BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ



Nguyễn Thanh Sơn

NGHIÊN CỨU PHÁT TRIỂN MỘT SỐ MÔ HÌNH PHÂN CỤM BÁN GIÁM SÁT DỰA TRÊN MẠNG NƠRON MIN-MAX MỜ

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:

1. Người hướng dẫn 1: PGS. TS. Lê Bá Dũng, Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam
2. Người hướng dẫn 2: TS. Vũ Đình Minh, Trường Công nghệ thông tin và Truyền thông, Đại học Công nghiệp Hà Nội.

Phản biện 1: PGS.TS. Nguyễn Văn Núi, Trường Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên

Phản biện 2: PGS.TS. Phạm Văn Hải, Trường Công nghệ thông tin và Truyền thông, Đại học Bách khoa Hà Nội

Phản biện 3: PGS.TS. Trần Thị Ngân, Trường Quốc tế, Đại học Quốc gia Hà Nội

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 14 giờ 00, ngày 08 tháng 9 năm 2026

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

1. Lý do chọn đề tài

Trong những năm gần đây, phân cụm bán giám sát thu hút sự quan tâm ngày càng lớn của cộng đồng nghiên cứu, xuất phát từ nhu cầu khai thác hiệu quả các tập dữ liệu có quy mô lớn, độ phức tạp cao trong bối cảnh thông tin nhãn khan hiếm và chi phí gán nhãn ngày càng đắt đỏ. Việc kết hợp đồng thời dữ liệu có nhãn và chưa có nhãn cho phép mô hình tận dụng tốt hơn cấu trúc tiềm ẩn của dữ liệu, từ đó cải thiện chất lượng phân cụm so với các phương pháp không giám sát thuần túy. Tuy nhiên, các phương pháp hiện có thường gặp hạn chế khi dữ liệu có nhiễu, phân bố không đồng đều, ranh giới cụm không rõ ràng hoặc khi khả năng diễn giải của mô hình là yêu cầu quan trọng, đặc biệt trong các bài toán hỗ trợ ra quyết định và y sinh.

Phân cụm mờ cho phép mỗi đối tượng thuộc về nhiều cụm với các mức độ khác nhau, nên phù hợp với dữ liệu có vùng chồng lấn và ranh giới mơ hồ. Trong số các mô hình mờ, mạng nơron Min-Max mờ (Fuzzy Min-Max Neural Network - FMN) được quan tâm nhờ khả năng biểu diễn dữ liệu bằng các siêu hộp mờ, học gia tăng và kết xuất luật IF-THEN có tính diễn giải cao. Một số công trình đã mở rộng FMN theo hướng học bán giám sát như GFMM, RFMN, SS-FMM, SSL-FMM, SCFMN và MSCFMN. Mặc dù vậy, các mô hình FMN bán giám sát hiện có vẫn còn ba hạn chế chính:

- Thứ nhất, quá trình phân cụm phụ thuộc nhiều vào chất lượng của các mẫu hoặc siêu hộp mang nhãn ban đầu, trong khi các đối tượng này chưa được đánh giá đầy đủ về mức độ đại diện và độ tin cậy.

- Thứ hai, cơ chế suy luận nhãn chủ yếu dựa trên quan hệ lân cận (độ thuộc) giữa các siêu hộp, chưa xem xét đầy đủ độ thuần nhất, mật độ phân bố và vai trò đại diện của siêu hộp, dẫn đến nguy cơ lan truyền lỗi nhãn.

- Thứ ba, các mô hình còn gặp khó khăn với các mẫu ở vùng biên hoặc vùng chồng lấn, nơi mức độ không chắc chắn cao ảnh hưởng đến quá trình suy luận và lan truyền nhãn.

Xuất phát từ các vấn đề trên, luận án nghiên cứu các hướng tiếp cận mới cho bài toán phân cụm bán giám sát dựa trên mạng nơron Min-Max

mờ, nhằm cải thiện cơ chế suy luận nhãn, giảm lan truyền lỗi và nâng cao chất lượng phân cụm trong điều kiện thông tin giám sát hạn chế.

2. Mục tiêu nghiên cứu

Mục tiêu tổng quát của luận án là đề xuất và phát triển các mô hình phân cụm bán giám sát dựa trên mạng nơron Min-Max mờ nhằm nâng cao độ tin cậy của quá trình suy luận nhãn, hạn chế lan truyền lỗi và cải thiện chất lượng phân cụm khi thông tin giám sát hạn chế. Các mục tiêu cụ thể gồm:

- Nghiên cứu cơ chế đánh giá chất lượng của các mẫu hoặc siêu hộp mang nhãn ban đầu, từ đó lựa chọn các đối tượng có mức độ đại diện và độ tin cậy cao làm nguồn thông tin giám sát.

- Đề xuất cơ chế suy luận nhãn mới cho các siêu hộp chưa biết nhãn, xem xét thêm độ thuần nhất, mật độ phân bố dữ liệu và vai trò đại diện của siêu hộp trong từng cụm.

- Xây dựng các chiến lược bỏ phiếu nhãn dựa trên mẫu và/hoặc siêu hộp nhằm tăng độ ổn định của quyết định gán nhãn khi dữ liệu có mật độ không đồng đều, cụm mất cân bằng hoặc ranh giới phức tạp.

- Thực nghiệm và đánh giá các mô hình đề xuất trên các tập dữ liệu chuẩn và dữ liệu thực tế bằng các chỉ số đánh giá phù hợp.

3. Đối tượng, phạm vi và nội dung nghiên cứu

Đối tượng nghiên cứu là các phương pháp phân cụm bán giám sát dựa trên mạng nơron FMN, với trọng tâm là (i) cơ chế hình thành và hiệu chỉnh siêu hộp mờ trong quá trình học, (ii) cơ chế suy luận/lan truyền nhãn trong điều kiện nhãn khan hiếm, và (iii) chiến lược lựa chọn hạt giống nhằm dẫn dắt quá trình học bán giám sát. Luận án tập trung vào ba nội dung: cơ chế suy luận nhãn dựa trên bỏ phiếu từ các siêu hộp lân cận (FMN-Voting); cơ chế suy luận nhãn dựa trên các mẫu mang nhãn lân cận (FMN-SAL); và chiến lược lựa chọn hạt giống chủ động/thích nghi (FA-Seed). Phạm vi đánh giá gồm các tập dữ liệu chuẩn (benchmark) từ UCI và CS, và bộ dữ liệu y tế ILPD về chẩn đoán bệnh gan.

4. Phương pháp nghiên cứu

Luận án sử dụng phương pháp nghiên cứu lý thuyết kết hợp với thực nghiệm. Trên cơ sở khảo sát, phân tích các nghiên cứu về phân cụm bán giám sát và mạng nơron Min-Max mờ, luận án đề xuất các cải tiến về cơ

chế suy luận nhãn và lựa chọn hạt giống; các mô hình đề xuất được cài đặt, thực nghiệm và đánh giá trên các tập dữ liệu được lựa chọn thông qua các độ đo phù hợp.

5. Ý nghĩa khoa học và thực tiễn

Về khoa học, luận án bổ sung các cơ chế đánh giá và sử dụng thông tin cấu trúc của siêu hộp vào quá trình suy luận nhãn và lựa chọn hạt giống của FMN bán giám sát. Về thực tiễn, việc kiểm chứng trên bộ dữ liệu xét nghiệm máu ILPD cho thấy khả năng áp dụng các mô hình trong bài toán hỗ trợ chẩn đoán khi dữ liệu có nhãn hạn chế.

6. Những đóng góp mới của luận án

- Đề xuất mô hình FMN-Voting cho phân cụm bán giám sát dựa trên mạng FMN với cơ chế bỏ phiếu đa số nhằm tăng độ ổn định suy luận nhãn.
- Đề xuất mô hình FMN-SAL với chiến lược bỏ phiếu dựa trên k mẫu mang nhãn lân cận, giảm ảnh hưởng của các siêu hộp mất cân bằng/độ lệch tâm lớn/chỉ số sử dụng thấp trong quá trình lan truyền nhãn.
- Đề xuất mô hình FA-Seed với chiến lược lựa chọn hạt giống chủ động/thích nghi dựa trên phân hoạch siêu hộp và đánh giá độ tin cậy, nâng cao chất lượng tập hạt giống và cải thiện hiệu quả học bán giám sát khi nhãn hạn chế.
- Thiết kế và thực hiện thực nghiệm trên các tập dữ liệu benchmark và bộ dữ liệu y tế về các chỉ số sinh hóa máu nhằm đánh giá hiệu năng, độ ổn định, khả năng khai thác dữ liệu ít nhãn và chi phí tính toán của các mô hình đề xuất.

7. Bố cục của luận án

Ngoài phần Mở đầu và Kết luận, luận án gồm ba chương. Chương 1 trình bày tổng quan về phân cụm bán giám sát và mạng nơron Min-Max mờ, phân tích các vấn đề còn tồn tại và xác định hướng nghiên cứu. Chương 2 trình bày hai mô hình FMN-Voting và FMN-SAL, gồm thuật toán, phân tích độ phức tạp và đánh giá thực nghiệm. Chương 3 trình bày mô hình FA-Seed với chiến lược lựa chọn hạt giống dựa trên cấu trúc siêu hộp, kết hợp thông tin phản hồi từ chuyên gia, cùng các thực nghiệm đánh giá.

CHƯƠNG 1. TỔNG QUAN VỀ PHÂN CỤM BÁN GIÁM SÁT SỬ DỤNG MẠNG NƠRON MIN-MAX MỜ

1.1. Tổng quan về bài toán phân cụm dữ liệu và ứng dụng

Phân cụm là kỹ thuật quan trọng trong khai phá dữ liệu, thuộc nhóm học không giám sát, nhằm nhóm các đối tượng tương đồng vào cùng một cụm. Phân cụm rõ gán mỗi điểm vào duy nhất một cụm, còn phân cụm mờ cho phép một điểm thuộc nhiều cụm qua giá trị độ thuộc trong khoảng [0, 1], nên phù hợp hơn với dữ liệu có ranh giới không rõ ràng hoặc chồng lấn. Phân cụm bán giám sát mờ mở rộng phân cụm mờ bằng cách khai thác một phần thông tin biết trước (nhãn, ràng buộc must-link/cannot-link hoặc độ thuộc xác định trước) để định hướng quá trình phân cụm.

Các hướng tiếp cận tiêu biểu gồm: (1) phân cụm bán giám sát mờ dựa trên mẫu hạt giống; (2) dựa trên ràng buộc cặp; (3) dựa trên mô hình nguyên mẫu; và (4) dựa trên mạng nơron mờ. Phương pháp dựa trên hạt giống định hướng tốt quá trình học nhưng phụ thuộc mạnh vào chất lượng hạt giống; phương pháp ràng buộc linh hoạt nhưng gặp khó khăn về chi phí và tính nhất quán của ràng buộc; phương pháp nguyên mẫu có khả năng giải thích nhưng phụ thuộc vào chất lượng nguyên mẫu. So với các hướng trên, FMN biểu diễn dữ liệu bằng các siêu hộp mờ, mô hình hóa linh hoạt các vùng lõi, vùng biên và vùng chồng lấn, có cơ chế học gia tăng và khả năng diễn giải tốt; vì vậy luận án lựa chọn FMN làm nền tảng phát triển các mô hình FMN-Voting, FMN-SAL và FA-Seed.

1.2. Phân cụm bán giám sát dựa trên mạng nơron Min-Max mờ

1.2.1. Siêu hộp mờ và hàm thuộc

Mô hình FMN do Simpson đề xuất cho phân lớp (1992) và mở rộng cho phân cụm (1993). Tri thức trong FMN được biểu diễn bằng các siêu hộp mờ; mỗi siêu hộp B_j được xác định bởi cặp vectơ biên dưới V_j và biên trên W_j:

$$B_j = \{x \in \mathbb{R}^n \mid v_{ji} \le x_i \le w_{ji},\ \forall i = 1, \ldots, n\} \tag{1.2}$$

Mức độ thuộc của điểm dữ liệu x_r vào siêu hộp B_j được tính theo hàm thuộc:

$$\mu(x_r, B_j) = \frac{1}{n} \sum_{i=1}^{n} \left[1 - f(x_{ri} - w_{ji}, \gamma) - f(v_{ji} - x_{ri}, \gamma)\right] \tag{1.3}$$

trong đó γ là tham số mờ điều chỉnh tốc độ suy giảm độ thuộc khi điểm nằm ngoài siêu hộp và $f(\cdot,\cdot)$ là hàm hai tham số:

$$f(s,z) = \begin{cases} 1, & sz > 1, \\ sz, & 0 \le sz \le 1, \\ 0, & sz < 0. \end{cases} \tag{1.4}$$

1.2.2. Phân hoạch dữ liệu với các siêu hộp mờ

Quá trình huấn luyện gồm ba thao tác cơ bản: mở rộng, kiểm tra chồng lấn và điều chỉnh chồng lấn (co lại). Khi mở rộng B_j để bao phủ x_r, các biên được cập nhật $v_{ji} \leftarrow \min(v_{ji}, x_{ri})$, $w_{ji} \leftarrow \max(w_{ji}, x_{ri})$; phép mở rộng chỉ được chấp nhận khi kích thước siêu hộp không vượt quá δ^{max}. Nếu hai siêu hộp khác lớp chồng lấn, FMN xét bốn trường hợp chồng lấn cơ bản và điều chỉnh biên trên chiều gây chồng lấn với mức thay đổi nhỏ nhất. Việc gán mẫu tuân theo các quy tắc bao chứa ($\mu(x, B_j) = 1$), mở rộng và tạo mới siêu hộp $B_{new} = (x, x)$.

1.2.3. Cấu trúc mạng và thuật toán học FMN

Với học không giám sát, mạng FMN gồm hai lớp: lớp vào F_A (n nơron) và lớp siêu hộp F_B. Với học có giám sát, mạng gồm ba lớp F_A, F_B và lớp ra F_C; liên kết giữa F_B và F_C là ma trận nhị phân U với $u_{jg} = 1$ nếu $\ell(B_j) = g$ và bằng 0 trong trường hợp ngược lại, đầu ra lớp g là $c_g = \max_j(b_j u_{jg})$. Luận án trình bày ba thuật toán học FMN: có giám sát, không giám sát và bán giám sát; trong đó thuật toán bán giám sát gán nhãn cho siêu hộp mới tạo từ mẫu chưa nhãn dựa trên siêu hộp gần nhất thỏa mãn ngưỡng β.

1.2.4. Một số vấn đề cần tiếp tục nghiên cứu

Nhiều hướng cải tiến FMN đã được đề xuất về cấu trúc mạng, thuật toán học, xử lý chồng lấn, loại bỏ siêu hộp có độ tin cậy thấp, cùng các mô hình bán giám sát GFMM, RFMN, SS-FMM, SSL-FMM, SCFMN và MSCFMN. Tuy vậy, các phương pháp hiện có còn hạn chế: (i) xem các siêu hộp có vai trò tương đương dù chất lượng khác nhau; (ii) siêu hộp có độ tin cậy thấp vẫn tham gia lan truyền nhãn; (iii) xử lý chồng lấn dựa trên quy tắc heuristic áp dụng đồng đều; (iv) chưa tách bạch học cấu trúc dữ liệu và học nhãn.

1.3. Hạn chế, vấn đề đặt ra và định hướng nghiên cứu

Từ đó, luận án định hướng: đánh giá chất lượng và độ tin cậy của siêu hộp để lựa chọn các siêu hộp có tính đại diện cao tham gia lan truyền nhãn; thiết kế cơ chế học bán giám sát có kiểm soát; và tách biệt giai đoạn học cấu trúc dữ liệu với giai đoạn học nhãn.

1.4. Kết luận chương 1

Chương 1 đã trình bày cơ sở lý thuyết về phân cụm bán giám sát mờ với trọng tâm là mạng nơron FMN (siêu hộp mờ, hàm thuộc, cấu trúc mạng, thuật toán học), tổng hợp các nghiên cứu từ GFMM, RFMN đến các mô hình FMN bán giám sát hiện đại và chỉ ra những hạn chế trong khai thác dữ liệu chưa gán nhãn, đánh giá độ tin cậy và kiểm soát lan truyền lỗi nhãn. Một phần nội dung chương đã được công bố trong công trình [CT1].

CHƯƠNG 2. PHÂN CỤM BÁN GIÁM SÁT DỰA TRÊN MẠNG NƠRON MIN-MAX MỜ VỚI BỎ PHIẾU ĐA SỐ

2.1. Học bán giám sát dựa trên siêu hộp mờ

FMN phân hoạch không gian dữ liệu thông qua tập siêu hộp mờ, cho phép biểu diễn một cụm bằng một hoặc nhiều vùng con, do đó mô hình hóa được các cụm không lồi, có hình dạng bất kỳ và mật độ không đồng đều. Để đánh giá mật độ và phân phối dữ liệu trong siêu hộp, luận án sử dụng độ lệch tâm $e(B_j)$ giữa tâm hình học $g(B_j)$ và tâm dữ liệu $d(B_j)$:

$$e(B_j) = \|g(B_j) - d(B_j)\|_2,$$
$$g(B_j) = \frac{V_j + W_j}{2}, \qquad d(B_j) = \frac{1}{|T_j|}\sum_{r \in T_j} x_r \tag{2.1)-(2.2}$$

Siêu hộp được xem là thuần khiết nếu độ lệch tâm thỏa mãn điều kiện (2.3); chỉ số sử dụng $c(B_j)$ phản ánh mức độ siêu hộp được khai thác khi biểu diễn và phân loại dữ liệu (2.4), trong đó U_j là tỷ lệ mẫu thuộc B_j trong nhóm chứa nó và A_j là tỷ lệ mẫu được phân loại chính xác bởi B_j:

$$e(B_j) \le \varepsilon, \quad \varepsilon > 0 \tag{2.3}$$

$$c(B_j) = (1 - \varphi)\, U_j + \varphi\, A_j \tag{2.4}$$

Trong FMN, siêu hộp là đơn vị biểu diễn cơ bản: các điểm nằm hoàn toàn trong một siêu hộp có độ thuộc đầy đủ và được xem là tương đương. Vì vậy, cơ chế giám sát được nâng từ mức điểm dữ liệu lên mức siêu hộp:

khi một điểm đại diện được gán nhãn, nhãn được lan truyền nhất quán tới toàn bộ các điểm có độ thuộc đầy đủ với siêu hộp đó. Cách tổ chức này giúp giảm chi phí gán nhãn, hạn chế ảnh hưởng của nhiễu và các điểm biên, đồng thời tăng khả năng diễn giải. Mỗi siêu hộp có nhãn $B_j = (V_j, W_j)$ với $\ell(B_j) = c$ có thể chuyển trực tiếp thành luật R_j: IF $x_1 \in [v_{j1}, w_{j1}]$ AND … AND $x_n \in [v_{jn}, w_{jn}]$ THEN $y = c$, với độ tin cậy được đánh giá qua số mẫu bao phủ $|T_j|$, độ lệch tâm $e(B_j)$ và chỉ số sử dụng $c(B_j)$.

2.2. Học bán giám sát dựa trên bỏ phiếu đa số trong FMN

Phương pháp bỏ phiếu đa số dựa trên mẫu (như kNN) giả định các mẫu gần nhau có xu hướng cùng lớp; nhãn của mẫu chưa biết được chọn là nhãn chiếm ưu thế trong k lân cận gần nhất. Kỹ thuật này trực quan và thích nghi tốt với phân bố phi tuyến, nhưng phụ thuộc mạnh vào chất lượng dữ liệu, nhạy với tham số k và kém ổn định khi bỏ phiếu trực tiếp trên các điểm riêng lẻ ở vùng chồng lấn. Trong FMN, cơ chế bỏ phiếu có thể được định nghĩa dựa trên (1) các đại diện hình học của siêu hộp - hướng tiếp cận của FMN-Voting [CT3], phát triển từ MSCFMN; hoặc (2) các mẫu dữ liệu lân cận mang nhãn - hướng tiếp cận của FMN-SAL [CT2], kế thừa từ SS-FMM.

2.3. Mô hình FMN-Voting

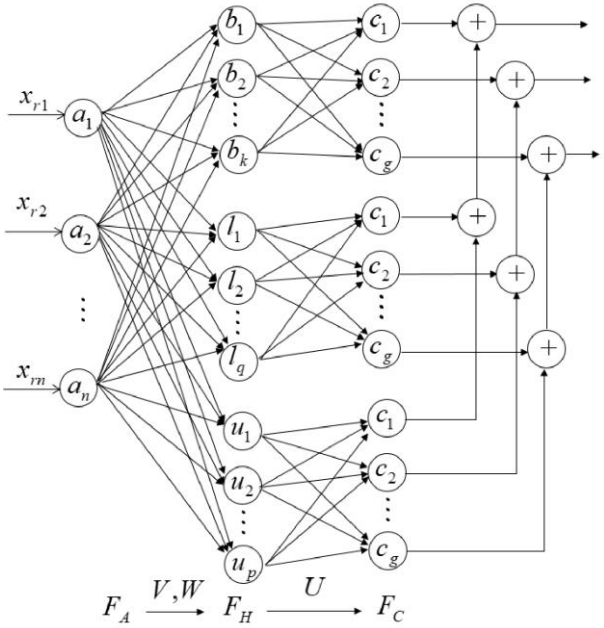
2.3.1. Mô tả bài toán và ý tưởng

Cho tập dữ liệu với thông tin nhãn hạn chế và tập siêu hộp được hình thành bởi FMN, bài toán đặt ra là khai thác cấu trúc siêu hộp để tự xác định thông tin bổ trợ, sau đó dùng thông tin này để lan truyền nhãn tới các vùng chưa xác định nhãn. Các mô hình SS-FMM, SCFMN, MSCFMN phụ thuộc vào ngưỡng chất lượng toàn cục β và cần nhiều vòng duyệt để lan truyền nhãn. Ý tưởng của FMN-Voting là chuyển sang cơ chế suy luận nhãn cục bộ: nhãn của siêu hộp mới không được quyết định bởi một siêu hộp dẫn dắt đơn lẻ mà bởi bỏ phiếu đa số từ các siêu hộp lân cận đã có nhãn, dựa trên độ thuộc. Thuật toán gồm hai giai đoạn: (1) xây dựng tập hạt giống - học không giám sát để tạo tập siêu hộp và gán nhãn bổ trợ cho các mẫu có độ thuộc đầy đủ; (2) huấn luyện bán giám sát với bỏ phiếu đa số dựa trên siêu hộp.

2.3.2. Kiến trúc mạng nơron FMN-Voting

Kiến trúc gồm ba lớp F_A, F_H, F_C (Hình 2.3). Lớp siêu hộp F_H được tổ chức thành các nhóm siêu hộp đã có nhãn, trung gian và chưa gán nhãn; lớp F_C gồm g nơron đại diện cho các cụm, tín hiệu từ nhiều siêu hộp được tổng hợp qua nút (+). Với mẫu chưa có nhãn x_r không thể được bao phủ hoặc mở rộng bởi siêu hộp phù hợp, mô hình xác định tập lân cận $N(x_r)$ gồm các siêu hộp có nhãn có độ thuộc cao nhất và thỏa mãn ngưỡng β; mỗi siêu hộp trong tập lân cận đóng góp một phiếu:

$$\hat{\ell}(x_r) = \arg\max_{c \in \mathcal{C}} \sum_{L_i \in N(x_r)} \mathbb{I}\big(\ell(L_i) = c\big) \tag{2.9}$$



Hình 2.3: *Kiến trúc mạng nơron FMN-Voting*

2.3.3. Thuật toán học trong FMN-Voting

Giai đoạn 1 (Thuật toán 2.1) thực hiện học không giám sát để xây dựng tập siêu hộp B, sau đó gán nhãn bổ trợ cho các mẫu có độ thuộc đầy đủ đối với một siêu hộp; độ phức tạp $T_1 = O(nMK)$. Giai đoạn 2 (Thuật toán 2.2) xây dựng tập siêu hộp có nhãn L rồi suy luận nhãn bằng bỏ phiếu đa số trên $k = 2C + 1$ siêu hộp lân cận, trong đó C là số lớp/cụm; k là số lẻ, lớn

hơn 2C để tạo tập lân cận đủ lớn và hạn chế quyết định phụ thuộc vào quá ít siêu hộp. Khi số siêu hộp có nhãn khả dụng nhỏ hơn 2C + 1, mô hình dùng số lân cận thực tế:

$$N(x_r) = \underset{\min(2C+1,\,|\mathcal{L}|)}{\arg\operatorname{top}} \big\{\mu(x_r, L_j) \mid L_j \in \mathcal{L},\ \mu(x_r, L_j) \geq \beta\big\} \tag{2.17}$$

$$NoC_\ell = \sum_{L_j \in N(x_r)} \mathbb{I}\big(\ell(L_j) = \ell\big),\ \ \ell = 1, \ldots, C;$$

$$\ell(L_{\text{new}}) = \arg\max_{\ell} NoC_\ell \tag{2.18)-(2.19}$$

Thuật toán 2.2 Huấn luyện bán giám sát với bỏ phiếu đa số dựa trên siêu hộp

Input: $\mathcal{B}$; $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$; $\delta_2^{\max}$; γ; β; $R_{\max}$;
Output: $\mathcal{L}$; $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$; U;

$\mathcal{L} \leftarrow \{B_j \in \mathcal{B} \mid \ell(B_j) \neq 0\}$; $X^1 \leftarrow \varnothing$; $r \leftarrow 0$;
repeat
  Chọn dữ liệu vào $(x_r, \ell_{x_r}) \in X$;
  if $\exists L_j \in \mathcal{L}$ sao cho $\mu(x_r, L_j) = 1$ **then**
    if $\ell_{x_r} \neq 0$ **then**
      $\ell_{x_r} \leftarrow \ell(L_j)$; $X^1 \leftarrow X^1 \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    end if
  else
    if $\exists L_j \in \mathcal{L}$: $\mu(x_r, L_j) = \max_{L_k \in \mathcal{L}} \mu(x_r, L_k)$ và $\delta(x_r, L_j) \leq \delta_2^{\max}$ **then**
      if $\ell_{x_r} \neq 0$ **then**
        $\ell_{x_r} \leftarrow \ell(L_j)$; $X^1 \leftarrow X^1 \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
      end if
      Mở rộng L_j; điều chỉnh chồng lấn giữa L_j và mọi $L_k \in \mathcal{L}$ có $\Omega(L_j, L_k) > 0$, $\ell(L_j) \neq \ell(L_k)$;
    else
      Xác định tập lân cận $N(x_r)$ theo (2.17);
      if $|N(x_r)| \geq 1$ **then**
        Tạo mới L_{new} chứa x_r; tính số phiếu theo (2.18); gán nhãn $\ell(L_{\text{new}})$ theo (2.19); $\mathcal{L} \leftarrow \mathcal{L} \cup \{L_{\text{new}}\}$; $X^1 \leftarrow X^1 \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
      end if
    end if
  end if
  $r \leftarrow r + 1$;
until $X = \varnothing$ **or** $r = R_{\max}$
$X \leftarrow X^1$; cập nhật ma trận liên kết lớp U dựa trên $\mathcal{L}$ theo (1.36);
return $\mathcal{L}$; X; U;

Mệnh đề 2.2 chỉ ra Thuật toán 2.2 có độ phức tạp $T_2 = O(R_{\max} nMK)$, với $R_{\max}$ là số lượt duyệt tối đa của vòng lặp repeat-until. Tổng chi phí của FMN-Voting là:

$$T = T_1 + T_2 = O(nMK) + O(R_{\max} MnK) = O(R_{\max} nMK) \quad (2.22)$$

FMN-Voting phù hợp khi các siêu hộp có nhãn lân cận phản ánh tốt phân bố cục bộ. Tuy nhiên, mô hình xem các siêu hộp lân cận là các nguồn bỏ phiếu tương đương; khi dữ liệu phân bố không đều bên trong siêu hộp, tâm hình học có thể lệch so với tâm phân bố thực, làm giảm độ tin cậy của bỏ phiếu. Ngoài ra, với dữ liệu có cụm không hình cầu, quá trình hình thành siêu hộp có thể nhạy với thứ tự mẫu huấn luyện.

2.4. Mô hình FMN-SAL

2.4.1. Mô tả bài toán và ý tưởng

FMN-SAL khai thác trực tiếp thông tin từ các mẫu đã có nhãn trong tập huấn luyện để xác định nhãn cho các siêu hộp, hình thành tập siêu hộp có nhãn làm nguồn giám sát ban đầu, rồi lan truyền nhãn tới các siêu hộp và mẫu chưa có nhãn bằng cơ chế bỏ phiếu. Giả định các siêu hộp lân cận có mức đại diện tương đương của FMN-Voting không luôn đúng: siêu hộp có thể bị mở rộng về một phía (mất cân bằng), có tâm hình học lệch so với tâm dữ liệu, hoặc có chỉ số sử dụng thấp. Vì vậy, FMN-SAL giữ mục tiêu suy luận nhãn nhưng thay đổi nguồn bỏ phiếu: thay vì các siêu hộp lân cận, các mẫu mang nhãn lân cận được dùng làm đơn vị bỏ phiếu, giảm sự phụ thuộc vào tâm hình học và mức độ đại diện của từng siêu hộp.

2.4.2. Kiến trúc và thuật toán học trong FMN-SAL

Kiến trúc FMN-SAL gồm lớp vào F_A, lớp siêu hộp F_B (gồm các siêu hộp đã gán nhãn $b_1, \ldots, b_q$ và các siêu hộp cần suy luận nhãn $u_1, \ldots, u_p$) và lớp ra F_C. Với mẫu x_r, tập lân cận gồm k mẫu có nhãn có độ thuộc lớn nhất được xác định theo (2.23), số phiếu và nhãn của siêu hộp mới theo (2.24)-(2.25):

$$N(x_r) = \underset{\min(k,|X'|)}{\arg\operatorname{top}} \left\{ \mu(x_r, x_i) \mid x_i \in X',\ \mu(x_r, x_i) \geq \beta \right\} \quad (2.23)$$

$$NoC_\ell = \sum_{x_i \in N(x_r)} \mathbb{I}(\ell_{x_i} = \ell); \qquad \ell(B_{\text{new}}) = \arg\max_\ell NoC_\ell \quad (2.24)\text{-}(2.25)$$

Mệnh đề 2.3: Thuật toán 2.3 có độ phức tạp thời gian trong trường hợp xấu nhất là

$$T_{SAL} = O\big(nMK + nM^2\big) = O\big(nM(K+M)\big) \tag{2.26}$$

FMN-SAL giảm sự phụ thuộc vào mức độ đại diện của siêu hộp nhưng phụ thuộc nhiều hơn vào chất lượng và phân bố của các mẫu mang nhãn; do đó mô hình không mặc nhiên cho kết quả tốt hơn trên mọi tập dữ liệu.

Thuật toán 2.3 Thuật toán học bán giám sát FMN-SAL

Input: $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$, $\delta^{\max}$, β
Output: $\mathcal{B}$, $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$, U

Khởi tạo tập $\mathcal{U}$ gồm p siêu hộp theo (1.33); $\mathcal{B} \leftarrow \varnothing$; $X' \leftarrow \varnothing$;
Tạo mới siêu hộp B_{new} theo (1.32); $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$;
repeat
  Chọn dữ liệu vào $(x_r, \ell_{x_r}) \in X$;
  if $\exists B_j \in \mathcal{B}$: $\mu(x_r, B_j) = \max_{B_k \in \mathcal{B}} \mu(x_r, B_k)$ và $\delta(x_r, B_j) \le \delta^{\max}$ **then**
    Điều chỉnh B_j theo (1.6), (1.7); **if** $\ell_{x_r} = 0$ **then** $\ell_{x_r} \leftarrow \ell(B_j)$;
    $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    Điều chỉnh chồng lấn giữa B_j và B_k nếu $\ell(B_j) \neq \ell(B_k)$;
  else
    if $\ell_{x_r} \neq 0$ **then**
      Tạo B_{new}; $\ell(B_{\text{new}}) \leftarrow \ell_{x_r}$; $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$; $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    else
      Xác định tập lân cận $N(x_r)$ theo (2.23);
      if $|N(x_r)| = k$ **then**
        Tạo B_{new}; tính số phiếu theo (2.24); gán nhãn $\ell(B_{\text{new}})$ theo (2.25); $\ell_{x_r} \leftarrow \ell(B_{\text{new}})$; $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$; $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
      end if
    end if
  end if
until $X = \varnothing$
$X \leftarrow X'$; cập nhật U dựa trên $\mathcal{B}$ theo (1.36);
return $\mathcal{B}, X, U$;

2.5. Thực nghiệm và đánh giá

2.5.1. Thiết lập thực nghiệm

Thực nghiệm nhằm đánh giá hiệu năng của FMN-Voting và FMN-SAL so với các biến thể FMN và các phương pháp liên quan, phân tích tác động của các chiến lược suy luận nhãn, ảnh hưởng của tham số ($\delta^{\max}$, β, γ) và chi phí tính toán. Các thí nghiệm dùng kiểm tra chéo 10-fold; trong mỗi fold, tập kiểm tra chỉ dùng để đánh giá sau khi huấn luyện và chọn tham số hoàn tất, các tham số tiền xử lý (kể cả giá trị trung bình thay thế giá trị

thiếu) được xác định chỉ từ tập huấn luyện. Kết quả báo cáo dưới dạng trung bình và độ lệch chuẩn; luận án chưa thực hiện kiểm định giả thuyết thống kê nên chênh lệch được dùng để mô tả xu hướng. Tham số chính: $\gamma = 10$, β ban đầu 0,99, hệ số giảm $\varphi = 0{,}9$ và $\beta_{\min} \geq 0{,}8$. Môi trường: MATLAB R2014a 64-bit, Intel Core i5-6300HQ, RAM 16GB.

Thực nghiệm sử dụng 13 tập dữ liệu (Bảng 2.1 của luận án) có từ 150 đến 7200 mẫu, 2 đến 60 thuộc tính và 2 đến 15 nhóm. Các tập Aggregation, Flame, Spiral, Jain, R15, Pathbased, Thyroidnew lấy từ kho CS; Iris, Wine, PID, Thyroid, Sonar, ILPD lấy từ kho UCI. ILPD gồm 583 hồ sơ bệnh nhân (416 mắc bệnh gan, 167 không mắc bệnh) với 10 thuộc tính lâm sàng và sinh hóa máu. Vì các mô hình thực hiện phân cụm bán giám sát, nhãn cụm được ánh xạ sang nhãn lớp chỉ trong giai đoạn đánh giá theo lớp xuất hiện nhiều nhất trong cụm (2.32); các độ đo sử dụng gồm Accuracy, Precision, Recall, F1-Score, Specificity, RI, ARI và NMI.

$$\pi(G_j) = \arg\max_{c \in \mathcal{C}} \sum_{x_i \in G_j} \mathbb{I}(y_i = c) \tag{2.32}$$

2.5.2. Kết quả trên các tập dữ liệu chuẩn

Tham số $\delta^{\max}$ ảnh hưởng rõ rệt đến chất lượng phân cụm: khi $\delta^{\max}$ nhỏ hoặc trung bình thấp, mô hình thường đạt Accuracy cao và ổn định hơn; khi $\delta^{\max}$ tăng, siêu hộp mở rộng quá mức làm mờ ranh giới cụm và giảm độ chính xác. Số lượng siêu hộp của FMN-SAL giảm khi $\delta^{\max}$ tăng, ví dụ trên Aggregation giảm từ 423,1 ($\delta^{\max} = 0{,}01$) xuống 13,3 ($\delta^{\max} = 0{,}80$). Bảng 2.2 cho thấy trên các tập có kết quả của cả hai mô hình, FMN-SAL đạt Accuracy cao hơn FMN-Voting; ký hiệu "-" là trường hợp FMN-Voting biến động mạnh theo thứ tự mẫu huấn luyện, Accuracy thấp hơn 0,5 nên không báo cáo giá trị đại diện.

Bảng 2.2: *Giá trị độ đo Accuracy của FMN-Voting và FMN-SAL*

Tập dữ liệu	FMN-Voting	FMN-SAL
Flame	0,9622	0,9798
Iris	0,9588	0,9685
R15	0,9780	0,9984
Wine	0,9356	0,9450
Jain	-	0,9879
Spiral	-	0,9943
Pathbased	-	0,9781
PID	-	0,7004
Thyroid	-	0,9675
Aggregation	-	0,9911

So sánh với các phương pháp trong tài liệu [101] theo chỉ số ARI (Bảng 2.4 của luận án), FMN-SAL đạt 0,9984 trên R15 và 0,9798 trên Flame, là các giá trị cao nhất trong nhóm so sánh; trên Aggregation và Jain, FMN-SAL (0,9911 và 0,9879) gần với GCAE (0,9920 và 1,0000). Do kết quả của các phương pháp đối sánh được kế thừa từ [101], không thực nghiệm lại với cùng quy trình, các số liệu này chủ yếu dùng làm tham chiếu.

Bảng 2.5 so sánh Accuracy (%) của SCFMN, MSCFMN và FMN-SAL. FMN-SAL đạt giá trị cao hơn trên phần lớn các tập, đáng chú ý trên Spiral đạt 99,43%. Trên PID và Sonar, SCFMN đạt 70,57% và 73,91%, cao hơn nhẹ so với 70,04% và 73,58% của FMN-SAL; điều này cho thấy lợi ích của cơ chế bỏ phiếu dựa trên mẫu phụ thuộc vào đặc điểm dữ liệu.

Bảng 2.5: *Giá trị Accuracy (%) của ba mô hình trên các tập dữ liệu*

Tập dữ liệu	δ^{max}	SCFMN	MSCFMN	FMN-SAL
Iris	0,06	94,12	94,67	96,85
Thyroid	0,07	92,65	92,63	96,43
Wine	0,06	93,33	93,86	96,78
PID	0,06	70,57	70,12	70,04
Sonar	0,06	73,91	73,54	73,58
Spiral	0,06	61,05	63,43	99,43
Flame	0,05	97,93	97,95	97,98
Jain	0,05	98,72	98,54	98,78
R15	0,05	99,50	99,64	99,84

2.5.3. Kết quả trên tập dữ liệu bệnh gan ILPD

Trên ILPD (lớp bệnh gan là lớp dương), cả FMN-SAL và FMN-Voting đạt kết quả tốt nhất tại $\delta^{max} = 0{,}08$. Tại giá trị này, FMN-SAL đạt Accuracy 0,9039, Recall 0,9423, Specificity 0,8084, Precision 0,9245 và F1-Score 0,9333 trên lớp bệnh gan; FMN-Voting đạt Accuracy 0,8576, Recall 0,9135, Specificity 0,7186, Precision 0,8899 và F1-Score 0,9015. Khi δ^{max} quá lớn (0,90), Accuracy giảm còn khoảng 0,51 và ARI, NMI xấp xỉ 0.

Bảng 2.12 và Bảng 2.13 so sánh các mô hình đề xuất với K-means++, EM, DBSCAN (thực hiện trên Weka) và SCFMN. Precision, Recall và F1-Score được tổng hợp theo trung bình có trọng số theo số mẫu của từng lớp; với bài toán hai lớp đơn nhãn, Recall trung bình có trọng số bằng Accuracy.

Bảng 2.12: *Kết quả đánh giá tổng thể trên tập dữ liệu ILPD (trích)*

Mô hình	Accuracy	Specificity	Precision	F1-Score
K-means++	0,6252 ± 0,0222	0,4463 ± 0,0932	0,6148 ± 0,0437	0,6154 ± 0,0240
EM	0,5647 ± 0,0351	0,7353 ± 0,1299	0,7589 ± 0,0544	0,5648 ± 0,0249
DBSCAN	0,6416 ± 0,0022	0,4411 ± 0,0048	0,6229 ± 0,0021	0,6306 ± 0,0019
SCFMN	0,7780 ± 0,0184	0,6442 ± 0,0330	0,7718 ± 0,0185	0,7736 ± 0,0186
FMN-SAL	0,8806 ± 0,0043	0,8001 ± 0,0082	0,8786 ± 0,0044	0,8788 ± 0,0043
FMN-Voting	0,8045 ± 0,0366	0,7029 ± 0,0490	0,8021 ± 0,0365	0,8030 ± 0,0364

Recall trung bình có trọng số trùng với Accuracy nên không lặp lại.

Bảng 2.13: *So sánh FMN-SAL, FMN-Voting với các phương pháp khác trên ILPD*

Mô hình	RI	ARI	NMI
K-means++	0,5314 ± 0,0111	0,0189 ± 0,0336	0,0107 ± 0,0110
EM	0,5097 ± 0,0146	−0,0130 ± 0,0144	0,0926 ± 0,0343
DBSCAN	0,5393 ± 0,0013	0,0347 ± 0,0030	0,0066 ± 0,0011
SCFMN	0,6546 ± 0,0208	0,2772 ± 0,0432	0,1570 ± 0,0334
FMN-SAL	0,7894 ± 0,0066	0,5606 ± 0,0135	0,4077 ± 0,0137
FMN-Voting	0,6870 ± 0,0458	0,3499 ± 0,0942	0,2189 ± 0,0793

Kết quả cho thấy FMN-SAL đạt Accuracy 0,8806 ± 0,0043 và RI, ARI, NMI cao nhất trong nhóm so sánh, với độ lệch chuẩn nhỏ; FMN-Voting cao hơn SCFMN và các phương pháp trên Weka trên phần lớn các chỉ số. Các phương pháp K-means++ và DBSCAN có RI xấp xỉ 0,53 nhưng

ARI, NMI rất thấp, cho thấy cấu trúc cụm chưa tương ứng tốt với hai lớp bệnh gan và không bệnh gan.

2.6. Kết luận chương 2

Chương 2 đã đề xuất hai mô hình FMN-Voting và FMN-SAL. FMN-Voting khai thác bỏ phiếu dựa trên các siêu hộp lân cận, hạn chế sự phụ thuộc vào một siêu hộp đơn lẻ; FMN-SAL suy luận nhãn dựa trên các mẫu mang nhãn lân cận, giảm phụ thuộc vào tâm hình học và mức độ đại diện của siêu hộp. Thuật toán học được trình bày kèm phân tích độ phức tạp. Thực nghiệm cho thấy hai mô hình đạt kết quả cạnh tranh và cải thiện hiệu năng trên nhiều tập dữ liệu, tuy mức cải thiện phụ thuộc vào đặc điểm từng tập; kết quả trên PID và Sonar cho thấy không có cơ chế suy luận nhãn duy nhất chiếm ưu thế trên mọi tập dữ liệu. Kết quả của chương đã được công bố trong các công trình [CT2], [CT3], [CT4] và [CT5].

CHƯƠNG 3. HỌC BÁN GIÁM SÁT TRONG MẠNG NƠRON FMN VỚI CHIẾN LƯỢC CHỌN HẠT GIỐNG CHỦ ĐỘNG

3.1. Giới thiệu về mô hình FA-Seed

Hiệu quả của các mô hình ở Chương 2 vẫn phụ thuộc vào chất lượng của các hạt giống ban đầu. Trong luận án, hạt giống (seed) là mẫu dữ liệu đại diện được lựa chọn và xác định nhãn để cung cấp thông tin giám sát ban đầu; độ tin cậy của một siêu hộp phản ánh mức độ có thể dùng siêu hộp đó làm nguồn thông tin để chọn hạt giống, được xem xét qua mật độ dữ liệu, độ lệch giữa tâm hình học và tâm dữ liệu, và độ thuộc của các mẫu bên trong; tính đại diện phản ánh mức độ siêu hộp hoặc mẫu đặc trưng cho phân bố cục bộ của vùng dữ liệu nó bao phủ.

Các mô hình FMN bán giám sát hiện có hoặc chọn hạt giống ngẫu nhiên (GFMM, SSL-FMM, SS-FMM), hoặc chọn hạt giống dựa trên siêu hộp với giả thiết "mỗi siêu hộp đại diện cho một cụm" (SCFMN, MSCFMN). Giả thiết này chỉ phù hợp khi cụm gần cầu và kích thước tương đối đồng đều; với các tập có cấu trúc phức tạp như Aggregation, Flame, Pathbased, siêu hộp phải mở rộng để bao phủ cả cụm, dẫn đến chồng lấn nhiều cụm và giảm tính thuần nhất. Từ đó, luận án đề xuất mô hình FA-Seed [CT6] kết hợp chiến lược lựa chọn hạt giống chủ động với khả năng phân hoạch thích nghi của FMN: biểu diễn một cụm bằng nhiều siêu hộp, đánh giá độ tin cậy của hạt giống và lan truyền nhãn có kiểm soát.

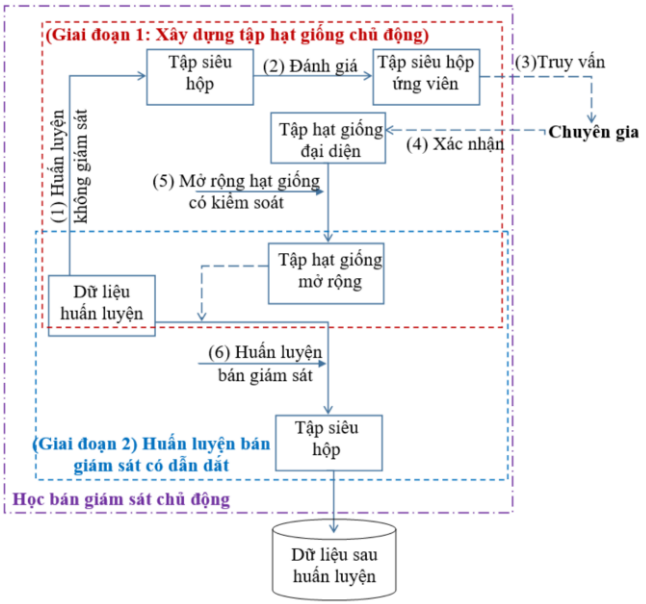
3.2. Mô hình FA-Seed: Lựa chọn hạt giống chủ động dựa trên FMN

3.2.1. Mô tả bài toán

Cho tập dữ liệu X gồm M mẫu, mỗi x_r có thể được ánh xạ vào C cụm; tập siêu hộp $B = L \uplus U$ gồm tập siêu hộp có nhãn L và chưa có nhãn U. Bài toán của FA-Seed là phân cụm bán giám sát khi dữ liệu huấn luyện ban đầu hầu như chưa có nhãn, hướng tới ba mục tiêu: (i) lựa chọn tập hạt giống có chất lượng và tính đại diện cao; (ii) giảm số lượng truy vấn chuyên gia; và (iii) lan truyền nhãn có kiểm soát nhằm nâng cao chất lượng phân cụm. Từ tập dữ liệu chưa gán nhãn, FA-Seed xác định tập hạt giống mục tiêu nhỏ nhưng có tính đại diện cao thông qua đánh giá ứng viên và truy vấn chủ động chuyên gia, sau đó mở rộng có kiểm soát thành tập hạt giống mở rộng để dẫn dắt quá trình huấn luyện bán giám sát.

3.2.2. Ý tưởng và kiến trúc của FA-Seed

Ở giai đoạn xác định hạt giống, FA-Seed sử dụng giới hạn kích thước siêu hộp nhỏ hơn để phân hoạch không gian dữ liệu thành các vùng cục bộ chi tiết hơn (ví dụ 41 siêu hộp trên Aggregation, 11 trên Flame và 19 trên Pathbased). Ràng buộc này làm nổi bật biến động cục bộ về mật độ, giúp nhận diện vùng biên, vùng chồng lấn và điểm bất thường. Kiến trúc mô hình gồm hai giai đoạn (Hình 3.7): (i) xây dựng tập hạt giống chủ động - huấn luyện không giám sát FMN, đánh giá các siêu hộp ứng viên theo mật độ và phân phối dữ liệu, truy vấn chuyên gia để xác nhận nhãn và mở rộng có kiểm soát; (ii) huấn luyện bán giám sát có dẫn dắt bằng tập hạt giống mở rộng. Trong ví dụ trên tập Aggregation, 15 trong số 25 siêu hộp được chọn để sinh hạt giống.



Hình 3.7: *Kiến trúc mô hình phân cụm bán giám sát mờ với học chủ động FA-Seed*

Mạng nơron FA-Seed gồm lớp đầu vào F_A; lớp đánh giá siêu hộp F_B gồm K nút ứng với K siêu hộp tạo ra ở Giai đoạn 1; lớp đánh giá và lan truyền nhãn F_H chia thành tập siêu hộp đã gán nhãn và chưa gán nhãn, áp dụng cơ chế cổng và quy tắc kế thừa nhãn để chuyển dần các siêu hộp sang tập có nhãn; và lớp đầu ra F_C gồm các nút cụm.

3.2.3. Thuật toán học của mô hình FA-Seed

Giai đoạn xây dựng tập hạt giống với huấn luyện chủ động: mô hình chủ động chọn một số mẫu có tính đại diện từ các siêu hộp thỏa mãn tiêu chí để chuyên gia cung cấp nhãn. Các siêu hộp được đánh giá theo độ lệch tâm (2.1); các siêu hộp không thuần khiết bị loại; tập ứng viên hạt giống và tập hạt giống được xác định bởi:

$$S = \bigcup_{j \in J_Q} \{x_r \in X \mid \mu(x_r, B_j) \geq \tau\},$$

$$J_Q = \{j \mid B_j \in \mathcal{B},\ e(B_j) \leq \varepsilon\}, \qquad Q = \{d(B_j) \mid j \in J_Q\} \qquad \text{(3.9)-(3.12)}$$

Thuật toán 3.1 Xây dựng tập hạt giống với huấn luyện chủ động

Input: X ($M = |X|$); $\delta^{\max}$; ε

Output: Tập dữ liệu đã sắp xếp lại X', tập siêu hộp $\mathcal{B}$

Bước 1: Sinh tập siêu hộp $\mathcal{B}$
Duyệt một lần qua tập X để xây dựng tập siêu hộp mờ $\mathcal{B}$ dựa trên $\delta^{\max}$;
Bước 2: Lọc siêu hộp (xác định các điểm dữ liệu cần truy vấn)
for mỗi siêu hộp $B_j \in \mathcal{B}$ **do**
    Gọi T_j là tập các mẫu được gán cho B_j (3.10); V_j, W_j là các đỉnh min, max của B_j;
    Tính tâm dữ liệu $d(B_j)$ theo (2.2) và tâm hình học $g(B_j)$ theo (1.40);
    if điều kiện (2.3) không được thỏa mãn **then**
        Loại bỏ siêu hộp B_j: $\mathcal{B} \leftarrow \mathcal{B} \setminus \{B_j\}$;
    end if
end for
Bước 3: Xây dựng tập hạt giống, phân hoạch và sắp xếp lại dữ liệu
Xây dựng tập hạt giống Q từ tâm dữ liệu của các siêu hộp thuần khiết theo (3.12) và truy vấn nhãn từ chuyên gia;
Xây dựng tập hạt giống S từ các mẫu nằm hoàn toàn trong ít nhất một siêu hộp thuần khiết theo (3.9);
Xây dựng hoán vị π của $\{1, \ldots, M\}$ sao cho các phần tử của S được xếp trước, tập $S' = X \setminus S$ được xếp sau;
Tạo tập dữ liệu đã sắp xếp lại $X' = (X_{\pi(1)}, \ldots, X_{\pi(M)})$;
return $\mathcal{B}$ (tập siêu hộp còn tồn tại), X' (tập dữ liệu đã sắp xếp lại)

Các mẫu thuộc các siêu hộp bị loại không được bao phủ bởi siêu hộp thuần khiết nên thuộc tập *S′* và được xử lý ở giai đoạn sau. Thuật toán 3.1 có độ phức tạp $T_1 = O(nMK)$.

Giai đoạn huấn luyện bán giám sát có dẫn dắt: tập hạt giống đã được xác định nhãn được dùng làm thông tin giám sát ban đầu để định hướng quá trình hình thành siêu hộp và lan truyền nhãn; việc lan truyền được kiểm soát bởi ràng buộc kích thước siêu hộp và ngưỡng độ thuộc β:

$$j^* = \arg\max_{j \in \mathcal{J}} \mu(x_r, B_j), \qquad \mathcal{J} = \left\{ j \mid \delta(x_r, B_j) \le \delta_2^{\max} \right\} \tag{3.19)-(3.20}$$

$$k^* = \arg\max_{k \in J_L} \mu\big(g(B_j), B_k\big), \qquad J_L = \left\{ k \mid \mu\big(g(B_j), B_k\big) \ge \beta \right\} \tag{3.22)-(3.23}$$

Thuật toán 3.2 có độ phức tạp $T_2 = O(nMK)$. Ngoài ra, FA-Seed phát sinh chi phí truy vấn nhãn $O(q)$ cho q mẫu được chọn, với q nhỏ nên không làm thay đổi bậc độ phức tạp. Tổng chi phí tính toán của FA-Seed là:

$$T = T_1 + T_2 = O(nMK) + O(nMK) = O(nMK) \tag{3.27}$$

Thuật toán 3.2 Huấn luyện bán giám sát có dẫn dắt

Input: X'; β; $\delta_2^{\max}$
Output: $\mathcal{B}$

$\mathcal{L} = \varnothing; \mathcal{U} = \varnothing$;
for mỗi $x_r \in X'$ **do**
  Xác định tập $\mathcal{J}$ theo (3.20);
  if $\mathcal{J} \neq \varnothing$ **then**
    Xác định j^* theo (3.19); mở rộng B_{j^*} theo (3.21);
    Điều chỉnh B_{j^*} và B_k theo (1.15)–(1.26) nếu chồng lấn với $\ell(B_{j^*}) \neq \ell(B_k)$;
  else
    Tạo B_{new} với $V_{\text{new}} = W_{\text{new}} = x_r$;
    if $\ell_{x_r} \neq \varnothing$ **then** $\ell(B_{\text{new}}) = \ell_{x_r}$; $\mathcal{L} = \mathcal{L} \cup \{B_{\text{new}}\}$ **else** $\mathcal{U} = \mathcal{U} \cup \{B_{\text{new}}\}$;
  end if
end for
repeat
  for mỗi $B_j \in \mathcal{U}$ **do**
    Tính trọng tâm $g(B_j)$ theo (1.40);
    if $\exists k^*$ với $B_{k^*} \in \mathcal{L}$ thỏa mãn (3.22) **then**
      $\ell(B_j) = \ell(B_{k^*})$; $\mathcal{L} = \mathcal{L} \cup \{B_j\}$; $\mathcal{U} = \mathcal{U} \setminus \{B_j\}$;
    end if
  end for
  Giảm β;
until ($\mathcal{U} = \varnothing$) **or** ($\beta \leq$ ngưỡng tin cậy)
if $\mathcal{U} \neq \varnothing$ và $\mathcal{L} \neq \varnothing$ **then**
  Với mỗi $B_j \in \mathcal{U}$: $k^* \leftarrow \arg\max_k \mu(g(B_j), B_k)$, $B_k \in \mathcal{L}$; $\ell(B_j) = \ell(B_{k^*})$;
end if
$\mathcal{B} = \mathcal{L} \cup \mathcal{U}$
return $\mathcal{B}$

3.3. Thực nghiệm và đánh giá

Thực nghiệm tuân theo cùng thiết lập ở Chương 2; ngoài các tập ở Bảng 2.1, chương này sử dụng thêm các tập Ecoli, Soybean và Zoo từ kho UCI. Trên các tập Iris và Wine, so với phương pháp FC2 ở các mức tỷ lệ nhãn thấp (dưới khoảng 20%), FA-Seed đạt RI trung bình cao hơn và biên độ dao động nhỏ hơn; khi tỷ lệ nhãn tăng, hiệu năng của hai phương pháp dần tương đương. Tại cùng số lượng truy vấn, FA-Seed thu được số lượng hạt giống cao hơn MSCFMN nhờ phân hoạch siêu hộp tinh hơn. So với SSGC, SSDBSCAN, SSK-Means, K-Means và MSCFMN theo chỉ số RI, FA-Seed nhìn chung tốt hơn hoặc tương đương trên hầu hết các tập; trên Soybean, Zoo và Iris đạt RI gần như hoàn hảo. Về Accuracy, FA-Seed tương đương hoặc cao hơn MSCFMN trên hầu hết các tập, với mức cải

thiện rõ hơn trên các tập có phân bố lớp mất cân bằng như Thyroid, Pathbased, Zoo.

Trên tập ILPD, FA-Seed đạt kết quả tốt nhất tại $\delta^{max} = 0{,}08$ với Accuracy 0,9211, Recall 0,9567 và F1-Score 0,9454 trên lớp bệnh gan; RI, ARI, NMI tương ứng là 0,8544 ± 0,0263, 0,6972 ± 0,0549 và 0,5516 ± 0,0577. So sánh với MSCFMN và các phương pháp trên Weka (Bảng 3.6, Bảng 3.7), FA-Seed cải thiện Accuracy từ 0,7976 ± 0,0332 lên 0,9009 ± 0,0165, F1-Score từ 0,7941 ± 0,0323 lên 0,8998 ± 0,0167, đồng thời FP Rate giảm từ 0,3250 ± 0,0383 xuống 0,1638 ± 0,0255.

Bảng 3.6: *So sánh FA-Seed với MSCFMN và các phương pháp trên Weka (trích)*

Mô hình	Accuracy	Specificity	Precision	F1-Score
K-means++	0,6252 ± 0,0222	0,4463 ± 0,0932	0,6148 ± 0,0437	0,6154 ± 0,0240
EM	0,5647 ± 0,0351	0,7353 ± 0,1299	0,7589 ± 0,0544	0,5648 ± 0,0249
DBSCAN	0,6416 ± 0,0022	0,4411 ± 0,0048	0,6229 ± 0,0021	0,6306 ± 0,0019
MSCFMN	0,7976 ± 0,0332	0,6750 ± 0,0383	0,7929 ± 0,0329	0,7941 ± 0,0323
FA-Seed	0,9009 ± 0,0165	0,8362 ± 0,0255	0,8995 ± 0,0169	0,8998 ± 0,0167

Recall trung bình có trọng số trùng với Accuracy nên không lặp lại.

Bảng 3.7: *So sánh FA-Seed với MSCFMN và các phương pháp trên Weka: RI, ARI, NMI*

Mô hình	RI	ARI	NMI
K-means++	0,5314 ± 0,0111	0,0189 ± 0,0336	0,0107 ± 0,0110
EM	0,5097 ± 0,0146	−0,0130 ± 0,0144	0,0926 ± 0,0343
DBSCAN	0,5393 ± 0,0013	0,0347 ± 0,0030	0,0066 ± 0,0011
MSCFMN	0,6783 ± 0,0402	0,3279 ± 0,0812	0,1993 ± 0,0665
FA-Seed	0,8215 ± 0,0263	0,6285 ± 0,0549	0,4779 ± 0,0577

Về chi phí tính toán, FA-Seed có độ phức tạp $O(nMK)$; khi K nhỏ hơn nhiều so với M, chi phí tăng gần tuyến tính theo số mẫu, trong khi SSGC có độ phức tạp $O(n^2)$. Phân tích độ phức tạp và thời gian thực thi hiện tại mới cho thấy tiềm năng về hiệu quả tính toán, chưa đủ cơ sở để khẳng định khả năng mở rộng trên dữ liệu quy mô lớn.

3.4. Kết luận chương 3

Chương 3 đã trình bày mô hình FA-Seed với chiến lược lựa chọn hạt giống chủ động, xem xét đồng thời mức độ đại diện và độ tin cậy của từng hạt giống, giúp giảm nguy cơ lan truyền lỗi nhãn ở vùng chồng lấn, nhiễu hoặc phân bố không đồng đều, với chi phí tính toán phù hợp. Các thực

nghiệm cho thấy chiến lược này có tiềm năng cải thiện chất lượng phân cụm. Kết quả của chương đã được công bố trong các công trình [CT6], [CT7] và [CT8].

KẾT LUẬN

1. Kết quả đạt được

Luận án nghiên cứu bài toán phân cụm bán giám sát dựa trên mạng nơron min-max mờ trong điều kiện thông tin nhãn hạn chế, tập trung vào hai vấn đề: độ tin cậy của quá trình suy luận và lan truyền nhãn, và chất lượng của tập hạt giống ban đầu.

2. Những đóng góp mới của luận án

(1) Đề xuất mô hình FMN-Voting, trong đó nhãn của mẫu và siêu hộp mới được suy luận bằng bỏ phiếu đa số từ $k = 2C + 1$ siêu hộp có nhãn lân cận thay vì một siêu hộp dẫn dắt đơn lẻ; xây dựng thuật toán học hai giai đoạn và phân tích độ phức tạp $O(R_{\max} nMK)$. Kết quả đã được công bố trong công trình [CT3].

(2) Đề xuất mô hình FMN-SAL, chuyển nguồn bỏ phiếu từ siêu hộp sang các mẫu mang nhãn lân cận nhằm giảm ảnh hưởng của các siêu hộp lệch tâm hoặc có chỉ số sử dụng thấp. Trên 7/9 tập dữ liệu chuẩn, FMN-SAL đạt Accuracy cao hơn SCFMN và MSCFMN (Bảng 2.5); trên tập ILPD, FMN-SAL đạt Accuracy 0,8806 ± 0,0043 và các chỉ số RI, ARI, NMI cao nhất trong nhóm so sánh. Kết quả đã được công bố trong các công trình [CT2], [CT4] và [CT5].

(3) Đề xuất mô hình FA-Seed với chiến lược lựa chọn hạt giống chủ động dựa trên phân hoạch siêu hộp với giới hạn kích thước nhỏ, đánh giá siêu hộp theo độ lệch tâm và truy vấn chuyên gia cho các mẫu đại diện, sau đó huấn luyện bán giám sát có dẫn dắt. Trên tập ILPD, FA-Seed đạt Accuracy 0,9009 ± 0,0165 và RI 0,8215 ± 0,0263, so với 0,7976 ± 0,0332 và 0,6783 ± 0,0402 của MSCFMN. Kết quả đã được công bố trong các công trình [CT6], [CT7] và [CT8].

3. Ý nghĩa khoa học và thực tiễn

Luận án bổ sung các cơ chế đánh giá và sử dụng thông tin cấu trúc của siêu hộp (độ lệch tâm, chỉ số sử dụng, độ thuộc) vào suy luận nhãn và lựa chọn hạt giống, kèm phân tích độ phức tạp; các mô hình được kiểm chứng

trên dữ liệu bệnh gan ILPD và cấu trúc siêu hộp cho phép kết xuất các luật IF-THEN có thể diễn giải.

4. Hạn chế và hướng nghiên cứu tiếp theo

Luận án còn một số hạn chế: (i) các chênh lệch kết quả mới được báo cáo dưới dạng trung bình và độ lệch chuẩn, chưa được kiểm định ý nghĩa thống kê; trên PID và Sonar, FMN-SAL chưa vượt SCFMN; (ii) FMN-Voting chưa cho kết quả ổn định trên một số tập dữ liệu có cấu trúc cụm không hình cầu do phụ thuộc vào thứ tự mẫu huấn luyện; (iii) hiệu năng các mô hình phụ thuộc vào giới hạn kích thước siêu hộp δ^{max}; (iv) thực nghiệm mới tiến hành trên dữ liệu dạng bảng với tập lớn nhất 7200 mẫu, chưa đối sánh với các phương pháp phân cụm bán giám sát dựa trên học sâu; (v) FA-Seed cần sự tham gia của chuyên gia ở giai đoạn truy vấn nhãn.

Hướng nghiên cứu tiếp theo: (i) kết hợp với các mô hình học sâu hoặc biểu diễn đặc trưng để xử lý dữ liệu có cấu trúc phức tạp; (ii) phát triển cơ chế suy luận nhãn thích nghi hơn dựa trên độ tin cậy và động học của siêu hộp; (iii) nghiên cứu chiến lược song song và phân tán để nâng cao khả năng mở rộng; (iv) tiếp tục khai thác khả năng kết xuất luật mờ để tăng tính diễn giải trong các hệ thống hỗ trợ ra quyết định; và (v) kiểm định ý nghĩa thống kê các kết quả so sánh và đánh giá trên các tập dữ liệu có quy mô lớn hơn.

DANH MỤC CÔNG TRÌNH CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN

[CT1] Nguyễn Thanh Sơn, Vũ Thị Dương, Nguyễn Anh Thái, Hoàng Quang Huy, Vũ Đình Minh. "Tổng quan về mạng nơron min-max mờ và ứng dụng", Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông, Hội thảo Quốc gia lần thứ XXV, Hà Nội, ngày 8-9/12/2022, Nhà xuất bản Khoa học và Kỹ thuật, 2022, tr. 238-244.

[CT2] Dinh-Minh Vu, Thanh-Son Nguyen, Trung-Nghia Phung, Trinh-Hoang Vu. "Choosing Data Points to Label for Semi-supervised Learning Based on Neighborhood", Advances in Information and Communication Technology: Proceedings of the 2nd International Conference ICTA 2023, Volume 1, Lecture Notes in Networks and Systems, vol. 847, Springer Cham, 2023, tr. 134-141. DOI: 10.1007/978-3-031-49529-8_15.

[CT3] Dinh-Minh Vu, Thanh-Son Nguyen, Van Tinh Nguyen, Duc-Luu Nguyen. "FMN-Voting: A Semi-supervised Clustering Technique Based on Voting

Methods Using FMM Neural Networks", Advances in Information and Communication Technology: Proceedings of the 3rd International Conference ICTA 2024, Lecture Notes in Networks and Systems, vol. 1205, Springer Cham, 2025, tr. 137-150. DOI: 10.1007/978-3-031-80943-9_15.

[CT4] Thanh-Son Nguyen, Dinh-Minh Vu, Duc-Luu Nguyen. "FMN-SSL: A Semi-supervised Approach for Lung Tumor Segmentation", Advances in Information and Communication Technology: Proceedings of the 4th International Conference ICTA 2025, Volume 3, Lecture Notes in Networks and Systems, vol. 1833, Springer Cham, 2026.

[CT5] Dinh-Minh Vu, Thanh-Son Nguyen, Trinh-Hoang Vu, Thi-Duong Vu, Duc-Luu Nguyen, Quang-Huy Hoang, Ba-Dung Le. "FMN-OE: Evaluation and Elimination of Hyperbox Overlap", Intelligent Systems and Data Science: Third International Conference, ISDS 2025, Can Tho City, Vietnam, October 18-19, 2025, Proceedings, Part II, Communications in Computer and Information Science, vol. 2714, Springer, Singapore, 2026, tr. 328-342. DOI: 10.1007/978-981-95-3358-9_28.

[CT6] Dinh-Minh Vu, Thanh-Son Nguyen. "FA-Seed: Flexible and Active Learning-Based Seed Selection". Information, 2025, 16, 884. DOI: 10.3390/info16100884. (ESCI, Scopus).

[CT7] Nguyễn Thanh Sơn, Vũ Đình Minh, Lê Bá Dũng, Nguyễn Văn Thanh. "Phân cụm bán giám sát với phương pháp chọn hạt giống dựa trên trọng số siêu hộp trong mạng nơron min-max mờ", Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông, Hội thảo Quốc gia lần thứ XXIV, Thái Nguyên, ngày 13-14/12/2021, Nhà xuất bản Khoa học và Kỹ thuật, 2021, tr. 433-438.

[CT8] Nguyễn Thanh Sơn, Vũ Đình Minh. "Integrating Active Learning and Controlled Label Propagation in Fuzzy Min-Max Neural Networks", Journal of Computer Science and Cybernetics.