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

TÓM TẮT LUẬN ÁN TIẾN SĨ MÁY TÍNH

Ngành: Hệ thống thông tin

Mã số:9 48 01 04

Hà Nội - 2026

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ,
Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học:
Người hướng dẫn 1: PGS.TS. Hà Hải Nam, Trường Đại học Điện lực
Người hướng dẫn 2: PGS.TS. Nguyễn Long Giang, Viện Công nghệ thông tin

Phản biện 1:
Phản biện 2:
Phản biện 3:

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi ……. giờ ……, ngày …….. tháng …….. năm

Có thể tìm hiểu luận án tại:
1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

Mở đầu

Tính cấp thiết của đề tài. Phân cụm dữ liệu là kỹ thuật nền tảng để khai thác tri thức trong bối cảnh dữ liệu lớn và đa dạng (web, văn bản, ảnh), với nhu cầu ngày càng tăng về mô hình hoá cấu trúc ẩn và giảm chiều/khám phá cụm [1]. Các phương pháp phân cụm rõ ràng (crisp) gặp hạn chế khi dữ liệu chồng lấn; Fuzzy C-Means (FCM) kế thừa ý niệm tập mờ của Zadeh [2] và trở thành phương pháp kinh điển nhờ gán độ phụ thuộc mềm [3], được phát triển mạnh mẽ trong các nghiên cứu gần đây [4]. Dẫu vậy, FCM vẫn dễ sai ở ranh giới cụm, nhạy khởi tạo và kém ổn định với nhiễu. Phân cụm bán giám sát mờ (Semi-supervised fuzzy clustering - SSFC) bổ sung tri thức ngoài (nhãn , cặp ràng buộc, độ phụ thuộc cho trước) để cải thiện chất lượng [5, 6], song dễ bị ảnh hưởng bởi nhãn không đáng tin cậy, đặc biệt ở biên cụm. Học tích cực (Active Learning) cho phép truy vấn có chọn lọc các điểm mơ hồ/biên, giảm chi phí gán nhãn và tăng độ tin cậy [7, 8], phù hợp cho SSFC và các bài toán phân đoạn ảnh (y tế, viễn thám, giám sát) vốn mơ hồ tại ranh giới vùng [9, 10]. Các tiếp cận SSFC hiện hữu còn thiếu cơ chế xác định và hiệu chỉnh nhãn sai ở biên cũng như khai thác tri thức miền có hệ thống và chiến lược truy vấn tối ưu để hạn chế lan truyền sai. Chính vì khoảng trống khoa học này, luận án hướng tới việc nghiên cứu nâng cao chất lượng phân cụm trong điều kiện dữ liệu nhiễu, cấu trúc cụm phức tạp và chồng lấn dựa vào việc áp dung học tích cực trong phân cụm bán giám sát mờ.

Mục tiêu nghiên cứu. (1) Khảo sát toàn diện SSFC/Active Learning hiện đại, thực nghiệm trên bộ dữ liệu chuẩn để định vị ưu—nhược—điểm. (2) Đề xuất các phương pháp mới cho *SSFC dựa vào học tích cực* tập trung xử lý vùng biên và nhãn không tin cậy. (3) Ứng dụng vào phân đoạn ảnh, đặc biệt bài toán phát hiện vùng lũ.

Đối tượng, phạm vi và phương pháp. Luận án nghiên cứu SSFC, SSFC mờ, và SSFC mờ tích cực trên dữ liệu nhiễu, mơ hồ vùng biên; sử dụng bộ dữ liệu chuẩn (UCI) và dữ liệu ảnh. Đánh giá bằng các chỉ số chuẩn hoá chất lượng phân cụm và hiệu quả. Phương pháp tiếp cận gồm: (i) tổng quan và phân tích lý thuyết; (ii) Tối ưu, cải tiến thuật toán, đề xuất thuật toán mới và đánh giá thực nghiệm.

Đóng góp chính. (1) Đề xuất phương pháp *phân cụm bán giám sát mờ tích cực dựa vào biên cụm*, xác định và tinh chỉnh điểm biên để nâng cao độ chính xác phân cụm vùng biên nơi khả năng phân cụm sai nhiều nhất cũng như điều chỉnh tậm cụm dựa vào tâm vùng biên làm co nhỏ vùng biên khiến hiệu quả phân cụm được chính xác hơn. (2) Đề xuất phương pháp *phân cụm bán giám sát mờ tích cực an toàn* với ràng buộc cặp theo biên, cân bằng giữa cải thiện độ chính xác và kiểm soát rủi ro lan truyền sai. Cả hai đều nhằm tăng độ chính xác, độ vững trong bối cảnh cụm chồng lấn và dữ liệu nhiễu, với chi phí gán nhãn tiết kiệm nhờ truy vấn chọn lọc.

Cấu trúc luận án. *Chương 1* trình bày tổng quan phân cụm, SSFC/SSFC mờ, học tích cực và thước đo đánh giá; nêu vấn đề và định vị khoảng trống khoa học. *Chương 2* đề xuất và đánh giá phương pháp SSFC mờ tích cực dựa vào biên cụm trên dữ liệu

UCI và ảnh. *Chương 3* phát triển biến thể an toàn với ràng buộc cặp theo biên và thực nghiệm phát hiện vùng lũ. Kết luận tổng hợp kết quả và hướng nghiên cứu tiếp theo.

Chương 1

Tổng quan về phân cụm bán giám sát mờ, học chủ động và bài toán ứng dụng phân cụm bán giám sát trong thực tiễn

1.1 Tổng quan phân cụm bán giám sát mờ

1.1.1 phân cụm mờ

Phân cụm mờ (Fuzzy/Soft Clustering) là mở rộng phân cụm "cứng" bằng cách gán cho mỗi điểm một *vector độ phụ thuộc* vào c cụm, với các thành phần trong $[0, 1]$ và thường chuẩn hoá tổng bằng 1. Biểu diễn mềm này phản ánh tốt hơn thực tế khi *ranh giới cụm mơ hồ*, *cụm chồng lấn* hoặc dữ liệu *nhiễu*, vì một điểm có thể đồng thời liên quan đến nhiều cụm với mức độ khác nhau.

1.1.2 phân cụm bán giám sát

Phân cụm bán giám sát (SSC) kết hợp dữ liệu không nhãn quy mô lớn với một lượng nhỏ *thông tin giám sát* nhằm nâng cao chất lượng phân cụm, giảm chi phí gán nhãn và tăng ổn định mô hình so với học thuần không giám sát. Nguồn giám sát phổ biến gồm: *(i) nhãn ít* cho một số điểm [9]; *(ii) ràng buộc cặp must-link/cannot-link* nêu quan hệ thuộc–không thuộc cùng cụm [11]; và *(iii) độ phụ thuộc cho trước* trong khung phân cụm mờ [12, 13].

1.1.3 Phân cụm bán giám sát mờ

Trong phần nghiên cứu tổng quan này, chúng tôi xem FCM như nền tảng của phân cụm mờ: thuật toán tối ưu luân phiên độ phụ thuộc các điểm và tâm cụm, thuật toán đã chứng minh tính hiệu quả trên nhiều lĩnh vực (phân đoạn ảnh, nhận dạng, y khoa. . .) nhưng vẫn nhạy cảm với khởi tạo, dễ bị nhiễu/ngoại lai và giả định hình học cụm gần đẳng phương [14]. Nhiều biến thể đã được đề xuất để gia tăng độ bền và linh hoạt—từ rough/kernel/robust, Wasserstein, trọng số không gian, lựa chọn fuzzifier, lai tối ưu, tới ràng buộc giá trị riêng và các hạng phạt chống đơn điệu [15]. Khi nhãn khan hiếm nhưng cần phù hợp ngữ nghĩa, phân cụm bán giám sát mờ (SSFC) đưa tri thức miền vào mục tiêu thông qua nhãn ít, ràng buộc ML/CL và độ phụ thuộc cho trước; các nhánh Kernel SSFC (kể cả đa nhân) tăng khả năng mô hình hoá phi tuyến nhưng khó hiệu chỉnh tham số, nhánh thích nghi xử lý dữ liệu dòng/đa chiều, và nhánh "an toàn" giảm tác động nhãn sai bằng trọng số tin cậy/đồ thị cục bộ [16–18]. Đáng chú ý, 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 và vùng biên nhiễu [12]. Ở hướng kết

hợp học sâu, các mô hình neuro-fuzzy/Deep SSFC học biểu diễn (AE/CNN) đồng thời với mục tiêu FCM có ràng buộc, cho kết quả tốt trên dữ liệu phức tạp nhưng nhạy siêu tham số và chi phí huấn luyện lớn [19, 20]. Cuối cùng, học chủ động cho SSFC chọn truy vấn/cặp tại vùng biên không chắc chắn—AFCC (FHV), K-GBS^{3}FCM (đồ thị KNN "an toàn"), SFCM-PM (prior+entropy)—giúp đạt chất lượng cao với rất ít nhãn, song còn phụ thuộc tham số và chi phí xây đồ thị [5, 8, 21]. Những quan sát này chỉ ra các khoảng trống cần giải quyết: tích hợp chặt chẽ chọn cặp chủ động với cơ chế "an toàn" chống nhãn sai, tối ưu đồng thời biểu diễn–ràng buộc–chọn mẫu, cùng các thiết kế ít nhạy tham số và mở rộng tốt trên dữ liệu lớn.

1.2 Học chủ động

Học chủ động (Active Learning) nhằm giảm chi phí gán nhãn bằng cách *chủ động* chọn những mẫu giàu thông tin nhất để hỏi chuyên gia, thay vì thu thập/ghi nhãn ngẫu nhiên như *học thụ động* (passive learning). Giả thiết cốt lõi: nếu bộ học được phép chọn mẫu để hỏi, có thể đạt hiệu năng tương đương (hoặc tốt hơn) với *ít nhãn hơn*. Cách tiếp cận này đặc biệt hữu ích khi dữ liệu thô dồi dào nhưng nhãn khan hiếm/đắt đỏ; đã chứng minh hiệu quả trong phân loại văn bản, trích xuất thông tin, ảnh–viễn thám–y sinh, lựa chọn cảm biến, học cấu trúc mạng và phân tích phần mềm [22].

Loại truy vấn & tiêu chí chọn. Ba dạng truy vấn thường dùng: (i) *nhãn điểm* (point-label); (ii) *ràng buộc cặp* ML/CL—phổ biến vì dễ hỏi [7]; (iii) *truy vấn cấu trúc/so sánh* (triplet, "peak" trong DPC) [23, 24]. Tiêu chí utility tiêu biểu: *bất định* (entropy/margin, bất đồng committee), *đại diện* (mật độ/kNN), *ảnh hưởng/kỳ vọng đổi mô hình* (expected model change), và *đa dạng* để tránh trùng lặp trong truy vấn theo lô; kết hợp nhiều tiêu chí thường hiệu quả hơn dùng đơn lẻ [25–27].

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) kết hợp dữ liệu chưa nhãn với thông tin bổ trợ như nhãn điểm, ràng buộc ML/CL hoặc độ phụ thuộc cho trước, qua đó cải thiện độ chính xác so với phân cụm mờ không giám sát [7, 27, 28]. Tuy nhiên, SSFC vẫn đối mặt nhiều thách thức: chi phí nhãn cao, sự nhiễu và các điểm biên gây sai lệch, tính nhạy cảm với tham số và số cụm, khả năng xuất hiện ràng buộc mâu thuẫn và chi phí tính toán lớn. Học chủ động được xem là giải pháp khả thi khi chỉ truy vấn "ít nhưng đúng chỗ", tập trung vào vùng bất định để củng cố ranh giới cụm và tăng độ bền vững mô hình [25, 26].

B. Chiến lược truy vấn

Trong SSFC, phản hồi có thể ở dạng nhãn trực tiếp (khi biết số cụm), ràng buộc ML/CL vốn thân thiện với người dùng [7, 8], hoặc cấu trúc so sánh như triplet/peak

phù hợp với mô hình dựa trên mật độ [24]. Việc chọn truy vấn chủ yếu dựa trên bốn tiêu chí: mức độ bất định của điểm, tính đại diện, mức độ đa dạng và tác động đối với mô hình [25].

C. Các nghiên cứu tiêu biểu

Các hướng nghiên cứu quan trọng gồm: AFCC (2008) sử dụng Fuzzy Hypervolume để chọn ràng buộc tại vùng biên, đạt độ chính xác cao nhưng chi phí lớn và nhạy tham số [8]; ASC (2017) tối ưu ràng buộc trên đồ thị KNN giúp xác định số cụm và giảm truy vấn dư thừa [29]; K-GBS3FCM (2024) lan truyền nhãn an toàn dựa trên KNN nhưng phụ thuộc mạnh vào lựa chọn K [21]; và SFCM-PM (2025) tích hợp độ phụ thuộc cho trước hữu ích cho dữ liệu mất cân bằng nhưng dễ nhiễu nếu prior thiếu tin cậy [5].

D. Hạn chế chung và khoảng trống nghiên cứu

Mặc dù các phương pháp trên đều giảm được chi phí truy vấn, phần lớn vẫn chưa xử lý hiệu quả vùng biên chồng lấn — nơi sai số tập trung và nơi mô hình mờ thể hiện mức bất định cao [27]. Đồng thời, việc tối ưu hoá truy vấn dựa trên cấu trúc độ phụ thuộc còn hạn chế [8], và hầu như chưa có phương pháp nào chú trọng mạnh đến tính bền vững trước nhiễu và điểm ngoại lai [25]. Đây là khoảng trống nghiên cứu đáng kể.

E. Định hướng và đóng góp của luận án

Luận án hướng tới khai thác cấu trúc mờ của cụm bằng cách xác định rõ vùng biên theo độ bất định và mật độ, từ đó thiết kế chiến lược truy vấn chủ động ưu tiên các điểm nằm tại ranh giới cụm — nơi phản hồi của chuyên gia mang giá trị cao nhất. Trên cơ sở đó, luận án đề xuất thuật toán phân cụm bán giám sát mờ chủ động dựa trên vùng biên, giúp giảm chi phí nhãn, tăng độ chính xác và tăng khả năng chống nhiễu, hướng đến hiệu quả cao hơn trong các ứng dụng thị giác máy tính thực tế.

1.3.2 Phân cụm bán giám sát mờ an toàn

A. Bối cảnh và động lực

Trong phân cụm bán giám sát mờ (SSFC), độ tin cậy của dữ liệu có nhãn và ràng buộc ML/CL đóng vai trò then chốt. Khi các nguồn thông tin này bị nhiễu, thiếu nhất quán hoặc mang tính thiên lệch, mô hình rất dễ khuếch đại sai lệch thông qua quá trình lan truyền nhãn, tối ưu sai hướng do gán trọng số cao cho điểm kém tin cậy, hoặc hội tụ cục bộ tại các vùng biên cụm. Vì vậy, học an toàn trong SSFC chú trọng vào việc nhận diện và điều tiết độ tin cậy ở cả mức điểm, mức cặp và mức cụm, nhằm duy trì sự ổn định của quá trình tối ưu.

B. Các phương pháp SSFC an toàn gần đây

Những nghiên cứu gần đây đã tích hợp cơ chế trọng số và thích ứng nhằm nâng cao tính ổn định cho mô hình. S^3FCM, LHC-S3FCM và CS3FCM [18, 30] sử dụng độ tin cậy kết hợp đồ thị lân cận để điều chỉnh mức đóng góp của nhãn. TS3FCM [12] tăng cường khả năng hội tụ và độ bền vững thông qua lọc nhãn và khai thác độ phụ thuộc cho trước, đặc biệt hiệu quả trong môi trường dữ liệu nhiễu.

C. Đánh giá đặc điểm và hạn chế

Mặc dù các phương pháp trên cải thiện đáng kể mức độ an toàn, chúng vẫn còn phụ

thuộc mạnh vào khởi tạo ngẫu nhiên và phân bố mật độ không đồng đều, xử lý chưa hiệu quả vùng biên chồng lấn hoặc điểm nhiễu, và khả năng kiểm soát rủi ro của ràng buộc còn hạn chế khi các cụm mờ giao nhau.

D. Khoảng trống nghiên cứu

Hiện vẫn thiếu cơ chế bảo đảm an toàn ở mức độ *cục bộ* cho vùng biên mơ hồ, nơi sai số tập trung cao nhất. Mức độ tin cậy của nhãn và ràng buộc cũng chưa được tích hợp một cách tự nhiên vào cấu trúc mờ, khiến mô hình khó cân bằng giữa an toàn và chi phí giám sát — đặc biệt trong các bài toán có độ nhiễu lớn.

E. Định hướng và đóng góp của luận án

Luận án hướng đến nhận diện và xử lý vùng biên mơ hồ dựa trên bất định và mật độ; thiết kế cơ chế tự điều chỉnh trọng số theo mức độ tin cậy để hạn chế tác động của nhiễu; và xây dựng mô hình SSFC an toàn, hiệu quả cho các ứng dụng thị giác máy tính, nơi yêu cầu về độ tin cậy và khả năng chịu nhiễu đặc biệt quan trọng. Mục tiêu cuối cùng là tăng độ bền vững trước nhiễu và cải thiện chất lượng phân cụm trong các kịch bản thực tế phức tạp.

1.4 Mô hình và dữ liệu đánh giá

Để kiểm chứng hiệu quả trong bối cảnh biên mờ, chồng lấn cao và nhiễu thực tế, chúng tôi sử dụng ba nguồn dữ liệu bổ trợ: (i) benchmark UCI (10 bộ dữ liệu kinh điển như Iris, Wine, BCW, Glass, Thyroid, Soybean(small), Haberman, Australian Credit, Spambase, Waveform-5000) để so sánh định lượng có đối sánh chuẩn [31]; (ii) dữ liệu tự sinh có kiểm soát chồng lấn (GEN2D) nhằm cô lập và điều khiển mức chồng lấn (tỷ lệ p) và độ mờ biên (hệ số trộn α) phục vụ phân tích độ nhạy theo độ nhiễu, số cụm và băng thông mật độ; (iii) dữ liệu ảnh gồm (a) Landsat-8 khu vực ven biển Thanh Hóa cho bài toán phân lớp phủ đất (6 lớp) với biên phức tạp và không đồng nhất phổ; (b) tập phân đoạn vùng ngập gồm khoảng 290 ảnh (mask nhị phân "ngập/không ngập") để kiểm định trong kịch bản đường biên mảnh và nhiễu phản xạ.

Độ đo đánh giá. Luận án sử dụng bộ đo gồm các chỉ số: RI, F1, NMI, DB, PC, PE, DB_fuzzy

1.5 Kết luận Chương 1

Chương 1 đã hệ thống hoá nền tảng cho phân đoạn ảnh: từ tập mờ, cơ chế sử dụng độ phụ thuộc và FCM, mở rộng tới phân cụm bán giám sát mờ (SSFC) với nhãn hạt giống, scribbles, ràng buộc ML/CL và thông tin lân cận. Chương làm rõ vai trò của học chủ động trong SSFC (chọn truy vấn điểm/cặp/cấu trúc ở vùng biên, đại diện, ảnh hưởng lớn) nhằm giảm mạnh chi phí nhãn, đồng thời kết nối với học sâu (FCN/U–Net/DeepLab, . . .) để vừa giữ ranh giới phức tạp vừa tiết kiệm gán nhãn.

Bộ độ đo được sử dụng thống nhất gồm RI, F1 theo cặp và NMI (các chỉ số ngoại tại), cùng với DB (chỉ số nội tại) để đánh giá đồng thời mức độ khớp nhãn và độ chặt–tách của các cụm. Với dữ liệu mờ, việc tính RI, F1 và NMI được thực hiện sau bước defuzzification bằng quy tắc tạo ra phụ thuộc lớn nhất, trong khi các thước đo

đặc trưng cho phân cụm mờ như PC và PE được dùng để phản ánh mức độ phân biệt giữa các cụm và mức độ mờ của toàn hệ thống. Chỉ số DB_fuzzy được sử dụng song song với DB nhằm mô tả độ chặt cụm trong không gian độ phụ thuộc, đặc biệt hữu ích khi các cụm có ranh giới mềm. Từ phần tổng quan, có thể thấy khoảng trống còn tồn tại liên quan đến việc bảo toàn vùng biên trong không gian mờ, nâng cao độ tin cậy của nhãn và phát triển cơ chế truy vấn chủ động quy mô lớn; các vấn đề này tạo động lực cho những phương pháp và mô hình được trình bày trong các chương tiếp theo.

Từ tổng quan, các khoảng trống về bảo toàn biên, độ tin cậy nhãn và truy vấn chủ động quy mô lớn được xác định cho các chương sau;

Chương 2

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

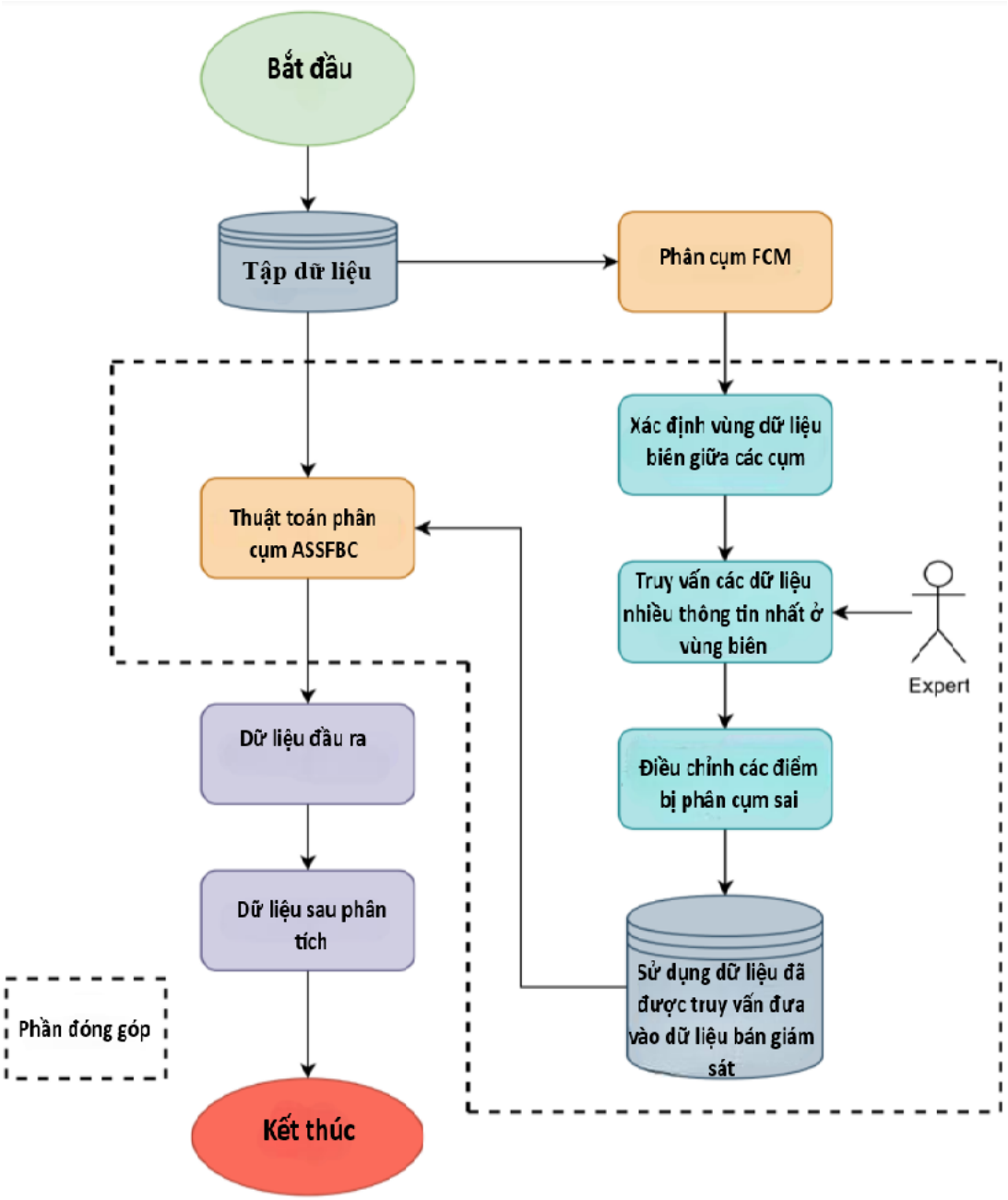
2.1 Biên giữa các cụm và vấn đề liên quan đến phân cụm bán giám sát mờ

Biên cụm là vùng độ phụ thuộc của dữ liệu mơ hồ, dễ gây gán nhãn sai—nhất là trong ảnh và sinh học—khiến SSFC dù có ràng buộc cặp vẫn nhạy khởi tạo và khó xử lý chồng lấn/nhiễu. Học chủ động có thể cải thiện bằng cách ưu tiên truy vấn các điểm "giàu thông tin" ngay tại biên và áp dụng ML/CL để cố định vùng biên với chi phí nhãn thấp. Cần cơ chế xác định biên đáng tin, tiêu chí truy vấn bền vững với nhiễu và mục tiêu mờ có "phạt mềm" để giảm tác động nhãn sai, hướng tới SSFC+AL theo chiến lược *biên–trọng tâm*.

2.2 Ý tưởng và chi tiết thuật toán

Thuật toán đề xuất gồm ba ý chính, tập trung vào *xử lý vùng biên* nơi độ phụ thuộc của dữ liệu mơ hồ: (i) chạy FCM ban đầu để thu được phân cụm mềm và *phát hiện biên*—các điểm có độ phụ thuộc phân tán giữa nhiều cụm; (ii) áp dụng **học chủ động** chỉ trên vùng biên: chọn lọc các điểm mơ hồ, truy vấn oracle để hiệu chỉnh nhãn/độ phụ thuộc, nhờ đó cố định vùng biên cục bộ với chi phí nhãn thấp; (iii) *tinh chỉnh lại* bằng một hàm mục tiêu phân cụm mới tận dụng các điểm/"tâm biên" đã hiệu chỉnh.

Quy trình của thuật toán gồm năm bước chính. Trước hết, mô hình được khởi tạo bằng cách chạy FCM trên toàn bộ dữ liệu để thu được độ phụ thuộc ban đầu $U^{(0)}$ và tập tâm cụm $V^{(0)}$, làm nền cho các bước bán giám sát tiếp theo. Sau đó, thuật toán phát hiện vùng biên bằng cách đo độ mơ hồ của từng điểm, chẳng hạn thông qua chỉ số độ chênh $\Delta_k = \min_{p \neq q} |u_{kp} - u_{kq}|$, và chọn ra N_q điểm có mức bất định nhỏ nhất làm các ứng viên biên. Tiếp theo, cơ chế học chủ động được áp dụng lên các điểm này: mô hình truy vấn oracle để thu nhãn hoặc thông tin prior, từ đó xây dựng ma trận độ phụ thuộc cho trước $\bar{U}$ phản ánh tri thức giám sát cục bộ đáng tin cậy. Dựa trên các điểm được giám sát tại biên, bước bán giám sát tiếp theo tối ưu hàm mục tiêu với điều chỉnh độ phụ thuộc sao cho tiệm cận các giá trị ưu tiên trong $\bar{U}$. Cuối cùng, thuật toán thực hiện quá trình lặp cập nhật U và V cho đến khi tâm hội tụ (sai số không vượt quá ε) hoặc đạt số vòng lặp tối đa.



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

Hàm mục tiêu

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

Cập nhật tham số

Membership. Với ràng buộc $\sum_j u_{kj} = 1$:

$$u_{kj} = \frac{\alpha \bar{u}_{kj}}{1+\alpha} + \frac{1 - \frac{\alpha}{1+\alpha} \sum_{i=1}^{C} \bar{u}_{ki}}{d_{kj}^2 \sum_{i=1}^{C} \frac{1}{d_{ki}^2}}. \tag{2.2}$$

Tâm cụm.

$$v_j = \frac{2 \sum_{k=1}^{N} u_{kj}^2 x_k + 2\alpha \sum_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 x_k + 2\beta \sum_{l \in N_{id}} (1 - \mu_{id}^l) T_{id}}{2 \sum_{k=1}^{N} u_{kj}^2 + 2\alpha \sum_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 + 2\beta \sum_{l \in N_{id}} (1 - \mu_{id}^l)}. \tag{2.3}$$

Thuật toán 1 ASSFBC

1: **Đầu vào**: Tập dữ liệu X, số điểm N, số cụm C, `tỷ_lệ_hạt_giống`, ngưỡng hội tụ ε, số vòng lặp tối đa $T_{\max}$
2: (FCM) Khởi tạo ma trận thành viên $U^{(0)}$ và tâm cụm $V^{(0)}$ bằng thuật toán FCM.
3: (Biên) Tính độ mơ hồ Δ_k và chọn $N_q = \lfloor \texttt{tỷ\_lệ\_hạt\_giống} \cdot N \rfloor$ điểm ở vùng biên.
4: (Học chủ động) Truy vấn chuyên gia $\Rightarrow$ tinh chỉnh $\bar{U}$ trên các điểm biên.
5: Đặt $t \leftarrow 1$.
6: **Lặp** khi $t \leq T_{\max}$:
7: Cập nhật $U^{(t)}$ theo công thức 2.2 .
8: cập nhật $V^{(t)}$ theo công thức tâm cụm 2.3.
9: **Nếu** $\max_j \|v_j^{(t)} - v_j^{(t-1)}\| \leq \varepsilon$ **thì** dừng lặp.
10: Tăng $t \leftarrow t + 1$.
11: **Kết thúc lặp**

Độ phức tạp

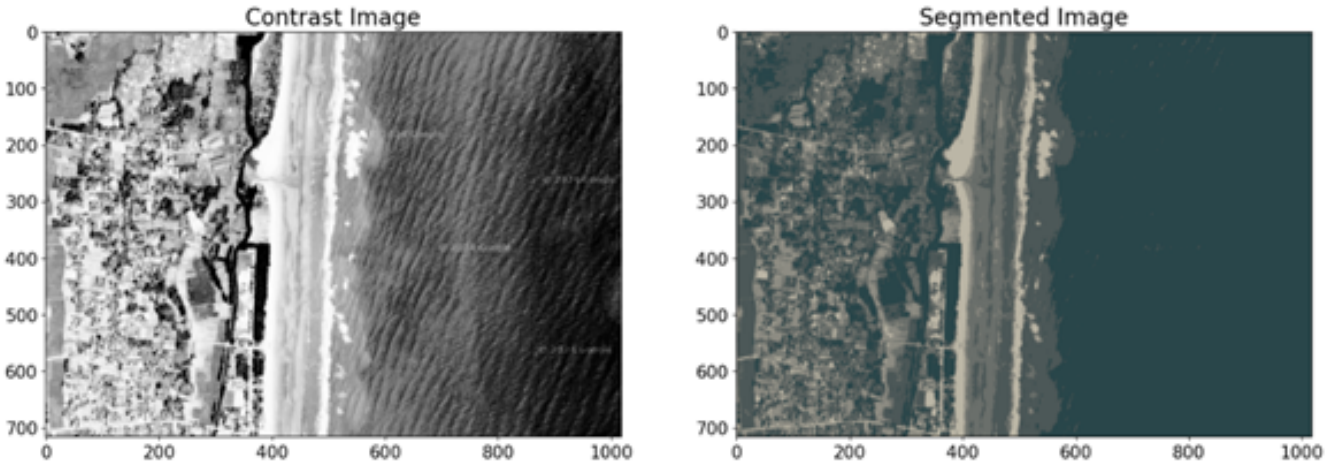
Với I_0 vòng FCM khởi tạo, T vòng tối ưu, $N_q \ll N$:

$$\boxed{\text{Time} \;=\; O\big(NC^2(I_0 + T) + O(N \log N)\big)}$$

2.3 Kết quả 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.

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. Các thí nghiệm được tiến hành để so sánh phương pháp đề xuất ASSFBC với các phương pháp liên quan như: FCM [4], SSFCM [29], eSFCM [44], AFFC [45], và AFFC [46]. Các chỉ số được sử dụng để đánh giá chất lượng phân cụm là: RI, F1, NMI, DB, PC, PE, DB_fuzzy. Sau khi tiến hành các thí nghiệm trên tập dữ liệu, kết quả được thể hiện chi tiết trong các **Hình 2.1, 2.2 và 2.3**, tương ứng với **tập dữ liệu UCI, dữ liệu tổng hợp, và dữ liệu ảnh**.



Hình 2.2: Minh hoạ phân đoạn lớp phủ đất trên ảnh Landsat-8

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
IRIS						
RI	0.8741	0.9018	0.9094	0.9253	0.9492	**0.9939**
F1	0.8364	0.8491	0.8614	0.9086	0.9304	**0.9898**
NMI	0.7381	0.7704	0.7825	0.8241	0.8463	**0.9744**
DB	**0.9414**	1.3615	1.4173	1.5471	1.6932	1.2294
PC	0.7803	0.8112	0.8324	0.8517	0.8831	**0.9418**
PE	0.4216	0.3897	0.3712	0.3431	0.3124	**0.2135**
DB_{fuzzy}	**0.3827**	0.4721	0.5834	0.5219	0.5028	0.4617
Breast						
RI	0.7594	0.7594	0.7641	0.7689	0.7731	**0.7848**
F1	0.7894	0.7894	0.7961	0.8044	0.7971	**0.8667**
NMI	0.4751	0.4751	0.4928	0.4974	0.5002	**0.5123**
DB	**0.7485**	0.8156	1.3481	2.3923	0.9658	0.9144
PC	0.7412	0.7427	0.7529	0.7638	0.7714	**0.7925**
PE	0.4829	0.4817	0.4698	0.4523	0.4427	**0.4235**
DB_{fuzzy}	0.7124	0.6942	**0.6028**	0.7729	0.8157	0.9526
Glass						
RI	0.8214	0.8192	0.8258	0.8310	0.8322	**0.8494**
F1	0.5983	0.5942	0.6029	0.6058	0.6091	**0.6154**
NMI	0.6891	0.6804	0.6875	0.6971	0.6582	**0.7584**
DB	**0.6383**	2.8294	1.9192	1.4598	2.1304	1.1964
PC	0.6921	0.6834	0.7019	0.7128	0.7256	**0.7517**
PE	0.5318	0.5524	0.5247	0.4928	0.4719	**0.4125**
DB_{fuzzy}	**0.5231**	1.8827	0.7216	1.0029	1.5324	2.2612

Bảng 2.1: So sánh tổng hợp các độ đo đánh giá trên các tập dữ liệu IRIS, Breast và Glass.

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
Wine						
RI	0.7199	0.7418	0.7367	0.7556	0.7583	**0.7689**
F1	0.5868	0.6234	0.6166	0.6417	0.6459	**0.6578**
NMI	0.4295	0.4537	0.4469	0.4604	0.4671	**0.4955**
DB	**0.8365**	1.6125	0.8490	0.8367	0.9608	2.2867
PC	0.7213	0.7328	0.7394	0.7517	0.7596	**0.7729**
PE	0.5074	0.4931	0.4989	0.4728	0.4627	**0.4382**
DB_{fuzzy}	**0.6934**	1.5142	0.7249	0.8157	0.9345	2.1479
Soybean						
RI	0.8234	0.8408	0.7865	0.8441	0.8464	**0.8677**
F1	0.6819	0.7124	0.6081	0.7168	0.7194	**0.7571**
NMI	0.7002	0.7544	0.5454	0.7568	0.7671	**0.7784**
DB	3.7219	3.7587	4.7462	**2.4579**	2.7805	2.7581
PC	0.8124	0.8317	0.7829	0.8368	0.8441	**0.8619**
PE	0.3881	0.3627	0.4786	0.3579	0.3425	**0.3287**
DB_{fuzzy}	3.4972	3.6045	4.5289	**2.3841**	2.6974	2.6812
Thyroid						
RI	0.6394	0.6513	0.6655	0.7175	0.7258	**0.7649**
F1	0.6314	0.6338	0.6485	0.6872	0.6988	**0.7367**
NMI	0.3135	0.2968	0.3301	0.3635	0.3671	**0.5284**
DB	**2.0871**	2.4405	4.8259	2.6492	2.5697	2.1271
PC	0.6329	0.6463	0.6617	0.7018	0.7163	**0.7524**
PE	0.5871	0.5783	0.5692	0.5393	0.5284	**0.4189**
DB_{fuzzy}	**2.0471**	3.3829	4.6985	2.7987	2.5234	2.4481

Bảng 2.2: So sánh tổng hợp các độ đo đánh giá trên các tập dữ liệu Wine, Soybean và Thyroid.

Dựa trên kết quả thực nghiệm, ASSFBC nhất quán vượt trội về các chỉ số RI, F1, NMI cũng như các độ đo mờ như PC và PE trên hầu hết bộ dữ liệu. Hai độ đo này cho thấy ma trận độ phụ thuộc trở nên sắc nét hơn (PC cao) và mức độ mờ giảm đáng kể (PE thấp), phản ánh việc sửa biên giúp giảm sự nhập nhằng ở các điểm khó phân loại. Đối với các độ đo nội tại như DB và DB_fuzzy, giá trị của ASSFBC cao hơn so với FCM. Điều này xuất phát từ việc FCM chỉ tối ưu khoảng cách điểm–tâm cụm nên tạo ra các cụm rất tròn, gọn và đối xứng, khiến DB và DB_fuzzy thấp một cách tự nhiên. Tuy vậy, so với các thuật toán phân cụm bán giám sát mờ khác, ASSFBC thường đạt DB và DB_fuzzy tốt hơn hoặc cạnh tranh, cho thấy việc điều chỉnh vùng biên không làm suy giảm cấu trúc cụm mà vẫn đảm bảo ổn định.

Trên IRIS, Breast, Glass, Wine, Soybean, Thyroid, Data2 và Satellite, ASSFBC dẫn đầu ba thước đo theo nhãn; riêng Thyroid còn đạt DB thấp nhất, chứng tỏ việc sửa biên vẫn giữ được cụm gọn. Ở một vài tập dữ liệu như Glass, Wine, Data2, Satellite và Breast, DB cao hơn phản ánh hiện tượng cụm hơi dãn khi thuật toán ưu tiên sửa biên để khớp nhãn trong bối cảnh chồng lấn lớn, nhưng điều này không ảnh hưởng đáng kể đến chất lượng theo nhãn. Tổng thể, ASSFBC ổn định và mạnh về các thước đo theo nhãn, trong khi PC cao, PE thấp và DB_fuzzy cạnh tranh cho thấy sự cân bằng tốt giữa độ rõ của độ phụ thuộc và mức độ chặt cụm.

Algorithm Index	FCM	SSFCM	eFCM	AFFC(2008)	AFFC(2017)	ASSFBC
Data1						
RI	0.9523	0.9725	0.8941	**1.0000**	**1.0000**	**1.0000**
F1	0.9320	0.9624	0.8927	**1.0000**	**1.0000**	**1.0000**
NMI	0.9013	0.9317	0.7094	**1.0000**	**1.0000**	**1.0000**
DB	1.8451	2.3049	3.0185	**0.9282**	0.9347	0.9433
PC	0.9321	0.9340	0.8875	0.9687	0.9714	**0.9751**
PE	0.1184	0.1127	0.1548	0.0417	0.0372	**0.0294**
DB_{fuzzy}	1.0500	1.2800	1.6500	**0.4800**	0.5100	0.5200
Data2						
RI	0.7194	0.7426	0.7892	0.8131	0.8378	**0.8725**
F1	0.6694	0.6825	0.7024	0.7529	0.7391	**0.8148**
NMI	0.4895	0.5731	0.6152	0.6418	0.6184	**0.6761**
DB	**0.7724**	1.0081	1.5849	3.6928	4.5824	4.3715
PC	0.7216	0.7429	0.7714	0.7948	0.8162	**0.8467**
PE	0.5174	0.4831	0.4518	0.4236	0.4379	**0.3927**
DB_{fuzzy}	**0.6800**	0.9100	1.4800	3.4500	4.3150	4.1800
Satellite Image						
RI	0.7068	0.7309	0.7162	0.8278	0.8367	**0.8589**
F1	0.6847	0.6968	0.7061	0.7953	0.8075	**0.8368**
NMI	0.6314	0.6456	0.6593	0.7289	0.7494	**0.7879**
DB	0.5978	0.6081	**0.5506**	0.7371	0.7268	0.7498
PC	0.7051	0.7283	0.7152	0.8257	0.8354	**0.8609**
PE	0.5258	0.5006	0.4909	0.3952	0.3805	**0.3457**
DB_{fuzzy}	0.6108	0.6204	**0.5653**	0.7457	0.7324	0.7608

Bảng 2.3: So sánh tổng hợp các độ đo đánh giá trên các tập dữ liệu tự sinh và dữ liệu ảnh.

2.4 Kết luận

Luận án đã giới thiệu một cách tiếp cận 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 thách thức về sự không chắc chắn trong các vùng biên cụm bằng cách tích hợp học chủ động để cải thiện độ chính xác của việc gán thành viên tại các điểm quan trọng. Cách tiếp cận này nâng cao độ tin cậy của kết quả phân cụm và đề xuất một mô hình kết hợp giữa phân cụm mờ và học chủ động để tinh chỉnh vùng biên cụm.

Các kết quả thực nghiệm trên các tập dữ liệu chuẩn cho thấy phương pháp đề xuất ASSFBC đạt hiệu năng vượt trội so với các thuật toán phân cụm mờ truyền thống và các phương pháp bán giám sát khác, thể hiện qua các chỉ số RI, F1 và NMI luôn cao hơn. Các độ đo mờ như PC và PE cũng được cải thiện: PC tăng cho thấy ma trận độ phụ thuộc sắc nét hơn, còn PE giảm chứng tỏ mức độ mờ được thu hẹp ở các điểm khó phân loại. Đối với các chỉ số nội tại như DB và DB_fuzzy, ASSFBC thường thấp hơn hoặc cạnh tranh so với đa số phương pháp bán giám sát, dù không thấp bằng FCM do chỉ tối ưu khoảng cách điểm–tâm. Tuy vậy, phương pháp vẫn đạt sự cân bằng tốt giữa độ chặt cụm và độ chính xác theo nhãn.

Kết quả của chương này được công bố trong công trình **CT1, CT2**.

Chương 3

Đề xuất phương pháp phân cụm bán giám sát an toàn tích cực với cặp liên kết dựa vào biên cụm

3.1 Ý tưởng thuật toán

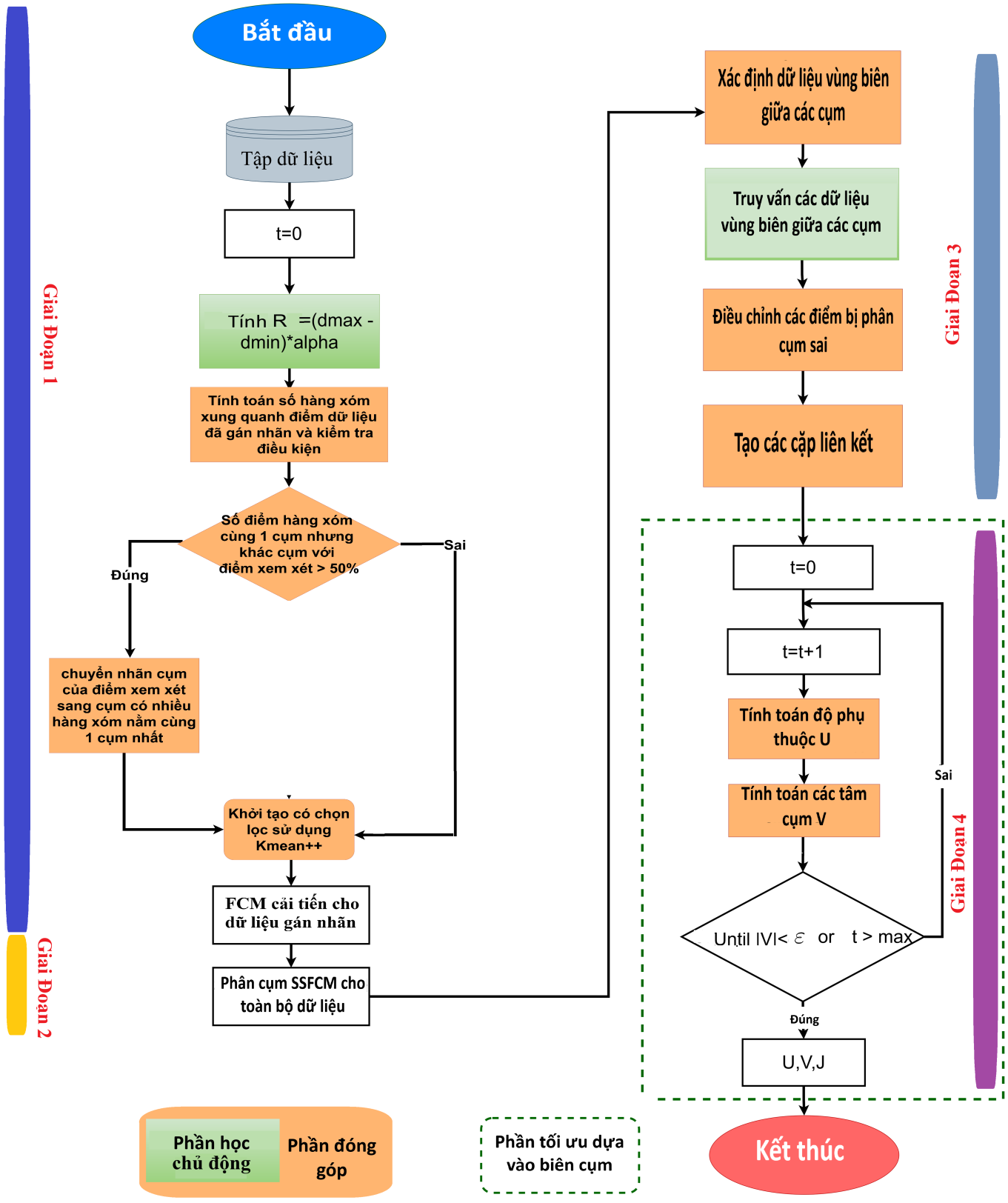
Mô hình kết hợp *phân cụm mờ* (biểu diễn bất định ở biên), *bán giám sát* (tận dụng ít nhãn/ràng buộc) và *học tích cực* (chọn truy vấn "đáng giá" tại vùng mơ hồ). Cốt lõi là dùng ràng buộc cặp ML/CL do *oracle* xác nhận ở biên cụm để hiệu chỉnh membership và cập nhật tâm, qua đó thu hẹp vùng chồng lấn, tăng tách biệt cụm và bền vững trước nhãn nhiễu. Quy trình lặp: FCM khởi tạo → phát hiện vùng biên → hỏi/áp dụng ML/CL → tối ưu lại phân cụm. Cách tích hợp này đạt độ chính xác cao khi nhãn khan hiếm và cấu trúc dữ liệu phức tạp, đồng thời giữ khả năng diễn giải tốt nhờ các ràng buộc miền.

3.2 Phương pháp hoạt động

Phương pháp **AS3FCPC** gồm bốn giai đoạn liên hoàn nhằm giảm bất định ở *vùng biên cụm* khi nhãn khan hiếm: (i) *khởi tạo học tích cực + FCM trên phần dữ liệu có nhãn*, (ii) *phân cụm bán giám sát mờ lặp* trên toàn bộ dữ liệu, (iii) *tinh chỉnh biên cụm bằng học tích cực và sinh ràng buộc cặp chọn lọc*, (iv) *tối ưu hoá lặp cho đến hội tụ*. Cốt lõi là *chỉ* truy vấn/điều chỉnh các điểm mơ hồ (membership sát nhau), sau đó *lan truyền* thông tin qua mục tiêu bán giám sát và ràng buộc ML/CL.

Thuật toán được triển khai qua bốn giai đoạn liên tiếp. Trong giai đoạn đầu, học tích cực được áp dụng lên phần dữ liệu có nhãn bằng cách xác định lân cận của mỗi điểm thông qua bán kính $R = (d_{\max} - d_{\min})\,\alpha$. Một điểm được xem là mơ hồ khi dưới 50% hàng xóm mang cùng nhãn; ngược lại, nếu đa số hàng xóm thuộc nhãn khác thì nhãn của điểm sẽ được hiệu chỉnh theo luật đa số. Sau bước lọc này, FCM được áp dụng để thu được U và V khởi tạo ổn định, làm nền cho giai đoạn bán giám sát.

Ở giai đoạn thứ hai, phân cụm bán giám sát mờ được thực hiện trên toàn bộ dữ liệu bằng cách sử dụng $\bar{U}$ từ giai đoạn trước như các neo giám sát. Các bước cập nhật U và V tuân theo nguyên lý FCM kết hợp với hạng phạt làm giảm độ lệch so với $\bar{U}$, giúp lan truyền thông tin nhãn đáng tin sang phần dữ liệu chưa được gán nhãn.



Hình 3.1: Mô hình phân cụm bán giám sát mờ an toàn tích cực với cặp liên kết dựa vào biên cụm (AS3FCPC).

Tiếp theo, giai đoạn thứ ba tập trung tinh chỉnh ranh giới cụm. Biên cụm được phát hiện dựa trên độ chênh membership nhỏ giữa hai cụm cạnh tranh; các điểm này được truy vấn oracle để xác nhận hoặc điều chỉnh nhãn, từ đó sinh ra các ràng buộc must-link và cannot-link một cách chọn lọc, chủ yếu tại các vùng chồng lấn — nơi mô

hình dễ nhầm lẫn nhất.

Cuối cùng, trong giai đoạn tối ưu và hội tụ, thuật toán tối thiểu hóa hàm mục tiêu tích hợp bao gồm thành phần mờ, thành phần bán giám sát và các ràng buộc cặp; quá trình lặp dừng lại khi sai lệch tâm cụm thoả $\|V^{(t)} - V^{(t-1)}\| \leq \epsilon$ hoặc khi đạt số vòng lặp tối đa.

Hàm mục tiêu

$$\begin{aligned}\min_{u,v}\ J(u,v) = &\sum_{i=1}^{N}\sum_{k=1}^{C} u_{ik}^2\, d_{ik}^2 + \alpha \sum_{i=1}^{L}\sum_{k=1}^{C}\Big(u_{ik} - \bar{u}_{ik}\Big)^2 \\ &+ \beta\left(\sum_{(x_i,x_j)\in\mathcal{M}}\sum_{k=1}^{C}\sum_{\substack{\ell=1\\ \ell\neq k}}^{C} u_{ik}\, u_{j\ell} + \sum_{(x_i,x_j)\in\mathcal{C}}\sum_{k=1}^{C} u_{ik}\, u_{jk}\right).\end{aligned} \tag{3.1}$$

ràng buộc $\sum_{k=1}^{C} u_{ik} = 1,\ u_{ik}, \bar{u}_{ik} \in [0,1]$ và $d_{ik}^2 = \|x_i - v_k\|^2$.
Ap dụng hệ số nhân Lagrange cho mỗi $k = 1, \ldots, N$, ta tính được:

$$\mathcal{L}(U,V,\lambda) = J(U,V) + \sum_{i=1}^{N}\lambda_i\Big(\sum_{k=1}^{C} u_{ik} - 1\Big) \tag{3.2}$$

Ta có độ phụ thuộc được tính như sau:

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

Cập nhật tâm cụm:

$$v_k = \frac{\displaystyle\sum_{i=1}^{N} u_{ik}^2 x_i}{\displaystyle\sum_{i=1}^{N} u_{ik}^2} \tag{3.4}$$

Phân cụm bán giám sát mờ an toàn tích cực với cặp liên kết dựa vào biên cụm - AS3FCPC

Thuật toán 2 AS3FCPC

Đầu vào:) $X = \{x_i\}_{i=1}^{N}$, số cụm C, số mẫu có nhãn $L(< N)$, neo $\bar{U}$, ngưỡng hội tụ ϵ, số vòng lặp tối đa $maxStep$, hệ số ràng buộc α.
Gđ.1 (Học tích cực + FCM trên phần có nhãn): phát hiện điểm mơ hồ bằng lân cận bán kính R, truy vấn/điều chỉnh nhãn; chạy FCM trên phần có nhãn $\Rightarrow$ thu U, V khởi tạo.
Gđ.2 (SSFCM lặp): cập nhật U, V trên toàn bộ dữ liệu với hạng bán giám sát $(u_{ik} - \bar{u}_{ik})^2 d_{ik}^2$; củng cố $\bar{U}$.
Gđ.3 (Biên + ràng buộc cặp): xác định biên qua $|u_{i(k_1)} - u_{i(k_2)}| < \varepsilon$; truy vấn oracle các điểm biên; sinh ràng buộc $\mathcal{M}$ (must-link), $\mathcal{C}$ (cannot-link) có chọn lọc.
Gđ.4 (Tối ưu hội tụ):
Đặt $t \leftarrow 0$.
Lặp các bước:
    $t \leftarrow t + 1$.
    Cập nhật ma trận độ phụ thuộc: tính u_{ik} theo 3.3
    Cập nhật tâm cụm: v_k theo 3.4
Cho đến khi $\|V^{(t)} - V^{(t-1)}\| \leq \epsilon$ **hoặc** $t > maxStep$

Độ phức tạp của thuật toán **AS3FCPC** :

$$\boxed{O\big(NC^2(I' + T) \; + \; N \log N\big)}$$

trong đó I' là số vòng lặp TS3FCM, T là vòng tinh chỉnh cuối, còn FCM trên tập có nhãn $L \ll N$ và điều chỉnh $N_q \ll N$ là không đáng kể ($O(LC^2I)$, $O(N_qC)$).
Vì $T \ll I'$, phần tăng chi phí do tinh chỉnh biên (học tích cực + ràng buộc cặp) là khiêm tốn so với lợi ích về độ chính xác/ổn định trên dữ liệu chồng lấn.
Tóm lại, chi phí tính toán chủ yếu giống TS3FCM về bậc, nhưng **AS3FCPC** đạt hiệu năng cao hơn nhờ sửa biên có định hướng.

3.3 Kết quả thực nghiệm

Hình 3.2: Hình ảnh mình họa phân đoạn ảnh xác định vùng ngập

Luận án đánh giá AS3FCPC trên ba nhóm dữ liệu: (i) các tập UCI đa dạng về số chiều và số lớp; (ii) ảnh vùng ngập có biên mềm và nhiễu cao; và (iii) dữ liệu tổng hợp được thiết kế với mức độ chồng lấn mạnh ở rìa cụm. Hiệu năng mô hình được đo bằng các chỉ số RI, F1, NMI (càng lớn càng tốt), DB (càng nhỏ càng tốt), cùng các

độ đo mờ PC, PE và DB_fuzzy nhằm phản ánh sắc nét của ma trận độ phụ thuộc và độ chặt cụm trong không gian mờ.

Method	IRIS	Wine	Australian	Spambase	Waveform	Breast	Haberman	Glass	Flood
RI									
FCM	0.812	0.632	0.498	0.514	0.531	0.691	0.471	0.788	0.819
SSFCM	0.864	0.671	0.503	0.527	0.546	0.736	0.503	0.796	0.832
CS3FCM	0.887	0.693	0.513	0.562	0.512	0.748	0.518	0.804	0.873
TS3FCM	0.899	0.703	0.518	0.551	0.528	0.755	0.524	0.816	0.884
AFFC (2017)	0.928	0.715	0.563	0.582	0.566	0.761	0.536	0.822	0.869
AS3FCPC	**0.955**	**0.742**	**0.598**	**0.618**	**0.589**	**0.772**	**0.559**	**0.833**	**0.917**
F1-score									
FCM	0.793	0.513	0.497	0.502	0.529	0.632	0.512	0.525	0.788
SSFCM	0.800	0.573	0.504	0.521	0.516	0.675	0.539	0.537	0.812
CS3FCM	0.829	0.623	0.514	0.542	0.507	0.714	0.530	0.558	0.861
TS3FCM	0.839	0.633	0.508	0.532	0.505	0.726	0.552	0.584	0.867
AFFC (2017)	0.874	0.645	0.552	0.573	0.560	0.739	**0.566**	0.609	0.860
AS3FCPC	**0.911**	**0.659**	**0.576**	**0.597**	**0.574**	**0.761**	0.559	**0.618**	**0.894**
NMI									
FCM	0.658	0.362	0.442	0.457	0.486	0.413	0.453	0.603	0.732
SSFCM	0.702	0.398	0.466	0.492	0.502	0.452	0.468	0.614	0.746
CS3FCM	0.728	0.422	0.473	0.511	0.491	0.474	0.482	0.623	0.753
TS3FCM	0.740	0.434	0.478	0.522	0.499	0.485	0.493	0.647	0.766
AFFC (2017)	0.801	0.451	0.528	0.547	0.538	0.501	0.508	0.662	0.854
AS3FCPC	**0.833**	**0.472**	**0.563**	**0.594**	**0.563**	**0.527**	**0.517**	**0.684**	**0.883**
DB									
FCM	**0.902**	**0.826**	**1.143**	**1.609**	3.513	**0.731**	1.258	**0.629**	**1.252**
SSFCM	1.348	1.513	1.234	1.943	4.552	0.843	1.372	2.318	1.944
CS3FCM	2.785	3.028	4.313	4.034	5.684	1.932	**1.198**	1.347	3.524
TS3FCM	2.864	3.194	3.512	3.613	6.228	1.884	1.330	1.436	3.628
AFFC (2017)	1.349	0.983	2.816	2.098	2.432	0.942	1.417	2.334	2.528
AS3FCPC	1.222	0.937	2.275	1.843	**2.118**	0.963	1.528	2.485	1.327

Bảng 3.1: So sánh tổng hợp hiệu năng các phương pháp theo các độ đo RI, F1, NMI và DB

Kết quả thực nghiệm cho thấy AS3FCPC nhất quán dẫn đầu RI, F1 và NMI trên hầu hết các bộ dữ liệu, vượt các phương pháp FCM, SSFCM, CS3FCM, TS3FCM và AFFC (2017). Mức cải thiện nổi bật trên IRIS, Waveform, Spambase, Glass (chồng lấn mạnh) và bộ ảnh Flood images (nhiễu cao). Các độ đo mờ PC và PE cũng cho thấy độ phụ thuộc của AS3FCPC sắc nét hơn: PC cao hơn và PE thấp hơn phần lớn đối thủ, khẳng định mô hình giảm mạnh mức độ mơ hồ tại các điểm biên. Về chỉ số nội tại DB và DB_fuzzy, phương pháp thường đạt mức tốt nhất so với các thuật toán bán giám sát mờ khác hoặc cạnh tranh, chỉ thấp hơn so với FCM.

Nhìn chung, cơ chế sửa biên chủ động kết hợp ràng buộc cặp chọn lọc giúp giảm nhầm lẫn giữa các cụm lân cận, duy trì cấu trúc phân tách và nâng cao các thước đo ngoại sinh. Trong các kịch bản yêu cầu khớp nhãn tin cậy trong điều kiện nhãn hạn chế và ranh giới mơ hồ (ví dụ phân tích ảnh viễn thám), AS3FCPC cho thấy hiệu năng ổn định và mạnh.

Method	IRIS	Wine	Australian	Spambase	Waveform	Breast	Haberman	Glass	Flood
PC									
FCM	0.780	0.720	0.611	0.605	0.626	0.741	0.591	0.691	0.705
SSFCM	0.811	0.736	0.629	0.618	0.640	0.765	0.615	0.706	0.728
CS3FCM	0.835	0.748	0.641	0.646	0.653	0.776	0.631	0.715	0.741
TS3FCM	0.851	0.761	0.655	0.659	0.670	0.782	0.646	0.725	0.755
AFFC (2017)	0.876	0.772	0.685	0.691	0.703	0.791	0.661	0.736	0.768
AS3FCPC	**0.910**	**0.795**	**0.706**	**0.721**	**0.726**	**0.805**	**0.686**	**0.761**	**0.811**
PE									
FCM	0.422	0.511	0.612	0.629	0.656	0.487	0.668	0.533	0.526
SSFCM	0.393	0.496	0.598	0.613	0.639	0.462	0.643	0.549	0.502
CS3FCM	0.371	0.487	0.583	0.598	0.617	0.454	0.628	0.523	0.476
TS3FCM	0.355	0.471	0.572	0.583	0.598	0.447	0.612	0.506	0.459
AFFC (2017)	0.329	0.460	0.556	0.566	0.584	0.440	0.589	0.492	0.433
AS3FCPC	**0.303**	**0.446**	**0.526**	**0.540**	**0.556**	**0.427**	**0.566**	**0.469**	**0.406**
DB_{fuzzy}									
FCM	**0.921**	**0.690**	**1.210**	**1.872**	2.413	**0.713**	1.246	**0.613**	**1.243**
SSFCM	1.430	1.513	1.384	2.105	4.563	1.028	1.383	2.318	1.973
CS3FCM	2.510	3.046	4.383	4.108	5.725	1.948	**1.213**	1.366	3.548
TS3FCM	2.620	3.216	3.546	3.658	6.266	1.918	1.343	1.453	3.673
AFFC (2017)	1.320	0.983	2.846	2.133	2.485	0.952	1.430	1.538	2.563
AS3FCPC	1.270	1.715	2.316	1.875	**2.158**	0.972	1.543	2.528	1.363

Bảng 3.2: So sánh tổng hợp hiệu năng các phương pháp theo các chỉ số PC, PE và DB_{fuzzy}

3.4 KẾT LUẬN

Tóm lại, AS3FCPC kết hợp phân cụm mờ, bán giám sát và học tích cực để xử lý hiệu quả các bài toán có ít nhãn và ranh giới cụm không rõ ràng. Thuật toán được khởi tạo một cách ổn định thông qua cơ chế truy vấn có chọn lọc các điểm dữ liệu bất định và lan truyền các ràng buộc cặp, qua đó định hình cấu trúc cụm ngay từ giai đoạn đầu. Trong các bước tối ưu tiếp theo, mô hình tiếp tục điều chỉnh độ phụ thuộc và cập nhật tâm cụm dưới tác động của các ràng buộc must-link/cannot-link, giúp cố định biên, giảm nhầm lẫn giữa các cụm lân cận và tăng mức độ tách biệt, đặc biệt hiệu quả trong các vùng có mức chồng lấn cao. Kết quả thực nghiệm trên nhiều bộ dữ liệu cho thấy các chỉ số ngoại sinh như RI, F1 và NMI của AS3FCPC luôn đạt giá trị cao hơn so với các phương pháp so sánh, trong khi các độ đo mờ PC và PE lần lượt đạt mức cao, phản ánh phân hoạch mờ sắc nét và mức độ bất định được kiểm soát tốt. Đồng thời, các chỉ số nội tại như DB và DB_fuzzy nhìn chung được cải thiện hoặc duy trì ở mức tốt so với các thuật toán phân cụm bán giám sát mờ khác, cho thấy cấu trúc cụm thu được vừa phù hợp với nhãn thực, vừa đảm bảo tính cô đọng và ổn định. Nhờ những đặc điểm này, AS3FCPC đặc biệt phù hợp cho các ứng dụng thực tiễn như phân vùng ngập trong ảnh vệ tinh và đã được trình bày trong công trình CT3.

KẾT LUẬN

A. Những kết quả chính của luận án

Luận án nghiên cứu phân cụm bán giám sát mờ, tổng hợp các nền tảng về tập mờ, FCM, bán giám sát và học chủ động, học an toàn, từ đó đề xuất hai phương pháp chính:

1. **ASSFBC**: tinh chỉnh vùng biên cụm bằng cơ chế truy vấn chủ động các điểm quan trọng, nâng độ chính xác gán thành viên và độ tin cậy phân cụm.
2. **AS3FCPC**: kết hợp phân cụm mờ, học bán giám sát an toàn và học chủ động; khai thác đồng thời dữ liệu có/không nhãn ngay từ khởi tạo, dùng ràng buộc cặp để cố định tâm và ranh giới cụm.

B. Những đóng góp mới của luận án

- Tích hợp học chủ động vào phân cụm bán giám sát mờ theo hướng *tập trung vùng biên*, cải thiện chất lượng phân cụm với ngân sách nhãn nhỏ.
- Mô hình **AS3FCPC** dung hoà mờ–bán giám sát an toàn–học chủ động, tận dụng ràng buộc cặp để tăng chính xác và ổn định.
- Xác thực thực nghiệm toàn diện, cho kết quả vượt trội về độ chính xác, độ tin cậy và hiệu quả mô hình.

C. Hướng phát triển tiếp theo

Các hướng nghiên cứu tiếp theo tập trung vào việc tối ưu hoá ASSFBC/AS3FCPC cho dữ liệu lớn, đồng thời khai thác ngữ cảnh theo miền (như y tế, viễn thám, mạng xã hội) nhằm tăng tính thích ứng của mô hình. Bên cạnh đó, cần phát triển chiến lược học chủ động động phù hợp với luồng dữ liệu trực tiếp và môi trường nhiễu, cũng như kết hợp các kỹ thuật học sâu để xây dựng mô hình phân cụm bán giám sát mờ chủ động mạnh hơn.

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN
LIÊN QUAN ĐẾN LUẬN ÁN

1. **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.

2. **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

3. **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