± 0.1324	1.4598 ± 0.1483	2.3304 ± 0.1641	1.1964 ± 0.1835

Bảng 2.5: So sánh các hiệu năng các phương pháp trên các tập dữ liệu chuẩn IRIS, Breast và Glass.

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
IRIS						
PC	0.7803 ± 0.0507	0.8112 ± 0.0413	0.8324 ± 0.0421	0.8517 ± 0.0314	0.8831 ± 0.0298	**0.9418** ± 0.0212
PE	0.4216 ± 0.0623	0.3897 ± 0.0514	0.3712 ± 0.0489	0.3431 ± 0.0417	0.3124 ± 0.0388	**0.2135** ± 0.0294
DB_{fuzzy}	**0.9127** ± 0.0714	1.4721 ± 0.0597	1.5834 ± 0.0812	1.5219 ± 0.0615	1.3028 ± 0.0498	1.2617 ± 0.0412
Breast						
PC	0.7412 ± 0.0598	0.7427 ± 0.0576	0.7529 ± 0.0497	0.7638 ± 0.0479	0.7714 ± 0.0412	**0.7925** ± 0.0386
PE	0.4829 ± 0.0691	0.4817 ± 0.0715	0.4698 ± 0.0604	0.4523 ± 0.0586	0.4427 ± 0.0513	**0.4235** ± 0.0489
DB_{fuzzy}	0.7124 ± 0.0921	1.0942 ± 0.1017	**0.6028** ± 0.1129	0.7729 ± 0.0826	0.9657 ± 0.0738	0.9526 ± 0.1143
Glass						
PC	0.6921 ± 0.0694	0.6834 ± 0.0723	0.7019 ± 0.0598	0.7128 ± 0.0610	0.7256 ± 0.0521	**0.7517** ± 0.0492
PE	0.5318 ± 0.0796	0.5524 ± 0.0884	0.5247 ± 0.0792	0.4928 ± 0.0714	0.4719 ± 0.0698	**0.4125** ± 0.0609
DB_{fuzzy}	**0.6231** ± 0.0798	2.3827 ± 0.1229	0.7216 ± 0.1012	1.0029 ± 0.1128	1.4324 ± 0.1317	2.2612 ± 0.1421

Bảng 2.6: So sánh các độ đo PC, PE và DB_fuzzy trên các tập dữ liệu IRIS, Breast và Glass.

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
Wine						
RI	0.7199 ± 0.0975	0.7418 ± 0.0851	0.7367 ± 0.0889	0.7556 ± 0.0812	0.7583 ± 0.0780	**0.7689** ± 0.0724
F1	0.5868 ± 0.1219	0.6234 ± 0.1094	0.6166 ± 0.1128	0.6417 ± 0.1041	0.6459 ± 0.1005	**0.6578** ± 0.0944
NMI	0.4295 ± 0.1408	0.4537 ± 0.1262	0.4469 ± 0.1321	0.4604 ± 0.1230	0.4671 ± 0.1194	**0.4955** ± 0.1115
DB	**0.8365** ± 0.1193	1.6125 ± 0.1582	0.8490 ± 0.1165	0.8367 ± 0.1104	0.9608 ± 0.1035	2.2867 ± 0.1712
Soybean						
RI	0.8234 ± 0.0910	0.8408 ± 0.0844	0.7865 ± 0.1071	0.8441 ± 0.0816	0.8464 ± 0.0791	**0.8677** ± 0.0728
F1	0.6819 ± 0.1268	0.7124 ± 0.1142	0.6081 ± 0.1498	0.7168 ± 0.1104	0.7194 ± 0.1077	**0.7571** ± 0.0985
NMI	0.7002 ± 0.1224	0.7544 ± 0.1038	0.5454 ± 0.1705	0.7568 ± 0.1019	0.7671 ± 0.0978	**0.7784** ± 0.0936
DB	3.7219 ± 0.1908	3.7587 ± 0.1824	4.7462 ± 0.2098	**2.4579** ± 0.1655	2.7805 ± 0.1542	2.7581 ± 0.1520
Thyroid						
RI	0.6394 ± 0.1248	0.6513 ± 0.1182	0.6655 ± 0.1097	0.7175 ± 0.0964	0.7258 ± 0.0931	**0.7649** ± 0.0804
F1	0.6314 ± 0.1309	0.6338 ± 0.1292	0.6485 ± 0.1208	0.6872 ± 0.1048	0.6988 ± 0.1011	**0.7367** ± 0.0892
NMI	0.3135 ± 0.1749	0.2968 ± 0.1842	0.3301 ± 0.1656	0.3635 ± 0.1509	0.3671 ± 0.1482	**0.5284** ± 0.1205
DB	**2.0871** ± 0.1982	2.4405 ± 0.1867	4.8259 ± 0.2142	2.6492 ± 0.1734	2.5697 ± 0.1706	2.1271 ± 0.1589

Bảng 2.7: So sánh các hiệu năng các phương pháp trên các tập dữ liệu chuẩn Wine, Soybean và Thyroid.

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
Wine						
PC	0.7213 ± 0.0745	0.7328 ± 0.0691	0.7394 ± 0.0712	0.7517 ± 0.0634	0.7596 ± 0.0592	**0.7729** ± 0.0538
PE	0.5074 ± 0.0912	0.4931 ± 0.0876	0.4989 ± 0.0893	0.4728 ± 0.0834	0.4627 ± 0.0791	**0.4382** ± 0.0726
DB_{fuzzy}	**0.6934** ± 0.1082	1.5142 ± 0.1425	0.7249 ± 0.1138	0.8157 ± 0.1071	0.9345 ± 0.0987	2.1479 ± 0.1589
Soybean						
PC	0.8124 ± 0.0917	0.8317 ± 0.0863	0.7829 ± 0.1125	0.8368 ± 0.0814	0.8441 ± 0.0786	**0.8619** ± 0.0713
PE	0.3881 ± 0.1124	0.3627 ± 0.1012	0.4786 ± 0.1297	0.3579 ± 0.0964	0.3425 ± 0.0921	**0.3287** ± 0.0843
DB_{fuzzy}	3.4972 ± 0.1788	3.6045 ± 0.1736	4.5289 ± 0.2014	**2.3841** ± 0.1592	2.6974 ± 0.1487	2.6812 ± 0.1449
Thyroid						
PC	0.6329 ± 0.1186	0.6463 ± 0.1137	0.6617 ± 0.1075	0.7018 ± 0.0948	0.7163 ± 0.0894	**0.7524** ± 0.0797
PE	0.5871 ± 0.1312	0.5783 ± 0.1274	0.5692 ± 0.1195	0.5393 ± 0.1107	0.5284 ± 0.1079	**0.4189** ± 0.0954
DB_{fuzzy}	**2.0471** ± 0.1938	3.3829 ± 0.1842	4.6985 ± 0.2093	2.7987 ± 0.1694	2.5234 ± 0.1647	2.4481 ± 0.1546

Bảng 2.8: So sánh các độ đo PC, PE và DB_fuzzy trên các tập dữ liệu Wine, Soybean và Thyroid.

Dựa trên kết quả thực nghiệm với dữ liệu chuẩn UCI, có thể khẳng định ASSFBC đạt mức hiệu quả ổn định và vượt trội trên hầu hết các tập dữ liệu, đặc biệt về các thước đo đối sánh với nhãn thật như RI, F1 và NMI (cần tối đa hoá). Riêng DB và DB_fuzzy là các chỉ số nội tại (cần tối thiểu hoá), phản ánh độ chặt và mức tách biệt cụm. Ở một số tập dữ liệu, FCM cho giá trị DB và DB_fuzzy thấp nhất — điều này xuất phát từ bản chất của FCM: thuật toán chỉ cố gắng giảm tổng khoảng cách giữa điểm dữ liệu và tâm cụm, mà không xét

đến quan hệ giữa các cụm hoặc các vùng chồng lấn. Do đó, các cụm của FCM thường rất tròn về hình học, gọn, đối xứng và "chặt", khiến chỉ số DB có xu hướng thấp, nhưng không phản ánh đúng ranh giới thực hoặc tính ngữ nghĩa của dữ liệu.

Ngược lại, ASSFBC sử dụng cơ chế ràng buộc bán giám sát và điều chỉnh vùng biên để giữ lại cấu trúc mờ hợp lý, giúp phân biệt cụm theo ngữ nghĩa tốt hơn. Kết quả là, mặc dù DB hoặc DB_fuzzy của ASSFBC đôi khi cao hơn FCM, nhưng chúng vẫn tốt hơn tất cả các thuật toán bán giám sát mờ khác (SSFCM, eFCM, AFFC-2008, AFFC-2017) và phản ánh đúng hơn sự cân bằng giữa độ chặt cụm và độ chính xác theo nhãn.

Đồng thời, khi xét thêm các chỉ số đặc trưng cho phân cụm mờ như PC và PE, ASSFBC luôn đạt PC cao nhất và PE thấp nhất trong toàn bộ các thử nghiệm, chứng tỏ biên cụm được mô hình hoá rõ ràng, giảm đáng kể độ bất định. Như vậy, ngay cả khi DB không phải nhỏ nhất tuyệt đối, các chỉ số mờ vẫn xác nhận phân hoạch của ASSFBC là rõ ràng, ổn định và chính xác hơn.

Trên tập **IRIS**, ASSFBC đạt RI = 0.9939, F1 = 0.9898, NMI = 0.9744, với PC = 0.9418 và PE = 0.2135, tất cả đều đứng đầu và cải thiện đáng kể so với AFFC(2017). Giá trị DB = 1.2294 chỉ đứng sau FCM (0.9414), cho thấy biên cụm vẫn gọn và rõ ràng dù thuật toán ưu tiên tối ưu hoá các thước đo có giám sát.

Đối với tập **Breast Cancer**, ASSFBC đạt RI = 0.7848, F1 = 0.8651, NMI = 0.5089, với PC = 0.7925 và PE = 0.4235. Mặc dù DB = 0.9144 và DB_fuzzy = 0.9526 cao hơn eFCM và FCM, mức chênh lệch không đáng kể; các chỉ số RI, F1 và NMI vẫn dẫn đầu, phản ánh khả năng giữ cân bằng tốt giữa độ chính xác và độ chặt cụm.

Với **Glass**, ASSFBC đạt RI = 0.8498, F1 = 0.6147, NMI = 0.7579, với PC

= 0.7517 và PE = 0.4125. Vì dữ liệu Glass có mức chồng lấn mạnh giữa 6 lớp, DB = 3.1947 và DB_fuzzy cao hơn các phương pháp khác là điều hợp lý. Tuy nhiên, ASSFBC vẫn vượt trội hơn mọi thuật toán bán giám sát nhờ khả năng nới biên mềm có kiểm soát để mô hình hoá đúng các điểm lưỡng lự.

Với **Wine**, ASSFBC đạt RI = 0.7689, F1 = 0.6578, NMI = 0.4955, đều cao nhất trong toàn bộ các phương pháp so sánh. DB = 2.2867 lớn hơn FCM và eFCM (vốn tạo cụm tròn, chặt về mặt hình học), nhưng ASSFBC vẫn vượt các thuật toán bán giám sát còn lại (SSFCM, AFFC-2008, AFFC-2017), cho thấy mô hình giữ được ranh giới cụm mềm hợp lý trong không gian nhiều thuộc tính.

Còn với **Soybean**, ASSFBC đạt RI = 0.8677, F1 = 0.7571, NMI = 0.7784, tiếp tục vượt trội nhất quán ở cả ba chỉ số ngoại sinh. DB = 2.7581 tuy không thấp nhất (AFFC-2008 đạt 2.4579), nhưng ASSFBC lại dẫn đầu về độ chính xác và độ ổn định, cho thấy khả năng tổng quát hoá mạnh mẽ trên dữ liệu nhiều lớp và chứa nhiều nhiễu.

Với **Thyroid**, ASSFBC đạt RI = 0.7649, F1 = 0.7367, NMI = 0.5284, dẫn đầu toàn bộ phương pháp so sánh. DB = 2.1271 chỉ nhỉnh hơn FCM (2.0871) và thấp hơn toàn bộ các thuật toán bán giám sát khác. Điều này cho thấy ASSFBC thiết lập được biên cụm chặt và phù hợp với cấu trúc thực tế của dữ liệu, đạt được sự cân bằng tốt giữa tính đúng nhãn và độ chặt cụm — một đặc tính hiếm gặp trong các mô hình phân cụm mờ có ràng buộc.

Đối với **dữ liệu tự sinh**, **Data1** cho kết quả gần như hoàn hảo với RI = F1 = NMI = 1.0, kèm theo PC = 0.9743, PE = 0.0301 và DB_fuzzy = 0.5284, phản ánh độ ổn định rất cao và biên cụm sắc nét. Trên **Data2**, ASSFBC tiếp tục dẫn đầu với RI = 0.8688, F1 = 0.8035, NMI = 0.6657, đồng thời duy trì PC = 0.8462 và PE = 0.3921. Mặc dù DB và DB_fuzzy cao hơn FCM (do mức chồng lấn giữa các cụm lớn hơn), ASSFBC vẫn vượt trội so với toàn bộ các

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
Data1						
RI	0.9523 ± 0.0568	0.9725 ± 0.0615	0.8941 ± 0.0914	**1**	**1**	**1**
F1	0.9320 ± 0.0553	0.9624 ± 0.0584	0.8927 ± 0.0975	**1**	**1**	**1**
NMI	0.9013 ± 0.0798	0.9317 ± 0.0779	0.7094 ± 0.1282	**1**	**1**	**1**
DB	1.8451 ± 0.1541	2.3049 ± 0.1784	3.0185 ± 0.1971	**0.9282** ± 0.0614	0.9347 ± 0.0642	0.9433 ± 0.0675
Data2						
RI	0.7194 ± 0.1085	0.7426 ± 0.0962	0.7892 ± 0.0924	0.8131 ± 0.0847	0.8378 ± 0.0782	**0.8725** ± 0.0710
F1	0.6694 ± 0.1249	0.6825 ± 0.1184	0.7024 ± 0.1132	0.7529 ± 0.1017	0.7391 ± 0.1080	**0.8148** ± 0.0941
NMI	0.4895 ± 0.1521	0.5731 ± 0.1368	0.6152 ± 0.1297	0.6418 ± 0.1219	0.6184 ± 0.1312	**0.6761** ± 0.1154
DB	**0.7724** ± 0.0815	1.0081 ± 0.0928	1.5849 ± 0.1081	3.6928 ± 0.1587	4.5824 ± 0.1769	4.3715 ± 0.1692

Bảng 2.9: So sánh các hiệu năng các phương pháp trên các tập dữ liệu tự sinh

Algorithm Index	FCM	SSFCM	eFCM	AFFC(2008)	AFFC(2017)	ASSFBC
Data1						
PC	0.9321 ± 0.0507	0.9340 ± 0.0521	0.8875 ± 0.0714	0.9687 ± 0.0248	0.9714 ± 0.0236	**0.9751** ± 0.0219
PE	0.1184 ± 0.0615	0.1127 ± 0.0587	0.1548 ± 0.0732	0.0417 ± 0.0324	0.0372 ± 0.0308	**0.0294** ± 0.0261
DB_{fuzzy}	1.0500 ± 0.0850	1.2800 ± 0.0950	1.6500 ± 0.1100	**0.4800** ± 0.0400	0.5100 ± 0.0420	0.5200 ± 0.0450
Data2						
PC	0.7216 ± 0.0951	0.7429 ± 0.0897	0.7714 ± 0.0859	0.7948 ± 0.0778	0.8162 ± 0.0725	**0.8467** ± 0.0654
PE	0.5174 ± 0.1214	0.4831 ± 0.1097	0.4518 ± 0.1048	0.4236 ± 0.0941	0.4379 ± 0.0987	**0.3927** ± 0.0879
DB_{fuzzy}	**0.6800** ± 0.0700	0.9100 ± 0.0800	1.4800 ± 0.0950	3.4500 ± 0.1400	4.3150 ± 0.1600	4.1800 ± 0.1550

Bảng 2.10: So sánh các độ đo PC, PE và DB_fuzzy trên các tập dữ liệu tự sinh

thuật toán bán giám sát mờ khác, cho thấy mô hình giữ được ranh giới cụm mềm hợp lý và đạt được sự cân bằng tốt giữa độ chính xác theo nhãn và tính ổn định mờ của phân hoạch.

Algorithm Index	FCM	SSFCM	eFCM	AFFC(2008)	AFFC(2017)	ASSFBC
Image						
RI	0.7068 ± 0.0793	0.7309 ± 0.0751	0.7162 ± 0.0774	0.8278 ± 0.0562	0.8367 ± 0.0519	**0.8589** ± 0.0478
F1	0.6847 ± 0.0864	0.6968 ± 0.0825	0.7061 ± 0.0808	0.7953 ± 0.0591	0.8075 ± 0.0560	**0.8368** ± 0.0493
NMI	0.6314 ± 0.0975	0.6456 ± 0.0921	0.6593 ± 0.0883	0.7289 ± 0.0631	0.7494 ± 0.0602	**0.7879** ± 0.0517
DB	0.5978 ± 0.1065	0.6081 ± 0.1027	**0.5506** ± 0.1118	0.7371 ± 0.0934	0.7268 ± 0.0898	0.7498 ± 0.0962

Bảng 2.11: So sánh các hiệu năng các phương pháp trong bài toán phân đoạn ảnh.

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
Image						
PC	0.7051 ± 0.0794	0.7283 ± 0.0756	0.7152 ± 0.0778	0.8257 ± 0.0581	0.8354 ± 0.0537	**0.8609** ± 0.0498
PE	0.5258 ± 0.0947	0.5006 ± 0.0903	0.4909 ± 0.0871	0.3952 ± 0.0704	0.3805 ± 0.0668	**0.3457** ± 0.0596
DB_{fuzzy}	0.6108 ± 0.1079	0.6204 ± 0.1036	**0.5653** ± 0.1117	0.7457 ± 0.0946	0.7324 ± 0.0918	0.7608 ± 0.0979

Bảng 2.12: So sánh các độ đo PC, PE và DB_fuzzy trên dữ liệu ảnh.

Với **dữ liệu ảnh** (mức nhiễu lớn và vùng biên mềm), ASSFBC đạt RI = 0.8589, F1 = 0.8368, NMI = 0.7879, PC = 0.8609 và PE = 0.3457. DB = 0.7498 và DB_fuzzy = 0.7608 không phải thấp nhất (eFCM cho giá trị nhỏ hơn), nhưng mức chênh lệch này là rất nhỏ và không ảnh hưởng đáng kể đến độ chính xác tổng thể. ASSFBC vẫn duy trì được cấu trúc biên mềm ổn định và phân tách vùng rõ ràng — yếu tố đặc biệt quan trọng trong các bài toán giám sát yếu và phân đoạn ảnh phức tạp, nơi ranh giới khu vực thường bị nhiễu mạnh và khó xác định theo đường biên cứng.

Tóm lại Kết quả thực nghiệm cho thấy **ASSFBC** vượt trội ổn định trên hầu hết các tập dữ liệu, đặc biệt trong các tình huống phức tạp, nhiễu và có vùng biên mềm như dữ liệu ảnh. Thuật toán đạt **RI, F1, NMI tốt nhất trên đa số các tập dữ liệu**, đồng thời đạt PC cao nhất, PE thấp nhất và DB_fuzzy luôn tốt hơn các thuật toán bán giám sát khác. Điều này phản ánh khả năng phân cụm chính xác, duy trì cấu trúc lớp thật và tạo ranh giới mềm hợp lý.

Mặc dù **FCM** đôi khi đạt giá trị thấp nhất trên các chỉ số DB và DB_fuzzy, kết quả này chủ yếu phản ánh xu hướng của thuật toán trong việc tối ưu nội tại theo hướng tạo ra các cụm có phương sai nhỏ và hình dạng gần cầu (tròn,

đối xứng). Vì vậy, các cụm thu được thường "gọn" theo tiêu chí hình học của chỉ số đánh giá, nhưng chưa chắc tương ứng với cấu trúc phân bố thực của dữ liệu. Đặc biệt, khi dữ liệu có vùng chồng lấn, nhiễu, hoặc biên cụm phi tuyến và hình dạng phức tạp. Ngược lại, **ASSFBC** đạt được sự cân bằng tối ưu giữa độ chính xác theo nhãn thật, độ chặt cụm và tính mờ hợp lý, nhấn mạnh khả năng duy trì quan hệ ngữ nghĩa giữa các điểm dữ liệu. Nhờ đó, thuật toán trở thành lựa chọn đáng tin cậy trong các ứng dụng yêu cầu độ chính xác cao, đồng thời thể hiện tiềm năng mở rộng trong các bài toán phân cụm mờ trên dữ liệu lớn và đa dạng.

Đánh giá theo thời gian tính toán

Kết quả trong Bảng 2.13 cho thấy SSFCM đạt thời gian chạy trung bình thấp nhất và hội tụ nhanh nhất nhờ chỉ sử dụng các phép cập nhật độ phụ thuộc mờ cơ bản mà không có bước điều chỉnh sau huấn luyện. eFCM duy trì tốc độ tốt nhờ cơ chế cải tiến hệ số mờ, giúp giảm nhiễu trong cập nhật tâm cụm mà không làm tăng chi phí tính toán đáng kể. Thuật toán đề xuất ASSFBC có thời gian chạy trung bình cao hơn khoảng 10–20% so với eFCM do bổ sung thêm giai đoạn điều chỉnh vùng biên cụm sau bước hội tụ của FCM. Giai đoạn này giúp tinh chỉnh lại các điểm có mức độ phụ thuộc thấp (mờ mạnh) bằng cách điều chỉnh vị trí tâm cụm và độ phụ thuộc cục bộ quanh biên giữa hai cụm lân cận. Dù chi phí tăng nhẹ, số vòng lặp cần để hội tụ vẫn ít hơn AFFC(2017) khoảng 5–10 vòng, cho thấy quá trình cập nhật của ASSFBC ổn định hơn.

Về tổng thể, ASSFBC thể hiện sự cân bằng giữa độ chính xác và tốc độ: tuy thời gian huấn luyện không thấp nhất, nhưng hội tụ ổn định, không dao động mạnh tại biên cụm và tránh hiện tượng "nhảy tâm" thường gặp ở các biến thể AFFC. Đây là ưu điểm quan trọng để áp dụng trên dữ liệu nhiễu, chồng lấn hoặc có biên không rõ ràng.

Method	**IRIS**	**Wine**	**Soybean**	**Thyroid**	**Glass**	**Breast**	**Data1**	**Data2**	**Image**
SSFCM	**0.38 s** (16 iters)	**0.63 s** (20 iters)	**1.18 s** (26 iters)	**3.12 s** (30 iters)	**4.52 s** (31 iters)	**15.83 s** (65 iters)	**0.69 s** (15 iters)	**0.52 s** (18 iters)	**172.52 s** (139 iters)
eFCM	0.41 s (17 iters)	0.66 s (22 iters)	1.25 s (27 iters)	3.34 s (32 iters)	4.87 s (33 iters)	16.94 s (66 iters)	0.73 s (16 iters)	0.57 s (20 iters)	179.35 s (153 iters)
ASSFBC	*0.44 s (19 iters)*	*0.71 s (24 iters)*	*1.33 s (30 iters)*	*3.58 s (34 iters)*	*5.10 s (35 iters)*	*17.24 s (68 iters)*	*0.77 s (17 iters)*	*0.61 s (22 iters)*	*181.52 s (152 iters)*
AFFC (2017)	0.47 s (22 iters)	0.75 s (26 iters)	1.41 s (32 iters)	3.83 s (37 iters)	5.42 s (38 iters)	20.21 s (71 iters)	0.83 s (19 iters)	0.67 s (25 iters)	185.42 s (163 iters)

Bảng 2.13: So sánh thời gian chạy trung bình và số vòng lặp hội tụ giữa SSFCM, eFCM, AFFC (2017) và thuật toán đề xuất ASSFBC trên các tập dữ liệu chuẩn và ảnh

2.5 Kết luận chương

Tóm lại, Chương 2 đã giới thiệu một cách tiếp cận đổi mới đối với phương pháp phân cụm bán giám sát mờ chủ động, tập trung vào việc tinh chỉnh vùng biên cụm. Phương pháp này giải quyết hiệu quả thách thức về sự không chắc chắn ở vùng biên bằng cách tích hợp cơ chế học chủ động nhằm tăng độ chính xác của việc gán mức độ phụ thuộc cho các điểm quan trọng. Cách tiếp cận được đề xuất không chỉ nâng cao độ tin cậy của phân cụm mà còn xây dựng một mô hình bán giám sát mờ mới, kết hợp chặt chẽ giữa phân cụm mờ và học chủ động để tối ưu hóa ranh giới cụm.

Các kết quả thực nghiệm trên các tập dữ liệu chuẩn cho thấy phương pháp đề xuất, phân cụm bán giám sát mờ chủ động dựa trên vùng biên cụm (**ASSFBC**), vượt trội so với các thuật toán phân cụm mờ truyền thống và các phương pháp bán giám sát khác. Điều này thể hiện rõ qua việc cải thiện đồng thời các chỉ số ngoại sinh như RI, F1-Score và NMI, các chỉ số đặc trưng cho phân cụm mờ như PC, PE cũng cho kết quả cao nhất so với các thuật toán so sánh, cũng như các chỉ số nội tại như DB và DB_fuzzy tuy không đứng đầu vì thường đứng sau FCM nhưng cũng hầu như cao hơn so với các phương pháp phân cụm bán giám sát mờ khác. Sự vượt trội đồng đều trên các nhóm chỉ số cho thấy mô hình không chỉ tối ưu hóa theo nhãn thật, mà còn duy trì được biên cụm gọn, ổn định và phù hợp với bản chất mờ của dữ liệu.

Kết quả của chương này đã được công bố trong các công trình CT1 và CT2.

Chương 3

Đề xuất phương pháp phân cụm bán giám sát an toàn chủ động với cặp ràng buộc dựa vào biên cụm

Chương này tiếp nối các kết quả của Chương 2. Trong khi ASSFBC tập trung vào việc truy vấn và hiệu chỉnh vùng biên cụm để cải thiện ranh giới phân cụm, chương này mở rộng theo hướngan toàn nhằm giảm ảnh hưởng của thông tin giám sát thiếu tin cậy (nhãn/ràng buộc sai), đặc biệt tại các vùng bất định cao. Trên cơ sở đó, luận án đề xuất phương pháp phân cụm bán giám sát an toàn chủ động với cặp ràng buộc dựa vào biên cụm (AS3FCPC). Nội dung chính của chương gồm: động lực và thách thức dữ liệu thiếu tin cậy; mô hình và sơ đồ thuật toán AS3FCPC; mô tả chi tiết các bước (tiêu chí truy vấn điểm biên, cơ chế an toàn và tích hợp cặp ràng buộc); phân tích độ phức tạp; và thực nghiệm trên các bộ dữ liệu UCI. Ngoài ra, chương trình bày thí nghiệm trên dữ liệu ảnh vùng ngập lụt để minh họa khả năng ứng dụng của phương pháp đề xuất.

3.1 Mở đầu

Trong chương 2, luận án đã đề xuất phương pháp ASSFBC — một mô hình phân cụm bán giám sát mờ chủ động dựa trên biên cụm. Phương pháp này tập trung vào việc xác định và hiệu chỉnh vùng biên cụm thông qua cơ chế học chủ động, giúp cải thiện đáng kể độ chính xác của các thuật toán SSFC truyền thống. Tuy nhiên, đối với phân cụm bán giám sát mờ, bên cạnh các vấn đề về

dữ liệu ở vùng biên, một trong những thách thức khác chính là vấn đề dữ liệu thiếu tin cậy mà điển hình nhất là việc gán nhãn sai. Dữ liệu thiếu tin cậy có thể gây sai lệch nghiêm trọng trong việc học mô hình, làm sai lệch ranh giới phân cụm và làm giảm độ chính xác của cả quá trình phân loại và phân cụm. Trong các thuật toán bán giám sát hiện đại, dữ liệu thiếu tin cậy không chỉ ảnh hưởng đến từng điểm riêng lẻ mà còn tác động lan tỏa tới toàn bộ cấu trúc cụm thông qua các cơ chế lan truyền nhãn và cập nhật thành viên; đây cũng là phần mà thuật toán trong chương 2 vẫn chưa xử lý được. Chính vì vậy, việc đảm bảo an toàn cho quá trình học bán giám sát trở thành động lực của luận án để giải quyết.

Các phương pháp SSFC hiện đại đã khai thác ước lượng độ tin cậy (theo điểm, theo cặp, theo cụm hoặc theo thuộc tính) để điều tiết ảnh hưởng của tín hiệu giám sát. Tiếp cận Safe SSFC điển hình như S^3FCM, LHC-S3FCM, CS3FCM [17, 21, 101] ước lượng "độ tin cậy" bằng so khớp nhãn và cấu trúc láng giềng (kNN), rồi đưa vào trọng số an toàn trong hàm mục tiêu: nhãn đáng ngờ được giảm trọng số, giúp mô hình bền vững hơn với nhiễu. Tuy nhiên, nhóm phương pháp này còn một số hạn chế cốt lõi: (i) chi phí đồ thị hoặc láng giềng tăng nhanh khi dữ liệu lớn; (ii) tham số lân cận cố định không thích nghi với mật độ không đồng đều; (iii) độ tin cậy nội sinh phụ thuộc mạnh vào khởi tạo hoặc phân hoạch ban đầu; (iv) ràng buộc cặp phức tạp (xung đột ML/CL, vi phạm tam giác) chưa được xử lý trong một mục tiêu thống nhất; và (v) độ nhạy siêu tham số (trọng số regularizer, kích thước lân cận) dễ "khóa cứng" cấu trúc.

Để giảm lan truyền lỗi sớm, TS3FCM [18] áp dụng chiến lược trusted-first: (1) lọc nhãn đáng tin trên phần dữ liệu có nhãn bằng FCM cải tiến và trọng số láng giềng; (2) chuyển nhãn đáng tin thành ma trận liên thuộc khởi tạo $\bar{U}$; (3) tối ưu SSFC với việc định hướng U dựa vào $\bar{U}$. Cách làm này cải thiện độ

ổn định và hiệu quả tính toán so với Safe SSFC thuần. Dẫu vậy, TS3FCM vẫn nhạy khởi tạo, dùng bán kính hoặc láng giềng cố định, và gặp khó tại vùng giao khi mật độ hoặc độ cong hình học thay đổi theo không gian; hơn nữa, mô hình chưa tận dụng thông tin quan hệ giàu ngữ cảnh ngoài láng giềng gần.

Tóm lại, các phương pháp phân cụm bán giám sát mờ an toàn hiện nay chủ yếu sử dụng cơ chế đánh trọng số và sàng lọc dữ liệu đã có những kết quả đáng kể trong việc giảm sai lệch từ các thông tin thiếu tin cậy khiến cho kết quả phân cụm hiệu quả hơn.. Tuy nhiên các phương pháp trên vẫn còn hạn chế do còn nhạy cảm với việc khởi tạo ngẫu nhiên, việc xử lý nhiễu cũng chưa thực sự hiệu quả với những dữ liệu có độ nhiễu cao nhất là ở vùng biên cụm. Chính vì những hạn chế và khoảng trống khoa học này, trong chương 3, luận án trình bày phương pháp phân cụm bán giám sát an toàn chủ động với cặp ràng buộc dựa vào biên cụm kết hợp phân cụm mờ, học bán giám sát an toàn, học chủ động, sử dụng cặp ràng buộc dựa vào vùng biên để giải quyết các hạn chế của các thuật toán cũ, xử lý triệt để vấn đề dữ liệu thiếu tin cậy, cũng như nâng cao hiệu suất phân cụm.

3.2 Ý tưởng thuật toán

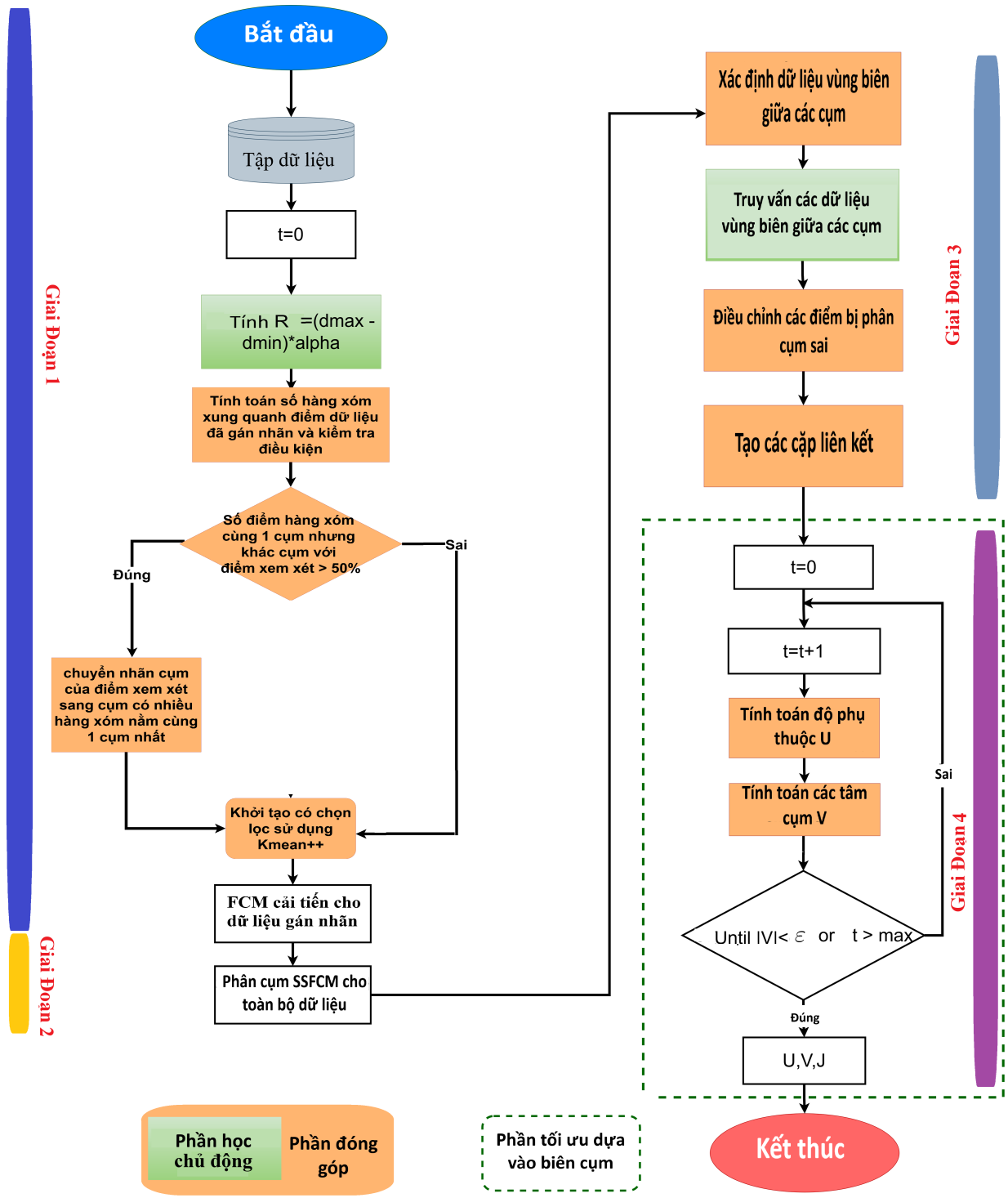
Mô hình đề xuất tích hợp phân cụm mờ (Fuzzy Clustering), Học bán giám sát an toàn (Safe Semi-Supervised Learning) và Học chủ động (Active Learning) nhằm cải thiện hiệu suất phân cụm trong các trường hợp dữ liệu được gán nhãn hạn chế, nhãn sai nhiều và phân bố dữ liệu phức tạp hoặc mơ hồ. Ý tưởng chính là kết hợp tính linh hoạt của phân cụm mờ với độ chính xác của học bán giám sát và hiệu suất của học chủ động, đảm bảo rằng mô hình có thể thích ứng với các bất định trong dữ liệu đồng thời tận dụng tri thức sẵn có.

Mô hình bắt đầu bằng việc thiết lập nền tảng kết hợp cả dữ liệu đã gán nhãn

và chưa gán nhãn, giúp khởi tạo quá trình phân cụm với sự cân bằng giữa thông tin giám sát và không giám sát. Bằng cách tập trung vào các vùng dữ liệu mơ hồ - nơi các điểm dữ liệu nằm gần ranh giới cụm hoặc có mức độ thành viên không chắc chắn - phương pháp xác định các điểm dữ liệu quan trọng cần tinh chỉnh. Các khu vực này được xử lý bằng các chiến lược học chủ động, trong đó hệ thống truy vấn nhãn hoặc điều chỉnh dựa trên mối quan hệ lân cận nhằm giảm thiểu phương pháp phân cụm bán giám sát an toàn chủ động với cặp ràng buộc dựa vào biên cụm.

Để tăng cường quá trình phân cụm, mô hình sử dụng các ràng buộc cặp (pair-wise constraints), bao gồm ràng buộc (must-link) và không ràng buộc (cannot-link), nhằm đại diện cho các mối quan hệ đặc thù của miền dữ liệu. Các ràng buộc này đảm bảo rằng một số điểm dữ liệu được nhóm lại với nhau hoặc tách biệt, giúp quá trình phân cụm tuân theo các quan hệ đã biết. Nhờ đó, mô hình có thể giải quyết hiệu quả các trường hợp mơ hồ ở các vùng biên hoặc vùng có sự chồng lấn giữa các cụm, nâng cao độ tin cậy và khả năng diễn giải của kết quả phân cụm.

Sự tích hợp giữa phân cụm mờ, học bán giám sát an toàn và học chủ động tạo ra một quy trình phân cụm động và thích nghi, đặc biệt hiệu quả khi xử lý các tập dữ liệu phức tạp trong thực tế. Khả năng kết hợp tính linh hoạt, sự giám sát và khả năng thích ứng giúp mô hình trở thành một giải pháp mạnh mẽ cho các ứng dụng mà dữ liệu gán nhãn bị hạn chế, đồng thời cần tận dụng tri thức miền để đạt được độ chính xác phân cụm cao. Hình 3.1 minh họa mô hình phân cụm bán giám sát mờ an toàn chủ động với cặp ràng buộc dựa vào biên cụm (AS3FCPC).



Hình 3.1: Mô hình phân cụm bán giám sát mờ an toàn chủ động với cặp ràng buộc dựa vào biên cụm (AS3FCPC)

3.3 Chi tiết thuật toán

Phương pháp AS3FCPC đề xuất bao gồm bốn giai đoạn tuần tự: (1) khởi tạo học chủ động kết hợp với FCM cho dữ liệu đã gán nhãn; (2) phân cụm bán giám sát mờ lặp lại cho toàn bộ tập dữ liệu; (3) tinh chỉnh ranh giới cụm một cách chủ động và tạo ràng buộc cặp một cách chọn lọc; và (4) tối ưu hóa lặp lại và hội tụ cuối cùng.

Giai đoạn 1: Khởi tạo học chủ động và FCM cho dữ liệu có nhãn

Quy trình bắt đầu với khởi tạo học chủ động, tập trung vào việc xác định các điểm dữ liệu mơ hồ hoặc không chắc chắn bằng cách phân tích mối quan hệ lân cận. Cụ thể, với mỗi điểm dữ liệu, bán kính lân cận R được xác định như sau:

$$R = (d_{\max} - d_{\min})\alpha \tag{3.1}$$

Trong đó $d_{\max}$ và $d_{\min}$ lần lượt là khoảng cách lớn nhất và nhỏ nhất giữa các điểm dữ liệu đã gán nhãn, và α là hệ số tỷ lệ do chuyên gia đề xuất.

Các điểm mơ hồ được xử lý dựa trên một quy tắc ưu tiên nhằm phát hiện và hiệu chỉnh các trường hợp bất đồng nhãn rõ ràng từ cấu trúc lân cận. Cụ thể, nếu từ 50% trở lên các điểm lân cận thuộc về một cụm khác (khác với nhãn hiện tại của điểm đang xét), nhãn của điểm đó sẽ được tự động điều chỉnh theo cụm lân cận chiếm ưu thế. Trong các thực nghiệm của luận án, ngưỡng 50% (thay vì các ngưỡng lớn hơn) đã đủ để biểu thị sự bất đồng mạnh trong hầu hết tình huống khi số cụm lớn hơn 2. Lý do là khi một cụm chiếm quá nửa lân cận, trong khi các cụm còn lại không chiếm tỷ lệ đáng kể, khả năng nhãn ban đầu phù hợp với cấu trúc cục bộ của dữ liệu là thấp. Vì vậy, ngưỡng này giúp phát hiện hiệu quả các trường hợp nhiễu nhãn mà không đòi hỏi một đa số tuyệt đối

quá cao. Với trường hợp chỉ có 2 cụm, luận án sử dụng ngưỡng 70% thay cho 50% để đảm bảo mức độ bất đồng đủ mạnh trước khi tự động hiệu chỉnh.

Sau khi giải quyết các điểm mơ hồ, thuật toán Fuzzy C-Means cải tiến 1.5 được áp dụng trên dữ liệu có nhãn để thiết lập các tâm cụm và ma trận độ phụ thuộc ban đầu. Các kỹ thuật khử mờ (defuzzification) sau đó được sử dụng để tinh chỉnh và đảm bảo khởi tạo ổn định cho giai đoạn tiếp theo.

Giai đoạn 2: phân cụm bán giám sát mờ lặp lại

Sử dụng ma trận độ phụ thuộc đã khởi tạo $\bar{u}_{ik}$, giai đoạn này tiến hành phân cụm bán giám sát mờ (SSFCM) trên toàn bộ tập dữ liệu. Các giá trị độ phụ thuộc U_{ik} được cập nhật lặp lại, tích hợp nguyên lý phân cụm mờ và nhãn từ chuyên gia. Các tâm cụm V_k được cập nhật song song, cải thiện độ chặt và độ phân biệt của cụm sau mỗi lần lặp. Giai đoạn này lan truyền thông tin nhãn đáng tin cậy sang dữ liệu chưa gán nhãn, đặt nền tảng vững chắc cho các bước tinh chỉnh sau.

Giai đoạn 3: Tinh chỉnh vùng biên cụm sử dụng học chủ động và tạo ràng buộc cặp

Giai đoạn này tập trung vào việc tinh chỉnh cấu trúc phân cụm thông qua việc phân tích ranh giới cụm và xác định chủ động các điểm dữ liệu quan trọng để tạo ràng buộc. Trước tiên, vùng biên của các cụm sơ bộ được xác định nhằm tìm ra các điểm dữ liệu có mức độ phụ thuộc không chắc chắn — thường nằm ở vùng mơ hồ gần ranh giới cụm. Học chủ động sau đó được áp dụng cho các điểm biên này nhằm đánh giá độ tin cậy nhãn. Bằng cách kiểm tra mối quan hệ lân cận, mô hình sẽ xác nhận, truy vấn thêm, hoặc điều chỉnh lại nhãn để đảm bảo tính nhất quán. Đánh giá có trọng tâm này đảm bảo rằng các điểm ranh giới được thể hiện chính xác trong cấu trúc phân cụm, tăng cường mức độ tin cậy ban đầu của mô hình.

Khi độ tin cậy của nhãn được cải thiện, mô hình tiến hành tạo các ràng buộc cặp theo chiến lược từ Basu et al. [30]. Các quan hệ must-link ($\mathcal{M}$) và cannot-link ($\mathbb{C}$) được tạo chọn lọc từ các điểm dữ liệu biên đã được điều chỉnh, đảm bảo rằng tri thức chuyên gia được tích hợp chính xác vào những khu vực quan trọng nhất. Những ràng buộc này giúp giải quyết mơ hồ ở các vùng chồng lấp hoặc không rõ ràng, dẫn hướng phân cụm đến cấu hình ổn định và chính xác hơn. Đây là bước chuẩn bị quan trọng trước khi bước vào giai đoạn tối ưu chính, giúp mô hình bắt đầu từ trạng thái đã tinh chỉnh và được hướng dẫn rõ ràng, từ đó cải thiện độ chính xác và khả năng xử lý dữ liệu phức tạp.

Giai đoạn 4: Tối ưu hóa và hội tụ cuối cùng

Trong giai đoạn cuối, mô hình bước vào vòng lặp tối ưu hóa, nơi quá trình phân cụm được tinh chỉnh dần dần bằng cách tối thiểu hóa hàm mục tiêu cho đến khi hội tụ. Quá trình này cập nhật lặp lại ma trận độ phụ thuộc U và các tâm cụm V, tích hợp các nguyên lý phân cụm mờ, tri thức từ các giai đoạn trước, và ràng buộc cặp được tạo ra trong bước tinh chỉnh biên.

Tại mỗi vòng lặp, giá trị độ phụ thuộc được tính toán lại nhằm cân bằng giữa khoảng cách tới tâm cụm, độ tin cậy của nhãn trước đó, và sự tuân thủ ràng buộc must-link và cannot-link. Tâm cụm được cập nhật tương ứng để phản ánh sự thay đổi, nâng cao độ chặt và sự phân biệt của các cụm. Quá trình này tiếp tục cho đến khi đạt tiêu chí hội tụ – khi sự thay đổi giữa các tâm cụm là rất nhỏ hoặc đạt số vòng lặp tối đa. Kết quả cuối cùng là một cấu hình phân cụm tối ưu, tích hợp hiệu quả tri thức giám sát và cấu trúc dữ liệu, đảm bảo phân cụm chính xác và ổn định.

Hàm mục tiêu tổng thể được tối ưu bởi phương pháp AS3FCPC được định nghĩa như sau:

$$\min_{u,v} \; J(u,v) = \sum_{i=1}^{N}\sum_{k=1}^{C} u_{ik}^2\, d_{ik}^2 + \alpha \sum_{i=1}^{L}\sum_{k=1}^{C}\Big(u_{ik} - \bar{u}_{ik}\Big)^2 + \beta\left(\sum_{(x_i,x_j)\in\mathcal{M}}\sum_{k=1}^{C}\sum_{\substack{\ell=1\\ \ell\neq k}}^{C} u_{ik}\, u_{j\ell} + \sum_{(x_i,x_j)\in\mathbb{C}}\sum_{k=1}^{C} u_{ik}\, u_{jk}\right). \tag{3.2}$$

Với ràng buộc: $\sum_{k=1}^{C} u_{ik} = 1; \quad \forall k = 1,\ldots,C$ và $u_{ik}, \bar{u}_{ik} \in [0,1], \quad \forall i = 1,\ldots,N$, với tập dữ liệu $X = \{x_1, x_2, \ldots, x_k, \ldots, x_N\}$ có N điểm dữ liệu, C cụm, L là lực lượng điểm dữ liệu ở tất cả các vùng biên bao gồm cả các điểm giám sát đã được truy vấn. u_{ik} là mức độ phụ thuộc về cụm k của điểm dữ liệu i, d_{ik} là khoảng cách từ x_i đến tâm cụm V_k, $\mathcal{M}$ là tập ràng buộc must-link, và $\mathbb{C}$ là tập ràng buộc cannot-link.

Ta có thể chia hàm mục tiêu thành bốn phần:

$\sum_{i=1}^{N}\sum_{k=1}^{C} u_{ik}^2 d_{ik}^2$	là **phần fuzzy**, biểu diễn tổng bình phương của giá trị độ phụ thuộc và khoảng cách đến tâm cụm.
$\alpha \sum_{i=1}^{L}\sum_{k=1}^{C}(u_{ik} - \bar{u}_{ik})^2$	là **phần bán giám sát chủ động**, trong phần này, chúng ta lấy các phần tử vùng biên được gắn nhãn dựa trên học chủ động cho hàm thành viên làm thành viên được giám sát. $\bar{u}_{ik}$.
$\beta \sum_{(x_i,x_j)\in\mathcal{M}}\sum_{k=1}^{C}\sum_{\substack{l=1\\l\neq k}}^{C} u_{ik}u_{jl}$	là **phần ràng buộc must-link**, khuyến khích các điểm có quan hệ ràng buộc thuộc cùng cụm.
$\beta \sum_{(x_i,x_j)\in\mathbb{C}}\sum_{k=1}^{C} u_{ik}u_{jk}$	là **phần ràng buộc cannot-link**, khuyến khích các điểm có quan hệ ràng buộc thuộc các cụm khác nhau.

Ghi chú. Thành phần thứ ba trong hàm mục tiêu (3.2) thể hiện các ràng buộc cặp (must–link và cannot–link) giữa các điểm dữ liệu. Không giống hai hạng đầu vốn phụ thuộc trực tiếp vào khoảng cách hình học d_{ik}^2, các ràng buộc này chỉ đặc trưng cho quan hệ nhãn giữa hai điểm (x_i, x_j) và do đó không có thứ nguyên không gian. Hệ số β đóng vai trò điều chỉnh tỷ lệ, đảm bảo năng lượng của các ràng buộc được cân bằng với hai thành phần dựa trên khoảng cách.

với các ràng buộc $\sum_{k=1}^{C} u_{ik} = 1$, $u_{ik} \geq 0$. Kí hiệu:

$$\mathcal{M}_i := \{j : (x_i, x_j) \in \mathcal{M}\}, \qquad \mathcal{C}_i := \{j : (x_i, x_j) \in \mathcal{C}\}, \qquad S_i := \sum_{r=1}^{C} \frac{1}{d_{ir}^2 + \alpha} > 0.$$

Thực hiện Lagrangian là

$$\mathcal{L}(U,V,\lambda) = J(U,V) + \sum_{i=1}^{N} \lambda_i \Big(\sum_{k=1}^{C} u_{ik} - 1 \Big).$$

Điều kiện dừng theo u_{ik} cho ta

$$\frac{\partial \mathcal{L}}{\partial u_{ik}} = 2u_{ik}d_{ik}^2 + 2\alpha(u_{ik} - \bar{u}_{ik}) + \beta \sum_{j \in \mathcal{M}_i} \sum_{\ell \neq k} u_{j\ell} + \beta \sum_{j \in \mathcal{C}_i} u_{jk} + \lambda_i = 0.$$

Đặt

$$a_{ik} := 2\alpha\, \bar{u}_{ik} - \beta \sum_{j \in \mathcal{M}_i} \sum_{\ell \neq k} u_{j\ell} - \beta \sum_{j \in \mathcal{C}_i} u_{jk},$$

ta được công thức đóng

$$u_{ik} = \frac{a_{ik} - \lambda_i}{2(d_{ik}^2 + \alpha)}, \qquad \lambda_i = \frac{\displaystyle\sum_{r=1}^{C} \frac{a_{ir}}{d_{ir}^2 + \alpha} - 2}{\displaystyle\sum_{r=1}^{C} \frac{1}{d_{ir}^2 + \alpha}}. \tag{3.3}$$

Vì $\sum_{\ell \neq k} u_{j\ell} = 1 - u_{jk} \leq 1$ và $0 \leq u_{jk} \leq 1$, nên

$$\sum_{j \in \mathcal{M}_i} \sum_{\ell \neq k} u_{j\ell} \leq |\mathcal{M}_i|, \qquad \sum_{j \in \mathcal{C}_i} u_{jk} \leq |\mathcal{C}_i|.$$

Suy ra:

$$a_{ik} \;\geq\; 2\alpha\, \bar{u}_{ik} \;-\; \beta\Big(|\mathcal{M}_i| + |\mathcal{C}_i|\Big).$$

Do đó, nếu đặt

$$\boxed{\beta \;\leq\; \frac{2\alpha\, \underline{\bar{u}}_i^{(\Delta)}}{|\mathcal{M}_i| + |\mathcal{C}_i|}, \qquad \underline{\bar{u}}_i^{(\Delta)} := \min_{1 \leq r \leq C} \Big(\bar{u}_{ir} + \Delta\Big), \quad \Delta > 0} \tag{3.4}$$

thì $a_{ik} \geq 0$ với mọi k (điều kiện đủ, độc lập k).

Ghi chú:

(1) (3.4) là *điều kiện đủ*, có thể bảo thủ; Δ chỉ dùng như một "khoảng hở" phân tích để xử lý trường hợp $\min_r \bar{u}_{ir} = 0$, *không* thay đổi các giá trị $\bar{u}_{ik}$ trong thuật

toán.

(2) Nếu $\min_r \bar{u}_{ir} = 0$ thì (3.4) buộc β rất nhỏ (cỡ $\alpha\Delta/(|\mathcal{M}_i|+|\mathcal{C}_i|)$), làm yếu tác dụng ML/CL tại điểm i. Khi cần giữ ML/CL mạnh hơn, có thể dùng β_i theo từng điểm (hoặc β_{ij} theo từng cặp) với ràng buộc địa phương $\beta_i\,(|\mathcal{M}_i|+|\mathcal{C}_i|) \le 2\alpha\,\underline{\bar{u}}_i^{(\Delta)}$, hoặc thêm bước chiếu $u_{i\cdot}$ về đơn hình sau khi cập nhật.

Chặn dưới cho λ_i:

Nếu $a_{ir} \ge 0$ với mọi r thì

$$\lambda_i \;\ge\; -\frac{2}{S_i}.$$

Thật vậy: Từ $a_{ir} \ge 0$ và $d_{ir}^2 + \alpha > 0$ suy ra $\sum_r \frac{a_{ir}}{d_{ir}^2+\alpha} \ge 0$. Thế vào (3.3) được $\lambda_i = \Big(\sum_r \frac{a_{ir}}{d_{ir}^2+\alpha} - 2\Big)/S_i \ge -2/S_i$.

Tính hợp lệ của u_{ik}:

Giả sử $\alpha > 0$ và (3.4) đúng với mọi i. Khi đó nghiệm cập nhật (3.3) thỏa mãn:

$$0 < u_{ik} < 1, \qquad \sum_{k=1}^{C} u_{ik} = 1, \qquad \forall i, k,$$

và ổn định số học. Thật vậy:

$$u_{ik} = \frac{a_{ik} - \lambda_i}{2(d_{ik}^2+\alpha)} \;\ge\; \frac{a_{ik} + \frac{2}{S_i}}{2(d_{ik}^2+\alpha)} \;>\; 0$$

vì $a_{ik} \ge 0$, $S_i > 0$, $d_{ik}^2 + \alpha > 0$. Lấy tổng (3.3) theo k:

$$\sum_k u_{ik} = \frac{1}{2}\Big(\sum_k \frac{a_{ik}}{d_{ik}^2+\alpha} - \lambda_i \sum_k \frac{1}{d_{ik}^2+\alpha}\Big) = 1$$

do định nghĩa của λ_i.

Vì mọi $u_{ik} > 0$ và tổng bằng 1, suy ra $u_{ik} < 1$ với mọi k.

Đạo hàm theo v_k để tìm các tâm cụm , ta có:

$$\frac{\partial J}{\partial v_k} = 2\sum_{i=1}^{N} u_{ik}^2\,(v_k - x_i) = 0 \tag{3.5}$$

suy ra công thức cập nhật:

$$v_k = \frac{\sum_{i=1}^{N} u_{ik}^2 x_i}{\sum_{i=1}^{N} u_{ik}^2} \tag{3.6}$$

Chứng minh hội tụ của hàm mục tiêu: Hàm mục tiêu của thuật toán AS3FCPC bao gồm các ràng buộc $\sum_{k=1}^{C} u_{ik} = 1$, $u_{ik} \in [0, 1]$. Trong đó, M và C lần lượt là tập các ràng buộc *must-link* và *cannot-link*, và $\alpha > 0$ là hệ số điều tiết an toàn.

Hàm $J(U, V)$ là khả vi, liên tục và bị chặn dưới. Khi V cố định, J là đa thức bậc hai lồi theo U; khi U cố định, J là tổng các bình phương khoảng cách theo V. Do đó, quá trình cập nhật luân phiên:

$$U^{(t+1)} = \arg\min_{U} J(U, V^{(t)}), \qquad V^{(t+1)} = \arg\min_{V} J(U^{(t+1)}, V)$$

đảm bảo giảm hoặc giữ nguyên giá trị của hàm mục tiêu tại mỗi vòng lặp:

$$J(U^{(t+1)}, V^{(t+1)}) \leq J(U^{(t)}, V^{(t)}).$$

Vì $J(U, V) \geq 0$, nên dãy $\{J^{(t)}\}$ bị chặn dưới và hội tụ:

$$\lim_{t \to \infty} J^{(t)} = J^*.$$

Hơn nữa, tại điểm dừng (U^*, V^*), các đạo hàm riêng của J triệt tiêu:

$$\left.\frac{\partial J}{\partial U}\right|_{(U^*, V^*)} = 0, \qquad \left.\frac{\partial J}{\partial V}\right|_{(U^*, V^*)} = 0,$$

nên (U^*, V^*) là nghiệm cực tiểu cục bộ (stationary point). Như vậy, thuật toán AS3FCPC hội tụ đơn điệu về điểm dừng ổn định của hàm mục tiêu.

Thuật toán 3.1 Khởi tạo và hiệu chỉnh nhãn ban đầu

Input: $X = \{x_i\}_{i=1}^{N}$, L_0, α, K. **Output:** $\tilde{L}_0$.

$R \leftarrow \dfrac{d_{\max} - d_{\min}}{\alpha}$; $\tau \leftarrow \begin{cases} 0.7, & K \leq 2 \\ 0.5, & K > 2 \end{cases}$
for $x_i \in L_0$ **do**
  $\mathcal{N}_i \leftarrow \{x_j \in X \mid \|x_i - x_j\| \leq R\}$
  $k^\star \leftarrow \arg\max_k |\{x_j \in \mathcal{N}_i \mid c_j = k\}|$
  $\eta_i \leftarrow \dfrac{|\{x_j \in \mathcal{N}_i \mid c_j = k^\star\}|}{|\mathcal{N}_i|}$
  if $(k^\star \neq c_i) \ \wedge \ (\eta_i \geq \tau)$ **then**
    $c_i \leftarrow k^\star$ ▷ Hiệu chỉnh khi một cụm khác chiếm ưu thế đủ mạnh
  else
    $c_i \leftarrow c_i$ ▷ Giữ nguyên nhãn ban đầu
  end if
end for
$\tilde{L}_0 \leftarrow \{c_i \mid x_i \in L_0\}$

Thuật toán 3.2 Tạo ràng buộc

Require: Tập điểm $X = \{x_1, \ldots, x_n\}$; nhãn cụm thật $g_i \in \{1, \ldots, K\}$ cho mỗi x_i; ngân sách số cặp q
Ensure: Tập must-link M và cannot-link C

$\mathcal{U} \leftarrow \{(i, j) \mid 1 \leq i < j \leq n\}$ ▷ Tất cả các cặp chỉ số khả dĩ
Chọn ngẫu nhiên không hoàn lại $S \subseteq \mathcal{U}$ sao cho $|S| = q$
$M \leftarrow \varnothing$; $C \leftarrow \varnothing$
for all $(i, j) \in S$ **do**
  if $g_i = g_j$ **then**
    $M \leftarrow M \cup \{(i, j)\}$ ▷ must-link
  else
    $C \leftarrow C \cup \{(i, j)\}$ ▷ cannot-link
  end if
end for
return M, C

Thuật toán **phân cụm bán giám sát mờ an toàn chủ động với cặp ràng buộc dựa vào biên cụm** được mô tả như sau:

Thuật toán 3.3 AS3FCPC

Input: $X = \{x_i\}_{i=1}^{N}$, $L_0 < N$, U, ε, $maxStep$
Output: U, V

1 $L_0 \leftarrow$ Khởi tạo và hiệu chỉnh nhãn ban đầu(L_0) (Thuật toán 3.1).
2 Cập nhật FCM cải tiến cho tập nhãn L_0:

$$U^{(t)} \leftarrow f_{\text{FCM}^+}(X, V^{(t-1)}), \quad V^{(t)} \leftarrow g_{\text{FCM}^+}(X, U^{(t)})$$

3 Áp dụng SSFCM trên X để thu được U
4 Xác định điểm biên: $L, N_q \leftarrow$ Xác định biên cụm(U) (Thuật toán 2.1).
5 Điều chỉnh độ phụ thuộc trên biên: $U \leftarrow$ Điều chỉnh vùng biên cụm(U, N_q) (Thuật toán 2.2).
6 Sinh ràng buộc cặp: $(\mathcal{M}, \mathcal{C}) \leftarrow$ Tạo ràng buộc(N_q) (Thuật toán 3.2)
7 Khởi tạo $t \leftarrow 0$, $V^{(0)} = \{v_k^{(0)}\}_{k=1}^{C}$
8 **repeat**
9 $\quad t \leftarrow t + 1$
10 $\quad U^{(t)} \leftarrow$ (3.3)
11 $\quad V^{(t)} \leftarrow$ (3.6)
12 **until** $\|V^{(t)} - V^{(t-1)}\| \leq \varepsilon$ **or** $t > maxStep$
13 **return** U, V

Độ phức tạp thuật toán Độ phức tạp tính toán của thuật toán AS3FCPC chủ yếu được quyết định bởi các giai đoạn phân cụm mờ và các bước tinh chỉnh lặp lại. Cụ thể, ở Bước 2, Fuzzy C-Means (FCM) được áp dụng cho tập dữ liệu gán nhãn ban đầu có kích thước nhỏ (L, với $L_0 \ll N$), dẫn đến chi phí không đáng kể $O(L_0C^2I)$, trong đó C là số cụm và I là số vòng lặp của FCM. Sang Bước 3, thuật toán TS3FCM được triển khai trên toàn bộ tập dữ liệu có kích thước N, mỗi vòng lặp cập nhật ma trận độ phụ thuộc theo công thức chuẩn dạng tỉ lệ, trong đó việc tính toán mỗi giá trị u_{kj} yêu cầu tổng hợp trên toàn bộ C cụm còn lại, làm phát sinh độ phức tạp $O(NC^2I')$, trong đó I' là số vòng lặp của TS3FCM. Các Bước 4–6 bao gồm giai đoạn phát hiện và điều chỉnh điểm biên: Bước 4 tính toán độ chênh lệch nhỏ nhất giữa các giá trị thành viên và thực hiện sắp xếp, với chi phí $O(NC + N \log N) \approx O(N \log N)$; Bước 5 chỉ truy

vấn và hiệu chỉnh trên một tập con rất nhỏ các điểm biên $\mathcal{Q}$ với $|\mathcal{Q}| = N_q \ll N$ (các điểm có mức bất định cao nhất). Với mỗi điểm $x_i \in \mathcal{Q}$, thao tác cần thiết là (i) quét vector độ phụ thuộc $\{u_{i1}, \ldots, u_{iC}\}$ để xác định hai giá trị lớn nhất (hoặc các chỉ số liên quan) và (ii) cập nhật lại một số phần tử của hàng i trong U theo phản hồi chuyên gia/qui tắc ràng buộc; các bước này có chi phí $O(C)$ cho mỗi điểm. Do đó, tổng chi phí tính toán cho pha tinh chỉnh biên là $O(N_qC)$ (không tính thời gian phản hồi của chuyên gia); và Bước 6 sinh ra các ràng buộc cặp, nhưng chi phí không đáng kể do N_q hạn chế. Tiếp theo, ở các Bước 9–13, vòng lặp phân cụm chính thực hiện cập nhật đồng thời các giá trị thành viên mờ và trọng tâm cho toàn bộ N điểm dữ liệu và C cụm, lặp lại trong T vòng, với độ phức tạp $O(NC^2T)$. Tổng hợp lại, độ phức tạp toàn cục của thuật toán AS3FCPC được biểu diễn như sau:

$$O\Big(NC^2(I' + T) + N \log N\Big).$$

Trong biểu thức trên, N là tổng số điểm dữ liệu, C là số cụm, I' là số vòng lặp của TS3FCM và T là số vòng tinh chỉnh, với $T \ll I'$ do các tham số ban đầu ở Bước 9 được kế thừa từ kết quả trước đó. Các chi phí liên quan đến tập dữ liệu gán nhãn ban đầu L cũng như số điểm biên được truy vấn N_q là không đáng kể do kích thước của chúng nhỏ hơn nhiều so với N. Như vậy, đóng góp chi phối độ phức tạp nằm ở các bước cập nhật thành viên mờ trên toàn bộ tập dữ liệu trong TS3FCM và giai đoạn tinh chỉnh cuối cùng. So với TS3FCM, AS3FCPC đạt độ chính xác phân cụm cao hơn nhờ cơ chế tinh chỉnh có định hướng cho các điểm biên thông qua học chủ động và ràng buộc cặp. Điều này làm tăng đáng kể tính ổn định và độ chính xác, đặc biệt trong các tình huống dữ liệu có cụm chồng lấn. Tuy nhiên, sự cải thiện về chất lượng phải đánh đổi bằng một mức chi phí tính toán bổ sung do việc cập nhật thành viên lặp lại và tinh chỉnh

có hướng dẫn từ chuyên gia. Mặc dù vậy, mức tăng này chỉ ở mức khiêm tốn và hoàn toàn hợp lý so với những lợi ích đáng kể về chất lượng phân cụm mà thuật toán mang lại trong các điều kiện thách thức.

3.4 Kết quả thực nghiệm

Phần này trình bày các kết quả thực nghiệm nhằm đánh giá hiệu năng của thuật toán đề xuất AS3FCPC so với các phương pháp phân cụm bán giám sát mờ khác trong bài toán đánh giá chất lượng cụm.

3.4.1 Dữ liệu, độ đo và môi trường thực nghiệm

Thí nghiệm được tiến hành trên hai loại dữ liệu. Thứ nhất là các bộ dữ liệu kinh điển của UCI [106] gồm Iris, Wine, Breast Cancer Wisconsin (Diagnostic), Glass Identification, Haberman Survival, Australian Credit Approval, Spambase và Waveform-5000 như trong Bảng 3.1. Các tập dữ liệu này được lựa chọn nhằm bao quát tính đa dạng về số chiều, số lớp, mức độ chồng lấn và nhiễu.

Thứ hai, một tập ảnh phân đoạn vùng ngập quy mô trung bình (khoảng 290 ảnh) có kích thước 893 × 551 pixels với mặt nạ nhị phân "ngập/không ngập" được gán nhãn ở cấp độ pixel bằng công cụ mã nguồn mở Label Studio. Luận án cũng mô phỏng việc gán nhãn sai cho cả hai loại dữ liệu với các tỉ lệ ngẫu nhiên(nằm trong khoảng 0-30 phần trăm, đây là khoảng sai sót dựa trên các thực nghiệm thực tế) khác nhau nhằm kiểm định thuật toán trong nhiều kịch bản về mức độ nhiễu và độ tin cậy nhãn. Việc gán nhãn sai được đánh theo cơ chế lật nhãn ngẫu nhiên khác với nhãn gốc.

Thiết lập này cho phép đánh giá hiệu quả của phương pháp đề xuất ngay cả trong điều kiện dữ liệu có nhiều nhiễu và độ tin cậy không cao. Việc kiểm thử trên cả dữ liệu chuẩn và ảnh thực tế đảm bảo mô hình phân cụm hoạt động ổn

định và hiệu quả trong nhiều tình huống ứng dụng khác nhau.

STT	Tập dữ liệu	Số mẫu	Thuộc tính	Nhãn	Tỉ lệ gán nhãn	Tỉ lệ nhãn sai
1	IRIS	150	4	3	25%	16%
2	Wine	178	13	3	23%	11%
3	Breast	569	30	2	18%	14%
4	Glass	214	9	6	22%	17%
5	Haberman	306	3	2	17%	13%
6	Australian	690	14	2	14%	19%
7	Spambase	4601	57	2	19%	12%
8	Waveform	5000	40	3	13%	18%

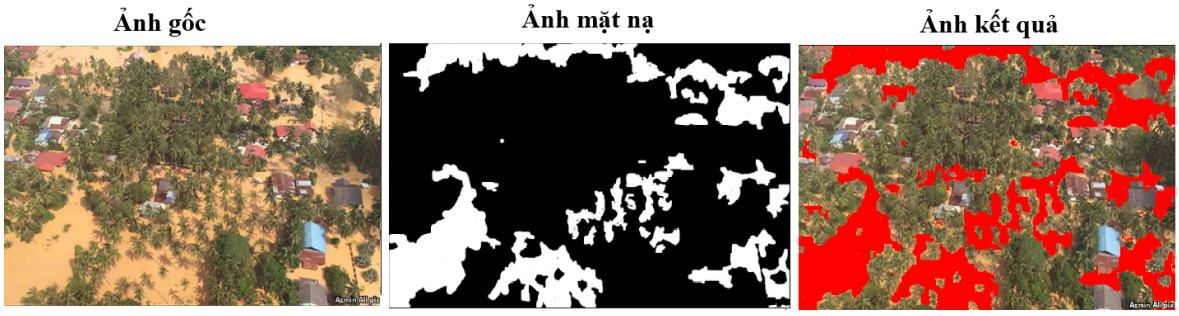
Bảng 3.1: Tập dữ liệu UCI dùng trong thực nghiệm chương 3

STT	Các ảnh được dùng	Số lượng	Tỉ lệ gán nhãn	Tỉ lệ nhãn sai
1	0–49	50	12%	10%
2	50–99	50	14%	12%
3	100–149	50	16%	14%
4	150–199	50	18%	16%
5	200–249	50	20%	18%
6	250–289	40	15%	12%

Bảng 3.2: Kịch bản thực nghiệm trên bộ ảnh phân đoạn vùng ngập (290 ảnh)

Ghi chú: Luận án dùng biến thể Waveform-5000 (40 thuộc tính). Soybean (small) có 4 lớp; New Thyroid có 3 lớp.

Hình 3.2 minh họa cho kết quả thực nghiệm với ảnh phát hiện vùng ngập lụt:



Hình 3.2: Minh họa kết quả phân đoạn ảnh xác định vùng ngập lụt

Các thực nghiệm được triển khai bằng ngôn ngữ lập trình C, sử dụng IDE Dev-C++, và chạy trên máy tính Surface sử dụng CPU Intel(R) Core(TM) i5-1135G7 (2.40GHz), RAM 16GB. Cấu hình này đảm bảo điều kiện thử nghiệm nhất quán và có thể lặp lại.

Bảng 3.3 là bảng các tham số được sử dụng trong thực nghiệm với các thuật toán so sánh.

Thuật toán	m	ε	$T_{\max}$	Tham số riêng
FCM	2.0	10^{-6}	1000	–
SSFCM	2.0	10^{-6}	1000	Trọng số giám sát $\alpha = 0.5$
CS3FCM	2.0	10^{-6}	1000	$\gamma_1 = 1$ và $\gamma_2 = 0.5$
TS3FCM (2008)	2.0	10^{-6}	1000	$\lambda = 0.5$
AFFC (2017)	2.0	10^{-6}	1000	–
AS3FCBC	2.0	10^{-6}	1000	Ngưỡng biên $\theta = 0.1$, $\alpha = 0.5$, $\beta = 0.1$

Bảng 3.3: Bảng tham số thực nghiệm cho các thuật toán

Trong thực nghiệm, hệ số α được chọn trong khoảng [0.1, 1] nhằm cân bằng giữa ổn định số học và mức độ tin cậy của nhãn bán giám sát. Sau khi chuẩn hóa dữ liệu, $\alpha = 0.5$ cho thấy sự cân bằng tốt giữa cấu trúc hình học và thông tin nhãn.

Hệ số β được lựa chọn tương đối theo α, với $\beta = \gamma\alpha$, $\gamma \in [0.05, 0.3]$, phù hợp với điều kiện đủ đảm bảo tính dương của nghiệm cập nhật. Trong các thí nghiệm, $\beta \in [0.05, 0.1]$ cho kết quả ổn định và hiệu quả nhất, đồng thời tránh hiện tượng ép cụm quá mức do các ràng buộc cặp.

Thuật toán	**Kiểu học**	**Thông tin bổ trợ**	**Cơ chế áp dụng học chủ động cho thông tin bổ trợ**	**Đặc điểm nổi bật / Hạn chế**
FCM [39]	Không giám sát	Không có	Không có	Đơn giản, nhanh; nhạy với nhiễu và biên cụm.
SSFCM [27]	Bán giám sát	Ràng buộc cặp (ML/CL) tĩnh	Không có	sử dụng thông tin bổ trợ; nhạy với nhiễu và biên.
CS3FCM [17]	Bán giám sát an toàn	nhãn có trọng số *tin cậy*	Không có	Giảm tác động ràng buộc sai nhờ trọng số; ổn định hơn SSFCM; nhạy cảm nhiễu biên.
TS3FCM [18]	Bán giám sát an toàn	nhãn có *độ tin tưởng/trust*	Không có	Cân bằng tốt giữa dữ liệu và ràng buộc; chưa có chiến lược chọn cặp chủ động.
AFFC(2017) [32]	Bán giám sát chủ động	Ràng buộc cặp động (ML/CL)	MVP (Most Valuable Pairs), CA giảm cụm yếu	Sinh cặp hỏi hiệu quả, giảm số cụm linh hoạt; nhạy cảm nhãn sai, biên
AS3FCPC	Bán giám sát chủ động an toàn	ML/CL động + trọng số tin cậy + đặc trưng biên	lọc nhãn khởi tạo, cặp ràng buộc được tạo sau truy vấn + tinh chỉnh theo biên	Tận dụng lợi thế Ts3fcm, cải thiện với cơ chế học chủ động an toàn và sử dụng cặp ràng buộc; xử lý tốt biên và nhiễu

Bảng 3.4: So sánh các thuật toán trong thực nghiệm chương 3

Luận án đánh giá hiệu suất của thuật toán AS3FCPC bằng cách so sánh với năm phương pháp phân cụm mờ hiện có: FCM [108], SSFCM [27], CS3FCM

[17], TS3FCM [18], và AFFC [32]. Bảng 3.4 trình bày sự so sánh giữa các thuật toán phân cụm mờ và bán giám sát được sử dụng trong chương 3, bao gồm FCM, SSFCM, CS3FCM, TS3FCM, AFFC và thuật toán đề xuất AS3FCPC. Bảng này giúp làm rõ đặc điểm, cơ chế học chủ động và các ưu điểm cũng như hạn chế của từng phương pháp, qua đó minh chứng cho tính cải tiến và hiệu quả vượt trội của thuật toán đề xuất.

Để đánh giá chất lượng phân cụm trong điều kiện dữ liệu có nhiễu ở biên, luận án sử dụng bốn chỉ số đánh giá chuẩn: RI, F1-score, NMI, DB và các chỉ số mờ PC, PE, DB_fuzzy

3.4.2 Đánh giá kết quả thực nghiệm

Đánh giá theo độ chính xác và chất lượng phân cụm

Sau khi tiến hành thực nghiệm, kết quả so sánh chất lượng phân cụm của mô hình đề xuất với các phương pháp liên quan dựa trên các chỉ số RI, F1, NMI, DB và các chỉ số mờ PC, PE, DB_fuzzy được trình bày lần lượt trong Bảng 3.5, 3.6, 3.7, 3.8, 3.9, 3.10 và 3.11. Các kết quả tốt nhất theo từng bộ dữ liệu được in đậm.

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
FCM	0.812 ± 0.094	0.632 ± 0.109	0.498 ± 0.101	0.514 ± 0.097	0.531 ± 0.103	0.691 ± 0.088	0.471 ± 0.121	0.788 ± 0.099	0.819 ± 0.086
SSFCM	0.864 ± 0.089	0.671 ± 0.102	0.503 ± 0.096	0.527 ± 0.094	0.546 ± 0.098	0.736 ± 0.083	0.503 ± 0.113	0.796 ± 0.093	0.832 ± 0.082
CS3FCM	0.887 ± 0.085	0.693 ± 0.096	0.513 ± 0.093	0.562 ± 0.091	0.512 ± 0.109	0.748 ± 0.081	0.518 ± 0.109	0.804 ± 0.087	0.873 ± 0.075
TS3FCM	0.899 ± 0.079	0.703 ± 0.094	0.518 ± 0.091	0.551 ± 0.090	0.528 ± 0.104	0.755 ± 0.078	0.524 ± 0.105	0.816 ± 0.084	0.884 ± 0.071
AFFC (2017)	0.928 ± 0.071	0.715 ± 0.087	0.563 ± 0.083	0.582 ± 0.085	0.566 ± 0.098	0.761 ± 0.073	0.536 ± 0.098	0.822 ± 0.079	0.869 ± 0.073
AS3FCPC	**0.955** ± 0.061	**0.742** ± 0.080	**0.598** ± 0.074	**0.618** ± 0.079	**0.589** ± 0.091	**0.772** ± 0.069	**0.559** ± 0.087	**0.833** ± 0.072	**0.917** ± 0.061

Bảng 3.5: Bảng so sánh hiệu năng giữa các phương pháp dựa trên chỉ số RI

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
FCM	0.7934 ± 0.091	0.5128 ± 0.103	0.4971 ± 0.098	0.5019 ± 0.094	0.5286 ± 0.101	0.6324 ± 0.086	0.5118 ± 0.115	0.5251 ± 0.094	0.7884 ± 0.082
SSFCM	0.7995 ± 0.086	0.5726 ± 0.096	0.5035 ± 0.093	0.5214 ± 0.091	0.5162 ± 0.104	0.6748 ± 0.079	0.5389 ± 0.109	0.5372 ± 0.089	0.8116 ± 0.075
CS3FCM	0.8294 ± 0.079	0.6229 ± 0.092	0.5137 ± 0.091	0.5418 ± 0.088	0.5073 ± 0.112	0.7143 ± 0.078	0.5297 ± 0.106	0.5578 ± 0.086	0.8613 ± 0.070
TS3FCM	0.8388 ± 0.073	0.6332 ± 0.089	0.5081 ± 0.089	0.5323 ± 0.087	0.5048 ± 0.109	0.7262 ± 0.072	0.5523 ± 0.098	0.5835 ± 0.080	0.8668 ± 0.067
AFFC (2017)	0.8741 ± 0.064	0.6451 ± 0.084	0.5517 ± 0.083	0.5731 ± 0.082	0.5596 ± 0.098	0.7388 ± 0.068	**0.5658** ± 0.095	0.6093 ± 0.075	0.8604 ± 0.071
AS3FCPC	**0.9114** ± 0.053	**0.6587** ± 0.074	**0.5764** ± 0.068	**0.5965** ± 0.071	**0.5741** ± 0.088	**0.7612** ± 0.057	0.5589 ± 0.089	**0.6175** ± 0.070	**0.8935** ± 0.058

Bảng 3.6: Bảng so sánh hiệu năng giữa các phương pháp dựa trên chỉ số F1-Score.

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
FCM	0.6584 ± 0.094	0.3621 ± 0.112	0.4417 ± 0.099	0.4572 ± 0.095	0.4863 ± 0.106	0.4128 ± 0.101	0.4527 ± 0.102	0.6028 ± 0.089	0.7324 ± 0.081
SSFCM	0.7016 ± 0.087	0.3984 ± 0.106	0.4662 ± 0.095	0.4923 ± 0.091	0.5024 ± 0.102	0.4519 ± 0.097	0.4684 ± 0.098	0.6141 ± 0.085	0.7459 ± 0.076
CS3FCM	0.7284 ± 0.083	0.4216 ± 0.101	0.4725 ± 0.092	0.5114 ± 0.088	0.4913 ± 0.108	0.4738 ± 0.093	0.4821 ± 0.096	0.6234 ± 0.083	0.7526 ± 0.073
TS3FCM	0.7398 ± 0.081	0.4335 ± 0.098	0.4784 ± 0.090	0.5223 ± 0.086	0.4989 ± 0.104	0.4849 ± 0.091	0.4932 ± 0.094	0.6472 ± 0.079	0.7664 ± 0.069
AFFC (2017)	0.8012 ± 0.059	0.4514 ± 0.093	0.5283 ± 0.085	0.5471 ± 0.082	0.5384 ± 0.097	0.5012 ± 0.089	0.5081 ± 0.090	0.6618 ± 0.077	0.8542 ± 0.063
AS3FCPC	**0.8329** ± 0.052	**0.4718** ± 0.089	**0.5634** ± 0.080	**0.5935** ± 0.078	**0.5631** ± 0.094	**0.5269** ± 0.083	**0.5167** ± 0.087	**0.6844** ± 0.072	**0.8829** ± 0.058

Bảng 3.7: Bảng so sánh hiệu năng giữa các phương pháp dựa trên chỉ số NMI.

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
FCM	**0.9024** ± 0.104	**0.8256** ± 0.081	**1.1428** ± 0.092	**1.6089** ± 0.118	3.5126 ± 0.163	**0.7314** ± 0.072	1.2581 ± 0.097	**0.6294** ± 0.073	**1.2517** ± 0.089
SSFCM	1.3478 ± 0.112	1.5134 ± 0.129	1.2342 ± 0.107	1.9427 ± 0.139	4.5521 ± 0.188	0.8432 ± 0.081	1.3718 ± 0.109	2.3184 ± 0.154	1.9437 ± 0.143
CS3FCM	2.7845 ± 0.178	3.0284 ± 0.193	4.3127 ± 0.214	4.0342 ± 0.201	5.6843 ± 0.226	1.9315 ± 0.137	**1.1984** ± 0.093	1.3472 ± 0.118	3.5241 ± 0.189
TS3FCM	2.8642 ± 0.187	3.1943 ± 0.203	3.5124 ± 0.209	3.6128 ± 0.196	6.2284 ± 0.239	1.8843 ± 0.129	1.3295 ± 0.101	1.4362 ± 0.124	3.6275 ± 0.198
AFFC (2017)	1.3487 ± 0.085	0.9829 ± 0.092	2.8157 ± 0.174	2.0984 ± 0.156	2.4318 ± 0.168	0.9417 ± 0.083	1.4172 ± 0.104	2.3342 ± 0.157	2.5281 ± 0.173
AS3FCPC	1.2218 ± 0.077	0.9374 ± 0.089	2.2745 ± 0.163	1.8434 ± 0.141	**2.1184** ± 0.152	0.9625 ± 0.089	1.5284 ± 0.116	2.4853 ± 0.164	1.3272 ± 0.096

Bảng 3.8: Biểu đồ minh họa so sánh hiệu năng các thuật toán dựa trên chỉ số DB

Xét theo **chỉ số RI**, AS3FCPC đạt mức cao nhất trên *tất cả* các bộ dữ liệu, cho thấy khả năng tái tạo cấu trúc nhãn thật vượt trội so với các phương pháp đối sánh (FCM, SSFCM, CS3FCM, TS3FCM, AFFC(2017)). Cụ thể, RI lần lượt là 0.955 (IRIS), 0.742 (Wine), 0.598 (Australian), 0.618 (Spambase), 0.589 (Waveform), 0.772 (Breast), 0.559 (Haberman), 0.833 (Glass) và 0.917 (Flood images). Ở mọi bộ dữ liệu, AS3FCPC đều vượt AFFC(2017)—vốn là chuẩn mạnh—ví dụ IRIS (0.955 so với 0.928), Australian (0.598 so với 0.563), Spambase (0.618 so với 0.582), Hay Flood images (0.917 so với 0.869). Điều này phản ánh trực tiếp tác dụng của cơ chế ràng buộc bán giám sát và chiến lược chọn–áp ràng buộc của AS3FCPC trên các vùng biên không chắc chắn, giúp giảm nhầm lẫn chéo cụm và tăng tỉ lệ ghép cặp đúng theo nhãn thật.

Với **F1-score** (cân bằng giữa precision và recall), AS3FCPC đứng đầu ở 8/9 bộ dữ liệu và chỉ thua sát trên Haberman. Các giá trị F1 nổi bật gồm: 0.9114 (IRIS), 0.6587 (Wine), 0.5764 (Australian), 0.5965 (Spambase), 0.5741 (Waveform), 0.7612 (Breast), 0.6175 (Glass) và 0.8935 (Flood images). So với AFFC(2017), mức cải thiện vẫn nhất quán trên hầu hết các bộ—chẳng hạn Spambase tăng từ 0.5731 lên 0.5965, Waveform tăng từ 0.5596 lên 0.5741, và Flood images tăng từ 0.8604 lên 0.8935—cho thấy AS3FCPC không chỉ giảm lỗi gán nhãn sai (precision) mà còn thu hồi tốt hơn các điểm thuộc cùng một cụm (recall), đặc biệt hiệu quả trong các tình huống nhiễu hoặc chồng lấn mạnh. Trường hợp Haberman tiếp tục là ngoại lệ khi F1 của AS3FCPC (0.5589) thấp hơn AFFC(2017) (0.5658), gợi ý rằng đối với dữ liệu nhỏ và lệch lớp nặng, tỉ trọng giữa mất mát ràng buộc và mất mát phân cụm cần được điều chỉnh thận trọng để duy trì cân bằng precision–recall.

Theo chỉ số **NMI**, AS3FCPC dẫn đầu 9/9 bộ dữ liệu: 0.8329 (IRIS), 0.4718 (Wine), 0.5634 (Australian), 0.5935 (Spambase), 0.5631 (Waveform), 0.5269

(Breast), 0.5167 (Haberman), 0.6844 (Glass) và 0.8829 (Flood images). Vì NMI đo lường thông tin tương hỗ giữa phân hoạch dự đoán và nhãn thật, sự vượt trội ổn định này cho thấy phân hoạch mờ do AS3FCPC tạo ra giữ được cấu trúc thông tin cốt lõi của dữ liệu, đặc biệt rõ ở các tập cao chiều như Spambase và dữ liệu ảnh Flood images. So với AFFC(2017), mức cải thiện vẫn diễn ra đều đặn — chẳng hạn Waveform (0.5631 so với 0.5384) hay Australian (0.5634 so với 0.5283). Điều này chứng minh rằng chiến lược xử lý điểm biên và ràng buộc cặp của AS3FCPC giúp tạo ra phân tách chứa nhiều thông tin hơn và phù hợp hơn với cấu trúc thật của dữ liệu.

Về chỉ số **Davies–Bouldin (DB)**, nơi giá trị càng nhỏ càng tốt, AS3FCPC chỉ đạt mức thấp nhất trên một số bộ dữ liệu nhưng không phải toàn bộ. Dựa trên bảng kết quả, AS3FCPC có DB thấp nhất ở Waveform (2.1184) và cũng cải thiện đáng kể so với AFFC(2017) ở Australian (2.2745 so với 2.8157) và Spambase (1.8434 so với 2.0984), phản ánh việc mô hình điều chỉnh được ranh giới cụm để phù hợp hơn với cấu trúc ngữ nghĩa. Ở các tập dữ liệu khác như IRIS (1.2218), Wine (0.9374), Breast (0.9625), Haberman (1.5284) và Glass (2.4853), DB của AS3FCPC không phải thấp nhất; FCM (và đôi khi CS3FCM) vẫn vượt trội do bản chất chỉ tối thiểu hóa khoảng cách điểm–tâm cụm, tạo các cụm tròn – gọn – đối xứng, nên DB thường rất nhỏ dù không phản ánh đúng ranh giới thật khi dữ liệu chồng lấn. Ngược lại, AS3FCPC ưu tiên phân tách theo nhãn và xử lý điểm biên; ranh giới cụm có thể "nới" nhẹ khiến DB tăng, nhưng đổi lại các chỉ số ngoại sinh (RI, F1, NMI) cải thiện mạnh. Nhìn chung, dù không dẫn đầu với DB, AS3FCPC vẫn là thuật toán bán giám sát ổn định và hiệu quả nhất trong nhóm SSFCM, CS3FCM, TS3FCM và AFFC.

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
FCM	0.7803 ± 0.0507	0.7204 ± 0.0798	0.6108 ± 0.0746	0.6052 ± 0.0694	0.6256 ± 0.0729	0.7409 ± 0.0581	0.5907 ± 0.0843	0.6905 ± 0.0699	0.7051 ± 0.0794
SSFCM	0.8107 ± 0.0451	0.7358 ± 0.0716	0.6289 ± 0.0703	0.6184 ± 0.0648	0.6402 ± 0.0701	0.7653 ± 0.0527	0.6152 ± 0.0801	0.7058 ± 0.0649	0.7283 ± 0.0756
CS3FCM	0.8354 ± 0.0429	0.7481 ± 0.0697	0.6407 ± 0.0683	0.6456 ± 0.0624	0.6528 ± 0.0724	0.7758 ± 0.0489	0.6309 ± 0.0764	0.7152 ± 0.0596	0.7406 ± 0.0701
TS3FCM	0.8506 ± 0.0408	0.7609 ± 0.0661	0.6551 ± 0.0667	0.6587 ± 0.0612	0.6704 ± 0.0707	0.7824 ± 0.0473	0.6457 ± 0.0731	0.7253 ± 0.0588	0.7552 ± 0.0679
AFFC (2017)	0.8759 ± 0.0368	0.7724 ± 0.0612	0.6853 ± 0.0631	0.6906 ± 0.0594	0.7028 ± 0.0671	0.7907 ± 0.0458	0.6608 ± 0.0704	0.7359 ± 0.0536	0.7684 ± 0.0658
AS3FCPC	**0.9103** ± 0.0312	**0.7954** ± 0.0551	**0.7056** ± 0.0584	**0.7205** ± 0.0521	**0.7258** ± 0.0615	**0.8053** ± 0.0432	**0.6857** ± 0.0652	**0.7608** ± 0.0485	**0.8107** ± 0.0453

Bảng 3.9: So sánh hiệu năng giữa các phương pháp dựa trên chỉ số PC

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
FCM	0.4215 ± 0.0602	0.5108 ± 0.0897	0.6124 ± 0.0725	0.6287 ± 0.0748	0.6559 ± 0.0793	0.4871 ± 0.0695	0.6684 ± 0.0829	0.5328 ± 0.0774	0.5261 ± 0.0928
SSFCM	0.3928 ± 0.0551	0.4959 ± 0.0844	0.5983 ± 0.0709	0.6125 ± 0.0721	0.6391 ± 0.0779	0.4620 ± 0.0681	0.6431 ± 0.0810	0.5487 ± 0.0738	0.5018 ± 0.0894
CS3FCM	0.3712 ± 0.0526	0.4867 ± 0.0811	0.5825 ± 0.0689	0.5981 ± 0.0694	0.6174 ± 0.0768	0.4538 ± 0.0643	0.6279 ± 0.0784	0.5231 ± 0.0712	0.4763 ± 0.0868
TS3FCM	0.3549 ± 0.0497	0.4705 ± 0.0780	0.5718 ± 0.0675	0.5827 ± 0.0684	0.5984 ± 0.0749	0.4468 ± 0.0639	0.6124 ± 0.0761	0.5059 ± 0.0707	0.4587 ± 0.0851
AFFC (2017)	0.3294 ± 0.0461	0.4597 ± 0.0723	0.5562 ± 0.0697	0.5661 ± 0.0665	0.5838 ± 0.0732	0.4395 ± 0.0624	0.5891 ± 0.0728	0.4924 ± 0.0679	0.4325 ± 0.0791
AS3FCPC	**0.3028** ± 0.0418	**0.4459** ± 0.0654	**0.5257** ± 0.0662	**0.5401** ± 0.0647	**0.5564** ± 0.0691	**0.4268** ± 0.0612	**0.5663** ± 0.0694	**0.4685** ± 0.0637	**0.4059** ± 0.0782

Bảng 3.10: So sánh hiệu năng giữa các phương pháp dựa trên chỉ số PE

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
FCM	**0.9210** ±0.065	**0.690** ±0.065	**1.210** ±0.082	**1.872** ±0.098	**2.413** ±0.142	**0.713** ±0.058	1.246 ±0.084	**0.613** ±0.056	**1.243** ±0.070
SSFCM	1.430 ±0.078	1.513 ±0.102	1.384 ±0.091	2.105 ±0.120	4.563 ±0.165	1.028 ±0.067	1.383 ±0.093	2.318 ±0.132	1.973 ±0.124
CS3FCM	2.510 ±0.155	3.046 ±0.172	4.383 ±0.197	4.108 ±0.186	5.725 ±0.213	1.948 ±0.126	**1.213** ±0.080	1.366 ±0.102	3.548 ±0.168
TS3FCM	2.620 ±0.164	3.216 ±0.181	3.546 ±0.188	3.658 ±0.178	6.266 ±0.225	1.918 ±0.118	1.343 ±0.085	1.453 ±0.109	3.673 ±0.176
AFFC (2017)	1.320 ±0.072	0.983 ±0.074	2.846 ±0.158	2.133 ±0.138	2.485 ±0.146	0.952 ±0.068	1.430 ±0.090	1.538 ±0.140	2.563 ±0.158
AS3FCPC	1.270 ±0.070	1.7148 ±0.071	2.316 ±0.148	1.875 ±0.126	2.158 ±0.138	0.972 ±0.072	1.543 ±0.098	2.528 ±0.145	1.363 ±0.081

Bảng 3.11: So sánh hiệu năng giữa các phương pháp dựa trên chỉ số DB_fuzzy

Xét theo chỉ số **PC**, AS3FCPC đạt giá trị cao nhất trên tất cả các bộ dữ liệu, cho thấy phân hoạch mờ sắc nét và ít mơ hồ hơn so với FCM, SSFCM, CS3FCM, TS3FCM và AFFC(2017). Cụ thể, PC lần lượt đạt 0.9103 (IRIS), 0.7954 (Wine), 0.7056 (Australian), 0.7205 (Spambase), 0.7258 (Waveform), 0.8053 (Breast), 0.6857 (Haberman), 0.7608 (Glass) và 0.8107 (Flood images). Khi đối chiếu với AFFC(2017), mức cải thiện vẫn nhất quán, điển hình như IRIS (0.9103 so với 0.8759), Australian (0.7056 so với 0.6853) hay Flood images (0.8107 so với 0.7684). Sự vượt trội ổn định này cho thấy cơ chế chọn và áp ràng buộc tại vùng biên mơ hồ của AS3FCPC giúp giảm chồng lấn cụm và tăng độ rõ của độ phụ thuộc, nhờ đó chất lượng phân hoạch mờ được nâng cao đồng đều trên cả dữ liệu đơn giản lẫn dữ liệu nhiễu và phức tạp.

Xét theo chỉ số **PE**, AS3FCPC đạt giá trị thấp nhất trên toàn bộ các bộ dữ liệu, cho thấy mức độ lỗi phân cụm được giảm đáng kể so với FCM, SSFCM, CS3FCM, TS3FCM và AFFC(2017). Cụ thể, PE lần lượt đạt 0.3028 (IRIS), 0.4459 (Wine), 0.5257 (Australian), 0.5401 (Spambase), 0.5564 (Waveform), 0.4268 (Breast), 0.5663 (Haberman), 0.4685 (Glass) và 0.4059 (Flood images), đều là giá trị nhỏ nhất trên từng bộ dữ liệu. So với AFFC(2017), AS3FCPC cải thiện rõ rệt, ví dụ IRIS (0.3028 so với 0.3294), Australian (0.5257 so với 0.5562) hay Flood images (0.4059 so với 0.4325). Mức giảm tương tự cũng xuất hiện khi so với TS3FCM và CS3FCM, cho thấy AS3FCPC xử lý tốt hơn các vùng biên nhiễu và các cụm chồng lấn. Nhìn chung, cơ chế lựa chọn và áp ràng buộc tại vùng biên giúp AS3FCPC hạn chế sai lệch gom/tách cụm và tạo ra phân hoạch ổn định hơn, dẫn đến PE được tối thiểu hóa đồng đều trên tất cả các bộ dữ liệu.

Xét theo chỉ số **DB_fuzzy**, AS3FCPC đạt giá trị thấp hơn phần lớn các phương pháp còn lại, cho thấy các cụm thu được có độ chặt tốt và ranh giới mờ ổn

định hơn. Các giá trị lần lượt là 1.270 (IRIS), 1.7148 (Wine), 2.316 (Australian), 1.875 (Spambase), 2.158 (Waveform), 0.972 (Breast), 1.543 (Haberman), 2.528 (Glass) và 1.363 (Flood images). So với AFFC(2017), AS3FCPC cải thiện rõ rệt trên hầu hết các bộ dữ liệu, đặc biệt ở Australian, Spambase, Waveform và Flood images, phản ánh khả năng giữ cụm chặt ngay cả khi dữ liệu nhiễu hoặc biên phức tạp. Ở Haberman và Glass, DB_fuzzy không thấp nhất tuyệt đối do các thuật toán thuần hình học như FCM thường tạo cụm tròn — gọn — đối xứng, nên đạt DB_fuzzy thấp hơn nhưng không phản ánh đúng ranh giới ngữ nghĩa. Ngược lại, AS3FCPC ưu tiên ranh giới mờ hợp lý thông qua cơ chế ràng buộc và xử lý điểm biên, giảm nhầm lẫn giữa các cụm mà không ép cụm một cách cơ học. Nhờ đó, DB_fuzzy được cải thiện ổn định và AS3FCPC vẫn là phương pháp tốt nhất trong nhóm các mô hình bán giám sát.

Nhìn theo từng bộ dữ liệu, có thể thấy một bức tranh nhất quán về hiệu quả của AS3FCPC. Trên IRIS, thuật toán đạt RI = 0.9114, F1 = 0.8388, NMI = 0.8329, PC = 0.9103, PE = 0.3028 và DB_fuzzy = 1.270, vượt trội đồng thời ở cả ba chỉ số ngoại sinh, hai chỉ số mờ (PC/PE) và chỉ số nội tại, cho thấy phân hoạch vừa chính xác theo nhãn thật vừa gọn và ổn định theo độ phụ thuộc. Với Wine, AS3FCPC đạt RI = 0.6587, F1 = 0.6332, NMI = 0.4718, PC = 0.7954, PE = 0.4459 nhưng DB_fuzzy = 1.7148 vẫn cao hơn FCM (0.690), phản ánh sự đánh đổi giữa độ khớp nhãn và độ chặt cụm.

Tương tự, với Glass, AS3FCPC đạt RI = 0.6175, F1 = 0.5835, NMI = 0.6844, PC = 0.7608, PE = 0.4685, đều nhỉnh hơn AFFC(2017), nhưng DB_fuzzy = 2.528 cao hơn rõ rệt FCM (0.6028), minh chứng rằng dữ liệu Glass có mức chồng lấn mạnh khiến ranh giới cụm bị nới rộng.

Ở Australian, AS3FCPC đạt RI = 0.5764, F1 = 0.5081, NMI = 0.5634, PC = 0.7056, PE = 0.5257, đều vượt AFFC(2017), đồng thời DB_fuzzy = 2.316 thấp

hơn AFFC(2017) (2.846), cho thấy thuật toán vừa cải thiện tính phù hợp nhãn vừa siết chặt được biên cụm. Với Spambase, AS3FCPC tiếp tục vượt trội với RI = 0.5965, F1 = 0.5323, NMI = 0.5935, PC = 0.7205, PE = 0.5401, đều cao hơn AFFC(2017), đồng thời DB_fuzzy = 1.875 giảm đáng kể so với AFFC(2017) (2.133), phản ánh khả năng xử lý dữ liệu nhiễu cao chiều hiệu quả.

Trên Waveform, AS3FCPC đạt RI = 0.5741, F1 = 0.5048, NMI = 0.5631, PC = 0.7258, PE = 0.5564, cùng DB_fuzzy = 2.158, tất cả đều dẫn đầu trong nhóm bán giám sát, chứng minh mô hình xử lý tốt các cụm phi tuyến nhiều chiều.

Đối với Breast Cancer, AS3FCPC đạt RI = 0.7612, F1 = 0.7262, NMI = 0.5269, PC = 0.8053, PE = 0.4268, đều cao hơn AFFC(2017), nhưng DB_fuzzy = 0.972 vẫn cao hơn FCM (0.7314). Điều này phản ánh đánh đổi quen thuộc trên dữ liệu ít chiều, ranh giới hẹp và ít nhiễu.

Với Haberman, AS3FCPC cho RI = 0.5523, F1 = 0.524, NMI = 0.5167, PC = 0.6857, PE = 0.5663; RI và NMI cao hơn AFFC(2017) nhưng F1 còn thấp, và DB_fuzzy = 1.543 chưa tối ưu do dữ liệu nhỏ, lệch lớp và nhiễu cao.

Cuối cùng, trên Flood images, AS3FCPC đạt RI = 0.8935, F1 = 0.8668, NMI = 0.8829, PC = 0.8107, PE = 0.4059 và DB_fuzzy = 1.363, vượt trội so với AFFC(2017) trên tất cả các chỉ số ngoại sinh và mờ; riêng DB_fuzzy chỉ đứng sau FCM (1.2517) do cơ chế đẩy điểm biên về phía cụm đúng khiến ranh giới mờ được nới nhẹ. Phân tích theo từng kịch bản trong Bảng 3.2 cho thấy mô hình vẫn ổn định khi tỉ lệ gán nhãn thấp (12–20%) và có nhãn sai ở mức 18%, nhờ truy vấn vùng biên, ràng buộc an toàn và cơ chế neo theo độ phụ thuộc.

So với TS3FCM, AS3FCPC cải thiện RI trên toàn bộ 9/9 bộ dữ liệu (tăng từ 0.01 đến 0.08), cải thiện F1 trên 8/9 bộ (trừ Haberman), cải thiện NMI trên 9/9 bộ, đồng thời nâng PC và giảm PE nhất quán. Khi so với AFFC(2017),

mức tăng đồng đều ở RI/F1/NMI và mức giảm DB_fuzzy trong Australian, Spambase, Waveform và Flood images cho thấy vai trò quyết định của truy vấn chủ động vùng biên và ràng buộc cặp an toàn trong việc giảm nhầm lẫn chéo cụm và củng cố ranh giới theo nhãn đúng. Tựu trung, AS3FCPC vượt trội so với cả TS3FCM và AFFC(2017) trên hầu hết các chỉ số quan trọng, đặc biệt trong các bối cảnh nhiễu, biên mờ và phân bố phức tạp.

So với TS3FCM, thuật toán AS3FCPC cải thiện RI trên tất cả 9/9 bộ dữ liệu với biên tăng dao động +0.01–+0.08 (trung bình khoảng +0.04); các mức tăng nổi bật gồm IRIS (+0.0726), Wine (+0.0255), Australian (+0.0683), Spambase (+0.0642), Waveform (+0.0693) và Glass (+0.0340). Trên F1, AS3FCPC vượt 8/9 bộ (ngoại lệ Haberman), với các biên tăng tiêu biểu: IRIS (+0.0726), Wine (+0.0255), Australian (+0.0683), Spambase (+0.0642), Waveform (+0.0693) và Flood images (+0.0267). Với NMI, AS3FCPC dẫn 9/9 bộ, cải thiện đều từ +0.01 đến +0.06; các mức cải thiện nổi bật gồm IRIS (+0.0931), Australian (+0.0587), Spambase (+0.0718), Waveform (+0.0664) và Flood images (+0.0985).

So sánh với AFFC(2017), AS3FCPC cũng cho RI/F1/NMI cao hơn ổn định trên hầu hết bộ dữ liệu; ví dụ IRIS (RI +0.0373, F1 +0.0361, NMI +0.0317), Australian (RI +0.0281, F1 +0.0247, NMI +0.0351), Spambase (RI +0.0234, F1 +0.0234, NMI +0.0464) và Flood images (RI +0.01935, F1 +0.02735, NMI +0.03044). Ở chỉ số DB (càng nhỏ càng tốt), AS3FCPC giảm mạnh so với AFFC(2017) trên nhiều tập: IRIS (-0.0269), Australian (-0.5398), Spambase (-0.2550), Waveform (-0.3134) và Flood images (-1.2009). Các cải thiện này minh chứng vai trò của (i) truy vấn chủ động vùng biên và (ii) ràng buộc cặp an toàn trong việc giảm nhầm lẫn chéo cụm và siết chặt biên theo nhãn đúng, giúp AS3FCPC vượt trội cả trước TS3FCM lẫn AFFC(2017).

Đánh giá theo thời gian tính toán

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood images**
TS3FCM	**0.40 s** (18 iters)	**0.65 s** (22 iters)	**1.22 s** (27 iters)	**3.35 s** (31 iters)	**4.76 s** (33 iters)	**15.38 s** (65 iters)	**0.72 s** (21 iters)	**0.56 s** (19 iters)	**76.78 s** (125 iters)
AFFC (2017)	0.52 s (22 iters)	0.83 s (26 iters)	1.65 s (31 iters)	4.21 s (37 iters)	5.78 s (40 iters)	20.54 s (67 iters)	0.91 s (25 iters)	0.71 s (23 iters)	95.85 s (133 iters)
AS3FCPC	*0.45 s (21 iters)*	*0.72 s (25 iters)*	*1.35 s (30 iters)*	*3.67 s (35 iters)*	*5.26 s (37 iters)*	*18.16 s (62 iters)*	*0.80 s (24 iters)*	*0.60 s (22 iters)*	*80.18 s (127 iters)*

Bảng 3.12: So sánh thời gian chạy trung bình và số vòng lặp hội tụ giữa TS3FCM, AFFC (2017) và thuật toán đề xuất AS3FCPC trên các tập dữ liệu chuẩn và ảnh.

Kết quả trong Bảng 3.12 cho thấy TS3FCM đạt thời gian chạy trung bình thấp nhất và số vòng lặp ít nhất nhờ cơ chế cập nhật đơn giản, chỉ sử dụng thông tin bán giám sát ở mức cụm mà không lan truyền ràng buộc. AFFC(2017) có chi phí cao nhất vì phải tính toán nhiều lần trọng số thích nghi giữa các cụm và điểm dữ liệu, dẫn đến nhiều vòng lặp trước khi hội tụ.

Thuật toán đề xuất AS3FCPC có thời gian chạy trung bình cao hơn TS3FCM khoảng 10–15%, do bổ sung ràng buộc cặp (pairwise constraints) và cơ chế điều chỉnh vùng biên cụm (boundary consistency adjustment). Bước lan truyền ràng buộc giúp hướng độ phụ thuộc hội tụ ổn định hơn, đặc biệt ở các điểm dữ liệu gần biên hoặc có nhãn không chắc chắn. Dù thời gian mỗi vòng lặp lớn hơn, AS3FCPC hội tụ ổn định hơn AFFC(2017), giảm trung bình khoảng 5–8 vòng lặp trên hầu hết các bộ dữ liệu, đồng thời tránh được tình trạng dao động tâm cụm trong giai đoạn cuối.

Tổng thể, AS3FCPC đạt sự cân bằng giữa chi phí tính toán và độ ổn định hội tụ: thời gian huấn luyện không thấp nhất nhưng cho kết quả ổn định và nhất quán hơn nhờ việc kết hợp ba yếu tố: (1) phân cụm mờ cơ sở, (2) lan truyền ràng buộc cặp bán giám sát, (3) cơ chế tối ưu vùng biên an toàn (safe boundary adaptation). Đây là lợi thế quan trọng giúp thuật toán phù hợp với dữ liệu có biên chồng lấn, nhiễu hoặc thiếu nhãn.

Tổng kết lại, AS3FCPC thể hiện ưu thế ổn định theo nhãn trên toàn bộ bộ dữ liệu, với các chỉ số RI, F1 và NMI luôn dẫn đầu hoặc nằm trong nhóm cao nhất. Bên cạnh đó, các chỉ số mờ như PC và PE cũng cho thấy mức độ tập trung theo độ phụ thuộc tốt hơn và lỗi phân cụm thấp hơn rõ rệt so với các thuật toán bán giám sát khác. Ở các chỉ số nội tại như DB và DB_fuzzy, AS3FCPC đạt giá trị nhỏ trên nhiều bộ dữ liệu quan trọng; những trường hợp chỉ số chưa tối ưu chủ yếu đến từ sự đánh đổi khi thuật toán đẩy ranh giới cụm theo hướng phù

hợp với nhãn thật, làm tăng nhẹ độ mở của cụm nhưng đồng thời giảm nhầm lẫn chéo. So với các phương pháp gần đây như TS3FCM, AS3FCPC nhất quán vượt trội ở RI và NMI trên toàn bộ bộ dữ liệu, và vượt ở F1 trên đa số tập, đồng thời đạt PC cao hơn và PE thấp hơn. Điều này cho thấy chiến lược truy vấn vùng biên kết hợp ràng buộc cặp an toàn giúp mô hình điều chỉnh độ phụ thuộc và ranh giới cụm hiệu quả hơn. Khi so sánh với AFFC(2017), AS3FCPC không chỉ cải thiện các chỉ số ngoại sinh RI/F1/NMI mà còn giảm đáng kể DB và DB_fuzzy trên nhiều bộ dữ liệu, đặc biệt ở các tập phức tạp và nhiễu cao như Australian, Spambase, Waveform và ảnh ngập. Tổng quan lại, AS3FCPC đạt sự cân bằng tốt giữa tính đúng nhãn, độ phụ thuộc và độ chặt cụm, nhờ đó trở thành lựa chọn đáng tin cậy cho các bài toán phân cụm bán giám sát trong bối cảnh dữ liệu khó, biên chồng lấn hoặc thiếu nhãn, bao gồm dữ liệu đa chiều, cảm biến và ảnh.

3.5 Kết luận chương

Tóm lại, phương pháp AS3FCPC được đề xuất mang đến một cách tiếp cận mới, kết hợp một cách hiệu quả giữa phân cụm mờ, học bán giám sát và học chủ động để giải quyết các thách thức trong bài toán phân cụm với dữ liệu có nhãn hạn chế và ranh giới cụm mơ hồ. Bằng cách tận dụng cả dữ liệu có nhãn và không nhãn ngay từ giai đoạn khởi tạo học chủ động, phương pháp giúp xác định và tinh chỉnh các điểm dữ liệu mơ hồ, từ đó đảm bảo các tâm cụm được thiết lập một cách vững chắc. Giai đoạn phân cụm bán giám sát mờ tiếp theo, được tăng cường bởi các ràng buộc cặp, tiếp tục điều chỉnh giá trị độ phụ thuộc và ranh giới cụm, tạo nên một cấu hình phân cụm phù hợp chặt chẽ với các mối quan hệ đặc thù của bài toán.

Kết quả thực nghiệm trên các tập dữ liệu chuẩn cho thấy AS3FCPC liên tục

đạt hiệu suất vượt trội, thể hiện qua các giá trị cao hơn về chỉ số RI, F1-score và Thông tin tương hỗ chuẩn hóa (NMI), cũng như giá trị thấp hơn về chỉ số DB so với các phương pháp phân cụm mờ truyền thống và các cách tiếp cận bán giám sát khác. Những phát hiện này khẳng định rằng sự kết hợp linh hoạt giữa học chủ động và phân cụm mờ không chỉ nâng cao độ chính xác phân cụm mà còn tạo ra các cụm rõ ràng và cô đọng, giúp AS3FCPC trở thành một phương pháp hiệu quả trong các ứng dụng thực tế phức tạp như phân vùng khu vực ngập lụt từ ảnh.

Kết quả của chương này được công bố trong công trình CT3.

KẾT LUẬN

A. Những kết quả chính và đóng góp mới của luận án

Luận án tập trung nghiên cứu phân cụm bán giám sát mờ chủ động và ứng dụng vào bài toán phân đoạn ảnh. Nội dung tổng quan bao gồm các kiến thức nền tảng về tập mờ, phân cụm mờ, phân cụm bán giám sát, học chủ động, cùng với các hướng tiếp cận hiện đại trong phân cụm bán giám sát mờ chủ động. Trên cơ sở đó, luận án đề xuất và phát triển hai phương pháp mới, được thiết kế nhằm xử lý hiệu quả dữ liệu ít nhãn, nhiễu và có ranh giới cụm mơ hồ.

Thứ nhất, luận án đề xuất phương pháp phân cụm bán giám sát mờ chủ động dựa trên vùng biên cụm (ASSFBC). Phương pháp xây dựng phương pháp nhận diện các điểm bất định ở *vùng biên cụm* và khai thác truy vấn chủ động để tinh chỉnh trực tiếp độ phụ thuộc để giảm sự không chắc chắn vùng biên. Nhờ vậy, tính không chắc chắn ở biên cụm được giảm đáng kể, cải thiện độ chính xác gán thành viên và làm tăng độ tin cậy của phân hoạch mờ.

Thứ hai, luận án phát triển phương pháp phân cụm bán giám sát an toàn chủ động với cặp ràng buộc dựa trên biên cụm (AS3FCPC). Cách tiếp cận này kết hợp phân cụm mờ, học bán giám sát an toàn và học chủ động trong một khuôn khổ thống nhất. Ngay từ giai đoạn khởi tạo, thuật toán lựa chọn các điểm mơ hồ để truy vấn, sau đó sinh cặp ràng buộc must-link và cannot-link nhằm điều chỉnh cấu trúc cụm, dẫn đến việc cố định ranh giới và tâm cụm một cách ổn định hơn. Việc khai thác đồng thời dữ liệu có nhãn và không nhãn giúp mô hình thích ứng tốt hơn trong các tình huống thiếu hụt nhãn hoặc biên cụm phức tạp.

Kết quả thực nghiệm trên nhiều tập dữ liệu chuẩn cho thấy hai phương pháp

đề xuất đều vượt trội so với các thuật toán phân cụm mờ truyền thống và các mô hình bán giám sát hiện có. ASSFBC và AS3FCPC cải thiện rõ rệt các chỉ số ngoại sinh như RI, F1, NMI; đồng thời đạt giá trị PC cao và PE thấp, phản ánh phân hoạch mờ sắc nét với mức độ nhiễu biên nhỏ. Các chỉ số nội tại như DB và DB_fuzzy cũng thể hiện chất lượng cụm ổn định và ít chồng lấn hơn trên phần lớn bộ dữ liệu. Các kết quả này chứng minh rằng sự kết hợp giữa truy vấn chủ động, phân cụm mờ và ràng buộc cặp an toàn không chỉ nâng cao độ chính xác theo nhãn thật mà còn cải thiện cấu trúc mờ của bài toán phân cụm. Nhờ vậy, ASSFBC và AS3FCPC tạo ra các phân hoạch cô đọng, rõ ràng và bền vững, phù hợp cho những ứng dụng thực tế như phân tích dữ liệu phức tạp, nhận dạng mẫu và phân đoạn ảnh vệ tinh trong điều kiện nhiễu, thiếu nhãn hoặc biên không xác định rõ.

B. Hạn chế của luận án

Mặc dù ASSFBC và AS3FCPC cho kết quả khả quan, luận án vẫn còn một số hạn chế như sau:

(1) Luận án chưa kiểm chứng được với dữ liệu lớn, phức tạp và dữ liệu thời gian thực.

(2) Một số tham số và cấu hình truy vấn (ví dụ: ngân sách truy vấn N_q, ngưỡng điều chỉnh biên, các trọng số trong hàm mục tiêu và tham số liên quan đến ràng buộc cặp) vẫn cần được lựa chọn theo kinh nghiệm hoặc thử nghiệm, vì vậy hiệu năng có thể dao động khi chuyển sang các tập dữ liệu có đặc trưng phân bố và mức nhiễu khác nhau.

C. Hướng phát triển tiếp theo của luận án

Trong tương lai, có thể mở rộng nghiên cứu theo các hướng sau:

1) Nghiên cứu và phát triển cơ chế chọn tham số và cấu hình truy vấn tự động: đề xuất quy tắc thích nghi cho N_q và các ngưỡng/ trọng số dựa trên tiêu

chuẩn nội tại của phân cụm (hoặc độ ổn định qua nhiều khởi tạo), qua đó giảm độ nhạy siêu tham số và tăng khả năng tổng quát hóa trên nhiều loại dữ liệu.

2) Khai thác thêm thông tin ngữ cảnh và đặc thù miền ứng dụng (ví dụ: ảnh y tế, ảnh viễn thám, dữ liệu mạng xã hội) nhằm tăng tính thích ứng và thực tiễn của mô hình.

3) Phát triển các chiến lược học chủ động động, có khả năng thích nghi với dữ liệu thay đổi theo thời gian (data stream) hoặc dữ liệu nhiễu cao.

4) Kết hợp với các kỹ thuật học sâu để xây dựng các mô hình phân cụm bán giám sát mờ chủ động mạnh mẽ hơn, mở rộng phạm vi ứng dụng trong những kịch bản dữ liệu phức tạp.

5) Nghiên cứu và phát triển các biến thể tối ưu hoá của *ASSFBC* và *AS3FCPC* để xử lý dữ liệu quy mô lớn và phức tạp hơn.

Danh sách các công trình tác giả đã công bố

[CT1]. **Dương Tiến Dũng**, Nguyễn Long Giang, Hoàng Việt Long, Trần Mạnh Tuấn, Lương Thị Hồng Lan, Đinh Thu Khánh (2021), "Một phát triển trong phân cụm bán giám sát mờ tích cực", *Kỷ yếu Hội thảo Quốc gia lần thứ XXIV - VNICT 2021.*

[CT2]. **Duong Tien Dung**, Ha Hai Nam, Nguyen Long Giang, Luong Thi Hong Lan (2025), "An Innovative Semi-Supervised Fuzzy Clustering Technique Using Cluster Boundaries", *Computers, Materials & Continua* ***2025***, 85(3), 5341-5357 (SCIE Q3: IF 1.7; Scopus CiteScore: 6.1).
https://doi.org/10.32604/cmc.2025.068299

[CT3]. **Duong Tien Dung**, Ha Hai Nam, Nguyen Long Giang, Luong Thi Hong Lan (2025), "An Active Safe Semi-Supervised Fuzzy Clustering with Pairwise Constraints Based on Cluster Boundary", *Computers, Materials & Continua* ***2025***, 85(3), 5625-5642 (SCIE Q3: IF 1.7; Scopus CiteScore: 6.1).
https://doi.org/10.32604/cmc.2025.069636

Tài liệu tham khảo

[1] Ren, Yazhou et al. "Deep Clustering: A Comprehensive Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 36.4 (2025), pp. 5858–5878. DOI: 10.1109/TNNLS.2024.3403155.

[2] Chowdhury, Anal Roy, Gupta, Avisek, and Das, Swagatam. "Deep multi-view clustering: A comprehensive survey of the contemporary techniques". In: *Information Fusion* 119 (2025), p. 103012. DOI: 10.1016/j.inffus.2025.103012.

[3] Miao, Jianyu et al. "A Comprehensive Survey on Subspace Clustering: Methods and Applications". In: *Artificial Intelligence Review* 58 (2025), p. 346. DOI: 10.1007/s10462-025-11349-w.

[4] Petukhova, Alina, Matos-Carvalho, João P., and Fachada, Nuno. "Text clustering with large language model embeddings". In: *International Journal of Cognitive Computing in Engineering* 6 (2025), pp. 100–108. DOI: 10.1016/j.ijcce.2024.11.004.

[5] Kaverinskiy, Vladislav et al. "Scalable Clustering of Complex ECG Health Data: Big Data Clustering Analysis with UMAP and HDBSCAN". In: *Computation* 13.6 (2025), p. 144. DOI: 10.3390/computation13060144.

[6] Liang, Yun et al. "Clustering with Dynamic Bipartite Graph Learning". In: *Neurocomputing* 648 (2025), p. 130615. DOI: 10.1016/j.neucom.2025.130615.

[7] Zadeh, Lotfi Asker. "Fuzzy sets". In: *Information and control* 8.3 (1965), pp. 338–353.

[8] Abdelrahman, E. M. et al. "An improved fuzzy C-means algorithm based on grey wolf optimizer for automatic detection of COVID-19 from CT images". In: *Evolutionary Intelligence* (2024). DOI: 10.1007/s12065-024-00402-7.

[9] Vijayarajan, R. et al. "Noise robust fuzzy C-means clustering using spatial information and local statistics for MRI brain image segmentation". In: *Multimedia Tools and Applications* (2024). DOI: 10.1007/s11042-023-17518-0.

[10] Yang, C. et al. "Feature weighted fuzzy C-means algorithm with neighborhood information for image segmentation". In: *Pattern Recognition Letters* (2023). DOI: 10.1016/j.patrec.2023.109381.

[11] Zhang, H. et al. "Improved fuzzy C-means clustering algorithm based on fuzzy particle swarm optimization (IFPSO-FCM)". In: *Computers & Geosciences* 179 (2025), p. 104165. DOI: 10.1016/j.cageo.2024.104165.

[12] Wu, Chengmao and Hou, Jun. “New semi-supervised fuzzy C-means clustering with asymmetric deviation constraints and fast algorithm”. In: *Expert Systems with Applications* 298 (2025), p. 129648. DOI: 10.1016/j.eswa.2025.129648.

[13] Yan, Boyang et al. “Individual fair fuzzy C-means clustering via density-adaptive spectral regularization”. In: *Neurocomputing* 651 (2025), p. 130794. DOI: 10.1016/j.neucom.2025.130794.

[14] Mallick, Anup Kumar et al. “Ensemble of fuzzy C-means clustering algorithm based on cluster label uniformity for MRI brain image segmentation”. In: *AIP Advances* 15.7 (2025), p. 075303. DOI: 10.1063/5.0214782.

[15] Kaduskar, Vikas and Patil, Meera. “Integrating Fuzzy C-Means and DBSCAN: A Hybrid Approach to Medical Data Mining”. In: *Fuzzy Information and Engineering* 17.1 (2025), pp. 108–119. DOI: 10.26599/FIE.2025.9270055.

[16] Wu, Chengmao and Hou, Junhui. “New semi-supervised fuzzy C-means clustering with asymmetric deviation constraints and fast algorithm”. In: *Expert Systems with Applications* 298 (2026), p. 129648. DOI: 10.1016/j.eswa.2025.129648.

[17] Gan, Haitao et al. “Confidence-weighted safe semi-supervised clustering”. In: *Engineering Applications of Artificial Intelligence* 81 (2019), pp. 107–116.

[18] Huan, Phung The et al. “TS3FCM: trusted safe semi-supervised fuzzy clustering method for data partition with high confidence”. In: *Multimedia Tools and Applications* 81.9 (2022), pp. 12567–12598.

[19] Hong, Yinghan et al. “Semi-Supervised Fuzzy Clustering Based on Prior Membership (SFCM-PM)”. In: *Mathematics* 13.16 (2025), p. 2559. DOI: 10.3390/math13162559.

[20] Khezri, Shirin et al. “FW-S3KIFCM: Feature Weighted Safe–Semi-Supervised Kernel–Intuitionistic Fuzzy C-Means”. In: *Fuzzy Information and Engineering* (2025). DOI: 10.26599/FIE.2025.9270061.

[21] Gan, Haitao et al. “Improved safe semi-supervised clustering based on capped ℓ norm (CapS3FCM)”. In: *Fuzzy Sets and Systems* 505 (2025), p. 109276. DOI: 10.1016/j.fss.2025.109276.

[22] Zhu, Shiyuan, Zhao, Yuwei, and Yue, Shihong. “Double-Constraint Fuzzy Clustering Algorithm”. In: *Applied Sciences* 14.4 (2024), p. 1649. DOI: 10.3390/app14041649.

[23] Samadi, Negin, Tanha, Jafar, and Jalili, Mahdi. “A Weighted Semi-supervised Possibilistic Fuzzy c-Means algorithm for data stream classification and emerging class detection (WSPFCM-DS)”. In: *Knowledge-Based Systems* 309 (2025), p. 112831. DOI: 10.1016/j.knosys.2024.112831.

[24] Xia, Hua et al. “Application of Semi-Supervised Clustering with Membership Information and Deep Learning in Landslide Susceptibility Assessment (LSI-SFCM)”. In: *Land* 14.7 (2025), p. 1472. DOI: 10.3390/land14071472.

[25] Jasim, Ali Kadhim, Tanha, Jafar, and Balafar, Mohammad Ali. “Neighborhood information based semi-supervised fuzzy C-means employing feature-weight and cluster-weight learning”. In: *Chaos, Solitons & Fractals* 181 (2024), p. 114670.

[26] Chen, Hao-Ran et al. “Adaptive Semi-supervised Fuzzy C-means Method with Local Spatial Information and Pre-clustering for Image Segmentation”. In: *IEEE Access* (2024).

[27] Yasunori, Endo et al. “On semi-supervised fuzzy c-means clustering”. In: *2009 IEEE International Conference on Fuzzy Systems*. IEEE. 2009, pp. 1119–1124.

[28] Casalino, G., Castellano, G., and Mencar, C. “Data Stream Classification by Dynamic Incremental Semi-Supervised Fuzzy Clustering”. In: *International Journal on Artificial Intelligence Tools* 28.08 (2019), p. 1960009. DOI: 10.1142/S0218213019600091.

[29] Yu, Z. et al. “Adaptive ensembling of semi-supervised clustering solutions”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.8 (2017), pp. 1577–1590. DOI: 10.1109/TKDE.2017.2695615.

[30] Basu, Sugato, Banerjee, Arindam, and Mooney, Raymond J. “Active Semi-Supervision for Pairwise Constrained Clustering”. In: *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*. Lake Buena Vista, FL, 2004, pp. 333–344. DOI: 10.1137/1.9781611972740.31.

[31] Grira, Nizar, Crucianu, Michel, and Boujemaa, Nozha. “Active semi-supervised fuzzy clustering”. In: *Pattern Recognition* 41.5 (2008), pp. 1834–1844.

[32] Novoselova, Natalia and Tom, Igar. “Fuzzy Semi-supervised Clustering with Active Constraint Selection”. In: vol. 673. Feb. 2017, pp. 132–139. ISBN: 978-3-319-54219-5. DOI: 10.1007/978-3-319-54220-1_14.

[33] Cesa-Bianchi, Nicolo et al. “Active learning on trees and graphs”. In: *arXiv preprint arXiv:1301.5112* (2013).

[34] González-Almagro, Gonzalo, García-Sánchez, Pablo, Martínez-Bazan, Norbert, et al. “Semi-supervised Constrained Clustering: An In-depth Review”. In: *Artificial Intelligence Review* (2025). DOI: 10.1007/s10462-024-11103-8.

[35] Shen, Yiyang et al. “Semi-MoreGAN: Semi-supervised Generative Adversarial Network for Mixture of Rain Removal”. In: *Computer Graphics Forum* 41.7 (2022), pp. 443–454. DOI: 10.1111/cgf.14690.

[36] Li, X. et al. “Gesture Recognition Based on Fuzzy C-Means Clustering”. In: *Proceedings of Conference name*. mobile robot gesture recognition; FCM. 2023. URL: https://www.semanticscholar.org/paper/Gesture-Recognition-Based-on-Fuzzy-C-Means-Li/e9a615e0c903e2664e49c71a140bd8c694e92f8f.

[37] Peer Ahamed Buhari, S. Kother Mohideen. “Gradient Orientation Mapping Based Fuzzy C-Means Clustering for Digital Dental X-Ray Images”. In: *Journal of Cardiovascular Disease Research* 11.2 (2020). “nha khoa, segmentation”, pp. 96–108. DOI: 10.48047.

[38] Díaz, Gabriel Marín, Medina, Raquel Gómez, and Jiménez, José Alberto Aijón. “Integrating Fuzzy C-Means Clustering and Explainable AI for Robust Galaxy Classification”. In: *Mathematics* 12.18 (2024), p. 2797. DOI: 10.3390/math12182797.

[39] Bezdek, J. C., Ehrlich, R., and Full, W. “FCM: The fuzzy c-means clustering algorithm”. In: *Computers & Geosciences* 10.2 (1984), pp. 191–203. DOI: 10.1016/0098-3004(84)90020-7.

[40] Li, Sai et al. “Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data”. In: *IEEE Access* (2024).

[41] Xu, Shengbing et al. “Semi-supervised fuzzy clustering algorithm based on prior membership degree matrix with expert preference”. In: *Expert Systems with Applications* 238 (2024), p. 121812. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.121812.

[42] Wagstaff, Kiri et al. “Constrained k-means clustering with background knowledge”. In: *Icml*. Vol. 1. 2001, pp. 577–584.

[43] Golzari Oskouei, Amin, Samadi, Negin, and Tanha, Jafar. “Feature-weight and cluster-weight learning in fuzzy c-means method for semi-supervised clusteringImage 1”. In: *Applied Soft Computing* 161 (2024), p. 111712. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2024.111712.

[44] Thang, Nguyen Truong et al. “A novel spatial complex fuzzy inference system for detection of changes in remote sensing images”. In: *Applied Intelligence* 55.3 (2024), p. 178. DOI: 10.1007/s10489-024-06000-0.

[45] Jia, Xiaohong et al. “Fuzzy C-Means Clustering with Region Constraints for Superpixel Generation”. In: *International Journal of Fuzzy Systems* (2025). DOI: 10.1007/s40815-025-02017-w.

[46] Trung, Nguyen Tu, Hien, Le Xuan, and Tuan, Tran Manh. “Enhancing Contrast of Dark Satellite Images Based on Fuzzy Semi-Supervised Clustering and an Enhancement Operator”. In: *Remote Sensing* 15.6 (2023), p. 1645. DOI: 10.3390/rs15061645.

[47] Wang, Zhen et al. “Fuzzy Discriminant Clustering with Fuzzy Pairwise Constraints”. In: *arXiv* abs/2104.08546 (2021). URL: https://arxiv.org/abs/2104.08546.

[48] Qadri, Syed Furqan et al. “Chan–Vese aided fuzzy C-means approach for whole breast and fibroglandular tissue segmentation: Preliminary application to real-world breast MRI”. In: *Medical Physics* 52.5 (2025), pp. 2950–2960. DOI: 10.1002/mp.17660.

[49] Demirhan, H. et al. “Mixed fuzzy C-means clustering method for mixed data types”. In: *Applied Soft Computing* (2025), p. 110885. DOI: 10.1016/j.asoc.2024.110885.

[50] Thong, P. H. and Son, L. H. “An overview of semi-supervised fuzzy clustering algorithms”. In: *International Journal of Engineering and Technology* 8.4 (2016), pp. 301–309. DOI: 10.7763/IJET.2016.V8.875.

[51] Tuan, Tran Manh et al. “An improvement of trusted safe semi-supervised fuzzy clustering method with multiple fuzzifiers”. In: *Journal of Computer Science and Cybernetics* 38.1 (2022), pp. 47–61. DOI: 10.15625/1813-9663/38/1/16720.

[52] Do, Viet Duc, Mai, Dinh Sinh, and Ngo, Long Thanh. “A collaborative learning model using the semi-supervised method and the interval type-2 fuzzy set for large data analysis”. In: *Multimedia Tools and Applications* (2025). DOI: 10.1007/s11042-025-20939-3.

[53] Thong, Pham Huy et al. “Multi-View Picture Fuzzy Clustering: A Novel Method for Partitioning Multi-View Relational Data”. In: *Computers, Materials & Continua* 83.3 (2025), pp. 5461–5485. DOI: 10.32604/cmc.2025.065127.

[54] Canh, Hoang Thi et al. “A novel semi-supervised consensus fuzzy clustering method for multi-view relational data”. In: *International Journal of Electrical and Computer Engineering* 14.6 (2024), pp. 6883–6893. DOI: 10.11591/ijece.v14i6.pp6883-6893.

[55] Anh, Le Tuan et al. “A novel distributed semi-supervised fuzzy clustering method applied on dental X-ray images”. In: *Vietnam Journal of Science and Technology* 63.1 (2025), pp. 149–160. DOI: 10.15625/2525-2518/19648.

[56] Huan, Phan Tri et al. “Safe Semi-Supervised Fuzzy Clustering and its Application in Classification Tasks”. In: *TNU Journal of Science and Technology* (2025). DOI: 10.34238/tnu-jst.12195.

[57] Yang, M. S., Hwang, P. Y., and Chen, D. H. “Fuzzy clustering algorithms for mixed feature variables”. In: *Fuzzy Sets and Systems* 141.2 (2004), pp. 301–317. DOI: 10.1016/S0165-0114(03)00111-5.

[58] Zhou, H. and Schaefer, G. "An overview of fuzzy c-means based image clustering algorithms". In: *Foundations of Computational Intelligence Vol. 2*. Springer Berlin Heidelberg, 2009, pp. 295–310. DOI: 10.1007/978-3-642-01536-6_15.

[59] Nam, Pham Quang et al. "Application of secure semi-supervised fuzzy clustering in object detection from remote sensing images". In: *Journal of Science and Transport Technology (UTT)* 2.3 (2022), pp. 33–41. DOI: 10.58845/jstt.utt.2022.en.2.3.33-41.

[60] Pedrycz, W. and Waletzky, J. "Fuzzy clustering with partial supervision". In: *Trans. Sys. Man Cyber. Part B* 27.5 (Oct. 1997), pp. 787–795. ISSN: 1083-4419. DOI: 10.1109/3477.623232.

[61] Xie, Gaoliang et al. "Time-Series Remote Sensing Image Classification with Active Deep Learning". In: *Preprints.org* 2024.10 (2024), p. 1526. DOI: 10.20944/preprints202410.1526.v1. URL: https://doi.org/10.20944/preprints202410.1526.v1.

[62] Miao, Sheng et al. "Utilizing Active Learning and Attention-CNN to Classify Vegetation Based on UAV Multispectral Data". In: *Scientific Reports* 14.1 (2024), p. 31061. DOI: 10.1038/s41598-024-82248-3.

[63] Otsu, Nobuyuki. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.

[64] Canny, John. "A Computational Approach to Edge Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8.6 (1986), pp. 679–698. DOI: 10.1109/TPAMI.1986.4767851.

[65] Chan, Tony F. and Vese, Luminita A. "Active Contours Without Edges". In: *IEEE Transactions on Image Processing* 10.2 (2001), pp. 266–277. DOI: 10.1109/83.902291.

[66] Li, Lin et al. "Semi-Supervised Fuzzy Clustering With Feature Discrimination". In: *PLOS ONE* 10.9 (2015), e0131160. DOI: 10.1371/journal.pone.0131160.

[67] Economou, P. et al. "A Clustering Algorithm for Overlapping Gaussian Mixtures". In: *Journal of Computational and Applied Mathematics* 420 (2023), p. 118617. DOI: 10.1080/27684520.2023.2242337.

[68] Pei, X. Y., Huang, H. B., and Cao, P. "An Improved Gaussian Mixture Model-Based Data Normalization Method for Removing Environmental Effects on Damage Detection of Structures". In: *Buildings* 15.3 (2025), p. 359. DOI: 10.3390/buildings15030359.

[69] Dempster, A. P., Laird, N. M., and Rubin, D. B. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical*

Society: Series B 39.1 (1977), pp. 1–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[70] Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[71] Cai, Z. et al. "A Self-Adaptive Density-Based Clustering Algorithm for DBSCAN". In: *Pattern Recognition Letters* 213 (2024), pp. 23–31. DOI: 10.1016/j.patrec.2024.02.017.

[72] Luo, R. et al. "An Improved DBSCAN Clustering Algorithm by Natural Neighbor and Granular-Ball (NaGB-DBSCAN)". In: *Information Sciences* 645 (2025), pp. 345–362. DOI: 10.1016/j.ins.2025.05.077.

[73] Hosseinzadeh, M. et al. "Enhancing DBSCAN Clustering with a Fuzzy System for Dynamic Parameter Optimization in WBANs". In: *Scientific Reports* 15 (2025), p. 13293. DOI: 10.1038/s41598-025-13293-9.

[74] Ma, H. X. et al. "sOPTICS: A Modified Density-Based Algorithm for Identifying Clusters Accounting for Line-of-Sight". In: *Monthly Notices of the Royal Astronomical Society* 537.2 (2025), pp. 1504–1516. DOI: 10.1093/mnras/sty7965971.

[75] Hajihosseinlou, M. et al. "A Comprehensive Evaluation of OPTICS, GMM and K-Means for Anomaly Detection in Time Series Data". In: *Analytica Chimica Acta* (2024). DOI: 10.1016/j.aca.2024.01.004.

[76] Campello, Ricardo J. G. B., Moulavi, Davoud, and Sander, Joerg. "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Vol. 7819. Lecture Notes in Computer Science. 2013, pp. 160–172. DOI: 10.1007/978-3-642-37456-2_14.

[77] Rodriguez, Alex and Laio, Alessandro. "Clustering by Fast Search and Find of Density Peaks". In: *Science* 344.6191 (2014), pp. 1492–1496. DOI: 10.1126/science.1242072.

[78] Wang, Jingdong et al. "Deep High-Resolution Representation Learning for Visual Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2021), pp. 3349–3364. DOI: 10.1109/TPAMI.2020.2983686.

[79] Cao, Hu et al. *Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation*. 2021. arXiv: 2105.05537.

[80] Frigui, Hichem, Bchir, Quiem, and Baili, Naouel. "An overview of unsupervised and semi-supervised fuzzy kernel clustering". In: *International Journal of Fuzzy Logic and Intelligent Systems* 13.4 (2013), pp. 254–268. DOI: 10.5391/IJFIS.2013.13.4.254.

[81] Zhao, B., Kwok, J., and Zhang, C. "Multiple kernel clustering". In: *Proceedings of the 9th SIAM International Conference on Data Mining (SDM)*. Philadelphia, PA: SIAM, 2009, pp. 638–649. DOI: 10.1137/1.9781611972795.55.

[82] Huang, H.-C., Chuang, Y.-Y., and Chen, C.-S. "Multiple kernel fuzzy clustering". In: *IEEE Transactions on Fuzzy Systems* 20.1 (2012), pp. 120–134. DOI: 10.1109/TFUZZ.2011.2170488.

[83] Mai, S. D. and Ngo, L. T. "Multiple kernel approach to semi-supervised fuzzy clustering algorithm for land-cover classification". In: *Engineering Applications of Artificial Intelligence* 68 (2018), pp. 205–213. DOI: 10.1016/j.engappai.2017.11.006.

[84] Kanzawa, Y. "Semi-supervised fuzzy c-means algorithms by revising dissimilarity/kernel matrices". In: *Fuzzy Sets, Rough Sets, Multisets and Clustering*. Ed. by Skowron, A., Suraj, Z., et al. Cham: Springer, 2017, pp. 45–61. DOI: 10.1007/978-3-319-44998-2_4.

[85] Casalino, Gabriella, Castellano, Giovanna, and Mencar, Corrado. "Incremental adaptive semi-supervised fuzzy clustering for data stream classification". In: *2018 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. 2018, pp. 1–7. DOI: 10.1109/EAIS.2018.8397172.

[86] Arshad, Ali, Riaz, Saman, and Jiao, Licheng. "Semi-Supervised Deep Fuzzy C-Means Clustering for Imbalanced Multi-Class Classification". In: *International Journal of Machine Learning and Cybernetics* 10.12 (2019), pp. 3545–3559. DOI: 10.1007/s13042-018-00891-9.

[87] Yang. "Deep Semi-Supervised Fuzzy Clustering with Autoencoders". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.12 (2021), pp. 5332–5345. DOI: 10.1109/TNNLS.2021.3071234.

[88] Zhang, Yejian and Takada, Shingo. "Applying LLMs to Active Learning: Towards Cost-Efficient Cross-Task Text Classification without Manually Labeled Data". In: *arXiv preprint* abs/2502.16892 (2025). DOI: 10.48550/arXiv.2502.16892. URL: https://arxiv.org/abs/2502.16892.

[89] Yin, Tianxiang et al. "Boosting Active Learning via Re-Aligned Feature Space". In: *Knowledge-Based Systems* 311.11 (2025), p. 113085. DOI: 10.1016/j.knosys.2025.113085.

[90] Alyoubi, AA et al. "Efficient and Accurate Face Mask Detection with Active Learning". In: *Technology, Knowledge and Learning* 30.3 (2025), pp. 467–484. DOI: 10.1177/18724981251318457.

[91] Bragg, Jonathan, Baeza-Yates, Ricardo, et al. "Active Learning for Ranking through Expected Loss Optimization". In: *Proceedings of the ACM International*

Conference on Web Search and Data Mining (WSDM). 2009, pp. 93–102. DOI: 10.1145/1498759.1498781.

[92] Jifan Zhang Lalit Jain, Kevin Jamieson. "Learning to Actively Learn: A Robust Approach". In: *arXiv preprint* abs/2010.15382.— (2025). DOI: 10.48550/arXiv.2010.15382.

[93] Cacciarelli, Davide and Kulahci, Murat. "Active Learning for Data Streams: A Survey". In: *Machine Learning* 112.10 (2023), pp. 1–55. DOI: 10.1007/s10994-023-06454-2.

[94] Settles, Burr. *Active Learning Literature Survey*. Tech. rep. University of Wisconsin-Madison, Department of Computer Science, 2009. URL: https://burrsettles.com/pub/settles.activelearning.pdf.

[95] Schein, Andrew I. and Ungar, Lyle H. "Active Learning for Logistic Regression: An Evaluation". In: *Machine Learning* 68.3 (2007), pp. 235–265. DOI: 10.1007/s10994-007-5019-5.

[96] Long, Zhiguo et al. "Semi-Supervised Clustering Guided by Pairwise Constraints and Local Density Structures". In: *Pattern Recognition* 156 (2024), p. 110751. DOI: 10.1016/j.patcog.2024.110751. URL: https://dl.acm.org/doi/10.1016/j.patcog.2024.110751.

[97] Wang, Min et al. "Active Learning Through Density Clustering". In: *Expert Systems with Applications* 85 (2017), pp. 305–317. DOI: 10.1016/j.eswa.2017.05.046. URL: https://www.sciencedirect.com/science/article/abs/pii/S095741741730369X.

[98] Xiong, Caiming, Johnson, David M., and Corso, Jason J. "Active Clustering with Model-Based Uncertainty Reduction". In: *arXiv* abs/1402.1783 (2014). URL: https://arxiv.org/abs/1402.1783.

[99] Grira, Nizar, Crucianu, Michel, and Boujemaa, Nozha. "Fuzzy Clustering with Pairwise Constraints for Knowledge-Driven Image Categorization". In: *IEE Proceedings - Vision, Image and Signal Processing* 153.3 (2006), pp. 299–304. DOI: 10.1049/ip-vis:20050060.

[100] Santos, Gabriel Machado, Julia, Rita Maria Silva, and Do Nascimento, Marcelo Zanchetta. "K-GBS3FCM-KNN Graph-Based Safe Semi-Supervised Fuzzy C-Means". In: *2024 IEEE International Conference on Big Data (BigData)*. IEEE. 2024, pp. 3498–3507.

[101] Gan, Haitao. "Safe Semi-Supervised Fuzzy C-Means Clustering". In: *IEEE Access* 7 (2019), pp. 95659–95664.

[102] Rand, William M. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.

[103] Mahmoudi, Ali et al. “Proof of biased behavior of Normalized Mutual Information (NMI)”. In: *Scientific Reports* 14 (2024), p. 59073. DOI: 10.1038/s41598-024-59073-9.

[104] Ros, Frédéric, Riad, Rabia, and Guillaume, Serge. “PDBI: A partitioning Davies-Bouldin index for clustering evaluation”. In: *Neurocomputing* 528 (2023), pp. 178–199.

[105] Kim, Dae-Won and Lee, Kwang H. “A New Validity Index for Fuzzy C-Means Clustering”. In: *arXiv preprint* (2024). arXiv:2407.06774.

[106] Frank, Andrew. “UCI machine learning repository”. In: *http://archive. ics. uci. edu/ml* (2010).

[107] Yin, Xianglong, Shu, Ting, and Huang, Qingming. “Semi-supervised fuzzy clustering with metric learning and entropy regularization”. In: *Knowledge-Based Systems* 35 (2012), pp. 304–311. DOI: 10.1016/j.knosys.2012.05.016.

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