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

LUẬN ÁN TIẾN SĨ MÁY TÍNH

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

Nguyễn Thanh Sơn

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

LUẬN ÁN TIẾN SĨ MÁY TÍNH

Ngành: Hệ thống thông tin

Mã số: 9 48 01 04

Xác nhận của Học viện Khoa học và Công nghệ

KT. GIÁM ĐỐC
PHÓ GIÁM ĐỐC

ĐẢNG CỘNG SẢN VIỆT NAM
HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ
VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM

Nguyễn Thị Trung

Người hướng dẫn 1

PGS.TS Lê Bá Dũng

Người hướng dẫn 2

TS. Vũ Đình Minh

Hà Nội – 2026

LỜI CAM ĐOAN

Tôi xin cam đoan luận án: **"Nghiên cứu phát triển một số mô hình phân cụm bán giám sát dựa trên mạng nơron min-max mờ"** là công trình nghiên cứu của chính mình dưới sự hướng dẫn khoa học của tập thể hướng dẫn. 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 06 tháng 10 năm 2026

Tác giả luận án



Nguyễn Thanh Sơn

LỜI CẢM ƠN

Trong quá trình nghiên cứu và hoàn thành luận án, nghiên cứu sinh đã nhận được sự định hướng, giúp đỡ, các ý kiến đóng góp quý báu và những lời động viên của các nhà khoa học, các thầy cô giáo, đồng nghiệp và gia đình.

Trước hết, nghiên cứu sinh xin bày tỏ lời cảm ơn tới các thầy PGS.TS. Lê Bá Dũng, TS. Vũ Đình Minh đã tận tình hướng dẫn và giúp đỡ trong quá trình nghiên cứu.

Nghiên cứu sinh xin chân thành cảm ơn các thầy cô giáo, các nhà khoa học của Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Học viện Khoa học và Công nghệ, Viện Công nghệ thông tin đã có các góp ý quý báu cho nghiên cứu sinh trong quá trình thực hiện luận án này.

Nghiên cứu sinh chân thành cảm ơn Ban Giám đốc, Phòng Đào tạo, Học viện Khoa học và Công nghệ đã tạo điều kiện thuận lợi để nghiên cứu sinh hoàn thành nhiệm vụ nghiên cứu.

Cuối cùng, nghiên cứu sinh xin bày tỏ lời cảm ơn tới các đồng nghiệp, gia đình, bạn bè đã luôn động viên, chia sẻ, ủng hộ và giúp đỡ nghiên cứu sinh vượt qua khó khăn để đạt được những kết quả nghiên cứu trong luận án này.

Hà Nội, ngày 06 tháng 10 năm 2026

Tác giả luận án



Nguyễn Thanh Sơn

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
MỤC LỤC 1
DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT 3
DANH MỤC BẢNG 6
DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ 6
MỞ ĐẦU 10
1. Lý do chọn đề tài 10
2. Mục tiêu nghiên cứu 11
3. Đối tượng, phạm vi và nội dung nghiên cứu 12
4. Phương pháp nghiên cứu 12
5. Cơ sở khoa học và thực tiễn của đề tài 13
6. Những đóng góp mới của luận án 14
7. Bố cục của luận án 15

CHƯƠNG 1. TỔNG QUAN VỀ PHÂN CỤM BÁN GIÁM SÁT SỬ DỤNG MẠNG NƠRON MIN-MAX MỜ **16**
1.1 Tổng quan về bài toán phân cụm dữ liệu và ứng dụng 16
1.2 Phân cụm bán giám sát dựa trên mạng nơron Min-Max mờ 19
1.2.1 Giới thiệu về mạng nơron Min-Max mờ 19
1.2.2 Hàm thuộc siêu hộp mờ 20
1.2.3 Cấu trúc mạng nơron FMN 27
1.2.4 Thuật toán học FMN 30
1.2.5 Một số vấn đề cần tiếp tục nghiên cứu để nâng cao chất lượng FMN cho phân cụm dữ liệu 31
1.3 Hạn chế, vấn đề đặt ra và định hướng nghiên cứu 36
1.4 Kết luận chương 1 36

CHƯƠNG 2. PHÂN CỤM BÁN GIÁM SÁT DỰA TRÊN MẠNG NƠRON MIN-MAX MỜ VỚI BỎ PHIẾU ĐA SỐ **38**
2.1 Học bán giám sát dựa trên siêu hộp mờ 38
2.1.1 Phân hoạch không gian dữ liệu bằng FMN 38
2.1.2 Đánh giá siêu hộp dựa trên mật độ và phân phối dữ liệu trong siêu hộp 38
2.1.3 Cơ chế giám sát bán giám sát dựa trên siêu hộp 40

2.2 Học bán giám sát dựa trên bỏ phiếu đa số trong FMN 42
2.2.1 Giới thiệu về phương pháp bỏ phiếu đa số 42
2.2.2 Kỹ thuật bỏ phiếu đa số trong mạng nơron FMN 44
2.3 Mô hình FMN-Voting 44
2.3.1 Mô tả bài toán 44
2.3.2 Ý tưởng của FMN-Voting 45
2.3.3 Kiến trúc mạng nơron FMN-Voting 46
2.3.4 Thuật toán học trong FMN-Voting 48
2.3.5 Độ phức tạp tính toán của FMN-Voting 52
2.4 Mô hình FMN-SAL 53
2.4.1 Mô tả bài toán 53
2.4.2 Ý tưởng của FMN-SAL 54
2.4.3 Kiến trúc mạng nơron FMN-SAL 54
2.4.4 Thuật toán học trong FMN-SAL 55
2.5 Thực nghiệm và đánh giá 58
2.5.1 Thiết lập thực nghiệm 58
2.5.2 Kết quả thực nghiệm và đánh giá 70
2.6 Kết luận chương 2 92

CHƯƠNG 3. HỌC BÁN GIÁM SÁT TRONG MẠNG NƠRON FMN VỚI CHIẾN LƯỢC CHỌN HẠT GIỐNG CHỦ ĐỘNG 94
3.1 Giới thiệu về mô hình FA-Seed 94
3.2 Mô hình FA-Seed: Lựa chọn hạt giống chủ động dựa trên FMN 96
3.2.1 Mô tả bài toán 96
3.2.2 Ý tưởng của FA-Seed 97
3.2.3 Kiến trúc mạng nơron FA-Seed 102
3.2.4 Thuật toán học của mô hình FA-Seed 103
3.3 Thực nghiệm và đánh giá 109
3.3.1 Thiết lập thực nghiệm 109
3.3.2 Kết quả thực nghiệm và đánh giá 110
3.4 Kết luận chương 3 125

KẾT LUẬN 127

DANH MỤC CÔNG TRÌNH CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN 129

DANH MỤC TÀI LIỆU THAM KHẢO 131

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

$\ell(B_j)$	Nhãn của siêu hộp B_j.
ℓ_{x_r}	Nhãn của mẫu dữ liệu x_r.
$\|\mathcal{B}\|$	Lực lượng của tập $\mathcal{B}$.
$\mu(x, B_j)$	Hàm xác định độ thuộc của mẫu x vào siêu hộp B_j.
$f(x, y)$	Hàm ngưỡng hai tham số x, y.
$\max(x, y)$	Hàm chọn giá trị lớn nhất.
$\min(x, y)$	Hàm chọn giá trị nhỏ nhất.
$\mathbb{R}^n$	Không gian Euclid n chiều.
V	Đỉnh min của siêu hộp.
W	Đỉnh max của siêu hộp.
α	Tham số chọn.
β	Tham số ngưỡng.
γ	Tham số mờ.
ε	Ngưỡng độ lệch tâm của siêu hộp.
$\delta(B_j)$	Kích thước của siêu hộp B_j.
δ^{max}	Giới hạn kích thước tối đa của siêu hộp.
$\delta(x, B_j)$	Kích thước của siêu hộp B_j khi kết nạp thêm mẫu x.
φ	Tham số giảm ngưỡng.
A2M	Protein huyết tương phân tử lớn (Alpha 2 Macroglobulin).
AIS	Hệ thống miễn dịch nhân tạo (Artificial Immune System).
ALP	Phosphatase kiềm (Alkaline Phosphatase).
ALT	Men gan ALT (Alanine Aminotransferase).
ANN	Mạng nơron nhân tạo (Artificial Neural Network).
APRI	Phương pháp phân loại xơ hóa gan dựa vào tiểu cầu (Aspartate Aminotransferase to Platelet Ratio Index).
ART	Lý thuyết cộng hưởng thích nghi (Adaptive Resonance Theory).
AST	Men gan AST (Aspartate Aminotransferase).
CCH	Siêu hộp bù chứa (Containment Compensation Hyperbox).
CDS	Hệ thống chẩn đoán xơ gan (Cirrhosis Diagnosis System).
CF	Hệ số đóng góp (Contribution Factor).
CFMN	FMN cải tiến dựa trên hệ số CF (Contribution-factor based Fuzzy Min-Max Neural Network).

CFMNN	FMN dựa trên tâm cụm (Centroid-based Fuzzy Min-Max Neural Network).
CN	Siêu hộp không chồng lấn (Classifying Neurons).
CSPA	Thuật toán phân cụm dựa trên phân vùng tương tự (Cluster-based Similarity Partitioning Algorithm).
DCFMN	Mạng FMN dựa trên tâm dữ liệu (Data-Core-Based Fuzzy Min-Max Neural Network).
ECT	Cây phân cụm (Ensemble of Clustering Trees).
EGWCA	Thuật toán phân cụm dựa trên khoảng cách Euclid (Euclidean Distance Generalized Weighted Cluster Aggregation).
EFMN	Mạng nơron FMN tăng cường (Enhanced Fuzzy Min-Max Neural Network).
eGWCA	Thuật toán phân cụm dựa trên khoảng cách hàm mũ (Exponential Distance Generalized Weighted Cluster Aggregation).
eSFCM	Phân cụm bán giám sát mờ dùng Entropy (Semi-supervised Entropy-Regularized Fuzzy Clustering).
FART	Lý thuyết cộng hưởng thích nghi mờ (Fuzzy Adaptive Resonance Theory).
FIB-4	Phương pháp phân loại xơ hóa gan (Fibrosis-4).
FMCN	FMN cải tiến sử dụng nút chồng lấn (Compensatory Neurons).
FMM	Lý thuyết min-max mờ (Fuzzy Min-Max).
FMN	Mạng nơron min-max mờ (Fuzzy Min-Max Neural Network).
GFMM	Mạng nơron min-max mờ tổng quát (General Fuzzy Min-Max).
GGT	Hoạt độ men GGT trong máu (Gamma Glutamyl Transferase).
GSOM	Bản đồ tự tổ chức tăng trưởng (Growing Self-Organizing Map).
HCF	Hệ số tin cậy của siêu hộp (Hyperbox Confidence Factor).
HE	Hệ số Entropy của siêu hộp (Hyperbox Entropy).
ID3	Thuật toán cây quyết định (Iterative Dichotomiser 3).
INR	Tỷ số bình thường hóa quốc tế (International Normalized Ratio).
IT	Kỹ thuật thông minh (Intelligent Techniques).
LCA	Thuật toán nhóm dẫn đầu (Leader-Cluster Algorithm).
MLF	Mạng FMN đa mức (Multi-Level Fuzzy Min-Max Neural Network).
MLP	Mạng nơron đa lớp (Multi-Layer Perceptron).
MRI	Chụp cộng hưởng từ (Magnetic Resonance Imaging).
NAFLD	Bệnh gan nhiễm mỡ không do rượu (Non-Alcoholic Fatty Liver Disease).

NMFC	Gom cụm dựa trên NMF (Non-negative Matrix Factorization based Consensus).
NoH	Số lượng siêu hộp (Number of Hyperboxes).
OCH	Siêu hộp chồng lấn bù (Overlapped Compensation Hyperbox).
OLN	Nút chồng lấn (Overlapping Neurons).
PLT	Tiểu cầu (Platelet Count).
PT	Thời gian đông máu (Prothrombin Time).
RFMN	FMN phản xạ (Reflex Fuzzy Min-Max Neural Network).
ROI	Vùng quan tâm (Regions of Interest).
SoL	Tỷ lệ mẫu có nhãn (Scale of Labeled Patterns).
SSC	Phân cụm bán giám sát mờ (Semi-Supervised Fuzzy Clustering).
SS-FMM	Học bán giám sát trong FMM (Semi-Supervised Fuzzy Min-Max).
UCI	Cơ sở dữ liệu máy học UCI (University of California, Irvine).
ULN	Giới hạn bình thường trên (Upper Limit of Normal).
WC	Phương pháp phân cụm đồng thuận có trọng số (Weighted Consensus).

DANH MỤC BẢNG

Bảng 2.1 Thông tin các tập dữ liệu thực nghiệm benchmark 63
Bảng 2.2 Giá trị độ đo $Accuracy$ của FMN-Voting và FMN-SAL 71
Bảng 2.3 Ảnh hưởng của δ^{max} đến số lượng siêu hộp hình thành trong FMN-SAL 78
Bảng 2.4 Giá trị ARI trên các tập dữ liệu thực nghiệm [101] 80
Bảng 2.5 Giá trị Accuracy của ba mô hình trên các tập dữ liệu 81
Bảng 2.6 Kết quả đánh giá trên lớp bệnh gan của FMN-SAL 82
Bảng 2.7 Đánh giá chất lượng phân cụm trung bình trên hai lớp của FMN-SAL khi thay đổi $\delta^{\max}$ 83
Bảng 2.8 Khả năng nhận diện lớp bệnh gan của mô hình FMN-Voting . . 86
Bảng 2.9 Kết quả các chỉ số đánh giá mô hình FMN-Voting khi thay đổi $\delta_{\max}$ 86
Bảng 2.10 Kết quả nhận diện lớp bệnh gan của các mô hình đối sánh 89
Bảng 2.11 Kết quả nhận diện lớp không bệnh gan của các mô hình đối sánh 90
Bảng 2.12 Kết quả đánh giá tổng thể của các mô hình đối sánh trên tập dữ liệu ILPD 91
Bảng 2.13 So sánh FMN-SAL, FMN-Voting với các phương pháp khác . . 92

Bảng 3.1 Đánh giá trên lớp bệnh gan của mô hình FA-Seed khi thay đổi $\delta^{\max}$ 116
Bảng 3.2 Kết quả đánh giá mô hình FA-Seed khi thay đổi $\delta^{\max}$ trên tập dữ liệu ILPD 118
Bảng 3.3 Kết quả đánh giá phân cụm của mô hình FA-Seed khi thay đổi $\delta^{\max}$ 120
Bảng 3.4 So sánh nhận diện lớp bệnh gan của FA-Seed với các phương pháp khác 121
Bảng 3.5 So sánh nhận diện lớp không bệnh của FA-Seed với các phương pháp khác 123
Bảng 3.6 So sánh FA-Seed với MSCFMN và các phương pháp phân cụm trên Weka: các độ đo phân loại 124
Bảng 3.7 So sánh FA-Seed với MSCFMN và các phương pháp phân cụm trên Weka: các chỉ số RI, ARI và NMI 125

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

Hình 1.1 *Minh họa phân cụm dữ liệu* 17
Hình 1.2 *Đồ họa minh họa về không gian siêu hộp với hai cụm* 20
Hình 1.3 *Siêu hộp B_j trong không gian 3D với hai đỉnh V_j và W_j* 21
Hình 1.4 *Đồ họa minh họa về sự biến động $\mu_(x_r, B_j)$ khi thay đổi γ* 22
Hình 1.5 *Đồ họa minh họa về sự biến thiên của $\mu(x_r, B_j)$* 22
Hình 1.6 Ví dụ về hoạt động mở rộng và điều chỉnh chồng lấn siêu hộp . . 23
Hình 1.7 *Chồng lấn giữa biên trên của B_j với biên dưới của B_k* 24
Hình 1.8 *Chồng lấn giữa biên trên của B_k với biên dưới của B_j* 24
Hình 1.9 *Chồng lấn dạng siêu hộp B_k bị bao (co lại) trong B_j* 24
Hình 1.10 *Chồng lấn dạng siêu hộp B_j bị bao (co lại) trong B_k* 24
Hình 1.11 *Cấu trúc mạng nơron FMN với học không giám sát* 28
Hình 1.12 *Cấu trúc mạng nơron FMN với học có giám sát* 28
Hình 1.13 *Đồ họa minh họa cấu tạo của nơron b_j* 29

Hình 2.1 Ví dụ phân hoạch không gian dữ liệu Aggregation thành các siêu hộp bằng FMN 39
Hình 2.2 Điểm dữ liệu mang nhãn và đơn vị nhãn 40
Hình 2.3 *Kiến trúc mạng nơron FMN-Voting* 48
Hình 2.4 B_1 mất cân bằng, B_2 cân bằng và có chỉ số sử dụng $c(B_2)$ cao, B_3 có chỉ số sử dụng thấp 55
Hình 2.5 *Kiến trúc mạng nơron FMN-SAL* 55
Hình 2.6 *Minh họa mức độ thuộc $\mu(x, B_j)$ trong không gian đặc trưng và ảnh hưởng đến độ tin cậy suy luận* 62
Hình 2.7 Đồ họa phân bố không gian dữ liệu thực tế của các tập dữ liệu thực nghiệm 65
Hình 2.8 Hiệu năng phân cụm theo chỉ số *Accuracy* trên tập dữ liệu Flame 72
Hình 2.9 Hiệu năng phân cụm theo chỉ số *Accuracy* trên tập dữ liệu Iris . 73
Hình 2.10 Hiệu năng phân cụm theo chỉ số *Accuracy* trên tập dữ liệu Wine 74
Hình 2.11 Hiệu năng phân cụm theo chỉ số *Accuracy* trên tập dữ liệu R15 . 75
Hình 2.12 Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Aggregation 75
Hình 2.13 Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Thyroid 76
Hình 2.14 Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu PID 76

Hình 2.15 Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Spiral 77
Hình 2.16 Hiệu năng phân cụm trung bình theo chỉ số *Accuracy* đạt được bởi FMN-SAL trên tập dữ liệu Jain 77
Hình 2.17 Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Pathbased 77
Hình 2.18 Ảnh hưởng của δ^{max} tới siêu hộp trong mạng FMN-SAL 79
Hình 2.19 Đánh giá độ đo trên lớp bệnh gan của FMN-SAL 83
Hình 2.20 Độ đo phân loại của mô hình FMN-SAL trên tập dữ liệu ILPD . 84
Hình 2.21 Chỉ số đánh giá phân cụm của mô hình FMN-SAL trên tập dữ liệu ILPD 85
Hình 2.22 Khả năng nhận diện lớp bệnh gan của mô hình FMN-Voting . . 87
Hình 2.23 Sự biến động giá trị độ đo của mô hình FMN-Voting 87
Hình 2.24 Giá trị độ đo RI, ARI, NMI của mô hình FMN-Voting 88
Hình 2.25 So sánh các độ đo đánh giá trên lớp bệnh gan của FMN-SAL, FMN-Voting và các phương pháp đối sánh 89
Hình 2.26 Kết quả so sánh FMN-SAL, FMN-Voting với các phương pháp khác 92

Hình 3.1 Phân hoạch siêu hộp bằng FMN trên tập dữ liệu Aggregation theo giả thiết mỗi siêu hộp đại diện cho một cụm 98
Hình 3.2 Phân hoạch siêu hộp bằng FMN trên tập dữ liệu Flame theo giả thiết mỗi siêu hộp đại diện cho một cụm 98
Hình 3.3 Phân hoạch siêu hộp bằng FMN trên tập dữ liệu Pathbased theo giả thiết mỗi siêu hộp đại diện cho một cụm 99
Hình 3.4 Phân hoạch siêu hộp trên tập dữ liệu Aggregation với giới hạn kích thước siêu hộp nhỏ của FA-Seed (41 siêu hộp) 99
Hình 3.5 Phân hoạch siêu hộp trên tập dữ liệu Flame với giới hạn kích thước siêu hộp nhỏ của FA-Seed (11 siêu hộp) 100
Hình 3.6 Phân hoạch siêu hộp trên tập dữ liệu Pathbased với giới hạn kích thước siêu hộp nhỏ của FA-Seed (19 siêu hộp) 100
Hình 3.7 *Kiến trúc mô hình phân cụm bán giám sát mờ với học chủ động FA-Seed* 101
Hình 3.8 Sự phân hoạch các siêu hộp, đánh giá và lựa chọn các siêu hộp ứng viên sinh hạt giống trên tập dữ liệu Aggregation 102
Hình 3.9 Kiến trúc mạng nơron FA-Seed 103

Hình 3.10 Hiệu năng phân cụm theo chỉ số RI đạt được bởi phương pháp đề xuất FA-Seed và FC2 trên tập dữ liệu Iris 110
Hình 3.11 Hiệu năng phân cụm theo chỉ số RI đạt được bởi phương pháp đề xuất FA-Seed và FC2 trên tập dữ liệu Wine 110
Hình 3.12 So sánh giá trị NoS giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu: (a) Ecoli, (b) Iris, (c) Soybean, (d) Thyroid, (e) Aggregation, (f) Zoo 112
Hình 3.13 So sánh giá trị RI giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu: (a) Ecoli, (b) Iris, (c) Soybean, (d) Thyroid, (e) Aggregation, (f) Zoo 113
Hình 3.14 Biểu đồ so sánh trực quan chỉ số $Accuracy$ giữa FA-Seed và MSCFMN 113
Hình 3.15 So sánh thời gian chạy (s) giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu thực nghiệm (Soybean, Zoo, Thyroid) 114
Hình 3.16 So sánh thời gian chạy (s) giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu thực nghiệm (Iris, Ecoli) 115
Hình 3.17 Đánh giá khả năng nhận diện lớp bệnh gan của FA-Seed 116
Hình 3.18 Sự biến động Accuracy của mô hình FA-Seed khi thay đổi giá trị δ^{max} 117
Hình 3.19 Sự biến động Precision của mô hình FA-Seed khi thay đổi giá trị δ^{max} 118
Hình 3.20 Sự biến động Specificity của mô hình FA-Seed khi thay đổi giá trị δ^{max} 118
Hình 3.21 Sự biến động F1-Score của mô hình FA-Seed khi thay đổi giá trị δ^{max} 119
Hình 3.22 Sự biến động của Accuracy, Precision và F1-Score của mô hình FA-Seed 119
Hình 3.23 Sự biến động của các độ đo mô hình FA-Seed khi thay đổi giá trị δ^{max} 120
Hình 3.24 Sự biến động độ đo mô hình FA-Seed qua các lần thực nghiệm . 121
Hình 3.25 So sánh các độ đo đánh giá khả năng nhận diện lớp bệnh gan của FA-Seed với MSCFMN và các phương pháp phân cụm trên Weka 125
Hình 3.26 So sánh các chỉ số RI, ARI và NMI của FA-Seed với MSCFMN và các phương pháp phân cụm trên Weka 125

MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, phân cụm bán giám sát đã 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 thường khan hiếm và chi phí gán nhãn ngày càng đắt đỏ. Việc kết hợp đồng thời dữ liệu có nhãn và không có nhãn cho phép các mô hình học 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 [1, 2, 3, 4]. Trên cơ sở đó, nhiều hướng tiếp cận phân cụm bán giám sát đã được đề xuất, bao gồm các phương pháp dựa trên ràng buộc cặp, các mô hình lan truyền nhãn trên đồ thị, các phương pháp phân cụm phổ bán giám sát, cũng như các mô hình học sâu bán giám sát cho phân cụm [1, 2, 3, 5, 6, 7]. Mặc dù đạt được những kết quả tích cực, các phương pháp này thường gặp hạn chế khi dữ liệu có nhiễu, phân bố không đồng đều hoặc khi yêu cầu khả năng diễn giải của mô hình là yếu tố quan trọng, đặc biệt trong các bài toán hỗ trợ ra quyết định và y sinh [4, 8].

Trong nhiều bài toán thực tế, ranh giới giữa các cụm thường không rõ ràng, dữ liệu có thể bị chồng lấn hoặc chịu ảnh hưởng của nhiễu và sự mất cân bằng phân bố. Do đó, các phương pháp phân cụm mờ, cho phép mỗi đối tượng thuộc về nhiều cụm với các mức độ khác nhau thông qua giá trị độ thuộc, được xem là phù hợp hơn so với các phương pháp phân cụm rõ [9, 10, 11, 12]. Khi kết hợp với học bán giám sát, phân cụm mờ cho phép khai thác hiệu quả thông tin giám sát hạn chế nhằm định hướng quá trình phân hoạch, đồng thời vẫn bảo toàn được tính linh hoạt cần thiết để mô hình hóa các vùng dữ liệu không chắc chắn [13, 14].

Trong số các mô hình phân cụm và phân lớp mờ, mạng nơron Min-Max mờ (Fuzzy Min-Max Neural Network - FMN) tiếp tục được quan tâm nhờ khả năng biểu diễn dữ liệu bằng các siêu hộp mờ, xử lý hiệu quả tính không chắc chắn và đặc biệt là khả năng kết xuất các luật quyết định dạng *IF–THEN* có tính diễn giải cao [8, 15, 16]. Gần đây, một số công trình đã mở rộng FMN theo hướng học bán giám sát nhằm khai thác thông tin nhãn hạn chế để cải thiện chất lượng học [17, 18, 19, 20, 21]. Có thể kể đến các mô hình như SCFMN và MSCFMN, trong đó thông tin giám sát được tích hợp trực tiếp vào quá trình hình thành và điều chỉnh siêu hộp, qua đó nâng cao khả năng phân biệt giữa các lớp, đặc biệt trong các vùng dữ liệu chồng lấn [20, 21].

Đáng chú ý, Liu [17] đã đề xuất mô hình SSL-FMM cho bài toán phân lớp bán giám sát, trong đó các mẫu có nhãn được sử dụng để khởi tạo và dẫn dắt quá trình học, còn các mẫu chưa gán nhãn được tích hợp dần thông qua cơ chế điều chỉnh siêu

hộp. Kết quả thực nghiệm cho thấy mô hình này cải thiện độ chính xác so với các biến thể FMN không giám sát và một số phương pháp bán giám sát truyền thống, qua đó khẳng định tiềm năng của FMN trong bối cảnh học bán giám sát.

Mặc dù đạt được những kết quả tích cực, các mô hình FMN bán giám sát hiện có vẫn còn tồn tại một số hạn chế đáng kể [17, 18, 19, 20, 21]. Cụ thể:

- Thứ nhất, quá trình phân cụm còn phụ thuộc nhiều vào chất lượng của các mẫu hoặc siêu hộp mang nhãn ban đầu. Tuy nhiên, các đối tượng này chưa được đánh giá đầy đủ về mức độ đại diện và độ tin cậy. Do đó, nếu chúng nằm ở vùng biên, vùng chồng lấn hoặc không phản ánh đúng cấu trúc cụm, quá trình lan truyền nhãn có thể bị sai lệch và làm giảm chất lượng phân cụm [1].
- Thứ hai, cơ chế suy luận nhãn chủ yếu dựa trên quan hệ lân cận (độ thuộc) giữa các siêu hộp, trong khi chưa xem xét đầy đủ độ thuần nhất, mật độ phân bố và vai trò đại diện của siêu hộp trong từng cụm. Việc sử dụng các siêu hộp có chất lượng thấp làm nguồn lan truyền nhãn có thể làm gia tăng nguy cơ lan truyền lỗi [17, 18, 19], đặc biệt với dữ liệu có mật độ không đồng đều, cụm mất cân bằng hoặc ranh giới phức tạp.
- Thứ ba, các mô hình hiện có còn gặp khó khăn đối với các mẫu vùng biên hoặc vùng chồng lấn, nơi mức độ không chắc chắn cao có thể ảnh hưởng đến quá trình suy luận và lan truyền nhãn. Đây là những điểm dữ liệu có độ không chắc chắn cao, dễ gây nhập nhằng trong quá trình mở rộng siêu hộp, suy luận nhãn và hình thành cấu trúc cụm [9, 10, 11, 12].

Xuất phát từ các vấn đề nêu trên, luận án tập trung nghiên cứu các hướng tiếp cận mới cho bài toán phân cụm bán giám sát dựa trên mạng nơron Min-Max mờ, với mục tiêu cải thiện cơ chế suy luận nhãn, giảm thiểu lan truyền lỗi và nâng cao chất lượng phân cụm trong điều kiện thông tin giám sát hạn chế.

2. Mục tiêu nghiên cứu

Xuất phát từ các khoảng trống nghiên cứu nêu trên, 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 trong điều kiện thông tin giám sát hạn chế.

Để đạt được mục tiêu tổng quát này, luận án tập trung vào các mục tiêu cụ thể sau:

- 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 để làm nguồn thông tin giám sát cho quá trình phân cụm bán giám sát.

- Đề xuất cơ chế suy luận nhãn mới cho các siêu hộp chưa biết nhãn, trong đó quá trình gán nhãn không chỉ dựa trên quan hệ lân cận hoặc độ thuộc giữa các siêu hộp, mà còn xem xét thêm các yếu tố như 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, đặc biệt trong các trường hợp dữ liệu có mật độ không đồng đều, cụm mất cân bằng hoặc ranh giới cụm phức tạp.
- Nghiên cứu các cơ chế nâng cao độ tin cậy của quá trình suy luận nhãn, hạn chế ảnh hưởng của các mẫu hoặc siêu hộp có tính đại diện thấp.
- Thực nghiệm và đánh giá các mô hình đề xuất trên các tập dữ liệu chuẩn và dữ liệu thực tế, sử dụng các chỉ số đánh giá phù hợp để chứng minh hiệu quả của các cải tiến về độ chính xác, độ ổn định và khả năng ứng dụng của mô hình.

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

Đối tượng nghiên cứu của luận án 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.

Trên cơ sở các mục tiêu đã xác định, luận án tập trung vào ba nội dung nghiên cứu chính: (i) nghiên cứu cơ chế suy luận nhãn dựa trên bỏ phiếu từ các siêu hộp lân cận trong mô hình FMN-Voting; (ii) nghiên cứu cơ chế suy luận nhãn dựa trên các mẫu mang nhãn lân cận trong mô hình FMN-SAL; và (iii) nghiên cứu chiến lược lựa chọn hạt giống chủ động/thích nghi trong mô hình FA-Seed.

Bên cạnh việc nghiên cứu các mô hình và thuật toán trên phương diện phương pháp, luận án còn tập trung kiểm chứng khả năng ứng dụng của các mô hình đề xuất trên nhiều kịch bản dữ liệu khác nhau. Cụ thể, phạm vi đánh giá không chỉ bao gồm các tập dữ liệu benchmark phổ biến dùng để đối sánh khách quan với các phương pháp liên quan, mà còn mở rộng sang bộ dữ liệu y tế về chẩn đoán bệnh gan. Việc đưa bộ dữ liệu chỉ số xét nghiệm máu vào thực nghiệm cho phép đánh giá rõ hơn khả năng thích nghi của các mô hình FMN bán giám sát trong điều kiện dữ liệu thực có nhiễu, cấu trúc phức tạp, chồng lấn giữa các lớp và tỷ lệ mẫu có nhãn hạn chế.

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. Cơ sở khoa học và thực tiễn của đề tài

Về cơ sở khoa học, đề tài được xây dựng trên nền tảng lý thuyết của phân cụm mờ, học bán giám sát và mạng nơron Min-Max mờ (FMN). Trong hướng tiếp cận này, dữ liệu được biểu diễn bởi các siêu hộp mờ, qua đó cho phép mô hình hóa linh hoạt ranh giới giữa các nhóm dữ liệu, xử lý tính không chắc chắn và hỗ trợ cơ chế học gia tăng. Đây là cơ sở lý thuyết quan trọng để áp dụng FMN cho các bài toán phân cụm và phân lớp trong điều kiện dữ liệu phức tạp, thiếu nhãn hoặc có mức độ chồng lấn cao [13, 15, 17, 18, 19].

Bên cạnh đó, các nghiên cứu về phân cụm bán giám sát cho thấy việc kết hợp thông tin giám sát hạn chế với cấu trúc nội tại của dữ liệu là một hướng tiếp cận hiệu quả. Các dạng tri thức giám sát như ràng buộc cặp điểm, hạt giống ban đầu hay cơ chế học chủ động đã được sử dụng để cải thiện chất lượng phân cụm và làm tăng tính định hướng cho quá trình học [1, 5, 6, 7]. Các công trình tổng quan gần đây cũng khẳng định phân cụm bán giám sát tiếp tục là một hướng nghiên cứu quan trọng, đặc biệt trong bối cảnh dữ liệu lớn, dữ liệu không đồng nhất và dữ liệu có ít nhãn [2, 3].

Đối với riêng khung mô hình FMN, các nghiên cứu đã chỉ ra tiềm năng lớn của phương pháp này trong học bán giám sát, song đồng thời cũng bộc lộ một số hạn chế như: độ tin cậy của quá trình suy luận và lan truyền nhãn còn phụ thuộc nhiều vào cấu trúc siêu hộp; việc xử lý dữ liệu chồng lấn, dữ liệu nhiễu và dữ liệu phân bố không đều vẫn còn khó khăn; chất lượng của tập hạt giống ban đầu ảnh hưởng mạnh đến kết quả cuối cùng; và cơ chế khai thác dữ liệu chưa gán nhãn chưa thật sự ổn định trong mọi tình huống [17, 18, 20, 21]. Những khoảng trống đó tạo nên cơ sở khoa học trực tiếp để nghiên cứu các giải pháp nâng cao chất lượng mô hình FMN theo hướng ổn định hơn, linh hoạt hơn và phù hợp hơn với bài toán phân cụm bán giám sát.

Về mặt thực tiễn, trong nhiều bài toán hiện nay, số lượng dữ liệu chưa gán nhãn thường rất lớn, trong khi dữ liệu có nhãn lại hạn chế do chi phí thu thập cao, phụ thuộc vào chuyên gia và mất nhiều thời gian xử lý. Thực trạng này xuất hiện phổ biến trong các lĩnh vực như y tế, nhận dạng, hỗ trợ chẩn đoán và phân tích dữ liệu thông minh, làm cho các phương pháp học bán giám sát trở thành lựa chọn phù hợp hơn so với học có giám sát thuần túy [2, 3, 4].

Ngoài ra, dữ liệu thực tế thường không thuần nhất mà có thể chứa nhiễu, thiếu giá trị, chồng lấn mạnh giữa các nhóm và phân bố không cân bằng. Trong những điều kiện như vậy, hiệu quả của mô hình phụ thuộc nhiều vào khả năng biểu diễn linh hoạt cấu trúc dữ liệu, vào chất lượng của hạt giống cũng như vào cơ chế suy luận nhãn.

Các nghiên cứu về seeding và active learning trong phân cụm mờ cho thấy việc lựa chọn các điểm đại diện tốt và khai thác thông tin phản hồi một cách hợp lý có thể cải thiện đáng kể chất lượng phân cụm [1, 7]. Đây là cơ sở thực tiễn quan trọng để đề tài xem xét các cơ chế lựa chọn hạt giống và tăng cường độ tin cậy cho quá trình phân cụm bán giám sát dựa trên FMN.

Mặt khác, tính thực tiễn của đề tài còn được thể hiện ở khả năng kiểm chứng trên các bộ dữ liệu chuẩn và các kịch bản dữ liệu khác nhau. Việc sử dụng các nguồn dữ liệu phổ biến như UCI và CS [22, 23], cùng với việc xem xét ảnh hưởng của dữ liệu thiếu hoặc dữ liệu phức tạp, cho phép đánh giá khách quan tính đúng đắn, độ ổn định và khả năng ứng dụng của các mô hình đề xuất [22, 23, 24]. Điều này cho thấy đề tài không chỉ có ý nghĩa về mặt lý thuyết mà còn có giá trị thực tiễn trong các bài toán phân tích và khai phá dữ liệu hiện nay. Bên cạnh các bộ dữ liệu chuẩn, tính thực tiễn của đề tài còn được thể hiện thông qua việc kiểm chứng trên bộ dữ liệu y tế ILPD. Đây là bộ dữ liệu về các chỉ số xét nghiệm máu dùng để đánh giá và phân loại bệnh gan trên người. Việc này đặt ra nhiều thách thức đối với các mô hình học bán giám sát như nhiễu thực tế, cấu trúc dữ liệu phức tạp, chồng lấn giữa các nhóm và tỷ lệ dữ liệu có nhãn hạn chế. Việc sử dụng bộ dữ liệu chỉ số xét nghiệm máu trong thực nghiệm cho phép kiểm chứng không chỉ hiệu năng của các mô hình đề xuất trên dữ liệu chuẩn, mà còn đánh giá khả năng thích nghi, độ ổn định và tính ứng dụng của chúng trên dữ liệu y tế thực, nơi yêu cầu về độ tin cậy của suy luận nhãn và khả năng xử lý các vùng dữ liệu không chắc chắn có ý nghĩa đặc biệt quan trọng.

Như vậy, đề tài có cơ sở khoa học dựa trên nền tảng của phân cụm mờ, học bán giám sát và mạng nơron Min-Max mờ; đồng thời có cơ sở thực tiễn từ nhu cầu khai thác dữ liệu trong điều kiện thông tin nhãn hạn chế.

6. Những đóng góp mới của luận án

Luận án có các đóng góp chính sau:

- 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 thực tế về các chỉ số sinh hóa máu nhằm đánh giá toàn diện 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, nội dung của luận án được tổ chức thành ba chương chính như sau:

- Chương 1 trình bày tổng quan về phân cụm bán giám sát và mạng nơron Min-Max mờ, bao gồm các khái niệm nền tảng, nguyên lý hoạt động, cơ chế học và các nghiên cứu liên quan; từ đó phân tích những vấn đề còn tồn tại và xác định hướng nghiên cứu của luận án.
- Chương 2 trình bày hai mô hình phân cụm bán giám sát dựa trên FMN là FMN-Voting và FMN-SAL. Trong đó, FMN-Voting sử dụng cơ chế bỏ phiếu từ các siêu hộp lân cận, còn FMN-SAL khai thác thông tin từ k mẫu mang nhãn lân cận để suy luận nhãn. Chương này đồng thời trình bày thuật toán, phân tích độ phức tạp và đánh giá thực nghiệm các mô hình đề xuất.
- Chương 3 trình bày mô hình FA-Seed với chiến lược lựa chọn hạt giống dựa trên cấu trúc siêu hộp và mức độ đại diện của dữ liệu, kết hợp thông tin phản hồi từ chuyên gia. Chương này trình bày mô hình, thuật toán và các thực nghiệm đánh giá hiệu quả của phương pháp đề xuất.

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

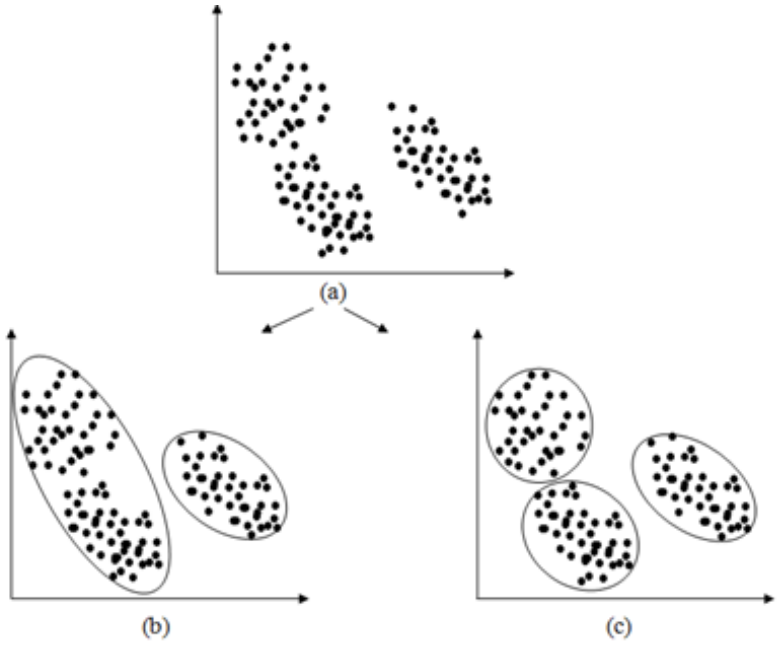
Nội dung Chương 1 cung cấp nền tảng lý thuyết liên quan đến phân cụm bán giám sát mờ và mô hình mạng nơron FMN. Thứ nhất, trình bày tổng quan về phân cụm dữ liệu, những hạn chế của phân cụm truyền thống và động lực xuất hiện của hướng tiếp cận bán giám sát. Thứ hai, trình bày về các phương pháp phân cụm bán giám sát mờ, làm rõ ưu điểm, nhược điểm và các thách thức trong thực tế. Phần cuối của chương giới thiệu khái quát về FMN, bao gồm nguyên lý hoạt động, cơ chế hình thành siêu hộp, hàm độ thuộc mờ và thuật toán học. Sự tổng hợp này nhằm cung cấp cái nhìn toàn diện về các vấn đề nghiên cứu hiện tại, chỉ ra khoảng trống nghiên cứu và đặt nền móng cho các cải tiến và thuật toán mới được trình bày trong Chương 2 và Chương 3.

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

Phân cụm là một kỹ thuật quan trọng trong khai phá dữ liệu, thuộc nhóm phương pháp học không giám sát trong học máy [25, 26, 27]. Mục tiêu của phân cụm là nhóm các đối tượng tương đồng vào cùng một cụm [15, 28]. Tuy nhiên, việc lựa chọn phương pháp phân cụm phụ thuộc vào đặc điểm dữ liệu, mục tiêu bài toán và sự đánh đổi giữa chất lượng cụm và chi phí tính toán. Do đó, không tồn tại một kỹ thuật hay tiêu chí đánh giá duy nhất phù hợp cho mọi bài toán phân cụm.

Hình 1.1 minh họa dữ liệu trong không gian hai chiều có thể được phân thành hai hoặc ba cụm tùy theo quan điểm phân tích. Điều này cho thấy kết quả phân cụm không hoàn toàn duy nhất mà phụ thuộc vào tiêu chuẩn phân cụm được lựa chọn.

Trong thực tế, phân cụm thường được chia thành hai nhóm chính: phân cụm rõ [29, 30, 31, 32] và phân cụm mờ [9, 33, 34, 35]. Phân cụm rõ gán mỗi điểm dữ liệu vào duy nhất một cụm, trong khi phân cụm mờ cho phép một điểm thuộc nhiều cụm thông qua các giá trị độ thuộc trong khoảng $[0, 1]$. Nhờ đó, phân cụm mờ phù hợp hơn với các bài toán có ranh giới cụm không rõ ràng hoặc có hiện tượng chồng lấn dữ liệu [10, 11, 12]. Nhiều thuật toán phân cụm mờ đã được đề xuất như FCM (Fuzzy C-Means), ε-FCM (ε-insensitive Fuzzy C-Means), FPCM (Fuzzy Possibilistic C-Means), song đa số đòi hỏi xác định trước số cụm, là tham số ảnh hưởng lớn đến kết quả.



Hình 1.1: *Minh họa phân cụm dữ liệu*

Phân cụm bán giám sát mờ (Semi-Supervised Fuzzy Clustering – SSC) là sự mở rộng của phân cụm mờ bằng cách khai thác một phần thông tin biết trước để định hướng quá trình phân cụm, từ đó nâng cao chất lượng và độ ổn định của kết quả [2, 3, 14]. Pedrycz đã chỉ ra rằng ngay cả một tỷ lệ nhỏ dữ liệu được gán nhãn cũng có thể cải thiện đáng kể hiệu quả phân cụm [13]. Các thông tin bổ trợ thường được biểu diễn dưới dạng ràng buộc *must-link/cannot-link*, nhãn hoặc độ thuộc xác định trước [8, 36, 37]. Một số mô hình tiêu biểu của hướng tiếp cận này gồm eSFCM [38], SSSFC [39], GSOM [40] và GFMM [15].

Phân cụm bán giám sát mờ kết hợp khả năng biểu diễn tính không chắc chắn của phân cụm mờ với một lượng nhỏ thông tin giám sát nhằm cải thiện kết quả học [41, 42]. Một số hướng tiếp cận tiêu biểu: (1) Phân cụm bán giám sát mờ dựa trên mẫu hạt giống; (2) Phân cụm bán giám sát mờ dựa trên ràng buộc cặp; (3) Phân cụm bán giám sát mờ dựa trên mô hình nguyên mẫu; và (4) Phân cụm bán giám sát mờ dựa trên mạng nơron mờ.

Các phương pháp dựa trên mẫu hạt giống sử dụng một tập nhỏ các mẫu đã gán nhãn để định hướng quá trình phân cụm [43, 44, 45]. Trong phân cụm mờ, thông tin này thường được tích hợp bằng cách cố định hoặc ưu tiên giá trị độ thuộc của các mẫu hạt giống vào cụm tương ứng [46, 47, 48]. Cách tiếp cận này giúp giảm phụ thuộc vào khởi tạo ngẫu nhiên và cải thiện độ ổn định.

Những nghiên cứu nền tảng cho hướng này có thể kể đến Bensaid và cộng sự [49] và Pedrycz [13]. Trên cơ sở đó, nhiều biến thể như Fuzzy C-Means khởi tạo bằng hạt giống (SFCM) [50, 51] hay FCM bán giám sát [14, 52, 53] đã được phát triển.

Gần đây, các phương pháp này còn được mở rộng theo hướng chủ động để lựa chọn hạt giống có giá trị thông tin cao hơn [42, 54, 55]. Tuy nhiên, hiệu quả của chúng vẫn phụ thuộc mạnh vào chất lượng, số lượng và tính đại diện của tập hạt giống [56, 57, 58].

Hướng tiếp cận dựa trên ràng buộc sử dụng các ràng buộc cặp dữ liệu dưới dạng *must-link/cannot-link* thay vì yêu cầu nhãn cụ thể cho từng mẫu [14, 59, 60]. Trong các mô hình mờ, các ràng buộc được tích hợp vào hàm mục tiêu như các điều khoản phạt mềm nhằm cân bằng giữa thỏa mãn ràng buộc và bảo toàn cấu trúc dữ liệu [61, 62, 63, 64]. Ưu điểm của cách tiếp cận này là linh hoạt và phù hợp khi khó thu thập nhãn đầy đủ, nhưng vẫn đối mặt với các thách thức về chi phí xây dựng ràng buộc, tính không nhất quán và khả năng mở rộng.

Trong các phương pháp dựa trên nguyên mẫu, mỗi cụm được biểu diễn bởi một hoặc nhiều nguyên mẫu như tâm cụm hoặc đại diện hình học [65, 66]. Thông tin bán giám sát được tích hợp thông qua việc ràng buộc vị trí hoặc độ thuộc của các nguyên mẫu dựa trên dữ liệu đã biết [67, 68, 69]. Nhóm phương pháp này có ưu điểm về khả năng giải thích và giảm độ nhạy với nhiễu, song hiệu quả phụ thuộc lớn vào cơ chế xác định, cập nhật và ràng buộc nguyên mẫu.

Các mô hình dựa trên mạng nơron mờ, đặc biệt là cấu trúc siêu hộp, cung cấp một cách biểu diễn hình học trực quan cho phân cụm bán giám sát [43, 58]. Trong đó, mỗi cụm hoặc vùng dữ liệu được biểu diễn bởi một hoặc nhiều siêu hộp, còn độ thuộc được xác định dựa trên vị trí tương đối của điểm dữ liệu với các siêu hộp. Thông tin giám sát có thể được tích hợp thông qua gán nhãn siêu hộp, ràng buộc mở rộng/chồng lấn hoặc điều chỉnh hàm thuộc [70, 71, 72]. Các mô hình này đặc biệt phù hợp với dữ liệu chồng lấn và có nhiễu [73, 74], nhưng vẫn gặp khó khăn trong việc kiểm soát số lượng siêu hộp và xử lý chồng lấn phức tạp.

Nhìn chung, các hướng tiếp cận phân cụm bán giám sát mờ đều khai thác một lượng nhỏ thông tin giám sát để cải thiện chất lượng phân cụm [2, 3]. Các phương pháp dựa trên hạt giống có ưu điểm định hướng tốt quá trình học nhưng phụ thuộc mạnh vào chất lượng hạt giống [1, 13, 50, 52, 58, 59]. Các phương pháp dựa trên ràng buộc cặp linh hoạt hơn trong khai thác thông tin giám sát, song gặp khó khăn về chi phí và tính nhất quán của ràng buộc [5, 6, 7, 36, 37, 60]. Trong khi đó, các phương pháp nguyên mẫu có khả năng giải thích tốt nhưng phụ thuộc vào chất lượng biểu diễn của nguyên mẫu [65, 66, 75].

So với các hướng tiếp cận trên, FMN là một lựa chọn phù hợp hơn cho bài toán

của luận án [15, 16, 17, 18, 19, 76]. FMN biểu diễn dữ liệu bằng các siêu hộp mờ, cho phép mô hình hóa linh hoạt các vùng lõi, vùng biên và vùng chồng lấn; đồng thời có cơ chế học gia tăng, hỗ trợ cập nhật trực tuyến mà không cần huấn luyện lại toàn bộ mô hình [15, 16, 76]. Ngoài ra, do tri thức được biểu diễn qua siêu hộp mờ, FMN cũng có khả năng diễn giải tốt trong các bài toán hỗ trợ quyết định [8, 20, 21, 77].

Tuy nhiên, các nghiên cứu hiện có cho thấy FMN vẫn tồn tại những hạn chế như phụ thuộc vào cấu trúc siêu hộp, nguy cơ lan truyền lỗi nhãn trong vùng chồng lấn và sự phụ thuộc mạnh vào chất lượng tập hạt giống ban đầu [17, 18, 19]. Đây chính là khoảng trống nghiên cứu để luận án lựa chọn FMN làm nền tảng phát triển các mô hình FMN-Voting, FMN-SAL và FA-Seed theo hướng nâng cao độ ổn định suy luận nhãn, giảm lan truyền lỗi và cải thiện chất lượng học bán giám sát trong điều kiện nhãn hạn chế.

Một phần nội dung tổng quan, phân tích các hướng tiếp cận phân cụm bán giám sát mờ và cơ sở lựa chọn mạng nơron Min-Max mờ làm nền tảng nghiên cứu của luận án đã được tác giả công bố trong công trình [CT1].

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

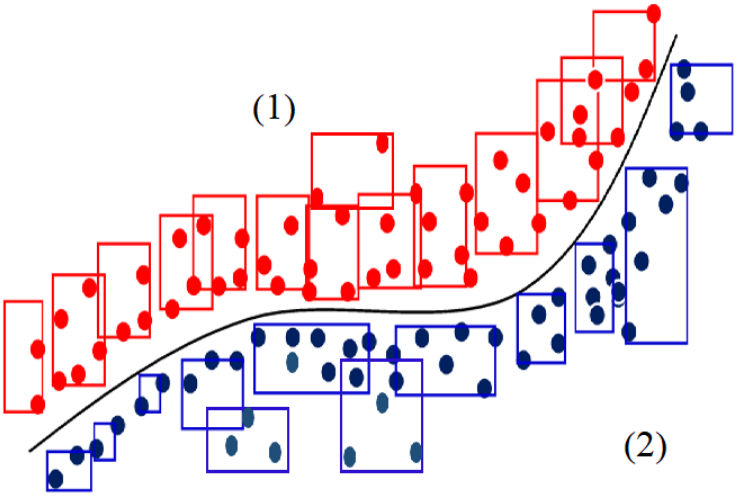
1.2.1. Giới thiệu về mạng nơron Min-Max mờ

Mô hình FMN do Patrick K. Simpson đề xuất cho bài toán phân lớp vào năm 1992 [16] và được mở rộng cho phân cụm vào năm 1993 [76]. FMN được phát triển dựa trên các nguyên lý của FART (Fuzzy Adaptive Resonance Theory) [78] và lý thuyết cộng hưởng thích nghi ART (Adaptive Resonance Theory) [79], cho phép mô hình vừa học ổn định vừa thích nghi với dữ liệu mới.

FMN là một mô hình mạng nơron học gia tăng, kết hợp logic mờ, mạng nơron nhân tạo và lý thuyết min-max mờ nhằm giải quyết hiệu quả các bài toán phân lớp và phân cụm. Tri thức trong FMN được biểu diễn dưới dạng các siêu hộp mờ, mỗi siêu hộp xác định một vùng con trong không gian đặc trưng thông qua hai vectơ biên min và max [16, 76]. Mỗi mẫu dữ liệu được phân lớp hoặc phân cụm dựa trên mức độ thuộc của nó đối với các siêu hộp tương ứng, nhờ đó FMN cung cấp các quyết định mềm thay vì ranh giới cứng. Hình 1.2 minh họa về FMN phân dữ liệu thành hai cụm dựa trên phân hoạch không gian dữ liệu thành các siêu hộp.

Thuật toán học của FMN dựa trên hai thao tác chính là mở rộng và co lại các siêu hộp để bao phủ dữ liệu huấn luyện, đồng thời hạn chế sự chồng lấn giữa các siêu hộp thuộc các lớp hoặc cụm khác nhau. Nhờ cơ chế học gia tăng, FMN đặc biệt phù

hợp với các bộ dữ liệu quy mô lớn và các bài toán dữ liệu động, trong đó mô hình có thể cập nhật tri thức từng bước mà không cần huấn luyện lại toàn bộ.



Hình 1.2: *Đồ họa minh họa về không gian siêu hộp với hai cụm*

FMN có cấu trúc và cơ chế hoạt động tương đối đơn giản, không yêu cầu xác định trước số lượng cụm, chỉ cần điều chỉnh một số ít tham số như kích thước tối đa của siêu hộp và tham số mờ. Đồng thời, FMN là một phương pháp phân cụm phi tuyến, có khả năng phát hiện các cụm với hình dạng và kích thước khác nhau trong không gian đặc trưng.

1.2.2. Hàm thuộc siêu hộp mờ

1.2.2.1. Siêu hộp mờ

Lý thuyết Min-Max mờ (Fuzzy Min-Max - FMM) mô hình hóa miền quyết định bằng các siêu hộp mờ trong không gian đặc trưng $\mathbb{R}^n$. Mỗi siêu hộp mờ được đặc trưng bởi cặp vectơ biên (ký hiệu là V và W) xác định miền $[V, W]$ bao phủ một phần dữ liệu thuộc cùng một khái niệm/lớp/cụm [16, 76].

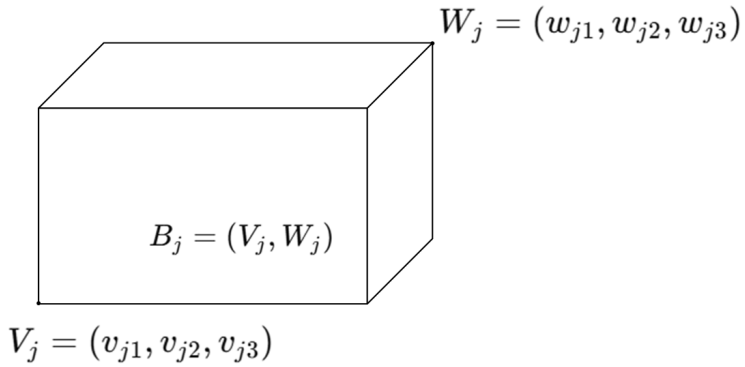
Cho $\mathcal{B}$ là tập các siêu hộp mờ được định nghĩa theo Công thức (1.1), khi đó siêu hộp mờ B_j ($B_j \in \mathcal{B}$) được định nghĩa theo Công thức (1.2):

$$\mathcal{B} = \{B_j\}_{j=1}^{K} \quad (\text{với } K = |\mathcal{B}|) \tag{1.1}$$

$$\begin{gathered} B_j = \{x \in \mathbb{R}^n \mid v_{ji} \leq x_i \leq w_{ji},\ \forall i = 1, \ldots, n\}, \\ V_j = (v_{j1}, \ldots, v_{jn}), \quad W_j = (w_{j1}, \ldots, w_{jn}), \\ v_{ji} \leq w_{ji},\ \forall i = 1, \ldots, n. \end{gathered} \tag{1.2}$$

Hình 1.3 là một ví dụ minh họa về một siêu hộp mờ B_j trong không gian 3D. Siêu hộp B_j được biểu diễn bởi hai vectơ cận dưới và cận trên, $V_j = (v_{j1}, v_{j2}, v_{j3})$ và $W_j = (w_{j1}, w_{j2}, w_{j3})$, trong đó v_{ji} và w_{ji} lần lượt là cận dưới và cận trên của siêu hộp j trên chiều đặc trưng thứ i.

Mỗi siêu hộp mờ B_j được gán một nhãn lớp $\ell(B_j) \in \{1, \ldots, C\}$, trong đó



Hình 1.3: *Siêu hộp B_j trong không gian 3D với hai đỉnh V_j và W_j*

$\ell(B_j) = k$ cho biết siêu hộp B_j thuộc về lớp thứ k.

1.2.2.2. Độ thuộc mờ

Giả sử tập dữ liệu X có M điểm dữ liệu ($X = \{x_r\}_{r=1}^{M},\ x_r \in \mathbb{R}^n$), với mỗi điểm x_r có thể ánh xạ vào C cụm dữ liệu. Thay vì gán nhãn cứng, FMM sử dụng hàm thành viên mờ (hay hàm thuộc mờ) đánh giá mức độ thuộc của điểm x_r vào siêu hộp mờ B_j, ký hiệu là $\mu(x_r, B_j)$, được định nghĩa theo Công thức (1.3) [76]:

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

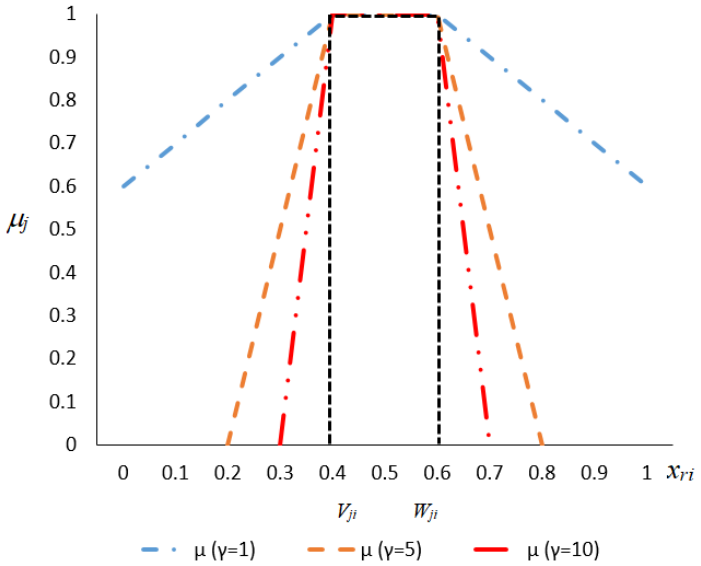
trong đó:

- $f(\cdot,\cdot)$ là hàm hai tham số dùng để điều chỉnh mức suy giảm của giá trị độ thuộc $\mu(x_r, B_j)$ dựa trên mối quan hệ giữa điểm dữ liệu x_r và biên của siêu hộp B_j, được định nghĩa theo Công thức (1.4);
- γ là tham số mờ dùng để điều chỉnh tốc độ suy giảm của giá trị hàm thuộc $\mu(x_r, B_j)$ khi một điểm dữ liệu nằm ngoài siêu hộp; giá trị γ càng lớn thì độ thuộc suy giảm càng nhanh, khiến mô hình phản ứng nhạy hơn đối với các vi phạm ranh giới. Hình 1.4 minh họa sự biến đổi của $\mu(x_r, B_j)$ khi thay đổi γ, Hình 1.5 minh họa sự biến đổi của $\mu(x_r, B_j)$ khi mẫu rời xa siêu hộp [15].

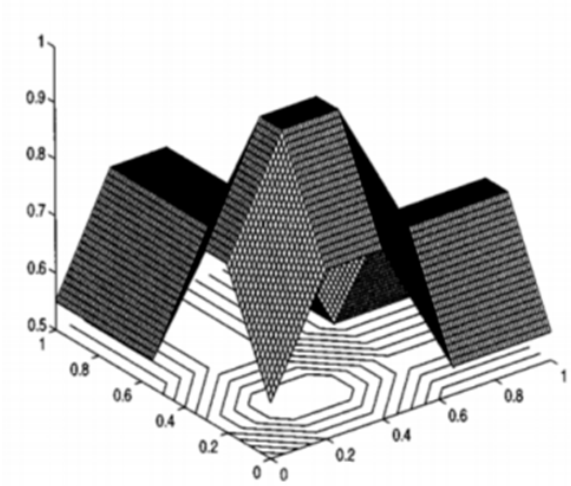
$$f(s,z) = \begin{cases} 1, & sz > 1, \\ sz, & 0 \le sz \le 1, \\ 0, & sz < 0. \end{cases} \tag{1.4}$$

Kích thước của siêu hộp B_j (ký hiệu là δ_{B_j}) là một đại lượng thuộc khoảng $[0, 1]$, phản ánh mức độ mở rộng trung bình của siêu hộp trên các chiều đặc trưng, được tính theo Công thức (1.5) [76]:

$$\delta_{B_j} = \frac{1}{n}\sum_{i=1}^{n}(w_{ji} - v_{ji}) \tag{1.5}$$



Hình 1.4: *Đồ họa minh họa về sự biến động $\mu_(x_r, B_j)$ khi thay đổi γ*



Hình 1.5: *Đồ họa minh họa về sự biến thiên của $\mu(x_r, B_j)$*

1.2.2.3. Phân hoạch dữ liệu với các siêu hộp mờ

Để đảm bảo khả năng thích nghi của mô hình FMM, các thao tác cơ bản được sử dụng trong quá trình huấn luyện siêu hộp mờ bao gồm (1) mở rộng, (2) kiểm tra chồng lấn, và (3) điều chỉnh chồng lấn (co lại) siêu hộp [16, 76].

Mở rộng là hoạt động cập nhật biên của siêu hộp mờ nhằm bao phủ thêm một điểm dữ liệu. Hình 1.6a và Hình 1.6b là đồ họa minh họa về hoạt động mở rộng siêu hộp B_j. Cụ thể với điểm dữ liệu $x_r \in \mathbb{R}^n$, với mỗi chiều đặc trưng i, biên dưới và biên trên của siêu hộp $B_j = (V_j, W_j)$ được cập nhật theo Công thức (1.6) và (1.7):

$$v_{ji} \leftarrow \min(v_{ji}, x_{ri}), \tag{1.6}$$

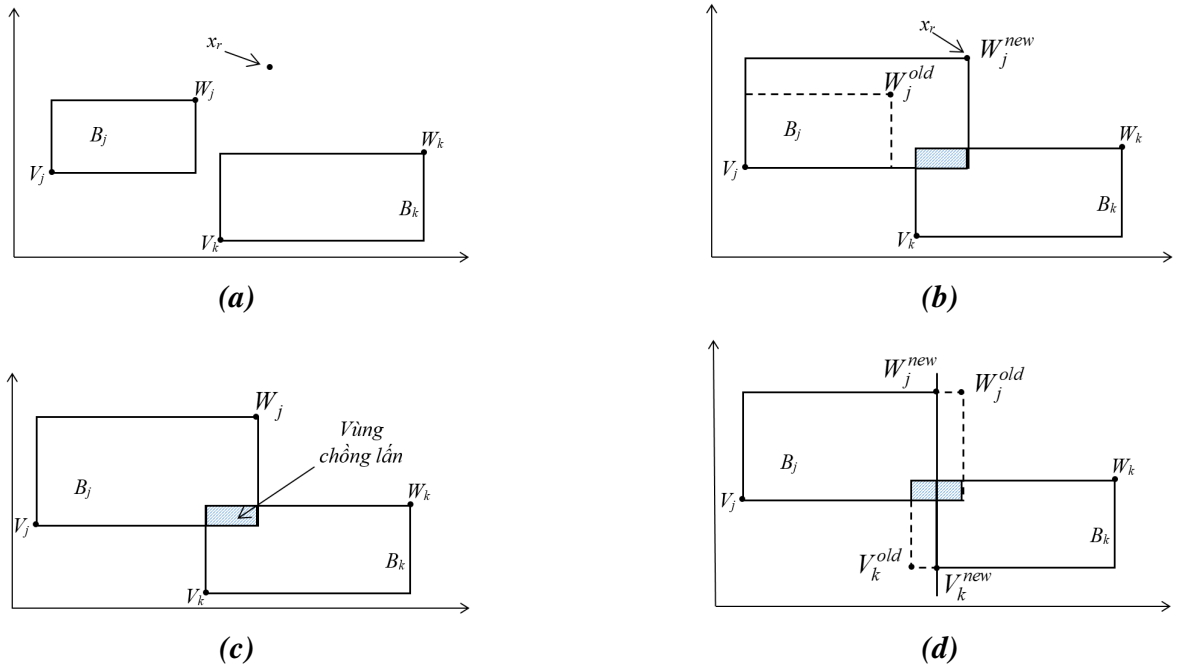
$$w_{ji} \leftarrow \max(w_{ji}, x_{ri}). \tag{1.7}$$

Phép mở rộng chỉ được chấp nhận nếu siêu hộp sau mở rộng vẫn thỏa mãn ràng buộc kích thước (được tính toán theo Công thức (1.8) hoặc (1.9)) nhằm tránh tạo ra

các siêu hộp quá lớn làm giảm khả năng phân biệt của mô hình [16, 76].

$$\max_i(w_{ji} - v_{ji}) \leq \delta^{max}. \tag{1.8}$$

$$\frac{1}{n}\sum_i(w_{ji} - v_{ji}) \leq \delta^{max}. \tag{1.9}$$



Hình 1.6: *Ví dụ về hoạt động mở rộng và điều chỉnh chồng lấn siêu hộp*

Hoạt động mở rộng có thể gây ra hiện tượng chồng lấn giữa các siêu hộp (thuộc lớp khác nhau). Kiểm tra chồng lấn nhằm xác định xem siêu hộp vừa được mở rộng B_j có giao với một siêu hộp khác lớp B_k ($\ell(B_j) \neq \ell(B_k)$) hay không. Trong FMN, chồng lấn được xét theo từng chiều đặc trưng: trên chiều i, nếu hai khoảng $[v_{ji}, w_{ji}]$ và $[v_{ki}, w_{ki}]$ giao nhau (thỏa mãn (1.10)),

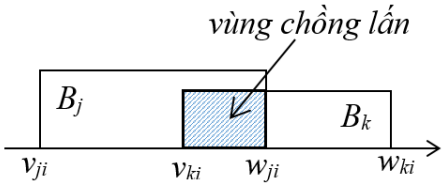
$$\max(v_{ji}, v_{ki}) < \min(w_{ji}, w_{ki}), \tag{1.10}$$

thì chiều đó được xem là gây chồng lấn. Hình 1.6a và Hình 1.6b là đồ họa minh họa về hoạt động mở rộng siêu hộp B_j khi tiếp nhận điểm dữ liệu x_r đã tạo ra vùng chồng lấn. Nếu tồn tại ít nhất một chiều gây chồng lấn, mô hình xác định các phương án điều chỉnh chồng lấn (cắt bớt biên) nhằm loại bỏ phần giao nhau với mức thay đổi nhỏ nhất. Bước loại bỏ chồng lấn được thực hiện khi hai siêu hộp B_j và B_k thuộc hai lớp khác nhau. Trong FMN, có bốn dạng chồng lấn cơ bản giữa hai siêu hộp B_j và B_k [16]:

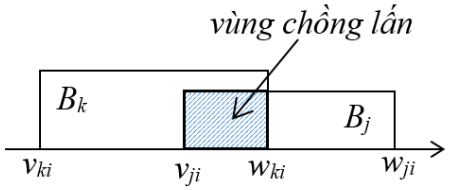
- Trường hợp 1. Biên trên của B_j chồng lấn với biên dưới của B_k:

$$v_{ji} < v_{ki} < w_{ji} < w_{ki}. \tag{1.11}$$

- Trường hợp 2. Biên dưới của B_j chồng lấn với biên trên của B_k:



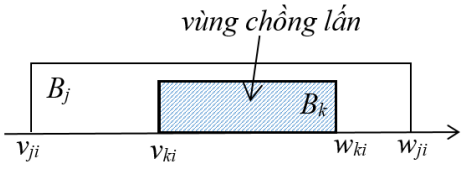
Hình 1.7: *Chồng lấn giữa biên trên của B_j với biên dưới của B_k*



Hình 1.8: *Chồng lấn giữa biên trên của B_k với biên dưới của B_j*

$$v_{ki} < v_{ji} < w_{ki} < w_{ji}. \tag{1.12}$$

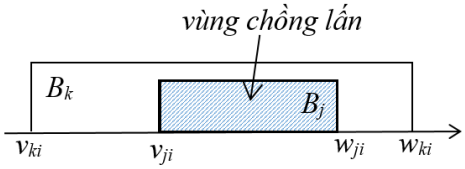
- Trường hợp 3. Siêu hộp B_k bị bao (co lại) trong B_j:



Hình 1.9: *Chồng lấn dạng siêu hộp B_k bị bao (co lại) trong B_j*

$$v_{ji} < v_{ki} \leq w_{ki} < w_{ji}. \tag{1.13}$$

- Trường hợp 4. Siêu hộp B_j bị bao (co lại) trong B_k:



Hình 1.10: *Chồng lấn dạng siêu hộp B_j bị bao (co lại) trong B_k*

$$v_{ki} < v_{ji} \leq w_{ji} < w_{ki}. \tag{1.14}$$

Điều chỉnh chồng lấn nhằm loại bỏ phần giao nhau, qua đó khôi phục ranh giới phân tách giữa các lớp [16]. Hình 1.6c và Hình 1.6d là đồ họa minh họa về hoạt động điều chỉnh chồng lấn giữa hai siêu hộp có chồng lấn. Khi hai siêu hộp B_j và B_k xảy ra chồng lấn trên chiều i, bốn trường hợp chồng lấn được xem xét đến như sau:

- Trường hợp 1. Biên trên của siêu hộp B_j chồng lấn với biên dưới của siêu hộp B_k: điều chỉnh biên dưới của B_k và biên trên của B_j theo (1.15) và (1.16):

$$v_{ki}^{\text{new}} = \frac{v_{ki}^{\text{old}} + w_{ji}^{\text{old}}}{2}, \tag{1.15}$$

$$w_{ji}^{\text{new}} = \frac{v_{ki}^{\text{old}} + w_{ji}^{\text{old}}}{2}. \tag{1.16}$$

- Trường hợp 2. Biên dưới của siêu hộp B_j chồng lấn với biên trên của siêu hộp B_k: điều chỉnh biên dưới của B_j và biên trên của B_k theo (1.17) và (1.18):

$$v_{ji}^{\text{new}} = \frac{v_{ji}^{\text{old}} + w_{ki}^{\text{old}}}{2}, \tag{1.17}$$

$$w_{ki}^{\text{new}} = \frac{v_{ji}^{\text{old}} + w_{ki}^{\text{old}}}{2}. \tag{1.18}$$

- Trường hợp 3. Siêu hộp B_k bị bao hoàn toàn trong siêu hộp B_j:
  + Nếu thỏa mãn điều kiện (1.19):

$$w_{ji} - v_{ki} \leq w_{ki} - v_{ji}, \tag{1.19}$$

  thì điều chỉnh biên dưới của B_j theo (1.20):

$$v_{ji}^{\text{new}} = w_{ki}^{\text{old}}. \tag{1.20}$$

  + Ngược lại, nếu thỏa mãn (1.21):

$$w_{ki} - v_{ji} > w_{ji} - v_{ki}, \tag{1.21}$$

  thì điều chỉnh biên trên của B_j theo (1.22):

$$w_{ji}^{\text{new}} = v_{ki}^{\text{old}}. \tag{1.22}$$

- Trường hợp 4. Siêu hộp B_j bị bao hoàn toàn trong siêu hộp B_k nếu thỏa mãn Công thức (1.14):
  + Nếu thỏa mãn điều kiện (1.23):

$$w_{ki} - v_{ji} \leq w_{ji} - v_{ki}, \tag{1.23}$$

  thì điều chỉnh biên trên của B_k theo (1.24):

$$w_{ki}^{\text{new}} = v_{ji}^{\text{old}}. \tag{1.24}$$

  + Ngược lại, nếu thỏa mãn điều kiện ràng buộc (1.25):

$$w_{ji} - v_{ki} > w_{ki} - v_{ji}, \tag{1.25}$$

  thì điều chỉnh biên dưới của B_k theo (1.26):

$$v_{ki}^{\text{new}} = w_{ji}^{\text{old}}. \tag{1.26}$$

Nhờ đặc trưng hình học này, FMM có thể học trực tuyến, cập nhật gia tăng mà không cần huấn luyện lại, đồng thời biểu diễn tự nhiên các ranh giới mơ hồ và vùng chuyển tiếp giữa các lớp/cụm.

1.2.2.4. Phân hoạch và gán mẫu thích nghi với các siêu hộp mờ

Quá trình phân hoạch và gán mẫu vào các siêu hộp mờ được thực hiện dựa trên độ thuộc của mẫu và các quy tắc bao chứa, mở rộng hoặc tạo mới siêu hộp [76].

Cho tập $\mathcal{B}$, mẫu dữ liệu $x \in X$, lựa chọn siêu hộp $B_j \in \mathcal{B}$ có độ thuộc lớn nhất đối với x ($\mu(x, B_j)$ được tính theo Công thức (1.3)). Không gian các siêu hộp sau đó được cập nhật theo các quy tắc sau:

(1) Quy tắc bao chứa: Nếu tồn tại B_j sao cho x có độ thuộc đầy đủ vào B_j thì gán trực tiếp x cho B_j, trong đó tính duy nhất được đảm bảo bởi điều kiện không giao nhau. Siêu hộp B_j và siêu hộp B_k được gọi là không giao nhau nếu $B_j \cap B_k = \varnothing$ với mọi $j \neq k$, ký hiệu là:

$$\Omega(B_j, B_k) \leq 0, \text{ với mọi } j \neq k, \tag{1.27}$$

trong đó $\Omega(B_j, B_k)$ là mức độ chồng lấn (phần giao) giữa hai siêu hộp B_j và B_k; $\Omega(B_j, B_k) > 0$ khi hai siêu hộp giao nhau.

Quy tắc bao chứa được sử dụng để gán trực tiếp một mẫu dữ liệu x cho siêu hộp mờ B_j nếu độ thuộc của mẫu đối với siêu hộp này đạt giá trị cực đại, tức là:

$$\mu(x, B_j) = 1. \tag{1.28}$$

$\mu(x, B_j) = 1$ thể hiện rằng x nằm hoàn toàn bên trong siêu hộp mờ B_j trên tất cả các chiều đặc trưng, hay $v_{ji} \leq x_i \leq w_{ji}, \forall i = 1, \ldots, n$. Điều này cho thấy mẫu x nằm hoàn toàn bên trong biên của B_j và không cần thực hiện các thao tác mở rộng hay tạo siêu hộp mờ mới.

Tính duy nhất của phép gán được đảm bảo bởi điều kiện (1.27). Theo đó, các siêu hộp mờ không chồng lấn nhau trong không gian đặc trưng, tức $B_j \cap B_k = \varnothing$. Nhờ vậy, mỗi mẫu dữ liệu chỉ có thể được bao chứa hoàn toàn bởi duy nhất một siêu hộp mờ, từ đó đảm bảo tính nhất quán và ổn định của mô hình.

(2) Quy tắc mở rộng: Thực hiện mở rộng tạm thời siêu hộp chiến thắng B_{j^*}, việc mở rộng B_{j^*} được thực hiện dựa trên việc tính toán lại các điểm min và max của siêu hộp B_{j^*} theo Công thức (1.29):

$$B_j^* = \big(\min(V_j, x), \max(W_j, x)\big). \tag{1.29}$$

Việc mở rộng được chấp nhận nếu B_{j^*} thỏa mãn ràng buộc (1.30) và (1.27) với mọi $j \neq k$:

$$\delta(x, B_j) \leq \delta^{\max}. \tag{1.30}$$

Kích thước siêu hộp sau mở rộng được tính bởi:

$$\delta(x, B_j) = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\big|\max(x_i, w_{ji}) - \min(x_i, v_{ji})\big|^2}. \tag{1.31}$$

Nếu mở rộng được chấp nhận, mẫu x được gán cho B_{j^*} đã cập nhật.

(3) Quy tắc tạo mới siêu hộp: Nếu việc mở rộng siêu hộp không được chấp

nhận, một siêu hộp mới được khởi tạo theo (1.32):

$$B_{\text{new}} = (x, x). \tag{1.32}$$

(4) Quy tắc khởi tạo siêu hộp: Trong mạng nơron FMN, siêu hộp mờ là đơn vị biểu diễn cơ bản của dữ liệu trong không gian đặc trưng. Việc khởi tạo siêu hộp đóng vai trò khởi đầu cho quá trình học, tạo nền tảng cho các thao tác mở rộng, kiểm tra và điều chỉnh siêu hộp ở các bước tiếp theo. Một cách khởi tạo hợp lý giúp đảm bảo tính nhất quán của biểu diễn ban đầu và hạn chế các ràng buộc không cần thiết trong quá trình phân hoạch không gian dữ liệu. Với siêu hộp mờ $B_j \in \mathcal{U}$, các vectơ biên dưới và biên trên được khởi tạo theo (1.33):

$$V_j = \mathbf{1}_n, \qquad W_j = \mathbf{0}_n, \tag{1.33}$$

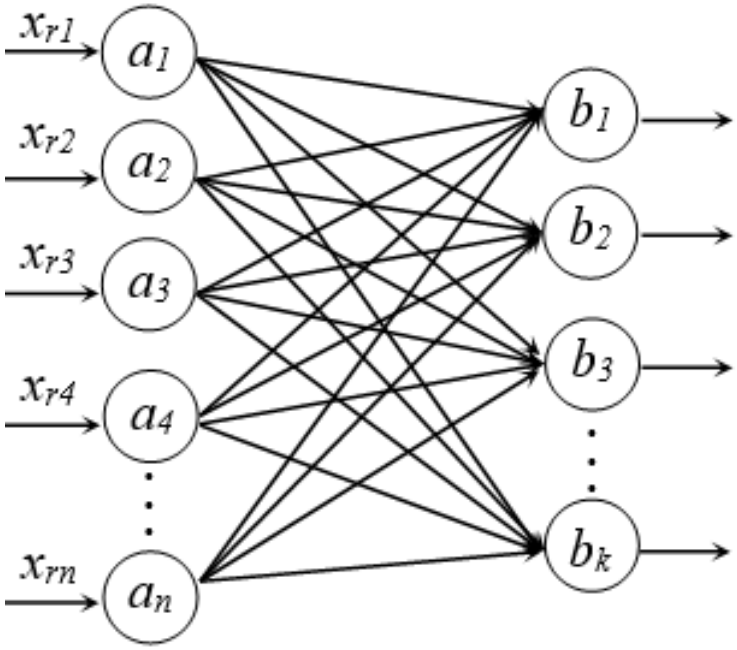
trong đó $\mathbf{1}_n$ và $\mathbf{0}_n$ lần lượt là các vectơ n chiều có tất cả phần tử bằng 1 và 0.

1.2.3. Cấu trúc mạng nơron FMN

Mạng nơron Min-Max mờ FMN là một mô hình Neuro-Fuzzy kết hợp giữa lý thuyết siêu hộp mờ trong FMM và kiến trúc của mạng nơron mờ (Fuzzy Neural Network – FNN), nhằm khai thác ưu điểm của cả hai hướng tiếp cận [80]. FMN vừa kế thừa khả năng biểu diễn tri thức dưới dạng các miền quyết định mờ, vừa tận dụng cấu trúc mạng nơron để thực hiện quá trình học và suy diễn một cách hiệu quả. Kiến trúc mạng nơron FMN trong bài toán học không giám sát được mô tả trên Hình 1.11 [76]. Mạng gồm hai lớp chính, lớp vào F_A bao gồm n nơron, mỗi nơron tương ứng với một thuộc tính của dữ liệu đầu vào; lớp ra F_C bao gồm q nơron, trong đó mỗi nơron đại diện cho một siêu hộp mờ, tương ứng với một cụm dữ liệu được hình thành trong quá trình học. Trong trường hợp này, quá trình học tập trung vào việc phát hiện cấu trúc cụm mà không cần thông tin nhãn lớp.

Đối với bài toán học có giám sát, kiến trúc mạng FMN gồm ba lớp, như minh họa trong Hình 1.12 [16]. Cụ thể, lớp vào F_A bao gồm n nơron tương ứng với các thuộc tính đầu vào; lớp F_B gồm q nơron, mỗi nơron tương ứng với một siêu hộp mờ được hình thành trong quá trình học; lớp ra F_C gồm K nơron, trong đó mỗi nơron đại diện cho một lớp, với K là tổng số lớp được xác định trước. Trong kiến trúc này, các siêu hộp mờ đóng vai trò trung gian kết nối giữa không gian đặc trưng và không gian nhãn lớp.

Mỗi nơron thứ j trong lớp F_B được kết nối với các nơron đầu vào thông qua một bộ trọng số kép, tương ứng với hai vectơ V_j và W_j (minh họa trên Hình 1.13). Cụ thể, kết nối giữa nơron thứ j trong lớp F_B và nơron thứ i trong lớp F_A được đặc

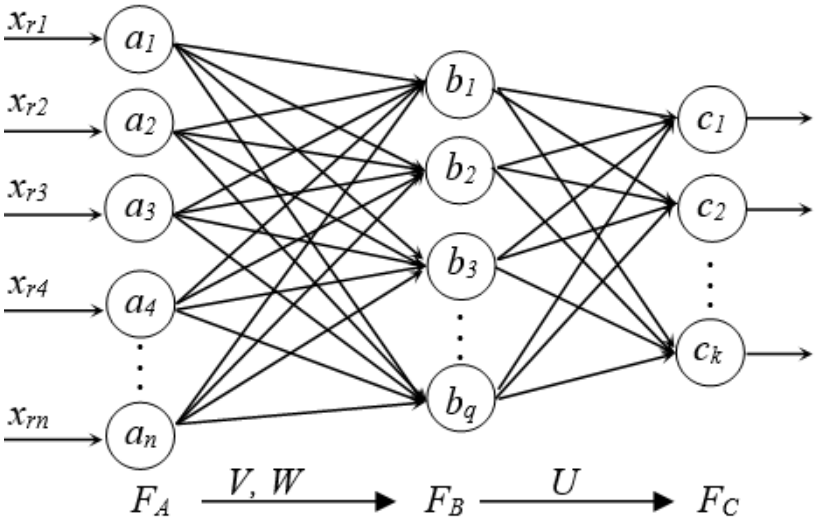


Hình 1.11: *Cấu trúc mạng nơron FMN với học không giám sát*

trưng bởi hai trọng số v_{ji} và w_{ji}, lần lượt biểu diễn giá trị biên dưới và biên trên của siêu hộp mờ trên chiều đặc trưng thứ i. Do đó, các biên của siêu hộp mờ B_j được biểu diễn dưới dạng hai vectơ V_j và W_j (được định nghĩa lần lượt theo Công thức (1.34) và (1.35)). Các vectơ này không chỉ đóng vai trò là trọng số của mạng nơron, mà còn mang ý nghĩa hình học rõ ràng, xác định trực tiếp hình dạng và vị trí của siêu hộp mờ trong không gian đặc trưng.

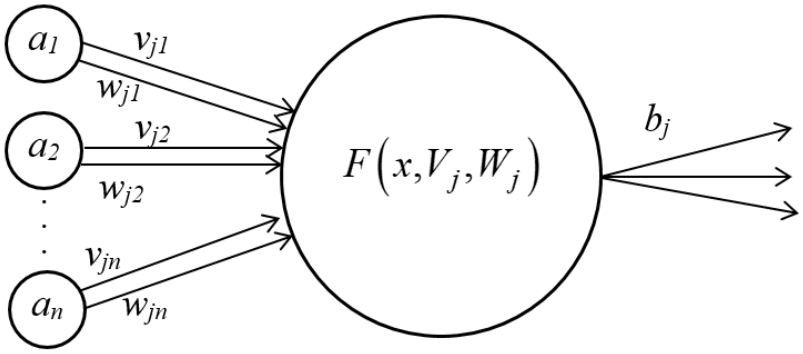
$$\mathbf{V}_j = (v_{j1}, v_{j2}, \ldots, v_{jn}) \tag{1.34}$$

$$\mathbf{W}_j = (w_{j1}, w_{j2}, \ldots, w_{jn}) \tag{1.35}$$



Hình 1.12: *Cấu trúc mạng nơron FMN với học có giám sát*

Trong kiến trúc mạng nơron FMN với học có giám sát, kết nối giữa mỗi nơron trong lớp siêu hộp F_B và các nơron trong lớp phân lớp F_C được biểu diễn bởi một ma



Hình 1.13: *Đồ họa minh họa cấu tạo của nơron* b_j

trận trọng số nhị phân U [16]. Mỗi phần tử $u_{jg} \in U$ phản ánh mối liên hệ giữa siêu hộp mờ B_j và lớp g, đồng thời cho biết siêu hộp đó có thuộc về lớp tương ứng hay không. Giá trị của u_{jg} được xác định theo Công thức (1.36):

$$u_{jg} = \begin{cases} 1, & \ell(B_j) = g, \\ 0, & \ell(B_j) \neq g, \end{cases} \tag{1.36}$$

trong đó b_j là nơron thứ j trong lớp F_B (tương ứng với siêu hộp mờ B_j), và c_g là nơron thứ g trong lớp F_C, đại diện cho lớp thứ g (C_g) trong bài toán phân lớp. Ma trận U do đó đóng vai trò như một cơ chế ánh xạ từ không gian siêu hộp sang không gian nhãn lớp.

Đầu ra của mỗi nơron trong lớp F_C biểu thị mức độ phù hợp của mẫu dữ liệu đầu vào x đối với lớp tương ứng. Cụ thể, giá trị kích hoạt của nơron lớp g được xác định dựa trên các mức độ kích hoạt của các siêu hộp mờ liên kết với lớp đó. Hàm truyền của nơron trong lớp F_C được định nghĩa theo Công thức (1.37):

$$c_g = \max_{j=1,\ldots,q} (b_j \, u_{jg}) \,, \tag{1.37}$$

trong đó b_j là giá trị đầu ra của nơron thứ j trong lớp F_B, biểu thị độ thuộc của mẫu dữ liệu đầu vào đối với siêu hộp mờ B_j; u_{jg} là trọng số liên kết giữa nơron b_j trong lớp F_B và nơron c_g trong lớp F_C, được xác định theo Công thức (1.36). Do $u_{jg} = 1$ khi siêu hộp B_j thuộc lớp C_g và $u_{jg} = 0$ trong trường hợp ngược lại, toán tử $\max(\cdot)$ lựa chọn giá trị kích hoạt lớn nhất trong số các nơron siêu hộp thuộc lớp C_g. Vì vậy, c_g biểu thị mức độ phù hợp của mẫu dữ liệu đầu vào đối với lớp C_g.

Tập mờ xác định lớp g, ký hiệu là C_g, được hình thành như hợp của tất cả các siêu hộp mờ thuộc về lớp tương ứng. Cụ thể, tập mờ này được định nghĩa theo Công thức (1.38):

$$C_g = \bigcup_{j \in G} B_j, \tag{1.38}$$

trong đó G là tập chỉ số của các siêu hộp mờ được gán cho lớp g. Định nghĩa này cho thấy mỗi lớp trong mô hình FMN được biểu diễn bởi một miền quyết định mờ, hình

thành từ sự kết hợp của nhiều siêu hộp mờ, qua đó cho phép mô hình mô tả linh hoạt các cấu trúc dữ liệu phức tạp trong không gian đặc trưng.

1.2.4. Thuật toán học FMN

Các thuật toán học của mạng nơron FMN được phát triển theo các cơ chế giám sát khác nhau. Đối với học có giám sát, thuật toán FMN được trình bày trong Thuật toán 1.1 [16], nhằm xây dựng tập các siêu hộp mờ đại diện cho các lớp dữ liệu thông qua cơ chế học gia tăng. Trên cơ sở đó, thuật toán học không giám sát của FMN được trình bày trong Thuật toán 1.2 [76], cho phép hình thành các siêu hộp mờ mà không cần thông tin nhãn lớp ban đầu. Mở rộng theo hướng kết hợp một phần thông tin giám sát, thuật toán học bán giám sát trong mạng nơron FMN được trình bày trong Thuật toán 1.3 [19], nhằm khai thác đồng thời dữ liệu có nhãn và chưa gán nhãn để nâng cao hiệu quả học và suy luận nhãn.

Thuật toán 1.1 Thuật toán học FMN có giám sát

Input: X ($M = |X|$), γ, $\delta^{\max}$
Output: $\mathcal{B}$, U

$\mathcal{B} \leftarrow \varnothing$
for $(x_r, \ell_{x_r}) \in X$ **do**
  Tìm tập các siêu hộp $B_j \in \mathcal{B}$ sao cho $\ell(B_j) = \ell_{x_r}$
  if $\exists\, B_j \in \mathcal{B}$ thỏa mãn $\mu(x_r, B_j)$ theo (1.39) và $\delta(x_r, B_j) \leq \delta^{\max}$ **then**
    Mở rộng B_j để bao phủ x_r theo (1.6) và (1.7)
    if phát hiện chồng lấn giữa B_j và B_k với $\ell(B_j) \neq \ell(B_k)$ **then**
      Thực hiện thao tác co lại siêu hộp B_j và B_k
    end if
  else
    Tạo siêu hộp mới B_{new} với $V_{\text{new}} = x_r$, $W_{\text{new}} = x_r$
    $\ell(B_{\text{new}}) \leftarrow \ell_{x_r}$
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$
  end if
end for
Cập nhật ma trận U dựa trên tập $\mathcal{B}$ (với u_{jg} được xác định theo (1.36))
return $\mathcal{B}, U$

$$\mu(x_r, B_{j^*}) = \max\left\{\mu(x_r, B_k) \mid k = 1, \ldots, K\right\}. \tag{1.39}$$

$$g(B_j) = \frac{V_j + W_j}{2} \tag{1.40}$$

Thuật toán 1.2 Thuật toán học FMN không giám sát

Input: $X = \{x_r\}_{r=1}^{M}$, γ, $\delta^{\max}$
Output: $\mathcal{B}$

Khởi tạo tập $\mathcal{U}$, với các điểm min và điểm max của U_j ($\forall U_j \in \mathcal{U}$) theo (1.33);
for $x_r \in X$ **do**
  if $\mathcal{B} \neq \varnothing$ **then**
    if $\exists B_j \in \mathcal{B}$ thỏa mãn $\mu(x_r, B_j)$ theo (1.39) và $\delta(x_r, B_j) \leq \delta^{\max}$ **then**
      Mở rộng B_j để bao phủ x_r theo (1.6) và (1.7);
      if $\exists B_k \in \mathcal{B}$, $k \neq j$ sao cho $\Omega(B_j, B_k) > 0$ **then**
        Co lại các siêu hộp B_j và B_k;
      end if
    else
      Tạo B_{new} bằng cách chuyển từ $U_j \in \mathcal{U}$ vào $\mathcal{B}$;
      Khởi tạo B_{new} với $V_{\text{new}} = x_r$, $W_{\text{new}} = x_r$;
    end if
  else
    Tạo B_{new} bằng cách chuyển từ $U_j \in \mathcal{U}$ vào $\mathcal{B}$;
    Khởi tạo B_{new} với $V_{\text{new}} = x_r$, $W_{\text{new}} = x_r$;
  end if
end for
return $\mathcal{B}$;

1.2.5. Một số vấn đề cần tiếp tục nghiên cứu để nâng cao chất lượng FMN cho phân cụm dữ liệu

1.2.5.1. Một số nghiên cứu cải tiến FMN gốc

Nhờ việc được xây dựng dựa trên sự kết hợp giữa logic mờ, mạng nơron nhân tạo và lý thuyết min-max mờ, mô hình FMN là một mô hình học gia tăng, cho phép xử lý hiệu quả các tập dữ liệu quy mô lớn [81]. Cơ chế học gia tăng giúp mô hình cập nhật và bổ sung tri thức chỉ với một lần duyệt dữ liệu, đồng thời vẫn bảo toàn thông tin đã học trước đó [82]. FMN thực hiện phân lớp/phân cụm dựa trên mức độ thuộc của mẫu dữ liệu đối với các siêu hộp tương ứng. Quá trình học bao gồm hai bước chính: mở rộng siêu hộp để bao phủ mẫu mới và co hẹp (nếu cần) nhằm xử lý hiện tượng chồng lấn giữa các lớp. Nhờ cấu trúc và cơ chế hoạt động đơn giản [76], FMN có nhiều ưu điểm nổi bật như: (i) không cần xác định trước số lượng cụm; (ii) dễ dàng mở rộng và điều chỉnh cụm thông qua phép so sánh và cập nhật biên; (iii) cung cấp quyết định mềm thông qua hàm thuộc; (iv) chỉ yêu cầu điều chỉnh một số ít tham số, chủ yếu là kích thước siêu hộp tối đa và tham số mờ; và (v) có khả năng học trực tuyến, thích hợp cho các bài toán có ranh giới lớp phi tuyến với hình dạng và kích thước đa dạng.

Mặc dù sở hữu nhiều ưu điểm, FMN vẫn tồn tại một số hạn chế ảnh hưởng đến

Thuật toán 1.3 Thuật toán học bán giám sát trong FMN

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

repeat
  Chọn dữ liệu vào $(x_r, \ell_{x_r}) \in X$;
  if $\exists B_j \in \mathcal{B}$ thỏa mãn $\mu(x_r, B_j) = \max_{B_k \in \mathcal{B}} \mu(x_r, B_k)$ và $\delta(x_r, B_j) \leq \delta^{\max}$ **then**
    Điều chỉnh B_j theo (1.6) và (1.7); $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    if phát hiện chồng lấn giữa B_j và B_k với $\ell(B_j) \neq \ell(B_k)$ **then**
      Điều chỉnh chồng lấn giữa các siêu hộp B_j và B_k;
    end if
  else
    if $\ell_{x_r} \neq 0$ **then**
      Tạo siêu hộp mới B_{new}; $\ell(B_{\text{new}}) \leftarrow \ell_{x_r}$; $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$
    else
      if $\exists B_j \in \mathcal{B}$ sao cho $\mu(x_r, g(B_j)) > \beta$ **then**
        Tạo siêu hộp mới B_{new}; $\ell(B_{\text{new}}) \leftarrow \ell(B_j)$; $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$;
      end if
      $\ell_{x_r} \leftarrow \ell(B_j)$;
    end if
    $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
  end if
until $X = \varnothing$
return $\mathcal{B}, \{(x_r, \ell_{x_r})\}_{r=1}^{M}$

hiệu quả và khả năng ứng dụng thực tiễn. Một trong những vấn đề đáng chú ý là cơ chế co hẹp siêu hộp nhằm xử lý chồng lấn có thể làm suy giảm hiệu suất phân loại của mạng [83]. Khi tham số giới hạn kích thước siêu hộp được đặt nhỏ, thuật toán có xu hướng tạo ra số lượng lớn siêu hộp, dẫn đến cấu trúc mạng phức tạp và nguy cơ quá khớp (overfitting). Ngược lại, nếu giới hạn kích thước được chọn lớn hơn để giảm số lượng siêu hộp, vùng chồng lấn giữa các lớp có thể gia tăng, từ đó làm giảm độ chính xác của mô hình [83]. Bên cạnh đó, việc lựa chọn cấu trúc và tham số mạng thường phải thực hiện thông qua quá trình thử–sai nhiều lần [76], gây tốn kém thời gian và công sức.

Trước những hạn chế trên, nhiều hướng cải tiến FMN đã được đề xuất. Các nghiên cứu chủ yếu tập trung vào việc điều chỉnh cấu trúc mạng, tối ưu hóa tham số, cải tiến hàm thuộc, giảm số lượng siêu hộp, hoàn thiện thuật toán học hoặc kết hợp với các phương pháp khác nhằm nâng cao hiệu năng tổng thể.

Theo hướng cải tiến cấu trúc, có thể kể đến mô hình FMCN do Nandedkar đề xuất năm 2007 [84], mô hình DCFMN của Zhang năm 2011 [85], và mô hình đa tầng MLF được đề xuất năm 2014 [83]. Ở hướng cải tiến thuật toán học, đáng chú ý là mô

hình GFMM của Gabrys năm 2000 [15] và biến thể RFMN do Nandedkar cải tiến năm 2008 [18].

Ngoài ra, nhiều công trình tập trung nâng cao hiệu suất thông qua các chiến lược điều chỉnh cụ thể. Ma đề xuất phương pháp xác định linh hoạt giới hạn kích thước siêu hộp tối đa năm 2012 [86]. Mohammed và cộng sự năm 2015 [87] cùng các nghiên cứu tiếp theo năm 2016 [19] bổ sung các luật kiểm tra và điều chỉnh chồng lấn trong FMN. Shinde [88], Quteishat [89] và Jin Wang [8] đề xuất loại bỏ các siêu hộp có độ tin cậy thấp nhằm tinh gọn mô hình. Năm 2017, Mohammed tiếp tục nghiên cứu giảm sinh siêu hộp ở vùng biên [90] và cơ chế lựa chọn siêu hộp chiến thắng [91]. Hou phát triển mô hình CFMN dựa trên hệ số đóng góp cụm năm 2018 [92]. Liu năm 2017 đề xuất sử dụng hệ số HE để đánh giá hiệu quả siêu hộp trong quá trình co hẹp [93]. Sonule xây dựng mô hình EFMN tối ưu hóa thuộc tính theo nhóm mẫu và sinh luật quyết định năm 2017 [94]. Seera phát triển mô hình cây phân cụm ECT nhằm nâng cao hiệu quả phân cụm trực tuyến năm 2018 [95].

Năm 2020, Dehariya và nhóm nghiên cứu đề xuất ứng dụng của thuật toán tối ưu hóa MFO (Moth-Flame Optimization) và mạng nơron FMN để phát triển hệ thống phân loại dữ liệu y tế [96]. Thuật toán MFO xem xét phần lớn các đặc trưng từ tập dữ liệu bệnh tật và tạo ra tập hợp các đặc trưng được tối ưu hóa dựa trên hàm đánh giá độ phù hợp. Nhờ khả năng tránh cực tiểu cục bộ, MFO tạo ra được tập đặc trưng tối ưu. Các đặc trưng được tối ưu hóa sau đó được chuyển đến FMN để phân loại các trường hợp ác tính và lành tính.

Năm 2024, Tao Hou cùng nhóm nghiên cứu của mình đã đề xuất mô hình QFMMNet (Quality-Aware Fuzzy Min-Max Neural Network) [97] cho phân tích mạng lưới não động đa khung nhìn trong chẩn đoán bệnh tâm thần phân liệt thực tế.

Bên cạnh đó, trong bối cảnh thực tế, dữ liệu có nhãn thường khan hiếm và chi phí gán nhãn cao. Do đó, các mô hình như GFMM [15], RFMN (Reflex FMN) [18], SSL-FMM [17, 98], SS-FMM [19], SCFMN [20], MSCFMN [21] được đề xuất nhằm kết hợp học có giám sát và không giám sát trong cùng một khung học. Hướng tiếp cận với học bán giám sát giúp tăng khả năng khai thác thông tin từ dữ liệu chưa nhãn.

Ngay từ những nghiên cứu nền tảng, mô hình GFMM [15] được phát triển nhằm thống nhất hai bài toán phân lớp và phân cụm trong cùng một kiến trúc. GFMM cho phép xử lý linh hoạt cả dữ liệu có nhãn và không có nhãn thông qua việc xây dựng các siêu hộp mờ đại diện cho cấu trúc dữ liệu, trong khi thông tin nhãn (nếu tồn tại) được sử dụng để cập nhật ma trận liên kết giữa siêu hộp và lớp. Cách tiếp cận này

được xem là tiền đề quan trọng cho học bán giám sát trong FMN, mặc dù GFMM chưa đưa ra cơ chế rõ ràng để kiểm soát ảnh hưởng của dữ liệu chưa nhãn đến chất lượng phân lớp.

Tiếp nối hướng nghiên cứu này, mô hình RFMN tập trung trực tiếp vào bài toán học bán giám sát [18]. RFMN giới thiệu cơ chế "phản xạ", cho phép mô hình tự điều chỉnh khi tiếp nhận các mẫu chưa gán nhãn, qua đó khai thác tốt hơn cấu trúc tiềm ẩn của dữ liệu. So với GFMM, RFMN nhấn mạnh vai trò của dữ liệu chưa nhãn trong quá trình học, đồng thời cố gắng giảm sự phụ thuộc vào số lượng mẫu có nhãn. Tuy nhiên, mô hình này vẫn đối mặt với nguy cơ lan truyền lỗi khi các siêu hộp hình thành từ dữ liệu chưa nhãn không đủ tin cậy.

Các nghiên cứu sau đó tập trung nhiều hơn vào bài toán phân cụm bán giám sát trong FMN. Vu và cộng sự (2016) đề xuất mô hình SS-FMM cho phân cụm bán giám sát dựa trên FMN, trong đó thông tin nhãn hạn chế được sử dụng để dẫn dắt quá trình hình thành cụm [19]. Thuật toán cho phép các mẫu chưa nhãn tham gia trực tiếp vào việc tạo siêu hộp, đồng thời sử dụng cơ chế lan truyền nhãn để gán nhãn cho các siêu hộp tương ứng. Cách tiếp cận này nhấn mạnh việc học cấu trúc dữ liệu trước, sau đó khai thác nhãn như một nguồn thông tin bổ trợ, phù hợp với các bài toán phân cụm có hình dạng dữ liệu phức tạp.

Trên cơ sở đó, năm 2019, Tran cùng với nhóm nghiên cứu của mình đề xuất mô hình SCFMN [20] kết hợp FMN với học bán giám sát nhằm giải quyết bài toán hỗ trợ chẩn đoán bệnh gan. SCFMN cho thấy khả năng ứng dụng thực tế của FMN bán giám sát, khi dữ liệu y sinh thường khan hiếm nhãn và chứa nhiều nhiễu. Việc tích hợp thông tin chưa nhãn giúp cải thiện hiệu quả chẩn đoán so với các mô hình học có giám sát thuần túy.

Năm 2020, Liu và cộng sự đề xuất mô hình SSL-FMM (Semi-supervised FMN), trong đó dữ liệu chưa nhãn được khai thác có hệ thống trong quá trình xây dựng và mở rộng siêu hộp mờ [17]. SSL-FMM cho phép các siêu hộp mở rộng bao phủ cả dữ liệu có nhãn và chưa nhãn, đồng thời sử dụng chiến lược gán nhãn gián tiếp dựa trên mức độ phù hợp. Cách tiếp cận này giúp cải thiện biên phân lớp và giảm hiện tượng quá khớp khi số lượng mẫu có nhãn hạn chế. Tương tự, mô hình SSL-FMM trong [98] đã kết hợp mô hình học hai giai đoạn trong FMN với cơ chế lựa chọn và đánh giá nhãn lớp dựa trên thuật toán kNN (k-Nearest Neighbors).

Năm 2022, Minh cùng nhóm nghiên cứu tiếp tục mở rộng FMN theo hướng kết hợp phân cụm và học bán giám sát nhằm trích xuất luật và hỗ trợ chẩn đoán, gọi

là MSCFMN [21]. Công trình này nhấn mạnh vai trò của việc tích hợp các kỹ thuật phân cụm với FMN để nâng cao chất lượng học trong các kịch bản dữ liệu phức tạp.

Mặc dù đã có nhiều nghiên cứu nhằm nâng cao chất lượng của mạng nơron FMN, các phương pháp hiện có vẫn còn tồn tại một số hạn chế đáng chú ý:

- Thứ nhất, phần lớn các mô hình FMN giả định rằng các siêu hộp mờ đóng vai trò tương đương trong quá trình học, trong khi trên thực tế chất lượng và độ ổn định của các siêu hộp có thể khác nhau đáng kể.

+ Một số siêu hộp có thể đại diện tốt cho cấu trúc dữ liệu, trong khi các siêu hộp khác lại dễ bị ảnh hưởng bởi nhiễu hoặc nằm gần biên phân lớp.
+ Việc không phân biệt vai trò của các siêu hộp này có thể làm suy giảm hiệu quả học, đặc biệt trong bối cảnh dữ liệu phức tạp.

- Thứ hai, mặc dù các mô hình như General FMN (GFMM) đã mở rộng FMN cho bài toán học bán giám sát, quá trình khai thác dữ liệu chưa gán nhãn trong các phương pháp này vẫn mang tính thụ động và thiếu cơ chế kiểm soát.

+ Cụ thể, các siêu hộp chưa được gán nhãn hoặc có độ tin cậy thấp vẫn có thể tham gia vào quá trình lan truyền nhãn, dẫn đến nguy cơ khuếch đại sai lệch nhãn trong quá trình huấn luyện.
+ Hiện tượng này đặc biệt nghiêm trọng trong các bài toán có tỷ lệ dữ liệu được gán nhãn rất thấp.

- Thứ ba, cơ chế xử lý chồng lấn trong FMN chủ yếu dựa trên các quy tắc heuristic, chẳng hạn như các trường hợp co lại được đề xuất trong mô hình FMN gốc.

+ Các quy tắc này thường được áp dụng đồng đều cho mọi siêu hộp, mà chưa xét đến mức độ ổn định hoặc tầm quan trọng của từng siêu hộp.
+ Do đó, việc co lại có thể làm phá vỡ cấu trúc cụm đã được hình thành ổn định, đặc biệt trong các tập dữ liệu có phân bố phức tạp hoặc nhiều nhiễu.

- Cuối cùng, hầu hết các mô hình FMN hiện nay chưa tách bạch rõ ràng giữa quá trình học cấu trúc dữ liệu và quá trình học nhãn.

+ Việc đồng thời vừa xây dựng siêu hộp vừa gán nhãn trong suốt quá trình huấn luyện khiến mô hình phụ thuộc mạnh vào nhãn.
+ Điều này làm giảm hiệu quả trong các kịch bản dữ liệu khan hiếm nhãn, vốn là tình huống phổ biến trong nhiều bài toán thực tế.

Nhìn chung, mạng nơron min-max mờ được xem là một hướng tiếp cận hiệu quả trong nhiều bài toán khai phá dữ liệu [99]. Một ưu điểm quan trọng của FMN là khả năng trích xuất các luật quyết định dạng IF–THEN một cách trực quan; mỗi siêu

hộp có thể được chuyển đổi thành một luật dựa trên các giá trị biên min-max của nó.

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

Từ tổng quan các hướng nghiên cứu có thể thấy rằng, mặc dù FMN và các biến thể đã đạt được nhiều cải tiến đáng kể, vẫn còn tồn tại một số vấn đề cần tiếp tục giải quyết. Trước hết, việc xây dựng cơ chế học hiệu quả trong điều kiện tỷ lệ mẫu có nhãn thấp vẫn là một thách thức quan trọng, đặc biệt trong bối cảnh học bán giám sát. Bên cạnh đó, số lượng siêu hộp sinh ra trong quá trình huấn luyện có thể tăng nhanh khi giới hạn kích thước bị thu hẹp hoặc khi dữ liệu có mức độ chồng lấn cao, dẫn đến cấu trúc mạng phức tạp và số lượng luật quyết định lớn. Ngoài ra, phần lớn các phương pháp hiện tại chủ yếu điều chỉnh hình học siêu hộp mà chưa xem xét đầy đủ phân bố nội tại của dữ liệu bên trong, điều này có thể làm lệch tâm cụm và suy giảm khả năng khái quát hóa. Trong trường hợp dữ liệu có chồng lấn mạnh, cấu trúc mạng cũng dễ trở nên thiếu ổn định do sự gia tăng số lượng siêu hộp hoặc nơron bù.

Những hạn chế trên cho thấy cần một hướng tiếp cận mới vừa bảo toàn ưu điểm học gia tăng của FMN, vừa giảm phụ thuộc vào tham số và tỷ lệ nhãn, đồng thời phản ánh chính xác hơn cấu trúc phân bố dữ liệu thực tế. Một định hướng tiềm năng là đánh giá chất lượng và độ tin cậy của các siêu hộp mờ, từ đó lựa chọn những siêu hộp có tính đại diện cao tham gia vào quá trình lan truyền nhãn, nhằm hạn chế ảnh hưởng của siêu hộp nhiễu và giảm thiểu lan truyền sai lệch. Đồng thời, việc thiết kế cơ chế học bán giám sát có kiểm soát, trong đó mức độ tham gia của dữ liệu chưa gán nhãn được điều chỉnh theo độ tin cậy của siêu hộp, có thể giúp tăng tính ổn định và độ chính xác của mô hình. Bên cạnh đó, các chiến lược xử lý chồng lấn thích nghi, cho phép điều chỉnh hoặc trì hoãn quá trình co hẹp dựa trên trạng thái hội tụ của siêu hộp, cũng là hướng nghiên cứu hứa hẹn nhằm bảo toàn cấu trúc cụm đã học. Cuối cùng, việc tách biệt giai đoạn học cấu trúc dữ liệu với giai đoạn học nhãn có thể giúp khai thác hiệu quả hơn thông tin từ dữ liệu chưa gán nhãn. Trên cơ sở đó, luận án tập trung đề xuất các mô hình FMN bán giám sát nhằm nâng cao chất lượng học thông qua cơ chế đánh giá độ tin cậy của siêu hộp và kiểm soát quá trình lan truyền nhãn trong huấn luyện.

1.4. Kết luận chương 1

Chương 1 đã trình bày cơ sở lý thuyết và các hướng tiếp cận liên quan đến bài toán phân cụm bán giám sát mờ, với trọng tâm là mạng nơron FMN. Chương đã làm rõ sự khác biệt giữa phân cụm rõ và phân cụm mờ, đồng thời chỉ ra hạn chế của các phương pháp truyền thống khi xử lý dữ liệu nhiễu, phân bố phức tạp và ranh giới cụm

không rõ ràng. Trên cơ sở đó, phân cụm bán giám sát mờ được xác định là hướng tiếp cận phù hợp nhằm khai thác đồng thời cấu trúc nội tại của dữ liệu và một lượng thông tin giám sát hạn chế. Chương cũng đã tổng hợp các nhóm phương pháp tiêu biểu như phương pháp dựa trên hạt giống, ràng buộc, mô hình nguyên mẫu và mạng nơron. Phần trọng tâm của chương tập trung trình bày về FMN, bao gồm cơ chế biểu diễn bằng siêu hộp mờ, hàm thuộc, cấu trúc mạng và các thuật toán học tương ứng. Thông qua việc tổng hợp các nghiên cứu từ GFMM, RFMN đến các mô hình FMN bán giám sát hiện đại, chương đã chỉ ra những hạn chế còn tồn tại trong khai thác dữ liệu chưa gán nhãn, đánh giá độ tin cậy và kiểm soát lan truyền lỗi nhãn. Một phần nội dung của Chương 1 đã được công bố trong kỷ yếu *Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông*, Hội thảo Quốc gia lần thứ XXV, Nhà xuất bản Khoa học và Kỹ thuật, năm 2022 [CT1]. Từ đó, Chương 1 đã xác định khoảng trống nghiên cứu và đặt nền tảng lý thuyết cho Chương 2 về mô hình và phương pháp đề xuất.

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

Trên cơ sở các phân tích và nhận định trong Chương 1, chương này tập trung nghiên cứu và đề xuất các mô hình phân cụm bán giám sát dựa trên mạng nơron FMN nhằm khắc phục những hạn chế của các biến thể FMN hiện có. Chương bắt đầu bằng việc trình bày các nguyên lý học bán giám sát trong FMN và vai trò của thông tin giám sát hạn chế trong quá trình hình thành và điều chỉnh các siêu hộp mờ.

Tiếp theo, các mô hình FMN-Voting, FMN-SAL được giới thiệu lần lượt, cùng với các thuật toán học tương ứng. Trọng tâm của chương là các cơ chế suy luận nhãn có kiểm soát và chiến lược bỏ phiếu nhằm nâng cao độ ổn định và độ tin cậy của quá trình lan truyền nhãn.

Nội dung Chương 2 được tổ chức theo trình tự: cơ sở và vấn đề nghiên cứu, mô tả bài toán, ý tưởng mô hình, kiến trúc, thuật toán học, thực nghiệm và đánh giá. Trong đó, Mục 2.1 và 2.2 trình bày cơ sở cho các mô hình đề xuất; Mục 2.3 và 2.4 lần lượt trình bày FMN-Voting và FMN-SAL; Mục 2.5 trình bày thiết lập, kết quả thực nghiệm và đánh giá.

2.1. Học bán giám sát dựa trên siêu hộp mờ

2.1.1. Phân hoạch không gian dữ liệu bằng FMN

Khác với các phương pháp phân cụm truyền thống, FMN thực hiện phân hoạch không gian dữ liệu thông qua tập các siêu hộp mờ. Cách tiếp cận này cho phép mô hình biểu diễn một cụm dữ liệu bằng một hoặc nhiều vùng con, tùy thuộc vào độ phức tạp của phân bố dữ liệu. Do đó, FMN có khả năng mô hình hóa các cụm không lồi, các cụm có hình dạng bất kỳ và các cụm có mật độ không đồng đều. Ví dụ về khả năng mô hình hóa các cụm không lồi, các cụm có hình dạng bất kỳ và các cụm có mật độ không đồng đều của FMN được minh họa trên Hình 2.1. Hình 2.1a biểu diễn nguyên mẫu dữ liệu, Hình 2.1b biểu diễn không gian các siêu hộp được phân hoạch bởi mạng nơron FMN.

2.1.2. Đánh giá siêu hộp dựa trên mật độ và phân phối dữ liệu trong siêu hộp

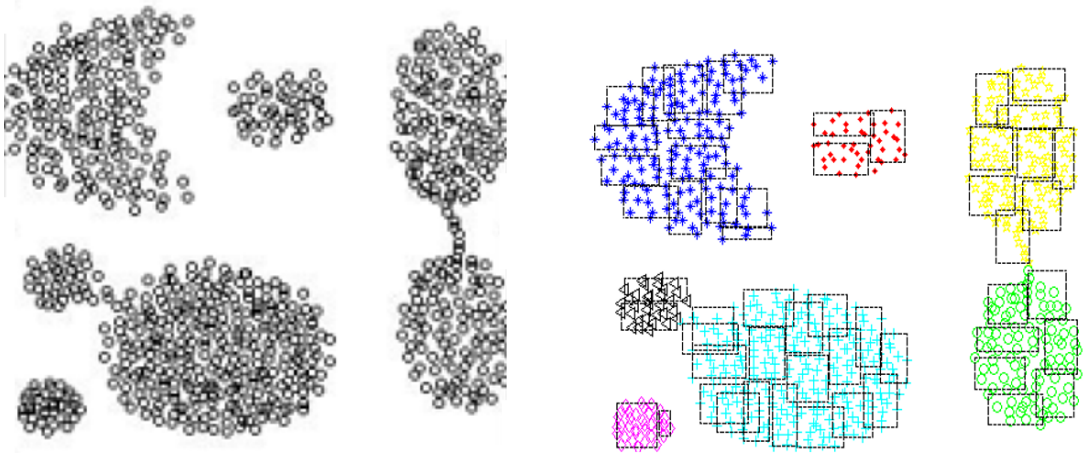
Trong mạng nơron FMN, việc đặc trưng vị trí đại diện của một siêu hộp là cần thiết nhằm đánh giá mức độ phù hợp của siêu hộp đối với các mẫu dữ liệu mà nó bao

phủ. Đối với mỗi siêu hộp B_j, mật độ và phân phối dữ liệu trong siêu hộp ảnh hưởng tới đặc trưng vị trí đại diện. Để đánh giá mật độ và phân phối dữ liệu trong siêu hộp B_j, luận án sử dụng chỉ số độ lệch tâm $e(B_j)$ (Công thức (2.1)) và chỉ số hiệu suất sử dụng siêu hộp $c(B_j)$ (Công thức (2.4) [8]).

$$e(B_j) = \|g(B_j) - d(B_j)\|_2 . \tag{2.1}$$

trong đó, $g(B_j)$ là tâm hình học của siêu hộp, được tính toán theo Công thức (1.40); $d(B_j)$ là tâm dữ liệu của siêu hộp, được tính toán theo Công thức (2.2)

$$d(B_j) = \frac{1}{|T_j|} \sum_{r \in T_j} x_r \tag{2.2}$$



***(a)** Phân bố dữ liệu* ***(b)** Phân hoạch siêu hộp*

***Hình 2.1:** Ví dụ phân hoạch không gian dữ liệu Aggregation thành các siêu hộp bằng FMN*

Giá trị $e(B_j)$ càng nhỏ cho thấy dữ liệu phân bố càng cân đối trong siêu hộp. Trong thực tế, một ngưỡng $\varepsilon > 0$ thường được sử dụng để ràng buộc độ lệch tâm của siêu hộp B_j. ε là tham số ngưỡng do người dùng xác định để đánh giá mức độ "*thuần khiết*" của siêu hộp B_j. Một siêu hộp mờ B_j được xem là thuần khiết nếu độ lệch tâm của nó thỏa mãn điều kiện ràng buộc (2.3):

$$e(B_j) \leq \varepsilon, \qquad \varepsilon > 0, \tag{2.3}$$

$c(B_j)$ được gọi là chỉ số sử dụng của siêu hộp B_j, phản ánh mức độ siêu hộp được khai thác trong quá trình biểu diễn và phân loại các mẫu dữ liệu. Giá trị $c(B_j)$ càng lớn cho thấy siêu hộp có mức độ sử dụng cao hơn đối với vùng dữ liệu tương ứng; ngược lại, giá trị $c(B_j)$ thấp cho thấy mức độ sử dụng của siêu hộp thấp hơn.

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

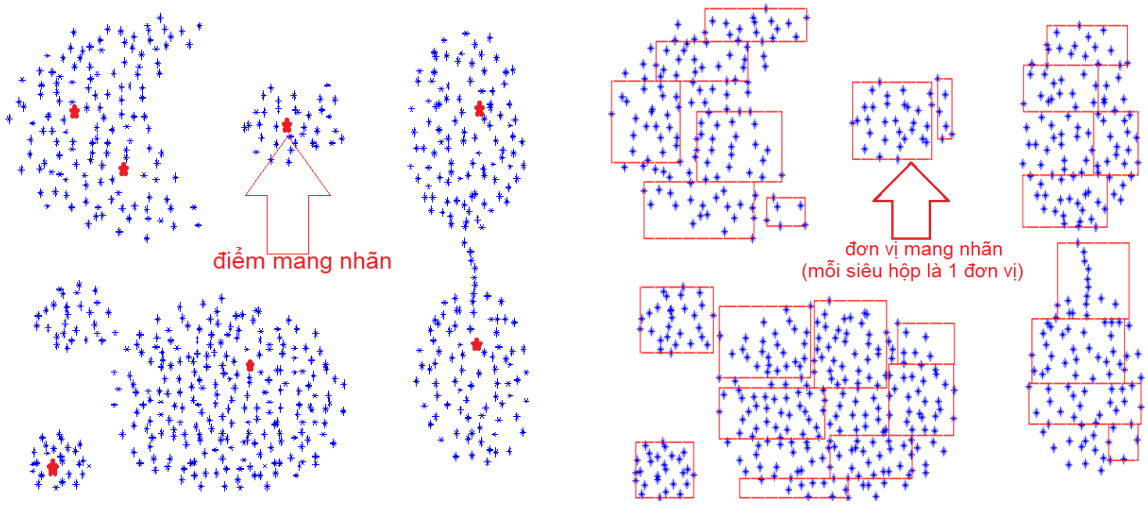
trong đó, U_j là tỷ số giữa các mẫu thuộc siêu hộp B_j trên tổng số các mẫu thuộc nhóm

chứa B_j. A_j là tỷ số giữa số lượng các mẫu được phân loại chính xác bởi siêu hộp B_j trên tổng số mẫu được phân loại chính xác trong cùng một nhóm tương ứng.

2.1.3. Cơ chế giám sát bán giám sát dựa trên siêu hộp

Trong học bán giám sát truyền thống, tập hạt giống thường là một tập con các mẫu đã được gán nhãn, trong đó mỗi điểm dữ liệu đóng vai trò khởi phát cho quá trình lan truyền nhãn sang các mẫu chưa gán nhãn. Cách tiếp cận này giả định rằng mỗi điểm mang nhãn là một thực thể độc lập và có khả năng đại diện cục bộ cho vùng lân cận của nó. Trong FMN, siêu hộp là đơn vị biểu diễn cơ bản, tương ứng với một vùng con có ý nghĩa hình học trong không gian đặc trưng. Các điểm dữ liệu nằm hoàn toàn trong cùng một siêu hộp có độ thuộc đầy đủ và được xem là tương đương về mức độ tương đồng. Vì vậy, siêu hộp không chỉ phân hoạch không gian dữ liệu mà còn thực hiện chức năng khái quát hóa, gom nhóm các điểm gần nhau thành một cấu trúc ổn định.

Do đó, trong FMN, hạt giống không còn gắn với từng điểm dữ liệu riêng lẻ mà gắn với từng siêu hộp mờ. Khi một điểm đại diện trong siêu hộp được gán nhãn, nhãn này được lan truyền nhất quán tới toàn bộ các điểm có độ thuộc đầy đủ với siêu hộp tương ứng. Khi đó, cơ chế giám sát được nâng từ mức điểm dữ liệu lên mức cấu trúc không gian, trong đó mỗi siêu hộp đóng vai trò như một *đơn vị mang nhãn* (Hình 2.2a, Hình 2.2b).



***(a)** Nhãn dựa trên điểm dữ liệu* ***(b)** Siêu hộp là một đơn vị nhãn*

***Hình 2.2:** Điểm dữ liệu mang nhãn và đơn vị nhãn*

Như minh họa trong Hình 2.2a, các phương pháp gán nhãn dựa trên điểm thực hiện giám sát ở mức mẫu, trong đó mỗi điểm mang nhãn được xem là một thực thể độc lập. Quá trình lan truyền nhãn vì vậy thường diễn ra từng bước giữa các điểm lân cận, dễ bị ảnh hưởng bởi nhiễu cục bộ và các điểm biên có độ không chắc chắn cao.

Khi nhãn ban đầu không đủ đại diện cho cấu trúc toàn cục của dữ liệu, các sai lệch nhỏ ở mức điểm có thể bị tích lũy và khuếch đại trong suốt quá trình lan truyền.

Ngược lại, như thể hiện trong Hình 2.2b, FMN thực hiện giám sát ở mức siêu hộp, trong đó mỗi siêu hộp đại diện cho một vùng dữ liệu tương đồng và đóng vai trò là đơn vị mang nhãn. Khi một điểm đại diện bên trong siêu hộp được gán nhãn, toàn bộ các điểm dữ liệu thuộc siêu hộp đó sẽ cùng chia sẻ nhãn tương ứng. Nhờ vậy, quá trình lan truyền nhãn được thực hiện ở mức cấu trúc ổn định hơn, phản ánh trực tiếp quan hệ hình học trong không gian đặc trưng thay vì phụ thuộc vào các mối quan hệ cục bộ giữa từng điểm dữ liệu riêng lẻ.

Sự khác biệt giữa hai cơ chế giám sát này mang lại nhiều lợi ích quan trọng. Việc gán nhãn ở mức siêu hộp giúp giảm chi phí gán nhãn vì chỉ cần một số điểm đại diện, đồng thời hạn chế ảnh hưởng của nhiễu và các điểm biên không chắc chắn do quyết định được đưa ra trên một vùng dữ liệu thay vì từng điểm đơn lẻ.

Trong FMN, các điểm có nhãn chỉ đóng vai trò kích hoạt thông tin giám sát ban đầu, còn siêu hộp mới là đơn vị thực sự tham gia vào quá trình học và lan truyền nhãn. Cách tổ chức này không chỉ nâng cao hiệu quả học bán giám sát mà còn tăng khả năng diễn giải của mô hình.

Đối với FMN bán giám sát, mỗi siêu hộp mờ cần gắn với một hạt giống đại diện để khởi tạo nhãn và liên kết thông tin từ các điểm dữ liệu có nhãn trong hoặc gần siêu hộp. Khi siêu hộp được xác định là đáng tin cậy, nó tạo thành một vùng nhãn cục bộ ổn định cho quá trình lan truyền nhãn tiếp theo.

Theo cách tiếp cận này, hạt giống gắn liền với cấu trúc siêu hộp thay vì tồn tại như một điểm đơn lẻ. Nhờ đó, siêu hộp có thể tổng hợp thông tin nhãn từ nhiều mẫu dữ liệu, giảm phụ thuộc vào một nhãn đơn lẻ và hạn chế nguy cơ lan truyền lỗi, đặc biệt trong các tập dữ liệu có nhãn hiếm hoặc phân bố không đồng đều.

Hơn nữa, việc lựa chọn hạt giống ở mức siêu hộp cho phép lan truyền nhãn diễn ra ở mức cấu trúc thông qua các quan hệ hình học như chồng lấn, lân cận và mức độ bao phủ. Vì vậy, bài toán lựa chọn hạt giống trong FMN bán giám sát cần được tiếp cận ở mức siêu hộp, kết hợp cả thông tin nhãn, đặc trưng hình học, độ ổn định và khả năng đại diện của siêu hộp trong không gian dữ liệu.

Bên cạnh vai trò là đơn vị nhãn, mỗi siêu hộp mờ có nhãn $B_j = (V_j, W_j)$ với $\ell(B_j) = c$ có thể được chuyển trực tiếp thành một luật quyết định dạng IF–THEN:

R_j : IF $x_1 \in [v_{j1}, w_{j1}]$ AND $x_2 \in [v_{j2}, w_{j2}]$ AND ... AND $x_n \in [v_{jn}, w_{jn}]$ THEN $y = c$.

Do siêu hộp là tập mờ, luật R_j được kích hoạt với mức độ $\mu(x, B_j)$ theo Công thức (1.3):

$\mu(x, B_j) = 1$ khi mẫu x nằm trong siêu hộp và giảm dần theo tham số γ khi x nằm ngoài biên. Khi nhiều luật cùng được kích hoạt, nhãn của x được xác định bởi luật có mức kích hoạt lớn nhất, tương ứng với toán tử $\max$ ở Công thức (1.37). Để luật ngắn gọn hơn, có thể bỏ các điều kiện trên những chiều mà khoảng $[v_{ji}, w_{ji}]$ bao toàn bộ miền giá trị đã chuẩn hóa $[0, 1]$, vì các điều kiện này không phân biệt giữa các lớp.

Độ tin cậy của luật R_j được đánh giá thông qua các đặc trưng của siêu hộp tương ứng đã trình bày ở trên: (i) số mẫu $|T_j|$ mà siêu hộp bao phủ, phản ánh độ hỗ trợ của luật; (ii) độ lệch tâm $e(B_j)$ theo Công thức (2.1), trong đó luật sinh từ siêu hộp thỏa mãn điều kiện (2.3) được xem là đáng tin cậy hơn; và (iii) chỉ số sử dụng $c(B_j)$ theo Công thức (2.4). Các luật có độ hỗ trợ thấp, độ lệch tâm lớn hoặc chỉ số sử dụng thấp có thể được loại bỏ để thu được tập luật gọn và đáng tin cậy. Ví dụ, với tập dữ liệu ILPD, một luật có dạng: IF *Age* $\in [v_{j,\mathrm{Age}}, w_{j,\mathrm{Age}}]$ AND *TB* $\in [v_{j,\mathrm{TB}}, w_{j,\mathrm{TB}}]$ AND $\ldots$ AND *A/G Ratio* $\in [v_{j,\mathrm{A/G}}, w_{j,\mathrm{A/G}}]$ THEN "mắc bệnh gan", trong đó các cận được lấy trực tiếp từ các đỉnh V_j, W_j của siêu hộp sau huấn luyện.

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

2.2.1. Giới thiệu về phương pháp bỏ phiếu đa số

Phương pháp bỏ phiếu đa số dựa trên mẫu là một trong những cơ chế ra quyết định đơn giản nhưng hiệu quả, được sử dụng rộng rãi trong các bài toán phân lớp và phân cụm, đặc biệt trong các mô hình học dựa trên khoảng cách như kNN (k-Nearest Neighbors) [100]. Ý tưởng của phương pháp này là giả định rằng các mẫu dữ liệu nằm gần nhau trong không gian đặc trưng có xu hướng thuộc cùng một lớp hoặc cùng một cấu trúc ngữ nghĩa. Do đó, nhãn của một mẫu chưa biết có thể được suy luận thông qua việc khảo sát tập các mẫu đã biết gần nhất và thực hiện bỏ phiếu để chọn nhãn chiếm ưu thế.

Với một mẫu truy vấn x chưa có nhãn, thuật toán kNN xác định tập k mẫu huấn luyện gần nhất:

$$\mathcal{N}_k(x) = \{x_1, x_2, \ldots, x_k\}, \tag{2.5}$$

dựa trên một hàm *khoảng cách* xác định trước (có thể theo Euclid (2.6) hoặc Manhattan (L1-norm) (2.7)).

$$d_{\mathrm{Euc}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{2.6}$$

$$d_{\text{Man}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^{n} |x_i - y_i|. \tag{2.7}$$

Mỗi mẫu x_i trong tập lân cận mang một nhãn $y_i \in \mathcal{C}$, trong đó $\mathcal{C} = \{1, 2, \ldots, C\}$ là tập các lớp khả dĩ và C là tổng số lớp.

Quy tắc bỏ phiếu đa số được áp dụng bằng cách đếm tần suất xuất hiện của từng nhãn trong $\mathcal{N}_k(x)$ và gán cho x nhãn có số phiếu cao nhất:

$$\hat{y} = \arg\max_{c \in C} \sum_{x_i \in \mathcal{N}_k(x)} \mathbb{I}(y_i = c), \tag{2.8}$$

trong đó $\mathbb{I}(\cdot)$ là hàm chỉ thị, tổng của nó cho biết số lần xuất hiện của từng nhãn lớp $c \in C$ trong tập lân cận $\mathcal{N}_k(x)$.

Ưu điểm của phương pháp bỏ phiếu đa số dựa trên mẫu nằm ở tính trực quan, dễ triển khai và khả năng thích nghi tốt với các phân bố dữ liệu phức tạp, phi tuyến. Do không áp đặt giả thiết về dạng biên quyết định, phương pháp này có thể tận dụng trực tiếp cấu trúc cục bộ của dữ liệu. Điều này đặc biệt hữu ích trong các bài toán mà ranh giới giữa các lớp không rõ ràng hoặc bị chồng lấn. Ngoài ra, cơ chế bỏ phiếu còn cho phép mở rộng tự nhiên sang các biến thể mềm, trong đó đóng góp của mỗi mẫu lân cận được điều chỉnh theo khoảng cách đến mẫu truy vấn.

Trong các thuật toán học bán giám sát dựa trên nhãn mẫu, bỏ phiếu đa số dựa trên mẫu thường được sử dụng như một cơ chế suy luận nhãn tạm thời. Khi số lượng mẫu có nhãn ban đầu hạn chế, nhãn của các mẫu chưa gán nhãn có thể được ước lượng dựa trên tập lân cận gồm các mẫu đã biết nhãn, từ đó mở rộng tập huấn luyện.

Tuy nhiên, kỹ thuật bỏ phiếu đa số dựa trên mẫu cũng tồn tại nhiều hạn chế:

- Thứ nhất, hiệu quả của phương pháp phụ thuộc mạnh vào chất lượng và sự phân bố của dữ liệu huấn luyện. Trong trường hợp dữ liệu nhiễu hoặc có mật độ không đồng đều, các mẫu lân cận được chọn có thể không mang tính đại diện, dẫn đến hiện tượng lan truyền sai nhãn;
- Thứ hai, việc lựa chọn tham số k có ảnh hưởng trực tiếp đến kết quả: giá trị k nhỏ khiến mô hình nhạy cảm với nhiễu, trong khi k lớn có thể làm mờ cấu trúc cục bộ và thiên lệch về các lớp chiếm ưu thế;
- Thứ ba, việc áp dụng bỏ phiếu đa số trực tiếp trên các mẫu điểm riêng lẻ có thể thiếu ổn định, đặc biệt khi dữ liệu có cấu trúc hình học phức tạp hoặc tồn tại các vùng chồng lấn lớn giữa các cụm;
- Ngoài ra, chi phí tính toán trong giai đoạn suy luận tăng nhanh theo kích thước

tập dữ liệu, do cần tính khoảng cách tới toàn bộ các mẫu huấn luyện.

2.2.2. Kỹ thuật bỏ phiếu đa số trong mạng nơron FMN

Trong mạng nơron FMN, dữ liệu được biểu diễn thành các siêu hộp mờ, với mỗi siêu hộp đại diện cho một vùng không gian đặc trưng có cùng nhãn lớp hoặc cùng cấu trúc ngữ nghĩa. Do đó, cơ chế bỏ phiếu đa số trong FMN không chỉ có thể được xây dựng dựa trên các mẫu dữ liệu riêng lẻ, mà còn có thể được định nghĩa dựa trên: (1) các đại diện hình học của siêu hộp, hoặc (2) cấu trúc của siêu hộp. Đây là các hướng tiếp cận giúp đánh giá các nhãn lớp dựa trên các điểm dữ liệu lân cận mang nhãn.

Trong hướng tiếp cận thứ nhất, nghiên cứu sinh đã đề xuất mô hình FMN-Voting nhằm kết hợp chiến lược chọn hạt giống trong học bán giám sát với phương pháp bỏ phiếu đa số [CT3]. FMN-Voting thực hiện lấy phiếu từ các điểm tâm hình học của siêu hộp thỏa mãn điều kiện nhận phiếu. Mô hình FMN-Voting được phát triển từ kỹ thuật bán giám sát MSCFMN [21].

Trong hướng tiếp cận thứ hai, nghiên cứu sinh đã đề xuất mô hình FMN-SAL với bỏ phiếu đa số sử dụng chiến lược chọn điểm lấy phiếu từ các mẫu dữ liệu lân cận mang nhãn [CT2]. Mô hình FMN-SAL được phát triển và kế thừa từ kỹ thuật bán giám sát SS-FMM [19].

2.3. Mô hình FMN-Voting

2.3.1. Mô tả bài toán

Xét tập dữ liệu đầu vào với thông tin nhãn hạn chế và tập các siêu hộp được hình thành bởi mạng nơron FMN. Trong quá trình học, các siêu hộp phản ánh sự phân bố và cấu trúc cục bộ của các mẫu dữ liệu trong không gian đặc trưng. Tuy nhiên, khi số lượng mẫu mang nhãn ban đầu còn hạn chế, thông tin giám sát chưa đủ để xác định trực tiếp nhãn cho toàn bộ các siêu hộp. Vì vậy, bài toán đặt ra là khai thác cấu trúc của các siêu hộp để tự xác định thông tin bổ trợ, sau đó sử dụng thông tin này nhằm hỗ trợ quá trình học bán giám sát và mở rộng thông tin nhãn tới các vùng dữ liệu chưa được xác định nhãn.

Để giải quyết bài toán trên, FMN-Voting thực hiện theo hai giai đoạn liên tiếp. Trong giai đoạn thứ nhất, mô hình khai thác cấu trúc dữ liệu được hình thành bởi mạng FMN để tự xác định thông tin bổ trợ dưới dạng thông tin nhãn cho một số siêu hộp. Các siêu hộp được xác định thông tin nhãn ở giai đoạn này đóng vai trò là nguồn thông tin giám sát ban đầu cho giai đoạn tiếp theo. Cách tiếp cận này cho phép mô

hình tận dụng thông tin tiềm ẩn trong cấu trúc dữ liệu, thay vì chỉ phụ thuộc vào lượng thông tin nhãn hạn chế được cung cấp ban đầu.

Trong giai đoạn thứ hai, thông tin nhãn đã xác định được sử dụng để thực hiện quá trình lan truyền nhãn. Đối với các siêu hộp chưa biết nhãn, FMN-Voting xem xét thông tin từ các siêu hộp mang nhãn lân cận và áp dụng cơ chế bỏ phiếu để xác định nhãn phù hợp. Việc sử dụng thông tin từ nhiều siêu hộp lân cận giúp quyết định gán nhãn không phụ thuộc vào một siêu hộp đơn lẻ, đồng thời khai thác được thông tin lân cận trong cấu trúc siêu hộp đã hình thành.

Như vậy, bài toán của FMN-Voting có thể được xem là quá trình tự xác định thông tin bổ trợ và khai thác thông tin bổ trợ đó để lan truyền nhãn trong mạng FMN. Kết quả của hai giai đoạn là tập các siêu hộp được bổ sung thông tin nhãn, từ đó hỗ trợ quá trình xác định cấu trúc cụm và hoàn thiện kết quả phân cụm bán giám sát.

2.3.2. Ý tưởng của FMN-Voting

Như đã trình bày trong mục 1.3, các mô hình học bán giám sát dựa trên mạng nơron FMN hiện nay vẫn tồn tại nhiều hạn chế cả về yêu cầu dữ liệu gán nhãn lẫn chi phí tính toán và khả năng thích nghi với cấu trúc dữ liệu phức tạp.

Các mô hình FMN cổ điển và biến thể như GFMM [15], RFMN [18] và SSL-FMM [17] đòi hỏi tỷ lệ mẫu huấn luyện có nhãn tương đối cao để đạt được chất lượng phân cụm chấp nhận được. Điều này làm giảm tính khả thi của các mô hình trong các bài toán thực tế, nơi dữ liệu gán nhãn thường khan hiếm hoặc chi phí lớn.

Mô hình SS-FMM [19] và các phương pháp kế thừa như SCFMN [20] đã mở rộng FMN sang bối cảnh học bán giám sát thông qua việc sử dụng ngưỡng đảm bảo chất lượng β để kiểm soát quá trình lan truyền nhãn. Tuy nhiên, các mô hình này phải thực hiện nhiều vòng duyệt trên tập dữ liệu huấn luyện nhằm thỏa mãn ràng buộc về β, dẫn đến chi phí tính toán lớn và khả năng mở rộng kém khi kích thước dữ liệu tăng. Mặc dù MSCFMN [21] đã cải thiện đáng kể hiệu quả tính toán so với SCFMN nhờ cơ chế lan truyền nhãn dựa trên siêu hộp, chất lượng phân cụm của mô hình này vẫn phụ thuộc vào ngưỡng đảm bảo chất lượng β. Nếu β nhỏ, điều kiện nhận nhãn từ siêu hộp lân cận dễ được thỏa mãn khi lan truyền nhãn; nếu β lớn, thuật toán cần nhiều lần duyệt hơn để lan truyền nhãn tới các siêu hộp chưa được gán nhãn.

Xuất phát từ các hạn chế nêu trên, ý tưởng của FMN-Voting được đề xuất là thay đổi cách thức dẫn dắt quá trình gán nhãn trong học bán giám sát của mạng nơron MSCFMN, từ việc phụ thuộc chủ yếu vào ngưỡng chất lượng toàn cục β sang một cơ chế suy luận nhãn cục bộ, dựa trên bỏ phiếu đa số giữa các đại diện hình học của siêu

hộp.

FMN-Voting xem mỗi siêu hộp như một thực thể đại diện cho một vùng không gian đặc trưng đã được học, và khai thác mối quan hệ hình học giữa các siêu hộp để suy luận nhãn cho các siêu hộp chưa được gán nhãn.

Cụ thể, FMN-Voting sử dụng tâm hình học của siêu hộp như một đại diện cho toàn bộ vùng dữ liệu mà siêu hộp bao phủ. Đối với một siêu hộp chưa có nhãn, nhãn của nó không được quyết định trực tiếp dựa trên một tiêu chí ngưỡng cứng, mà được suy luận thông qua bỏ phiếu đa số từ các siêu hộp lân cận đã có nhãn, dựa trên độ thuộc giữa các tâm siêu hộp trong không gian đặc trưng. Cách tiếp cận này cho phép FMN-Voting khai thác cấu trúc phân hoạch không gian đã được hình thành trong quá trình học FMN, đồng thời giảm đáng kể sự phụ thuộc vào giả thiết về hình dạng hoặc kích thước cụm. Đồng thời, cơ chế bỏ phiếu đa số giúp hạn chế lan truyền sai nhãn, do nhãn cuối cùng được quyết định bởi sự đồng thuận của nhiều đại diện lân cận, thay vì phụ thuộc vào một siêu hộp đơn lẻ. Nhờ đó, FMN-Voting có khả năng xử lý tốt hơn các tập dữ liệu có phân bố phức tạp, cụm không cân bằng hoặc chồng lấn mạnh, trong khi vẫn duy trì chi phí tính toán thấp và phù hợp với kịch bản học bán giám sát với số lượng mẫu gán nhãn hạn chế.

Thuật toán học trong FMN-Voting gồm hai giai đoạn chính:

Giai đoạn 1. Xây dựng tập hạt giống: tính toán thông tin bổ trợ cho các mẫu dữ liệu trong tập dữ liệu huấn luyện;

Giai đoạn 2. Huấn luyện bán giám sát với bỏ phiếu đa số dựa trên siêu hộp. Thuật toán học thực hiện lan truyền nhãn từ các mẫu có nhãn và thực hiện bỏ phiếu đa số dựa trên các siêu hộp lân cận (siêu hộp có độ thuộc cao nhất) với các mẫu dữ liệu chưa có nhãn.

2.3.3. Kiến trúc mạng nơron FMN-Voting

Kiến trúc mạng nơron FMN-Voting được xây dựng trên nền tảng FMN, gồm ba lớp chính là F_A, F_H và F_C, đồng thời bổ sung cơ chế *bỏ phiếu đa số* nhằm tăng độ ổn định và độ tin cậy trong suy luận nhãn. Hình 2.3 minh họa cấu trúc tổng thể của mô hình.

Lớp vào F_A gồm n nơron $a_1, a_2, \ldots, a_n$, mỗi nơron tương ứng với một thuộc tính của mẫu đầu vào $x_r = (x_{r1}, x_{r2}, \ldots, x_{rn})$. Lớp này chỉ thực hiện chức năng tiếp nhận và truyền dữ liệu.

Lớp siêu hộp F_H là tầng biểu diễn trung tâm, trong đó mỗi nơron biểu diễn một siêu hộp mờ trong không gian đặc trưng. Các siêu hộp được tổ chức thành các

tập con $\{b_1, \ldots, b_k\}$, $\{l_1, \ldots, l_q\}$ và $\{u_1, \ldots, u_p\}$, tương ứng với các nhóm siêu hộp đã có nhãn, trung gian và chưa được gán nhãn. Mỗi siêu hộp được xác định bởi hai vectơ V_j và W_j, lần lượt biểu diễn biên dưới và biên trên. Với một mẫu x_r, mỗi siêu hộp H_j tạo ra độ thuộc $\mu(x_r, H_j) \in [0, 1]$, phản ánh mức độ phù hợp của mẫu với vùng không gian mà siêu hộp biểu diễn.

Lớp F_C gồm g nơron $c_1, c_2, \ldots, c_g$, mỗi nơron đại diện cho một cụm mục tiêu. Các siêu hộp ở lớp F_H được nối với F_C thông qua các liên kết U, biểu diễn mối quan hệ giữa siêu hộp và cụm. Khác với FMN phân lớp truyền thống, trong FMN-Voting không tồn tại ánh xạ cứng một-một giữa siêu hộp và nhãn lớp; thay vào đó, mỗi nhãn có thể nhận kích hoạt từ nhiều siêu hộp khác nhau. Các tín hiệu này được tổng hợp thông qua nút $(+)$ và được dùng làm cơ sở cho quyết định phân cụm cuối cùng.

Trong FMN-Voting, cơ chế bỏ phiếu đa số được thực hiện trong quá trình lan truyền nhãn cho các mẫu chưa có nhãn dựa trên tập các siêu hộp đã có nhãn $\mathcal{L}$. Với mỗi mẫu chưa có nhãn x_r không thể được bao phủ hoặc mở rộng bởi một siêu hộp phù hợp, mô hình xác định tập $N(x_r)$ gồm các siêu hộp có nhãn trong $\mathcal{L}$ có độ thuộc cao nhất đối với x_r và thỏa mãn ngưỡng β. Các giá trị độ thuộc chỉ được sử dụng để lựa chọn tập lân cận; sau khi tập lân cận được xác định, mỗi siêu hộp đóng góp một phiếu cho lớp tương ứng.

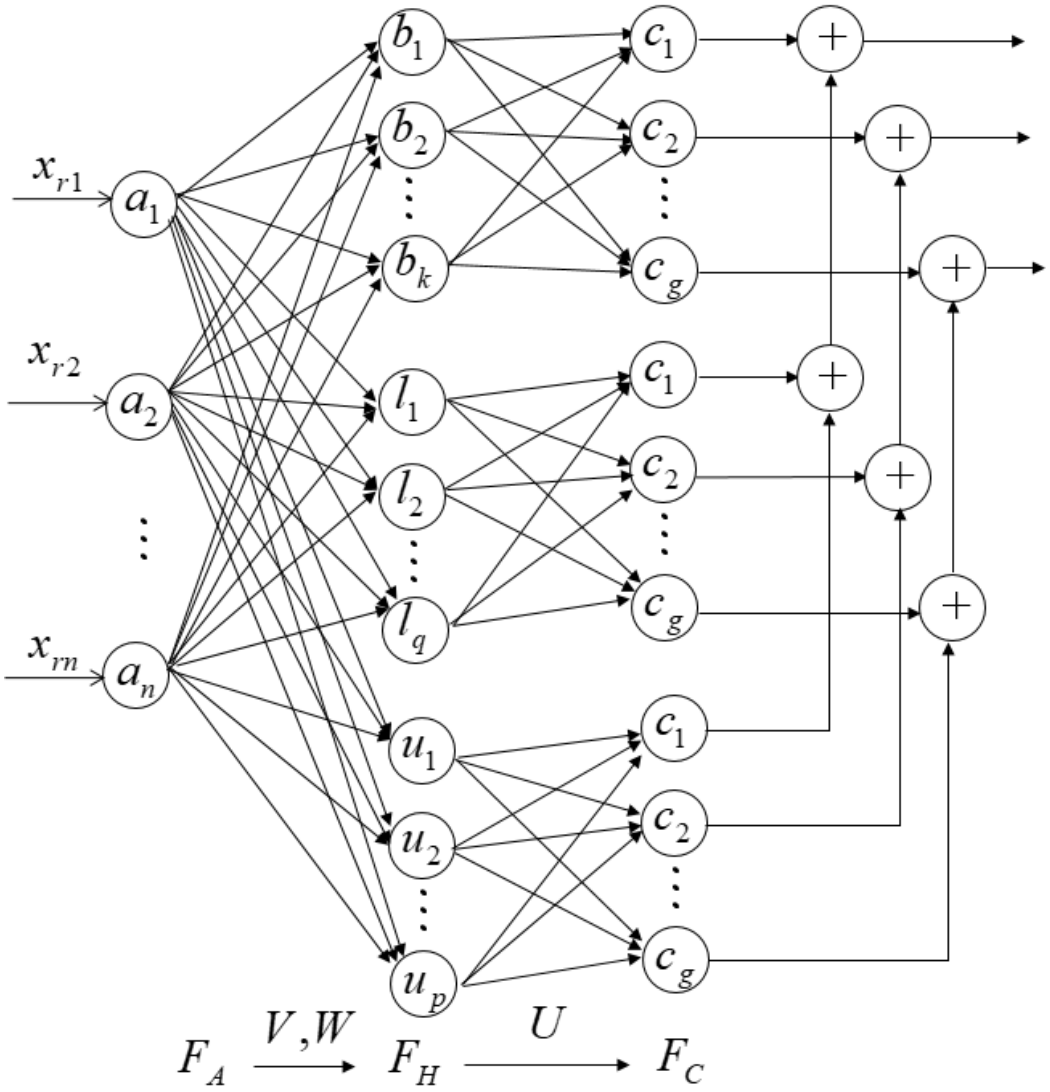
$$\hat{\ell}(x_r) = \arg\max_{c \in \mathcal{C}} \sum_{L_i \in N(x_r)} \mathbb{I}(\ell(L_i) = c), \tag{2.9}$$

trong đó, $\mathcal{C} = \{1, \ldots, C\}$ là tập các lớp khả dĩ; $N(x_r)$ là tập các siêu hộp có nhãn lân cận của x_r; $\ell(L_i)$ là nhãn của siêu hộp L_i và $\mathbb{I}(\cdot)$ là hàm chỉ thị, được xác định bởi

$$\mathbb{I}\big(\ell(L_i) = c\big) = \begin{cases} 1, & \text{nếu } \ell(L_i) = c, \\ 0, & \text{nếu } \ell(L_i) \neq c. \end{cases} \tag{2.10}$$

Như vậy, mỗi siêu hộp $L_i \in N(x_r)$ đóng góp một phiếu cho lớp tương ứng với nhãn của nó. Nhãn nhận được số phiếu lớn nhất được sử dụng để xác định nhãn cho siêu hộp mới chứa x_r, đồng thời cập nhật nhãn cho mẫu x_r.

Như vậy, trong FMN-Voting, nhãn của một siêu hộp mới không phụ thuộc vào một siêu hộp dẫn dắt đơn lẻ mà được suy luận từ một tập các siêu hộp đã gán nhãn trong $\mathcal{L}$. Cách tiếp cận này giúp khai thác tốt hơn cấu trúc cục bộ của không gian siêu hộp và tăng tính nhất quán của quá trình suy luận nhãn.



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

2.3.4. Thuật toán học trong FMN-Voting

2.3.4.1. Giai đoạn 1: Xây dựng tập hạt giống

Giai đoạn đầu tiên của thuật toán học xác định các hạt giống dựa trên việc phân hoạch không gian dữ liệu thành các siêu hộp mờ. Các hạt giống được gán nhãn giả dựa trên tập siêu hộp mờ. Các mẫu dữ liệu có độ thuộc đầy đủ sẽ mang nhãn của siêu hộp tương ứng.

Thuật toán 2.1 gồm hai bước chính: trước hết thực hiện học không giám sát để xây dựng tập siêu hộp B; sau đó khai thác cấu trúc siêu hộp đã hình thành để xác định thông tin nhãn bổ trợ cho các mẫu có độ thuộc đầy đủ đối với một siêu hộp.

Mệnh đề 2.1. *Thuật toán 2.1 có độ phức tạp thời gian là*

$$T_1 = \mathcal{O}(nMK), \tag{2.11}$$

trong đó M là số mẫu huấn luyện, n là số đặc trưng của mỗi mẫu và K là số siêu hộp

Thuật toán 2.1 Xây dựng tập siêu hộp và xác định thông tin nhãn bổ trợ

Input: $X = \{\mathbf{x}_r\}$, $M = |X|$, $\delta_1^{\max}$, γ, β
Output: $\mathcal{B}$; $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$;

1: Thực hiện giai đoạn huấn luyện không giám sát theo Thuật toán 1.2 trên tập dữ liệu X, thu được tập siêu hộp $\mathcal{B}$'
2: $X^1 \leftarrow \varnothing$; $X^2 \leftarrow \varnothing$;
3: **for** $x_r \in X$ **do**
4: **if** $\exists B_j \in \mathcal{B}$ sao cho $\mu(x_r, B_j) = 1$ **then**
5: $\ell_{x_r} \leftarrow \ell(B_j)$;
6: $X^1 \leftarrow X^1 \cup \{(x_r, \ell_{x_r})\}$;
7: **else**
8: $\ell_{x_r} \leftarrow 0$;
9: $X^2 \leftarrow X^2 \cup \{(x_r, \ell_{x_r})\}$;
10: **end if**
11: **end for**
12: $X \leftarrow X^1 \cup X^2$;
13: **return** $\mathcal{B}$;$X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$;

mờ được tạo ra trong quá trình huấn luyện.

Chứng minh. Xét Thuật toán 2.1, với $M = |X|$, n là số đặc trưng và $K = |B|$ là số siêu hộp mờ.

Tại dòng 1 của Thuật toán 2.1, thuật toán học không giám sát được thực hiện trên tập X để xây dựng tập siêu hộp $\mathcal{B}$. Với mỗi mẫu $x_r \in X$, trong trường hợp xấu nhất cần tính độ thuộc với tối đa K siêu hộp; mỗi phép tính trên n chiều có độ phức tạp $O(n)$. Do đó, chi phí của dòng 1 là:

$$\mathcal{O}(MKn). \tag{2.12}$$

Tại các dòng 2–10, thuật toán duyệt toàn bộ M mẫu để xác định thông tin nhãn bổ trợ. Trong trường hợp xấu nhất, đối với mỗi mẫu x_r, thuật toán cần kiểm tra độ thuộc với tối đa K siêu hộp. Do đó, chi phí của quá trình này là:

$$\mathcal{O}(MKn). \tag{2.13}$$

Các thao tác cập nhật X_1, X_2 và hợp nhất lại tập dữ liệu tại dòng 11 yêu cầu tối đa một số phép toán tuyến tính theo số mẫu, nên có chi phí:

$$\mathcal{O}(M). \tag{2.14}$$

Vì vậy, tổng độ phức tạp thời gian của thuật toán là

$$T_1 = \mathcal{O}(MKn) + \mathcal{O}(MKn) + \mathcal{O}(M). \tag{2.15}$$

Bỏ qua các hạng tử bậc thấp, suy ra

$$T_1 = \mathcal{O}(nMK). \tag{2.16}$$

□

2.3.4.2. Giai đoạn 2: Huấn luyện bán giám sát với bỏ phiếu đa số dựa trên siêu hộp

Giai đoạn 2 của thuật toán FMN-Voting tập trung vào việc suy luận và gán nhãn cho các siêu hộp cũng như các mẫu dữ liệu chưa được gán nhãn thông qua cơ chế bỏ phiếu đa số. Dựa trên tập siêu hộp đã được xây dựng ở Giai đoạn 1, thuật toán đồng thời khai thác thông tin cấu trúc hình học của các siêu hộp và thông tin giám sát sẵn có từ các mẫu mang nhãn nhằm thực hiện quá trình lan truyền nhãn một cách có kiểm soát. Quá trình này được thực hiện thông qua các bước chính sau:

(1) Xây dựng tập siêu hộp có nhãn $\mathcal{L}$ từ các mẫu dữ liệu đã được gán nhãn ($x_r \in X^1$);

(2) Đối với các siêu hộp được hình thành từ các mẫu chưa gán nhãn, nhãn của siêu hộp được suy luận thông qua cơ chế bỏ phiếu đa số dựa trên độ thuộc của nó đối với k siêu hộp lân cận trong tập $\mathcal{L}$, với $k = 2C + 1$, trong đó c là số cụm đầu vào. Trong FMN-Voting, số lượng siêu hộp lân cận được lựa chọn theo $k = 2C + 1$, trong đó C là số lớp/cụm. Giá trị k được chọn là một số lẻ và lớn hơn $2C$ nhằm tạo một tập lân cận đủ lớn để thực hiện bỏ phiếu đa số trên C lớp, đồng thời hạn chế trường hợp quyết định nhãn phụ thuộc vào một số lượng quá ít siêu hộp lân cận. Đây là tham số được lựa chọn theo nguyên tắc thiết kế của mô hình.

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

với $\mathcal{N}(x_r)$ là tập $2C + 1$ lân cận với điểm dữ liệu x_r được xác định bằng cách lựa chọn $2C + 1$ siêu hộp đã có nhãn $L_j \in \mathcal{L}$ có giá trị độ thuộc mờ $\mu(x_r, L_j)$ lớn nhất đối với mẫu truy vấn x_r ($\mu(x_r, L_j)$ được định nghĩa theo Công thức (1.3)).

$$\mathrm{NoC}_\ell = \sum_{L_j \in \mathcal{N}_{(x_r)}} \mathbb{I}(\ell(L_j) = \ell)\,, \qquad \ell = 1, \ldots, C \tag{2.18}$$

với $I(\cdot)$ là hàm chỉ thị, nhận giá trị bằng 1 khi điều kiện bên trong được thỏa mãn và bằng 0 trong trường hợp ngược lại; NoC_ℓ là số phiếu (số siêu hộp) có nhãn lớp là ℓ;

$$\ell(L_{\text{new}}) = \arg\max_{\ell} \mathrm{NoC}_\ell \tag{2.19}$$

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 $\exists L_j \in \mathcal{L}$ sao cho $\mu(x_r, L_j) = \max_{L_k \in \mathcal{L}} \mu(x_r, L_k)$ 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
      for all $L_k \in \mathcal{L}$ **do**
        if $\Omega(L_j, L_k) > 0$ **and** $\ell(L_j) \neq \ell(L_k)$ **then**
          Điều chỉnh chồng lấn giữa L_j và L_k;
        end if
      end for
    else
      Xác định tập lân cận $\mathcal{N}(x_r)$ theo (2.17)
      if $|\mathcal{N}(x_r)| \geq 1$ **then**
        Tạo mới L_{new} chứa x_r; Tính số phiếu cho từng nhãn lớp $L_i \in \mathcal{N}(x_r)$ theo (2.18); Gán nhãn cho $\ell(L_{\text{new}})$ 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 tập $\mathcal{L}$ (với u_{jk} được xác định theo Công thức (1.36))
return $\mathcal{L}$; X; U;

Mệnh đề 2.2. *Thuật toán 2.2 có độ phức tạp thời gian là*

$$T_2 = \mathcal{O}(R_{max} n M K), \tag{2.20}$$

trong đó M là số mẫu dữ liệu chưa gán nhãn, n là số đặc trưng của mỗi mẫu và K là số siêu hộp được tạo ra trong quá trình học, và R_{max} là số lượt duyệt tối đa của vòng lặp `repeat-until`.

Chứng minh. Trong mỗi lượt duyệt, một mẫu x_r được so sánh với tối đa K siêu hộp để tính độ thuộc, lựa chọn siêu hộp phù hợp, kiểm tra khả năng mở rộng, kiểm tra chồng lấn hoặc xác định tập lân cận phục vụ bỏ phiếu. Mỗi phép so sánh trên không gian n chiều có chi phí $O(n)$, do đó chi phí xử lý một mẫu bị chặn trên bởi $O(nK)$. Với tối đa M mẫu trong một lượt duyệt, chi phí là $O(nMK)$. Nếu vòng lặp `repeat-until` thực hiện tối đa R_{max} lượt thì độ phức tạp thời gian trong trường hợp xấu nhất là

$$T_2 = O(R_{max} M n K) = O(R_{max} n M K). \tag{2.21}$$

□

2.3.5. Độ phức tạp tính toán của FMN-Voting

Thuật toán học trong FMN-Voting được chia thành hai giai đoạn chính.

Giai đoạn 1: xây dựng tập hạt giống, với chi phí tính toán là $T_1 = \mathcal{O}(nMK)$; và giai đoạn 2: Huấn luyện bán giám sát với bỏ phiếu đa số dựa trên siêu hộp với chi phí tính toán $T_2 = \mathcal{O}(R_{max} M n K)$. Tổng chi phí tính toán của FMN-Voting được xác định bằng tổng chi phí của Giai đoạn 1 và Giai đoạn 2;

$$T = T_1 + T_2 = \mathcal{O}(nMK) + \mathcal{O}(R_{max} M n K) = \mathcal{O}(R_{max} M n K) \tag{2.22}$$

Trong trường hợp R_{max} được xem là một hằng số bị chặn bởi cơ chế dừng của thuật toán, độ phức tạp có thể rút gọn thành $O(nMK)$. Tuy nhiên, để phản ánh đầy đủ cấu trúc vòng lặp của giả mã, luận án sử dụng dạng tổng quát $O(R_{max} n M K)$.

FMN-Voting phù hợp khi các siêu hộp có nhãn lân cận phản ánh tương đối tốt phân bố cục bộ của dữ liệu, khi đó việc tổng hợp thông tin từ nhiều siêu hộp thông qua bỏ phiếu đa số giúp hạn chế sự phụ thuộc vào một siêu hộp đơn lẻ. Tuy nhiên, mô hình xem các siêu hộp được lựa chọn trong tập lân cận là các nguồn bỏ phiếu tương đương. Trong trường hợp dữ liệu phân bố không đồng đều bên trong siêu hộp, tâm hình học của siêu hộp có thể lệch đáng kể so với tâm phân bố thực tế của các mẫu. Khi đó, một siêu hộp có độ thuộc cao đối với mẫu truy vấn chưa chắc có tính đại diện cao, làm giảm độ tin cậy của quá trình bỏ phiếu. Hạn chế này rõ hơn tại các vùng có phân bố phức tạp hoặc chồng lấn giữa các cụm. FMN-Voting còn có hạn chế đối với các tập dữ liệu có cấu trúc cụm không hình cầu, mất cân bằng. Trong trường hợp này,

quá trình hình thành các siêu hộp có thể nhạy với thứ tự xuất hiện của các mẫu trong tập huấn luyện, dẫn đến sự khác biệt về cấu trúc và phân bố của các siêu hộp giữa các lần thực nghiệm. Sự thay đổi này ảnh hưởng đến tập siêu hộp lân cận tham gia bỏ phiếu và có thể làm kết quả suy luận nhãn biến động. Hạn chế này được quan sát trên một số tập dữ liệu thực nghiệm có cấu trúc cụm không hình cầu.

2.4. Mô hình FMN-SAL

2.4.1. Mô tả bài toán

Xét tập dữ liệu huấn luyện gồm một số mẫu đã được gán nhãn và các mẫu chưa có nhãn. Sau quá trình phân hoạch không gian dữ liệu bằng mạng nơron FMN, các mẫu được biểu diễn thông qua tập các siêu hộp phản ánh cấu trúc và phân bố cục bộ của dữ liệu. Tuy nhiên, các siêu hộp được hình thành ban đầu chưa phải tất cả đều có thông tin nhãn. Vì vậy, bài toán đặt ra đối với FMN-SAL là khai thác trực tiếp thông tin từ các mẫu đã có nhãn trong tập huấn luyện để xác định thông tin nhãn cho các siêu hộp, từ đó hình thành nguồn thông tin giám sát phục vụ quá trình lan truyền nhãn đến các siêu hộp và mẫu dữ liệu chưa có nhãn.

Để giải quyết bài toán trên, trước hết FMN-SAL sử dụng các mẫu mang nhãn trong tập huấn luyện để xác định nhãn cho các siêu hộp có chứa hoặc có quan hệ với các mẫu mang nhãn tương ứng. Qua quá trình này, một tập các siêu hộp có nhãn được hình thành từ thông tin giám sát sẵn có của dữ liệu huấn luyện. Các siêu hộp này đóng vai trò là nguồn thông tin giám sát ban đầu, làm cơ sở cho quá trình suy luận và lan truyền nhãn tiếp theo. Nhờ đó, thông tin nhãn ở mức mẫu được chuyển thành thông tin nhãn ở mức siêu hộp và được tích hợp vào cấu trúc của mạng FMN.

Trên cơ sở tập siêu hộp có nhãn đã được xây dựng, FMN-SAL tiếp tục thực hiện quá trình xác định nhãn cho các đối tượng chưa có nhãn. Đối với một mẫu hoặc siêu hộp chưa xác định nhãn, mô hình khai thác thông tin từ các mẫu mang nhãn lân cận và áp dụng cơ chế bỏ phiếu để xác định nhãn phù hợp. Việc sử dụng nhiều mẫu mang nhãn lân cận giúp quá trình suy luận khai thác trực tiếp phân bố thực tế của các mẫu dữ liệu, thay vì chỉ phụ thuộc vào đặc trưng hình học của một siêu hộp đơn lẻ.

Như vậy, FMN-SAL thực hiện bài toán theo quá trình khai thác thông tin nhãn có sẵn ở mức mẫu, xây dựng tập siêu hộp có nhãn, sử dụng cơ chế bỏ phiếu để lan truyền và xác định nhãn cho các đối tượng chưa có nhãn. Kết quả của quá trình này là thông tin nhãn từ một số lượng hạn chế các mẫu trong tập huấn luyện được khai thác và mở rộng tới các siêu hộp và mẫu chưa có nhãn, qua đó hỗ trợ hoàn thiện kết quả

phân cụm bán giám sát.

2.4.2. Ý tưởng của FMN-SAL

Trong mô hình FMN-Voting, nhãn của mẫu chưa gán nhãn được suy luận bằng cơ chế bỏ phiếu đa số từ các siêu hộp lân cận có độ thuộc cao. Cách tiếp cận này ngầm giả định rằng các siêu hộp lân cận đều có mức độ đại diện tương đương. Tuy nhiên, giả định này không luôn đúng, đặc biệt khi dữ liệu bên trong siêu hộp phân bố không đều hoặc bị ảnh hưởng bởi nhiễu.

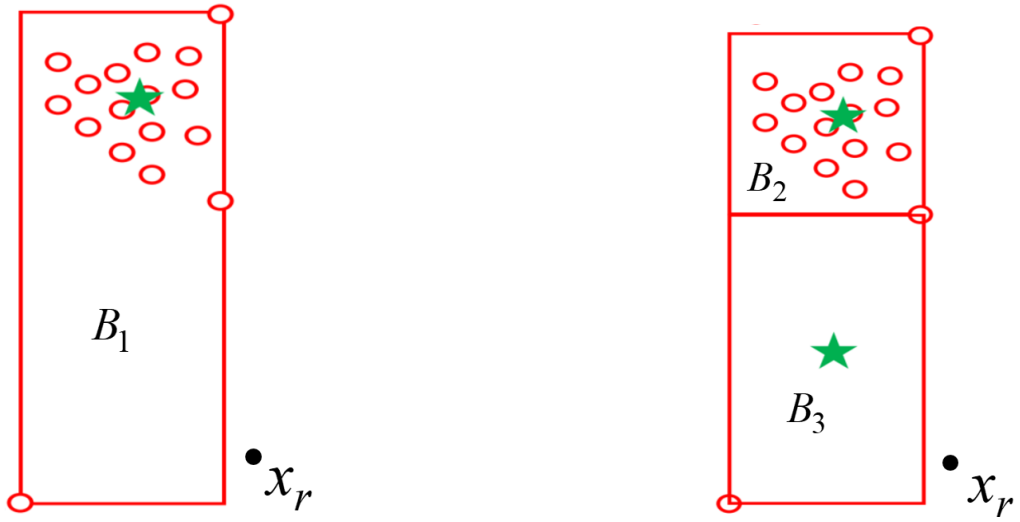
Trong quá trình học FMN, một siêu hộp có thể được mở rộng để bao phủ các mẫu phân bố lệch về một phía của không gian đặc trưng. Khi đó, tâm hình học của siêu hộp có thể sai khác đáng kể so với tâm dữ liệu thực, làm giảm độ tin cậy của siêu hộp khi tham gia bỏ phiếu. Vì vậy, việc xem các siêu hộp có độ thuộc cao như những nguồn bỏ phiếu ngang nhau có thể dẫn đến suy luận nhãn thiếu chính xác, nhất là tại các vùng biên hoặc chồng lấn giữa các cụm.

Như minh họa trên Hình 2.4, siêu hộp B_1 có phân bố dữ liệu lệch nên độ thuộc cao không đồng nghĩa với tính đại diện cao; siêu hộp B_2 phản ánh tốt cấu trúc cục bộ của dữ liệu; trong khi B_3 có chỉ số sử dụng thấp nên mức độ đại diện kém.

Để khắc phục một phần hạn chế này, mô hình FMN-SAL được đề xuất với chiến lược bỏ phiếu dựa trên các mẫu mang nhãn lân cận thay vì các siêu hộp lân cận. Theo đó, FMN-SAL không thay đổi mục tiêu suy luận nhãn của FMN-Voting mà thay đổi nguồn thông tin được sử dụng cho quyết định bỏ phiếu. Thay vì coi các siêu hộp lân cận là các đơn vị bỏ phiếu, FMN-SAL khai thác trực tiếp các mẫu mang nhãn trong tập huấn luyện. Cách tiếp cận này giảm sự phụ thuộc của quyết định nhãn vào tâm hình học và mức độ đại diện của từng siêu hộp, đặc biệt khi các mẫu bên trong siêu hộp có phân bố không đồng đều.

2.4.3. Kiến trúc mạng nơron FMN-SAL

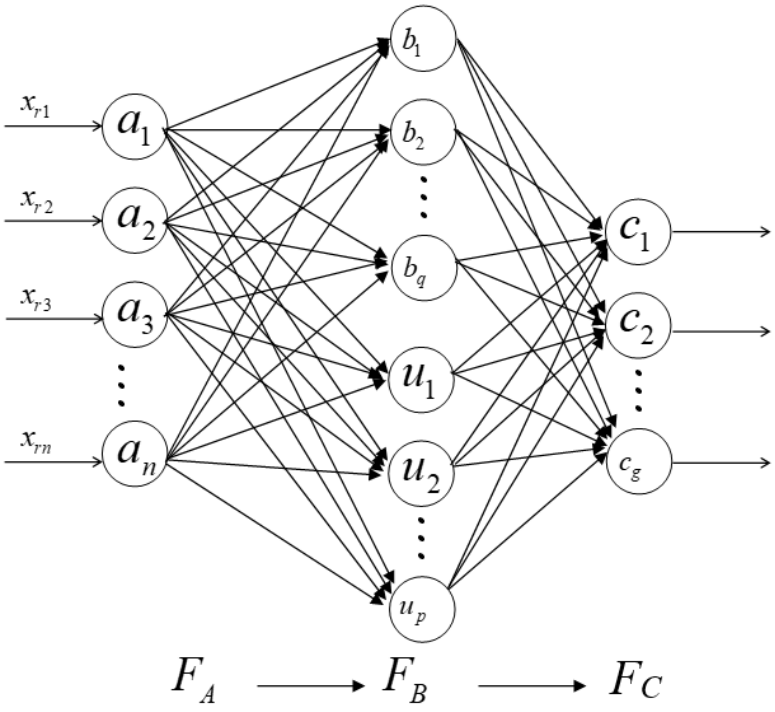
Kiến trúc mạng nơron FMN-SAL được minh họa trong Hình 2.5, gồm ba lớp chính: lớp vào F_A, lớp siêu hộp F_B và lớp ra F_C. Lớp vào F_A gồm n nơron $a_1, a_2, \ldots, a_n$, tương ứng với n thuộc tính của mẫu dữ liệu đầu vào $x_r = (x_{r1}, x_{r2}, \ldots, x_{rn})$. Lớp F_B gồm các nơron biểu diễn các siêu hộp mờ được hình thành trong quá trình học, trong đó có thể chia thành hai nhóm: nhóm các siêu hộp đã được gán nhãn, ký hiệu $b_1, b_2, \ldots, b_q$, và nhóm các siêu hộp chưa gán nhãn hoặc cần suy luận nhãn, ký hiệu $u_1, u_2, \ldots, u_p$. Các kết nối từ F_A đến F_B được xác định bởi hai ma trận V và W, lần lượt biểu diễn cận dưới và cận trên của các siêu hộp trong không gian đặc trưng.



(a) Phân bố dữ liệu không đều *(b) Phân bố dữ liệu đều*

Hình 2.4: *B_1 mất cân bằng, B_2 cân bằng và có chỉ số sử dụng $c(B_2)$ cao, B_3 có chỉ số sử dụng thấp*

Lớp ra F_C gồm g nơron $c_1, c_2, \ldots, c_g$, đại diện cho các lớp hoặc cụm đầu ra; quan hệ giữa các siêu hộp trong F_B và các nơron lớp ra trong F_C được biểu diễn thông qua ma trận U. Như vậy, trong kiến trúc FMN-SAL, các siêu hộp mờ đóng vai trò trung gian ánh xạ dữ liệu từ không gian đặc trưng sang không gian nhãn, đồng thời hỗ trợ quá trình suy luận nhãn cho các siêu hộp chưa được xác định lớp.



Hình 2.5: *Kiến trúc mạng nơron FMN-SAL*

2.4.4. Thuật toán học trong FMN-SAL

Thuật toán học bán giám sát trong FMN-SAL được trình bày trong Thuật toán 2.3. Quá trình học bán giám sát trong FMN-SAL dựa trên việc xác định các

mẫu lân cận có nhãn, tổng hợp số phiếu theo từng lớp và suy luận nhãn cho siêu hộp mới. Cụ thể, với mẫu dữ liệu x_r, tập lân cận $\mathcal{N}(x_r)$ được xác định theo Công thức (2.23):

$$\mathcal{N}(x_r) = \underset{\min(k,\,|X'|)}{\arg\operatorname{top}} \left\{ \mu(x_r, x_i) \;\middle|\; x_i \in X',\ \mu(x_r, x_i) \geq \beta \right\}. \tag{2.23}$$

với $\mathcal{N}(x_r)$ là tập gồm k lân cận với điểm dữ liệu x_r được xác định bằng cách lựa chọn k mẫu đã có nhãn $x_i \in X'$ có giá trị độ thuộc mờ $\mu(x_r, x_i)$ lớn nhất đối với mẫu truy vấn x_r.

$$\mathrm{NoC}_\ell = \sum_{x_i \in \mathcal{N}(x_r)} \mathbb{I}(\ell_{x_i} = \ell), \qquad \ell = 1, \ldots, c \tag{2.24}$$

với NoC_ℓ là số phiếu (số mẫu dữ liệu) có nhãn lớp là ℓ.

$$\ell(B_{\text{new}}) = \arg\max_{\ell} \mathrm{NoC}_\ell \tag{2.25}$$

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

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

trong đó M là số mẫu đầu vào, n là số chiều đặc trưng và K là số siêu hộp mờ được tạo ra trong quá trình huấn luyện.

Chứng minh. Xét Thuật toán 2.3 với $M = |X|$, n là số đặc trưng và $K = |\mathcal{B}|$ là số siêu hộp được tạo ra.

Với mỗi mẫu $x_r \in X$, thuật toán cần tìm siêu hộp phù hợp nhất và kiểm tra điều kiện mở rộng. Trong trường hợp xấu nhất, mẫu x_r phải được so sánh với tối đa K siêu hộp, mỗi phép tính hàm thuộc hoặc kích thước có độ phức tạp $\mathcal{O}(n)$. Vì vậy, chi phí của bước này là

$$\mathcal{O}(nMK). \tag{2.27}$$

Nếu một siêu hộp được mở rộng, thuật toán tiếp tục kiểm tra và điều chỉnh chồng lấn với các siêu hộp còn lại. Trong trường hợp xấu nhất, thao tác này cũng cần so sánh với tối đa K siêu hộp, nên có độ phức tạp bị chặn trên bởi

$$\mathcal{O}(nMK). \tag{2.28}$$

Khi không thể mở rộng siêu hộp cho một mẫu chưa gán nhãn, thuật toán xác định tập lân cận gồm các mẫu có nhãn dựa trên độ thuộc. Với cách cài đặt trực tiếp, mỗi lần truy vấn có chi phí $\mathcal{O}(nL)$ với $L \leq M$, và trong trường hợp xấu nhất tổng chi phí của bước này là

$$\mathcal{O}(nM^2). \tag{2.29}$$

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 Công thức (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 $\exists B_j \in \mathcal{B}$ sao cho $\mu(x_r, B_j) = \max_{B_k \in \mathcal{B}} \mu(x_r, B_k)$ và $\delta(x_r, B_j) \leq \delta^{\max}$ **then**
    Điều chỉnh B_j theo (1.6) và (1.7)
    if $\ell_{x_r} = 0$ **then**
      $\ell_{x_r} \leftarrow \ell(B_j)$;
    end if
    $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    if phát hiện chồng lấn giữa B_j và B_k với $\ell(B_j) \neq \ell(B_k)$ **then**
      Điều chỉnh chồng lấn giữa các siêu hộp B_j và B_k
    end if
  else
    if $\ell(x_r) \neq 0$ **then**
      Tạo siêu hộp mới 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 $\mathcal{N}(x_r)$ theo (2.23)
      if $|\mathcal{N}(x_r)| = k$ **then**
        Tạo siêu hộp mới B_{new};
        Tính số phiếu cho từng nhãn lớp của các mẫu $x_i \in \mathcal{N}(x_r)$ theo (2.24);
        Gán nhãn $\ell(B_{\text{new}})$ theo (2.25); $\ell_{x_r} \leftarrow \ell(B_{\text{new}})$;
        $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$; $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$; $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$;
      end if
    end if
  end if
until $X = \varnothing$
$X \leftarrow X'$
Cập nhật U dựa trên tập $\mathcal{B}$ (với u_{jk} được xác định theo (1.36));
return $\mathcal{B}$, X, U;

Do đó, độ phức tạp thời gian của thuật toán trong trường hợp xấu nhất là

$$T_{\text{SAL}} = \mathcal{O}(nMK) + \mathcal{O}(nMK) + \mathcal{O}(nM^2) = \mathcal{O}(nMK + nM^2).$$

Suy ra (2.26). □

Mặc dù giảm sự phụ thuộc vào mức độ đại diện của siêu hộp, FMN-SAL phụ thuộc nhiều hơn vào chất lượng và phân bố của các mẫu mang nhãn. Khi số mẫu mang nhãn ít hoặc các mẫu này chưa đại diện tốt cho cấu trúc cục bộ của dữ liệu, tập lân cận được lựa chọn có thể chưa cung cấp đủ thông tin cho quyết định bỏ phiếu. Vì vậy, FMN-SAL không mặc nhiên cho kết quả tốt hơn FMN-Voting hoặc các mô hình FMN bán giám sát khác trên mọi tập dữ liệu; hiệu quả của mô hình phụ thuộc vào cấu trúc dữ liệu và thông tin nhãn sẵn có.

2.5. Thực nghiệm và đánh giá

2.5.1. Thiết lập thực nghiệm

2.5.1.1. Mục tiêu thực nghiệm

Mục tiêu của các bài thực nghiệm là đánh giá hiệu năng của FMN-Voting và FMN-SAL trên các tập dữ liệu benchmark phổ biến được lấy từ kho dữ liệu học máy UCI [22], CS [23]. Thông qua các thực nghiệm, nghiên cứu sinh hướng tới việc phân tích khả năng học bán giám sát của từng mô hình trong điều kiện số lượng mẫu mang nhãn hạn chế, cũng như mức độ khai thác thông tin cấu trúc dữ liệu để lan truyền nhãn một cách hiệu quả và ổn định.

Cụ thể, các thí nghiệm được thiết kế nhằm:

(i) So sánh hiệu năng phân cụm/phân loại của các mô hình đề xuất (FMN-Voting, FMN-SAL) với các biến thể FMN truyền thống và các phương pháp học bán giám sát liên quan trên cùng tập dữ liệu benchmark;

(ii) Đánh giá tác động của các chiến lược suy luận nhãn khác nhau, bao gồm bỏ phiếu dựa trên siêu hộp FMN-Voting và bỏ phiếu dựa trên các mẫu dữ liệu lân cận FMN-SAL;

(iii) Phân tích hiệu năng và độ ổn định của các mô hình trên các tập dữ liệu có đặc điểm phân bố và mức độ chồng lấn khác nhau;;

(iv) Khảo sát ảnh hưởng của các tham số học quan trọng, bao gồm ngưỡng mở rộng $\delta^{\max}$, ngưỡng độ thuộc β, hệ số điều khiển độ dốc γ và các tham số liên quan khác, đến hiệu năng và độ ổn định của các mô hình đề xuất;

(v) Đánh giá chi phí tính toán và khả năng mở rộng của các mô hình khi áp dụng trên các tập dữ liệu có kích thước và độ phức tạp khác nhau.

Các bài thực nghiệm không chỉ nhằm chứng minh hiệu quả của từng mô hình đề xuất, mà còn làm rõ vai trò của từng cơ chế cải tiến trong việc nâng cao chất lượng học bán giám sát của mạng nơron FMN.

2.5.1.2. Phương pháp thực nghiệm

Để đảm bảo tính khách quan và độ tin cậy của kết quả đánh giá, các thí nghiệm được thực hiện theo phương pháp kiểm tra chéo k-fold cross-validation với $k = 10$. Cụ thể, quy trình thực nghiệm được tiến hành như sau:

- Tập dữ liệu ban đầu được chia ngẫu nhiên thành 10 tập con không giao nhau (k-fold) có kích thước xấp xỉ nhau;
- Tại mỗi lần lặp, một tập con được sử dụng làm tập kiểm thử, trong khi 9 tập con còn lại được dùng làm tập huấn luyện;
- Đối với mô hình học bán giám sát dựa trên mẫu, trong mỗi tập huấn luyện, một tỷ lệ nhỏ các mẫu được giữ lại làm tập mẫu mang nhãn ban đầu, các mẫu còn lại được coi là chưa gán nhãn;
- Các tham số của mô hình được giữ cố định trong từng thí nghiệm và được thay đổi có kiểm soát khi thực hiện phân tích ảnh hưởng tham số;
- Quá trình trên được lặp lại 10 lần, sao cho mỗi tập con đóng vai trò tập kiểm thử đúng một lần.

Trong mỗi fold, tập kiểm tra chỉ được sử dụng để đánh giá mô hình sau khi quá trình huấn luyện và lựa chọn tham số hoàn tất. Các tham số tiền xử lý và siêu tham số của mô hình được xác định từ dữ liệu huấn luyện, không sử dụng nhãn hoặc kết quả đánh giá trên tập kiểm tra để lựa chọn tham số.

Kết quả đánh giá cuối cùng được tính bằng giá trị trung bình của các chỉ số đo lường trên 10 lần thực nghiệm, nhằm giảm thiểu ảnh hưởng của việc phân chia dữ liệu ngẫu nhiên và phản ánh chính xác hơn hiệu năng tổng thể của các mô hình đề xuất.

Các kết quả thực nghiệm được trình bày dưới dạng giá trị trung bình và độ lệch chuẩn qua các lần thực nghiệm nhằm phản ánh hiệu quả trung bình và mức độ ổn định của mô hình. Luận án chưa thực hiện kiểm định giả thuyết thống kê đối với sự khác biệt giữa các phương pháp; do đó, các chênh lệch về giá trị trung bình được sử dụng để mô tả xu hướng thực nghiệm và không được diễn giải là khác biệt có ý nghĩa thống kê.

2.5.1.3. Thiết lập tham số thực nghiệm

Trong các bài thực nghiệm, các tham số của mô hình được thiết lập dựa trên khuyến nghị từ các nghiên cứu FMN trước đây, đồng thời được điều chỉnh thông qua thực nghiệm nhằm đảm bảo hiệu năng ổn định trên các tập dữ liệu benchmark khác nhau. Trừ khi có nêu rõ khác đi, cùng một tập tham số được sử dụng cho tất cả các mô hình để đảm bảo tính công bằng trong so sánh.

Đối với mạng nơron FMN, ngưỡng mở rộng siêu hộp $\delta^{\max}$ (bao gồm ${\delta_1}^{\max}$ và ${\delta_2}^{\max}$) được sử dụng để kiểm soát kích thước tối đa của siêu hộp trong quá trình học. Tham số này được lựa chọn trong một khoảng giá trị xác định trước và được giữ cố định khi so sánh các mô hình, đồng thời được thay đổi có kiểm soát trong các thí nghiệm phân tích ảnh hưởng tham số. Việc xác định giá trị $\delta^{\max}$ được thực hiện trong giai đoạn thiết lập mô hình trước khi đánh giá trên tập kiểm tra. Sau khi xác định giá trị phù hợp, $\delta^{\max}$ được giữ cố định trong quá trình đánh giá trên tập kiểm tra; dữ liệu kiểm tra không được sử dụng để lựa chọn hoặc điều chỉnh tham số này.

Hệ số γ được sử dụng để điều khiển độ dốc của hàm độ thuộc, ảnh hưởng trực tiếp đến mức độ suy giảm độ thuộc khi mẫu nằm xa biên của siêu hộp. Ngưỡng độ thuộc β được sử dụng nhằm loại bỏ các siêu hộp hoặc mẫu dữ liệu có mức ảnh hưởng thấp trong quá trình suy luận nhãn.

Số lượng lân cận tối đa trong FMN-Voting được thiết lập là $k = 2C + 1$, trong đó C là số lớp/cụm. Việc lựa chọn một số lẻ và phụ thuộc vào số lớp nhằm cung cấp một tập lân cận đủ lớn cho cơ chế bỏ phiếu đa số, đồng thời hạn chế sự phụ thuộc của quyết định nhãn vào một số lượng quá ít siêu hộp lân cận. Trong trường hợp số siêu hộp có nhãn khả dụng nhỏ hơn $2C + 1$, mô hình sử dụng số lượng lân cận thực tế có thể lựa chọn theo Công thức (2.17).

Trong các thí nghiệm phân tích ảnh hưởng tham số, từng tham số được thay đổi độc lập trong khi các tham số còn lại được giữ cố định ở giá trị mặc định. Cách thiết lập này cho phép đánh giá rõ ràng mối quan hệ giữa từng tham số học và hiệu năng của mô hình, đồng thời làm rõ mức độ nhạy cảm của các mô hình đề xuất đối với sự thay đổi của các tham số.

Trong đó, một số tham số được lựa chọn trong thực nghiệm bao gồm: tham số mờ $\gamma = 10$, ngưỡng ban đầu $\beta = 0{,}99$ và hệ số giảm ngưỡng $\varphi = 0{,}9$.

Việc chọn $\gamma = 10$ xuất phát từ cân bằng thực nghiệm giữa hai yêu cầu mâu thuẫn:

- Không quá nhỏ:

+ Nếu γ nhỏ, độ thuộc suy giảm chậm;
+ Nhiều mẫu ở xa vẫn có giá trị độ thuộc μ cao;
+ Ranh giới mờ trở nên quá "phẳng", dễ gây nhiễu trong quá trình suy luận nhãn.

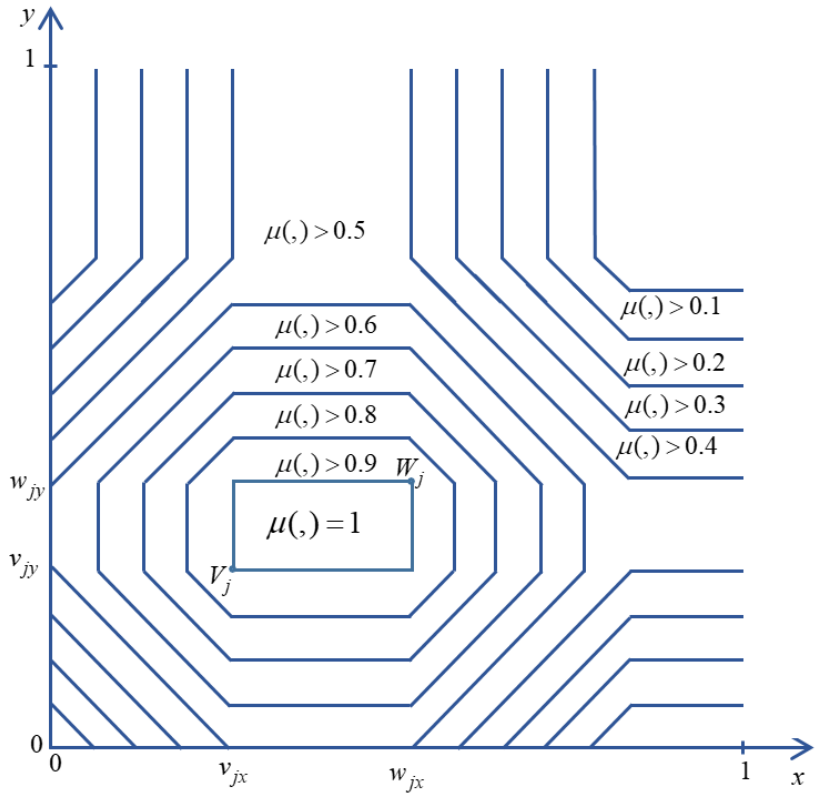
- Không quá lớn:
  + Nếu γ quá lớn, độ thuộc suy giảm rất nhanh;
  + Chỉ các mẫu nằm sát biên siêu hộp mới có $\mu > 0$;
  + Hành vi tiến gần tới phân loại cứng, làm mất ý nghĩa của mô hình mờ.

Điều này được minh họa trên Hình 1.4. Thực nghiệm trong nhiều nghiên cứu về FMN cho thấy $\gamma \approx 10$ tạo ra độ dốc vừa đủ để phân biệt tốt các vùng lân cận, đồng thời vẫn duy trì được tính mờ cần thiết. Do đó, $\gamma = 10$ được lựa chọn như một giá trị mặc định mang tính kinh nghiệm trong các mô hình FMN.

Ngưỡng độ thuộc β được sử dụng nhằm kiểm soát mức độ tin cậy của các siêu hộp hoặc mẫu dữ liệu tham gia vào quá trình suy luận nhãn. Trong các thí nghiệm, giá trị ban đầu của ngưỡng được thiết lập là $\beta = 0{,}99$, phản ánh yêu cầu khắt khe về mức độ phù hợp giữa mẫu truy vấn và cấu trúc đại diện (siêu hộp hoặc mẫu dữ liệu mang nhãn). Trong quá trình học, ngưỡng β được điều chỉnh giảm dần theo hệ số suy giảm $\varphi = 0{,}9$ để tăng dần số lượng phần tử lân cận được xem xét khi cần thiết. Tuy nhiên, giá trị của β không được phép giảm xuống dưới một ngưỡng tối thiểu. Trong luận án này, $\beta_{\min}$ được thiết lập không nhỏ hơn $0{,}8$. Lý do của lựa chọn này xuất phát từ quan sát rằng khi β quá nhỏ, các vùng có độ thuộc thấp vẫn được chấp nhận tham gia suy luận nhãn, dẫn đến việc đưa vào các siêu hộp hoặc mẫu dữ liệu có mức độ đại diện kém. Như minh họa trên Hình 2.6, các vùng tương ứng với giá trị $\mu(x, B_j)$ nhỏ nằm xa lõi của siêu hộp, nơi độ tin cậy của độ thuộc suy giảm rõ rệt, đặc biệt trong bối cảnh học không giám sát hoặc dữ liệu có nhiễu. Do đó, việc duy trì $\beta \geq 0{,}8$ giúp đảm bảo rằng chỉ các phần tử lân cận có mức độ phù hợp đủ cao mới được sử dụng trong quá trình lan truyền nhãn, từ đó nâng cao độ ổn định và độ tin cậy của kết quả học.

2.5.1.4. Tập dữ liệu thực nghiệm

Để đánh giá hiệu năng của thuật toán đề xuất, nghiên cứu sinh tiến hành thực nghiệm trên các tập dữ liệu chuẩn (*benchmark*) được lấy từ kho dữ liệu học máy UCI [22] và CS [23]. Bên cạnh đó, nghiên cứu còn sử dụng tập dữ liệu *ILPD*, gồm các chỉ số lâm sàng và xét nghiệm sinh hóa máu nhằm phục vụ bài toán phân biệt bệnh nhân mắc bệnh gan và không mắc bệnh gan. Tập dữ liệu này phản ánh đặc trưng dữ liệu y sinh thực tế với tính chất nhiễu, mất cân bằng lớp và tương quan đặc trưng



Hình 2.6: *Minh họa mức độ thuộc $\mu(x, B_j)$ trong không gian đặc trưng và ảnh hưởng đến độ tin cậy suy luận*

phức tạp, qua đó cho phép đánh giá khả năng tổng quát hóa của mô hình trong bối cảnh ứng dụng y tế.

Các tập dữ liệu từ CS bao gồm: *Aggregation*, *Flame*, *Spiral*, *Jain*, *R15*, *Path-based*, *Thyroidnew*.

Các tập dữ liệu từ UCI bao gồm: *Iris*, *Wine*, *PID* (Pima Indians Diabetes), *Thyroid*, *Sonar, ILPD*. Thông tin chi tiết của các tập dữ liệu được trình bày trong Bảng 2.1. Đồ họa trên Hình 2.7 minh họa phân bố không gian dữ liệu trên các tập dữ liệu thực nghiệm.

Bộ dữ liệu bệnh nhân gan Ấn Độ (*Indian Liver Patient Dataset* - ILPD) là một bộ dữ liệu chuẩn được công bố công khai trên Kho lưu trữ Học máy UCI (*UCI Machine Learning Repository*) và được phát hành theo giấy phép CC BY 4.0. Dữ liệu được thu thập từ các bệnh nhân tại khu vực Đông Bắc bang Andhra Pradesh, Ấn Độ, nhằm phục vụ việc xây dựng, huấn luyện và đánh giá các phương pháp học máy hỗ trợ chẩn đoán bệnh gan.

ILPD gồm tổng cộng 583 hồ sơ bệnh nhân, trong đó có 416 trường hợp mắc bệnh gan và 167 trường hợp không mắc bệnh gan. Mỗi mẫu dữ liệu được mô tả bởi 10 thuộc tính lâm sàng và sinh hóa, bao gồm:

- **Tuổi** (*Age*): Tuổi của bệnh nhân tại thời điểm thu thập dữ liệu, được biểu diễn

Bảng 2.1: *Thông tin các tập dữ liệu thực nghiệm benchmark*

TT	Tập dữ liệu	Số mẫu	Số thuộc tính	Số nhóm
1	Aggregation	788	2	7
2	Flame	240	2	2
3	Iris	150	4	3
4	Jain	373	2	2
5	R15	600	2	15
6	Spiral	312	2	3
7	Sonar	208	60	2
8	Pathbased	317	2	3
9	PID	768	8	2
10	Thyroid	7200	21	3
11	Thyroidnew	215	5	3
12	Wine	178	13	3
13	ILPD	583	10	2

bằng số năm. Đây là thuộc tính nhân khẩu học có thể liên quan đến sự thay đổi chức năng gan và nguy cơ mắc bệnh theo độ tuổi. Trong bộ dữ liệu ILPD, những bệnh nhân trên 89 tuổi được quy ước ghi nhận là 90 tuổi.

- **Giới tính** (*Gender*): Thuộc tính nhân khẩu học cho biết giới tính của bệnh nhân, gồm hai giá trị nam và nữ. Thuộc tính này được sử dụng để xem xét sự khác biệt về đặc điểm sinh học, chỉ số xét nghiệm và khả năng mắc bệnh gan giữa các nhóm giới tính.
- **Bilirubin toàn phần** (*Total Bilirubin* - TB): Chỉ số biểu thị tổng lượng bilirubin trong máu, bao gồm bilirubin trực tiếp và bilirubin gián tiếp. Bilirubin là sản phẩm được tạo ra trong quá trình phân hủy hồng cầu và được gan xử lý trước khi đào thải qua đường mật. Giá trị bilirubin toàn phần bất thường có thể liên quan đến rối loạn chức năng gan, tắc nghẽn đường mật hoặc sự gia tăng phân hủy hồng cầu.
- **Bilirubin trực tiếp** (*Direct Bilirubin* - DB): Phần bilirubin đã được gan liên hợp để có thể hòa tan trong nước và bài tiết vào dịch mật. Chỉ số này thường được xem xét cùng bilirubin toàn phần nhằm hỗ trợ đánh giá khả năng chuyển hóa và bài tiết bilirubin của gan. Sự gia tăng bilirubin trực tiếp có thể liên quan đến tổn thương tế bào gan hoặc cản trở lưu thông mật.
- **Phosphatase kiềm** (*Alkaline Phosphatase* - Alkphos): Enzyme hiện diện chủ yếu ở gan, đường mật và xương. Trong đánh giá chức năng gan, chỉ số này thường được sử dụng để hỗ trợ phát hiện các bất thường liên quan đến đường mật hoặc tình trạng ứ mật. Tuy nhiên, do phosphatase kiềm cũng có nguồn gốc

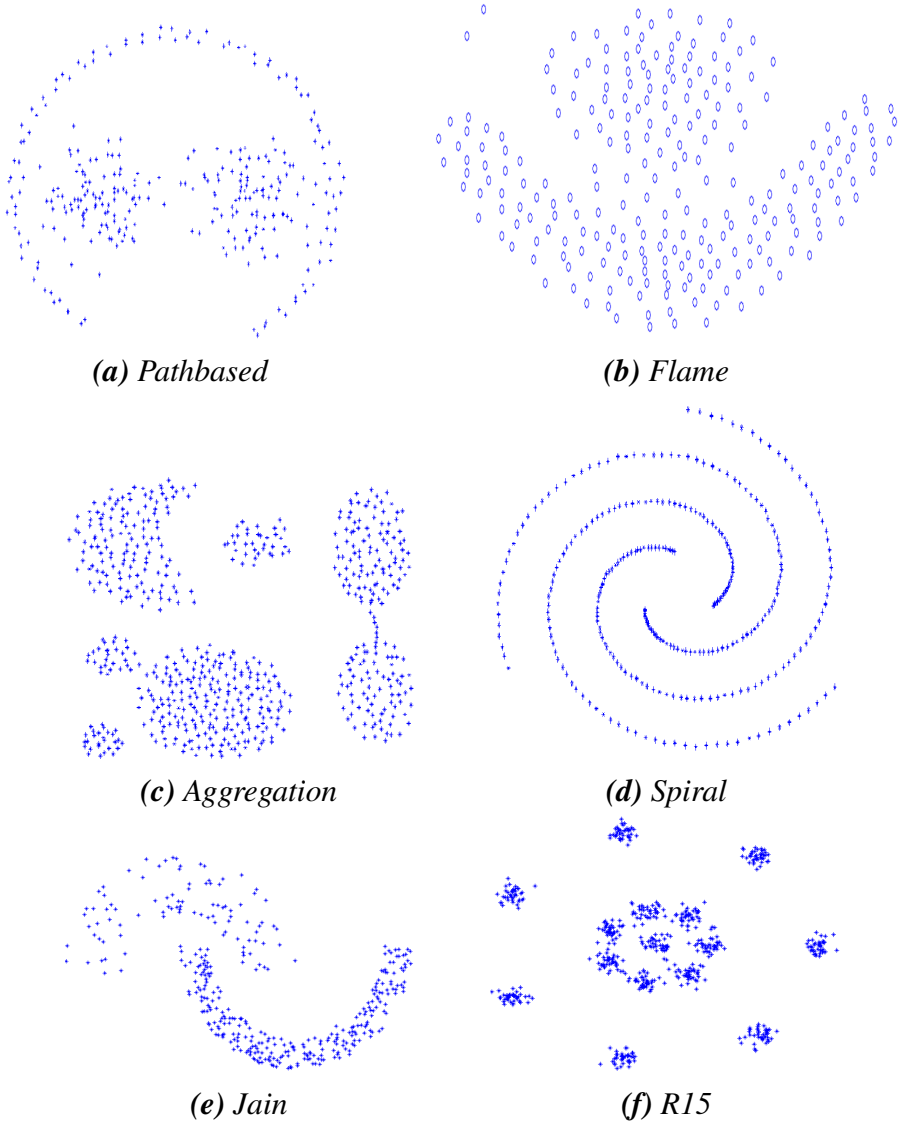
từ xương và một số mô khác, kết quả cần được xem xét kết hợp với các chỉ số xét nghiệm liên quan.

- **Alanine aminotransferase** (*Alanine Aminotransferase* - ALT/SGPT): Enzyme tập trung chủ yếu trong tế bào gan, còn được gọi theo tên cũ là serum glutamic-pyruvic transaminase (SGPT). Khi tế bào gan bị tổn thương, enzyme này có thể được giải phóng vào máu, làm tăng nồng độ ALT. Vì có mức độ đặc hiệu đối với gan tương đối cao, ALT thường được sử dụng để hỗ trợ đánh giá tổn thương tế bào gan.
- **Aspartate aminotransferase** (*Aspartate Aminotransferase* - AST/SGOT): Enzyme có trong gan, tim, cơ, thận và một số mô khác, còn được gọi theo tên cũ là serum glutamic-oxaloacetic transaminase (SGOT). Nồng độ AST có thể tăng khi các tế bào tại những cơ quan này bị tổn thương. Do AST không đặc hiệu hoàn toàn cho gan, chỉ số này thường được phân tích kết hợp với ALT và các xét nghiệm chức năng gan khác.
- **Protein toàn phần** (*Total Proteins* - TP): Chỉ số phản ánh tổng lượng protein trong huyết thanh, chủ yếu bao gồm albumin và globulin. Protein toàn phần cung cấp thông tin liên quan đến khả năng tổng hợp protein của gan, tình trạng dinh dưỡng, chức năng thận và hoạt động của hệ miễn dịch. Giá trị bất thường có thể xuất hiện trong các bệnh lý gan, thận hoặc tình trạng thiếu hụt dinh dưỡng.
- **Albumin** (*Albumin* - ALB): Protein chiếm tỷ lệ lớn trong huyết tương và được tổng hợp chủ yếu tại gan. Albumin có vai trò duy trì áp lực keo của máu, vận chuyển nhiều chất trong cơ thể và phản ánh một phần khả năng tổng hợp protein của gan. Nồng độ albumin thấp có thể liên quan đến suy giảm chức năng tổng hợp của gan, bệnh thận, tình trạng viêm hoặc suy dinh dưỡng.
- **Tỷ lệ albumin/globulin** (*Albumin/Globulin Ratio* - A/G Ratio): Tỷ số giữa nồng độ albumin và globulin trong huyết thanh. Chỉ số này thể hiện sự cân bằng giữa hai nhóm protein chính trong máu và thường được phân tích cùng protein toàn phần và albumin. Tỷ lệ A/G bất thường có thể liên quan đến những thay đổi trong chức năng gan, chức năng thận, tình trạng dinh dưỡng hoặc hoạt động của hệ miễn dịch.

Đây là những chỉ số thường được sử dụng để phản ánh chức năng gan và hỗ trợ đánh giá tình trạng tổn thương gan.

Thuộc tính mục tiêu của bộ dữ liệu "Selector" cho biết mỗi đối tượng thuộc

nhóm mắc bệnh gan hay nhóm không mắc bệnh gan. Do có cấu trúc dữ liệu rõ ràng, số lượng mẫu phù hợp và bao gồm nhiều chỉ số xét nghiệm liên quan đến chức năng gan, ILPD đã được sử dụng rộng rãi như một bộ dữ liệu chuẩn trong các nghiên cứu về dự đoán bệnh gan. Bộ dữ liệu này thường được dùng để đánh giá và so sánh hiệu quả của các phương pháp học máy, khai phá dữ liệu, phân loại và học bán giám sát.



***(a)** Pathbased* ***(b)** Flame*

***(c)** Aggregation* ***(d)** Spiral*

***(e)** Jain* ***(f)** R15*

***Hình 2.7:** Đồ họa phân bố không gian dữ liệu thực tế của các tập dữ liệu thực nghiệm*

Phạm vi thực nghiệm của luận án tập trung vào dữ liệu dạng bảng có quy mô nhỏ và vừa; tập dữ liệu lớn nhất được sử dụng là *Thyroid* với 7200 mẫu (Bảng 2.1). Các mô hình đề xuất được đối sánh với các biến thể FMN bán giám sát (SCFMN, MSCFMN), các thuật toán phân cụm phổ biến (K-means++, EM, DBSCAN) và một số phương pháp phân cụm bán giám sát (FC2, SSGC, SSDBSCAN, SSK-Means). Luận án chưa thực nghiệm trên các tập dữ liệu có quy mô lớn hơn, dữ liệu phi cấu trúc (ảnh, văn bản) và chưa đối sánh với các phương pháp phân cụm bán giám sát dựa trên học sâu. Vì vậy, các nhận định về hiệu năng và chi phí tính toán trong luận án được

giới hạn trong phạm vi các tập dữ liệu nêu trên; khả năng mở rộng trên dữ liệu lớn mới được phân tích ở mức độ phức tạp thuật toán và là hướng nghiên cứu tiếp theo. Trong mỗi fold, các tham số tiền xử lý và chuẩn hóa được xác định chỉ từ tập huấn luyện, sau đó được áp dụng không thay đổi cho tập kiểm tra tương ứng. Quy trình này bảo đảm tập kiểm tra không tham gia vào việc xác định các tham số tiền xử lý. Với các tập dữ liệu bị thiếu thông tin (*missing values*), các giá trị thiếu được xử lý tương tự như trong nghiên cứu của Batista [24]. Đối với mỗi fold của quá trình kiểm tra chéo, các giá trị thiếu được xử lý dựa trên tập huấn luyện của fold tương ứng. Cụ thể, với mỗi thuộc tính có giá trị thiếu, giá trị trung bình được tính chỉ trên các mẫu huấn luyện có giá trị quan sát của thuộc tính đó; giá trị trung bình này sau đó được sử dụng để thay thế các giá trị thiếu trong cả tập huấn luyện và tập kiểm tra của fold. Việc xử lý được thực hiện độc lập trên từng fold nhằm tránh sử dụng thông tin từ tập kiểm tra trong quá trình tiền xử lý. Cụ thể, giá trị trung bình của thuộc tính thứ i tại fold f được xác định theo Công thức (2.30):

$$\bar{x}_i^{(f)} = \frac{1}{\left|\mathcal{I}_{i,\text{obs}}^{(f)}\right|} \sum_{j \in \mathcal{I}_{i,\text{obs}}^{(f)}} x_{ji}, \tag{2.30}$$

trong đó $\mathcal{I}_{i,\text{obs}}^{(f)}$ là tập chỉ số các mẫu thuộc tập huấn luyện của fold f có giá trị quan sát tại thuộc tính thứ i.

Đối với mẫu x_r có giá trị bị thiếu tại thuộc tính thứ i, giá trị thiếu được thay thế bởi:

$$x_{ri} \leftarrow \bar{x}_i^{(f)}. \tag{2.31}$$

2.5.1.5. Môi trường thực nghiệm

Các thí nghiệm được triển khai trong môi trường MATLAB R2014a phiên bản 64-bit. Cấu hình phần cứng sử dụng bộ xử lý Intel Core i5-6300HQ với bốn lõi xử lý vật lý, không hỗ trợ công nghệ Hyper-Threading; bộ nhớ RAM 16GB. Toàn bộ quá trình thực nghiệm được thực hiện trên nền tảng CPU, không sử dụng tăng tốc phần cứng GPU.

2.5.1.6. Độ đo và tiêu chí đánh giá kết quả

Để đánh giá hiệu năng của các thuật toán và so sánh chất lượng kết quả phân cụm giữa các phương pháp khác nhau, nghiên cứu sử dụng tập hợp các chỉ số đánh giá phổ biến, bao gồm các chỉ số dựa trên nhãn thực và các chỉ số phản ánh mức độ bảo toàn cấu trúc dữ liệu.

Các chỉ số được sử dụng gồm Accuracy, Precision, Recall, F1-Score, Specificity, RI (Rand Index), ARI (Adjusted Rand Index) và NMI (Normalized Mutual Information).

Các mô hình đề xuất trong luận án thực hiện bài toán phân cụm bán giám sát, do đó đầu ra trực tiếp của mô hình là phân hoạch dữ liệu thành các cụm. Để sử dụng các độ đo dựa trên nhãn lớp như Accuracy, Precision, Recall, F1-Score và Specificity, nhãn cụm thu được cần được ánh xạ sang nhãn lớp thực tương ứng. Việc ánh xạ này chỉ được thực hiện trong giai đoạn đánh giá và không tham gia vào quá trình huấn luyện hay hình thành cụm của mô hình.

Cụ thể, với mỗi cụm G_j thu được, nhãn lớp tương ứng được xác định là lớp xuất hiện nhiều nhất trong cụm theo nhãn thực của tập đánh giá:

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

trong đó $\mathcal{C} = \{1, \ldots, C\}$ là tập các lớp, y_i là nhãn lớp thực của mẫu x_i và $\mathbb{I}(\cdot)$ là hàm chỉ thị. Sau khi xác định ánh xạ π, nhãn dự đoán của mỗi mẫu $x_i \in G_j$ được xác định bởi $\hat{y}_i = \pi(G_j)$. Các độ đo dựa trên nhãn lớp được tính bằng cách so sánh y_i và $\hat{y}_i$.

Độ đo $Accuracy$ phản ánh mức độ trùng khớp giữa nhãn dự đoán và nhãn thực của các mẫu dữ liệu. Giá trị Accuracy càng lớn thì kết quả phân cụm hoặc phân lớp càng chính xác. Giả sử x_i là mẫu dữ liệu, y_i là nhãn thực của x_i và $\bar{y}_i$ là nhãn dự đoán của x_i, Accuracy được xác định theo Công thức (2.33):

$$Accuracy = \frac{1}{n} \sum_{i=1}^{n} H(y_i = \bar{y}_i), \tag{2.33}$$

trong đó $H(\cdot)$ là hàm chỉ thị, $H(y_i = \bar{y}_i) = 1$ nếu hai nhãn trùng nhau và bằng 0 trong trường hợp ngược lại, n là tổng số mẫu trong tập dữ liệu kiểm tra.

Chỉ số RI đo lường mức độ nhất quán cặp đôi giữa phân hoạch thu được và phân hoạch thực. RI đánh giá xem các cặp mẫu có được gán cùng cụm hoặc khác cụm một cách nhất quán hay không. Giá trị RI càng gần 1 thì chất lượng phân cụm càng tốt. RI được xác định theo công thức:

$$RI = \frac{a + b}{a + b + c + d}, \tag{2.34}$$

trong đó a là số cặp mẫu được gom cùng cụm trong cả hai phân hoạch, b là số cặp mẫu được gom khác cụm trong cả hai phân hoạch, c và d lần lượt là số cặp mẫu không nhất quán giữa hai phân hoạch.

Chỉ số ARI là biến thể hiệu chỉnh của RI, được đề xuất nhằm loại bỏ ảnh hưởng của sự trùng khớp ngẫu nhiên giữa các phân hoạch. ARI có giá trị trong khoảng

$[-1, 1]$, trong đó giá trị càng gần 1 thể hiện mức độ tương đồng càng cao giữa kết quả phân cụm và nhãn thực; giá trị gần 0 tương ứng với mức trùng khớp ngẫu nhiên, trong khi giá trị âm cho thấy mức độ tương đồng thấp hơn ngẫu nhiên. ARI được sử dụng rộng rãi trong các bài toán đánh giá phân cụm trên các tập dữ liệu benchmark.

Giả sử $U = \{U_1, \ldots, U_r\}$ là phân hoạch thu được từ thuật toán phân cụm và $V = \{V_1, \ldots, V_s\}$ là phân hoạch chuẩn (nhãn thực). Ký hiệu $n_{ij} = |U_i \cap V_j|$ là số mẫu chung giữa cụm U_i và lớp V_j, $a_i = \sum_j n_{ij}$ và $b_j = \sum_i n_{ij}$. Khi đó, chỉ số ARI được xác định theo công thức:

$$ARI = \frac{\displaystyle\sum_{i,j} \binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\displaystyle\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}, \tag{2.35}$$

trong đó n là tổng số mẫu dữ liệu và $\binom{\cdot}{2}$ là toán tử tổ hợp.

NMI được sử dụng để đánh giá mức độ tương đồng giữa kết quả phân cụm và nhãn lớp thực tế. NMI được xác định như sau:

$$\mathrm{NMI}(U, V) = \frac{2I(U; V)}{H(U) + H(V)}, \tag{2.36}$$

trong đó $I(U; V)$ là lượng thông tin tương hỗ giữa hai phân hoạch:

$$I(U; V) = \sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} \log\left(\frac{p_{ij}}{p_i p_j}\right), \tag{2.37}$$

với

$$p_{ij} = \frac{n_{ij}}{n}, \qquad p_i = \frac{n_{i\cdot}}{n}, \qquad p_j = \frac{n_{\cdot j}}{n}. \tag{2.38}$$

Entropy của hai phân hoạch được xác định bởi:

$$H(U) = -\sum_{i=1}^{r} p_i \log p_i, \tag{2.39}$$

và

$$H(V) = -\sum_{j=1}^{s} p_j \log p_j. \tag{2.40}$$

Trong đó, n_{ij} là số mẫu đồng thời thuộc lớp thực tế U_i và cụm V_j; $n_{i\cdot}$ là tổng số mẫu thuộc lớp U_i; và $n_{\cdot j}$ là tổng số mẫu thuộc cụm V_j. Giá trị NMI nằm trong khoảng $[0, 1]$; giá trị càng lớn thể hiện mức độ tương đồng càng cao giữa kết quả phân cụm và nhãn lớp thực tế.

Việc kết hợp các độ đo đánh giá dựa trên nhãn lớp sau ánh xạ (*Accuracy, Precision, Recall, F1-Score, Specificity*) với các độ đo so sánh phân hoạch (*RI, ARI và NMI*) cho phép đánh giá kết quả phân cụm dưới các góc độ bổ sung nhau.

Với tập dữ liệu bệnh gan, kết quả phân cụm được đánh giá dựa trên nhãn lớp

thông qua các chỉ số *Accuracy, Precision, Recall, F1-Score, Specificity, RI, ARI,* và *NMI*. Chỉ số Accuracy được sử dụng trong cả hai nhóm tiêu chí nhằm đảm bảo tính nhất quán và khả năng so sánh giữa các phương pháp.

Việc lựa chọn các chỉ số trên trong đánh giá hiệu năng là cần thiết nhằm phản ánh đầy đủ các khía cạnh khác nhau của mô hình trong bối cảnh dữ liệu thực tế. Cụ thể, mặc dù *Accuracy* cung cấp cái nhìn tổng quát về tỷ lệ dự đoán đúng, chỉ số này có thể bị ảnh hưởng đáng kể khi dữ liệu mất cân bằng, vốn là đặc trưng phổ biến trong các bài toán y sinh như chẩn đoán bệnh gan.

Do đó, *Precision* được sử dụng để đánh giá mức độ chính xác của các dự đoán dương, giúp hạn chế các trường hợp dương giả (*FP*), trong khi *Recall* phản ánh khả năng phát hiện đầy đủ các mẫu dương, đặc biệt quan trọng nhằm giảm thiểu các trường hợp âm giả (*FN*) có thể gây bỏ sót bệnh. *F1-Score* đóng vai trò là thước đo tổng hợp, cân bằng giữa *Precision* và *Recall*, qua đó cung cấp đánh giá ổn định hơn trong các tình huống dữ liệu không cân bằng.

Bên cạnh đó, nhằm đánh giá khả năng nhận diện chính xác các mẫu âm, chỉ số *Specificity* được đưa vào để đo lường tỷ lệ dự đoán đúng các trường hợp không thuộc lớp quan tâm, góp phần giảm thiểu các cảnh báo sai.

Như vậy, việc kết hợp các chỉ số trên không chỉ cho phép đánh giá toàn diện hiệu năng của mô hình trên cả hai lớp dữ liệu, mà còn đặc biệt phù hợp với các bài toán phân cụm bán giám sát trong lĩnh vực y sinh, nơi yêu cầu đồng thời khả năng phát hiện chính xác đối tượng quan tâm và hạn chế tối đa các sai lệch trong dự đoán.

Sau khi thực hiện ánh xạ từ nhãn cụm sang nhãn lớp như mô tả ở trên, các giá trị TP, TN, FP và FN được xác định trên nhãn lớp thực y_i và nhãn dự đoán $\hat{y}_i$. Các chỉ số đánh giá được tính toán theo các Công thức từ (2.41) đến (2.46); trong bài toán hai lớp, Công thức (2.41) là dạng tương đương của Công thức (2.33):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.41}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.42}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.43}$$

$$F1\text{-}Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.44}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.45}$$

$$FP\ Rate = \frac{FP}{FP + TN} = 1 - Specificity \tag{2.46}$$

Các giá trị dự đoán TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative) được định nghĩa như sau:

- TP (True Positive – Dương thật): Dự đoán đúng là dương
- TN (True Negative – Âm thật): Dự đoán đúng là âm
- FP (False Positive – Dương giả): Dự đoán sai là dương
- FN (False Negative – Âm giả): Dự đoán sai là âm

2.5.2. Kết quả thực nghiệm và đánh giá

Các kết quả thực nghiệm thể hiện trong các hình từ Hình 2.8 đến Hình 2.17 và trên Bảng 2.2 và Bảng 2.3 cho thấy tham số $\delta^{\max}$ có ảnh hưởng rõ rệt đến chất lượng phân cụm của phương pháp FMN-Voting trên các tập dữ liệu benchmark. Nhìn chung, khi $\delta^{\max}$ được thiết lập ở mức nhỏ hoặc trung bình thấp, mô hình thường đạt giá trị *Accuracy* cao hơn và ổn định hơn. Ngược lại, khi $\delta^{\max}$ tăng lên, độ chính xác có xu hướng suy giảm, đồng thời độ dao động kết quả giữa các lần thực nghiệm cũng lớn hơn, thể hiện qua các thanh sai số mở rộng hơn ở vùng giá trị lớn của tham số này. Xu hướng đó cho thấy việc cho phép siêu hộp mở rộng quá mức dễ làm mờ ranh giới cụm, đặc biệt đối với các tập dữ liệu có phân bố không đều, cụm kéo dài, cụm phi lồi hoặc tồn tại nhiễu và vùng chồng lấn cục bộ. Khi đó, cơ chế bỏ phiếu của FMN-Voting khó duy trì được tính phân biệt giữa các cụm, dẫn đến suy giảm hiệu năng.

Đối với các trường hợp được ký hiệu "-" trong Bảng 2.2, FMN-Voting không cho kết quả đủ ổn định để sử dụng làm giá trị đại diện trong bảng so sánh. Qua thực nghiệm, hiện tượng này chủ yếu được quan sát trên các tập dữ liệu có cấu trúc cụm không hình cầu, trong đó kết quả của FMN-Voting phụ thuộc đáng kể vào thứ tự các mẫu trong tập huấn luyện và Accuracy thu được có thể thấp hơn 0,5. Do đó, luận án không báo cáo một giá trị Accuracy đại diện cho các trường hợp này.

Trên các tập dữ liệu có kết quả của cả hai mô hình, FMN-SAL đạt *Accuracy* cao hơn FMN-Voting, với mức chênh lệch khác nhau tùy theo tập dữ liệu. Kết quả này cho thấy việc khai thác trực tiếp các mẫu mang nhãn lân cận có thể mang lại lợi ích khi cấu trúc siêu hộp chưa phản ánh đầy đủ phân bố cục bộ của dữ liệu. Tuy nhiên, kết quả thực nghiệm này không hàm ý FMN-SAL luôn ưu thế trong mọi trường hợp, do hiệu quả của mô hình còn phụ thuộc vào số lượng, chất lượng và phân bố của các

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

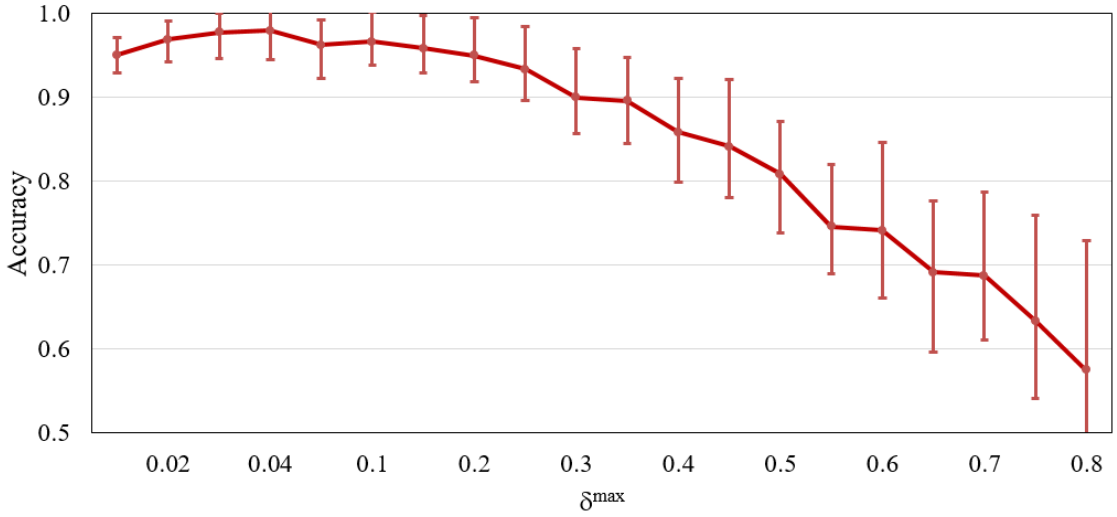
-: không báo cáo giá trị đại diện do kết quả FMN-Voting biến động mạnh theo thứ tự mẫu huấn luyện và Accuracy thấp hơn 0,5 trong các lần thực nghiệm tương ứng.

mẫu mang nhãn.

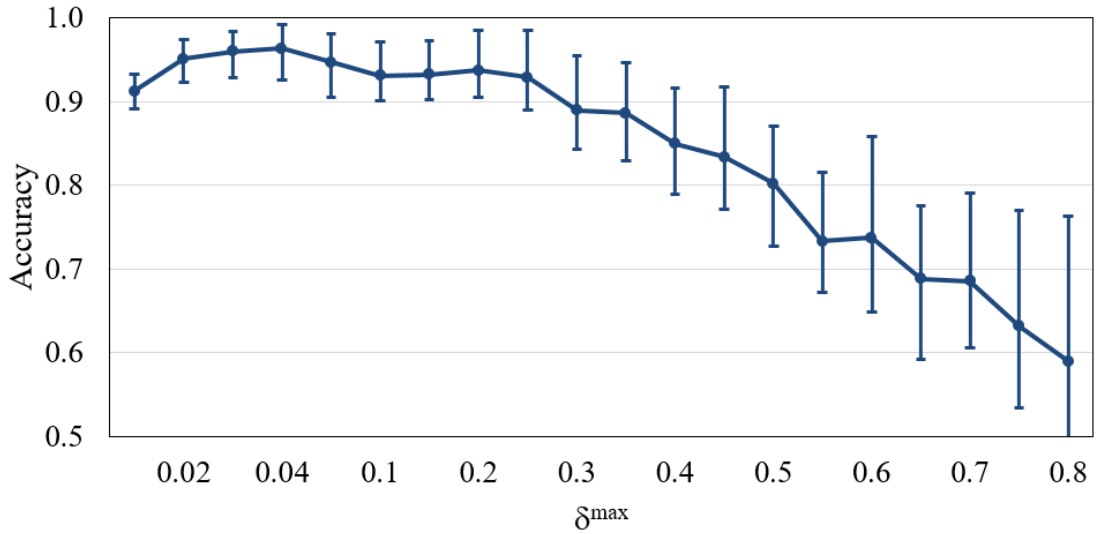
Trên tập dữ liệu *Flame* (Hình 2.8), cả hai mô hình FMN-SAL và FMN-Voting đều đạt *Accuracy* cao khi tham số $\delta^{\max}$ nhỏ, đặc biệt trong khoảng từ 0,02 đến 0.2. Điều này cho thấy khi kích thước siêu hộp được giới hạn chặt chẽ, mô hình có khả năng biểu diễn tốt cấu trúc cục bộ của dữ liệu và duy trì ranh giới phân cụm tương đối ổn định. Tuy nhiên, khi $\delta^{\max}$ tăng, hiệu năng của cả hai mô hình đều suy giảm rõ rệt. Nguyên nhân là do các siêu hộp được phép mở rộng lớn hơn, dẫn đến nguy cơ bao phủ đồng thời nhiều vùng dữ liệu thuộc các cụm khác nhau, đặc biệt trong trường hợp *Flame* có cấu trúc phi tuyến và vùng tiếp giáp giữa các cụm tương đối phức tạp.

Có thể quan sát rằng FMN-SAL duy trì độ chính xác cao hơn trong giai đoạn đầu khi $\delta^{\max}$ còn nhỏ, nhưng sau đó cũng giảm mạnh khi tham số này vượt quá ngưỡng phù hợp. Xu hướng tương tự xuất hiện ở FMN-Voting, tuy nhiên độ dao động có phần lớn hơn ở các giá trị $\delta^{\max}$ cao, thể hiện qua khoảng sai số mở rộng. Kết quả này cho thấy tập *Flame* khá nhạy với cơ chế mở rộng siêu hộp: các siêu hộp nhỏ giúp mô hình nắm bắt tốt hơn cấu trúc cục bộ, trong khi các siêu hộp quá lớn dễ làm mờ ranh giới cụm, tăng ảnh hưởng của nhiễu và làm giảm chất lượng suy luận nhãn. Do đó, việc lựa chọn $\delta^{\max}$ phù hợp có vai trò quan trọng đối với hiệu năng phân cụm của các mô hình FMN trên tập dữ liệu này.

Trên tập dữ liệu *Iris* (Hình 2.9), cả hai mô hình FMN-SAL và FMN-Voting đều đạt *Accuracy* cao khi $\delta^{\max}$ nằm trong vùng giá trị nhỏ. Khi kích thước siêu hộp nhỏ, mô hình có khả năng biểu diễn tốt các vùng dữ liệu cục bộ, từ đó hạn chế hiện tượng bao phủ sai sang các lớp lân cận. Tuy nhiên, khi $\delta^{\max}$ tăng, độ chính xác của cả



(a) *FMN-SAL*



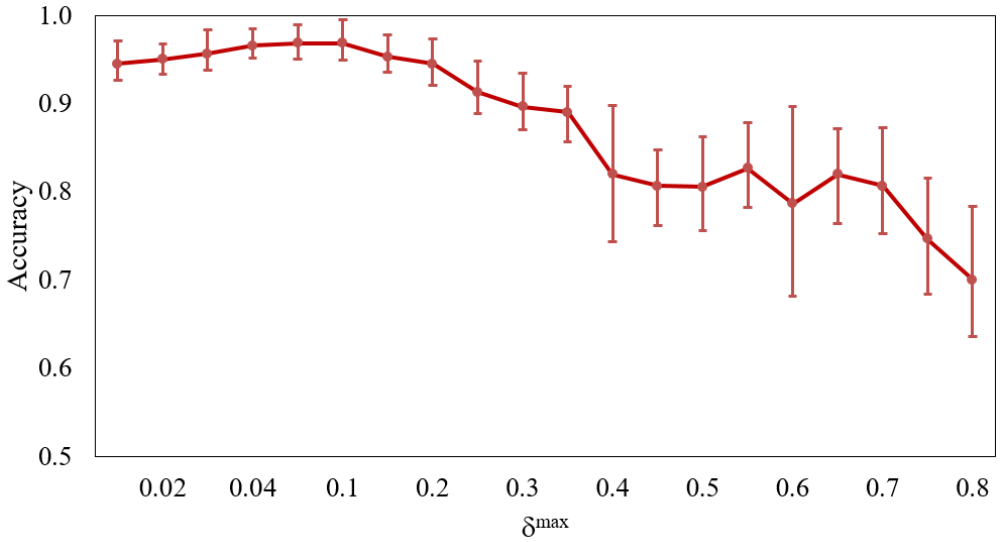
(b) *FMN-Voting*

Hình 2.8: *Hiệu năng phân cụm theo chỉ số Accuracy trên tập dữ liệu Flame*

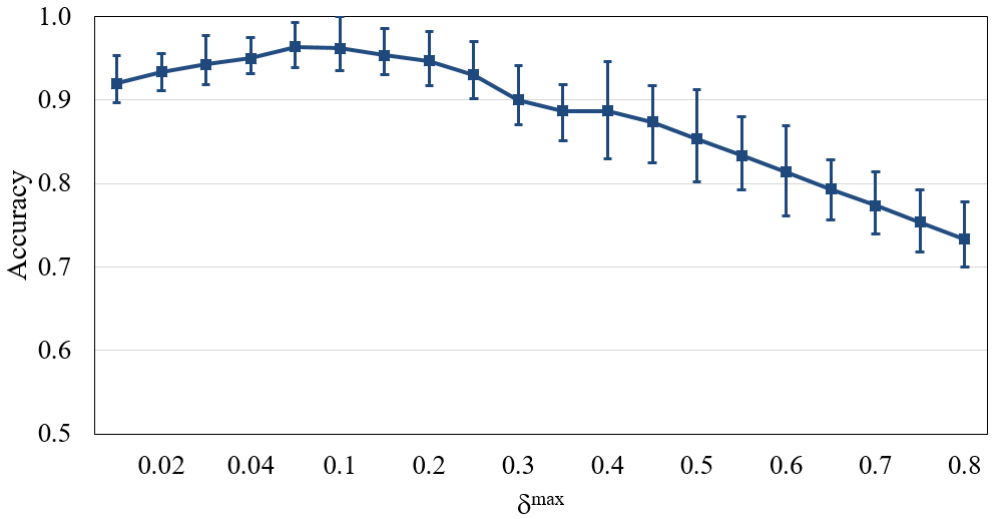
hai mô hình có xu hướng giảm dần. Nguyên nhân là do các siêu hộp được phép mở rộng lớn hơn, làm tăng khả năng chồng lấn giữa các vùng dữ liệu thuộc các lớp khác nhau. Điều này đặc biệt ảnh hưởng đến tập *Iris*, do một số lớp có đặc trưng gần nhau và không hoàn toàn tách biệt trong không gian thuộc tính. Kết quả trên cho thấy việc lựa chọn $\delta^{\max}$ phù hợp có vai trò quan trọng trong việc duy trì ranh giới giữa các lớp, đồng thời bảo đảm chất lượng suy luận nhãn của các mô hình FMN.

Trên tập *Wine* (Hình 2.10), FMN-Voting duy trì *Accuracy* cao trong một khoảng $\delta^{\max}$ rộng hơn. Kết quả này cho thấy cấu trúc phân bố của dữ liệu tương đối thuận lợi cho biểu diễn bằng siêu hộp và ít bị ảnh hưởng bởi nhiễu hoặc chồng lấn ở vùng tham số trung bình.

Đối với tập *R15* (Hình 2.11), hai mô hình đạt *Accuracy* tốt khi $\delta^{\max}$ nhỏ. Tuy nhiên, khi tham số tăng lên, hiệu năng giảm nhanh, cho thấy mặc dù các cụm tách biệt



***(a)** FMN-SAL*

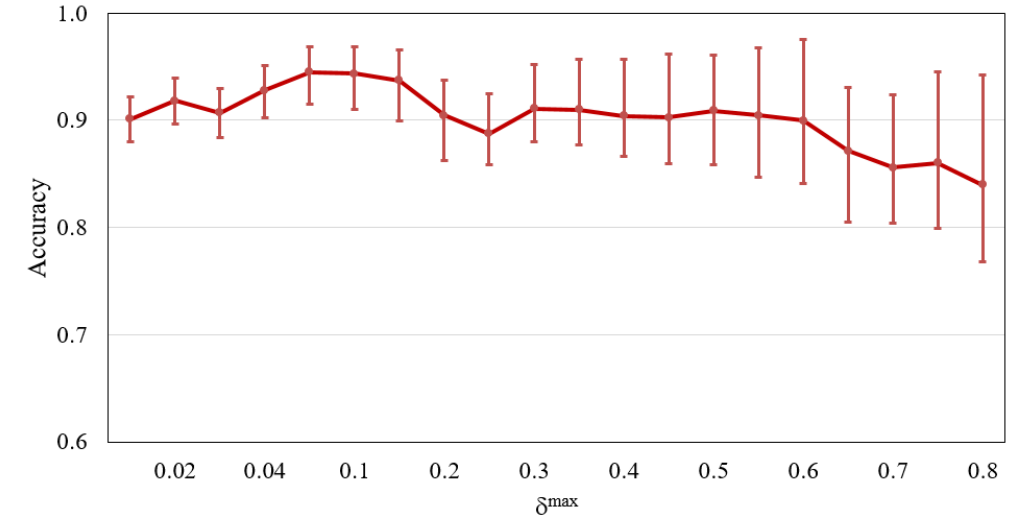


***(b)** FMN-Voting*

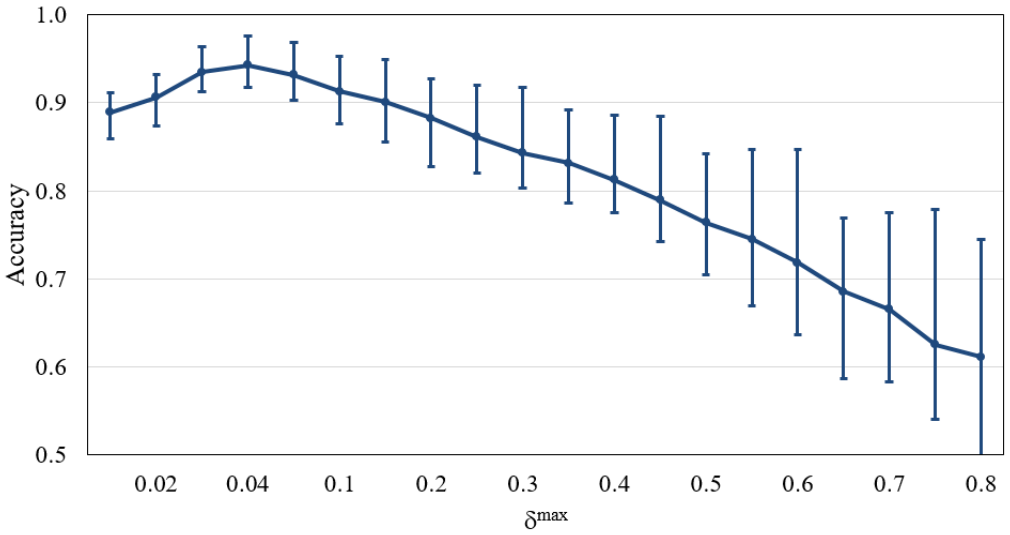
***Hình 2.9:** Hiệu năng phân cụm theo chỉ số Accuracy trên tập dữ liệu Iris*

tốt, mô hình vẫn nhạy với việc mở rộng siêu hộp quá mức, đặc biệt ở các vùng thưa hoặc gần biên cụm.

Trên tập *Aggregation* (Hình 2.12), FMN-SAL cho thấy hiệu năng tốt ở vùng $\delta^{\max}$ nhỏ và trung bình thấp. Điều này cho thấy việc giữ siêu hộp đủ nhỏ giúp mô hình bám sát cấu trúc cục bộ của dữ liệu và hạn chế nhầm lẫn giữa các cụm có hình dạng phức tạp. Khi $\delta^{\max}$ tăng, độ chính xác có xu hướng dao động và suy giảm, đặc biệt ở các giá trị lớn. Nguyên nhân là do các siêu hộp được mở rộng quá mức có thể bao phủ sang vùng dữ liệu lân cận, làm mờ ranh giới giữa các cụm. Với đặc trưng gồm nhiều cụm có mật độ và hình dạng khác nhau, tập *Aggregation* yêu cầu mô hình phải kiểm soát tốt kích thước siêu hộp. Kết quả này cho thấy $\delta^{\max}$ là tham số quan trọng ảnh hưởng trực tiếp đến chất lượng phân cụm của FMN-SAL trên dữ liệu có cấu trúc không đồng nhất.



***(a)** FMN-SAL*



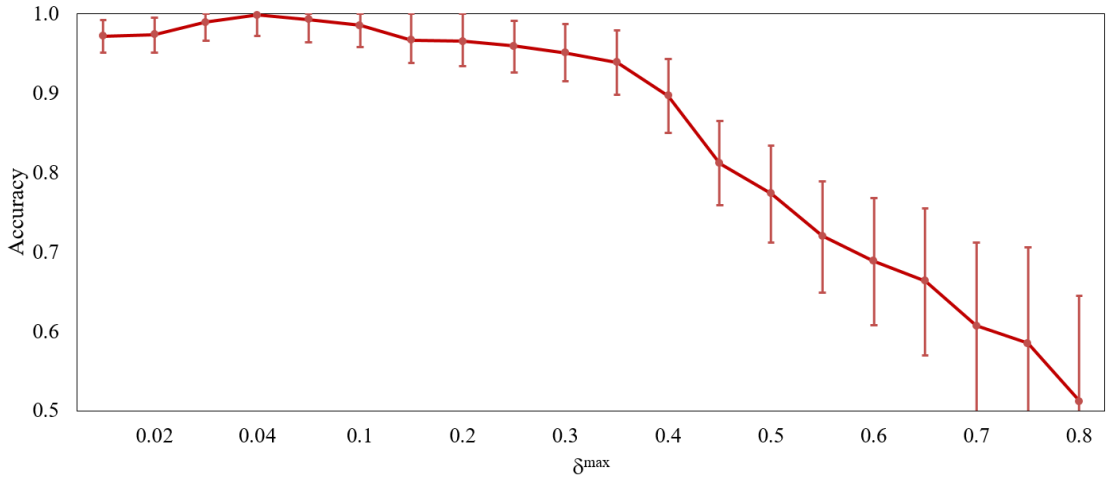
***(b)** FMN-Voting*

***Hình 2.10:** Hiệu năng phân cụm theo chỉ số Accuracy trên tập dữ liệu Wine*

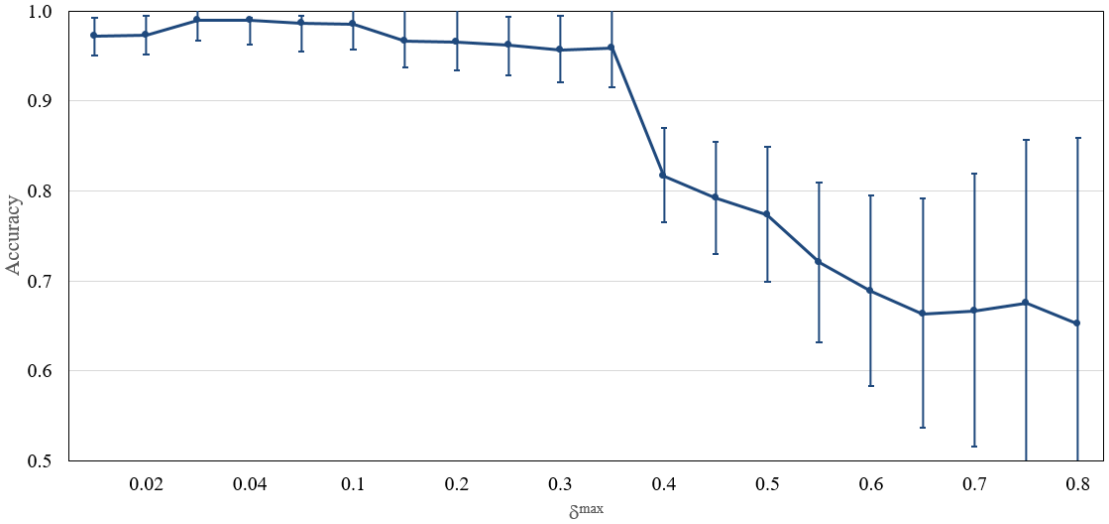
Trên tập *Thyroid* (Hình 2.13), độ chính xác giảm tương đối đều khi $\delta^{\max}$ tăng. Kết quả này phản ánh đặc điểm dữ liệu có phân bố phức tạp, mật độ không đồng đều và dễ chịu tác động của nhiễu hoặc các điểm biên.

Với tập *PID* (Hình 2.14), hiệu năng của FMN-SAL thấp hơn so với các tập dữ liệu có cấu trúc cụm rõ ràng và có xu hướng giảm dần khi $\delta^{\max}$ tăng. Điều này cho thấy dữ liệu *PID* có mức độ chồng lấn lớn, các cụm kém tách biệt và ảnh hưởng của nhiễu rõ hơn so với các tập dữ liệu khác. Khi siêu hộp được mở rộng quá mức, mô hình dễ bao phủ đồng thời các mẫu thuộc các vùng phân bố khác nhau, làm giảm khả năng suy luận nhãn.

Đối với tập *Spiral* (Hình 2.15), hiệu năng của FMN-SAL suy giảm rất mạnh khi $\delta^{\max}$ tăng. Nguyên nhân là dữ liệu có cấu trúc phi tuyến và đan xen, nên các siêu hộp lớn khó bám theo hình dạng cụm thực. Khi $\delta^{\max}$ nhỏ, các siêu hộp có khả năng

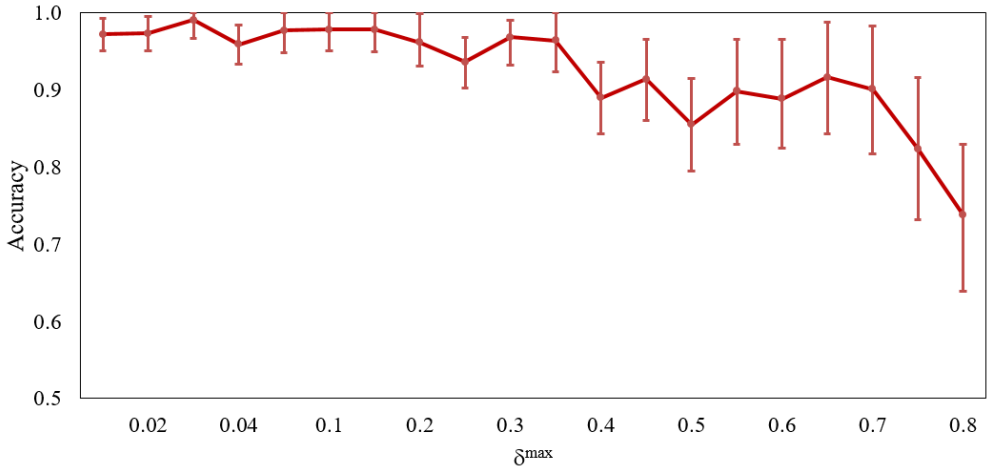


***(a)** FMN-SAL*

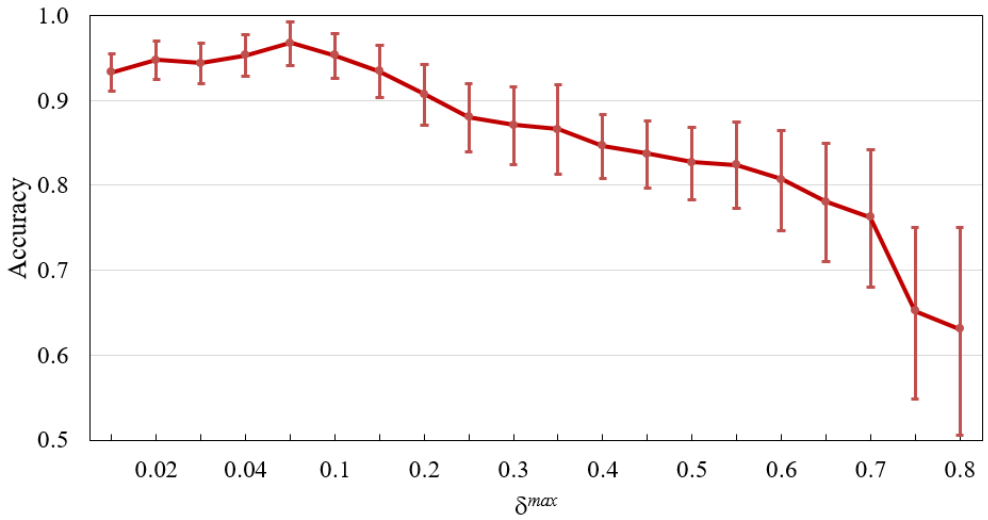


***(b)** FMN-Voting*

***Hình 2.11:** Hiệu năng phân cụm theo chỉ số Accuracy trên tập dữ liệu R15*

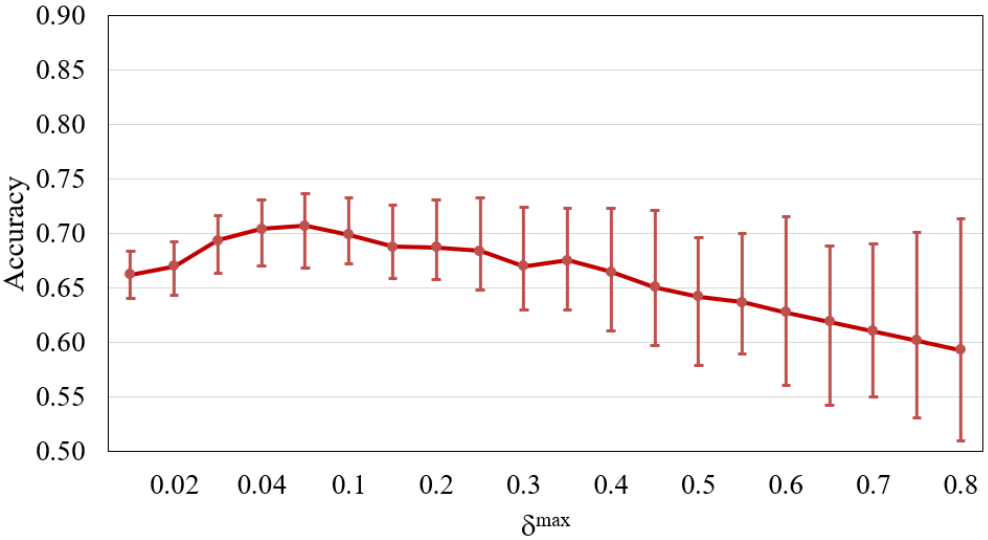


***Hình 2.12:** Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Aggregation*



Hình 2.13: *Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Thyroid*

mô tả tốt hơn các vùng cục bộ trên từng nhánh xoắn. Ngược lại, khi siêu hộp được mở rộng quá mức, chúng dễ bao phủ sang các nhánh lân cận, làm tăng chồng lấn giữa các cụm và dẫn đến suy giảm đáng kể độ chính xác phân cụm.

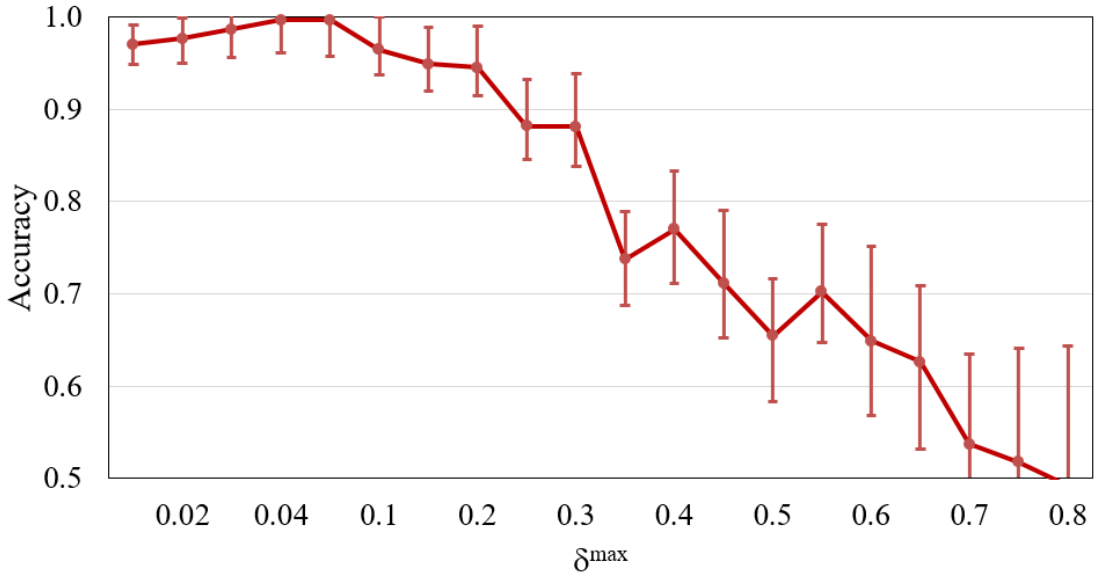


Hình 2.14: *Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu PID*

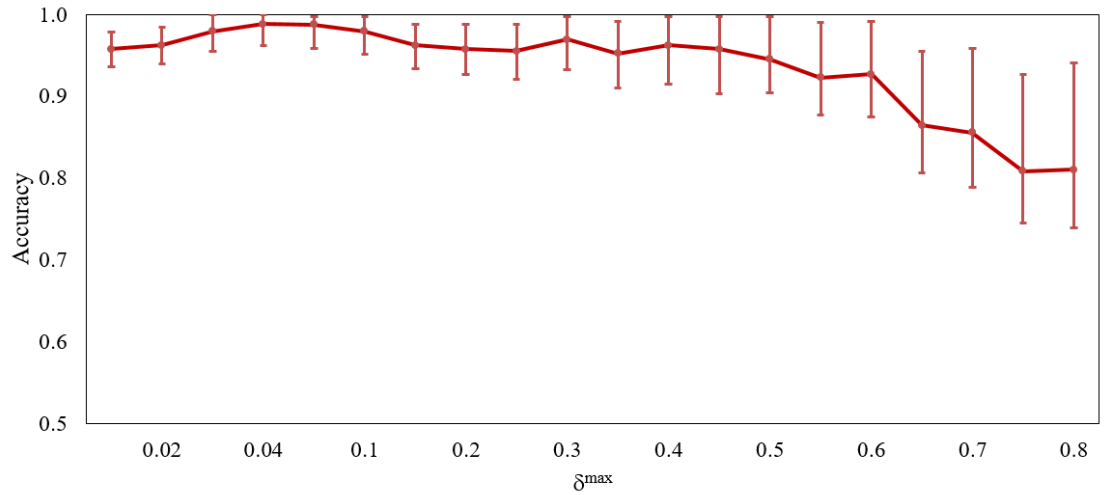
Trên tập *Jain* (Hình 2.16), FMN-SAL giữ được hiệu năng tốt trong một khoảng $\delta^{\max}$ khá rộng. Tuy vậy, ở vùng $\delta^{\max}$ lớn, độ chính xác vẫn giảm do siêu hộp bắt đầu bao phủ vượt quá ranh giới cụm thực.

Trên tập *Pathbased* (Hình 2.17), mô hình đạt hiệu năng cao khi $\delta^{\max}$ nhỏ, nhưng giảm khi tham số tăng. Kết quả này cho thấy việc kiểm soát kích thước siêu hộp là cần thiết để bảo toàn cấu trúc cụm phi lồi và hạn chế lan truyền nhãn sai.

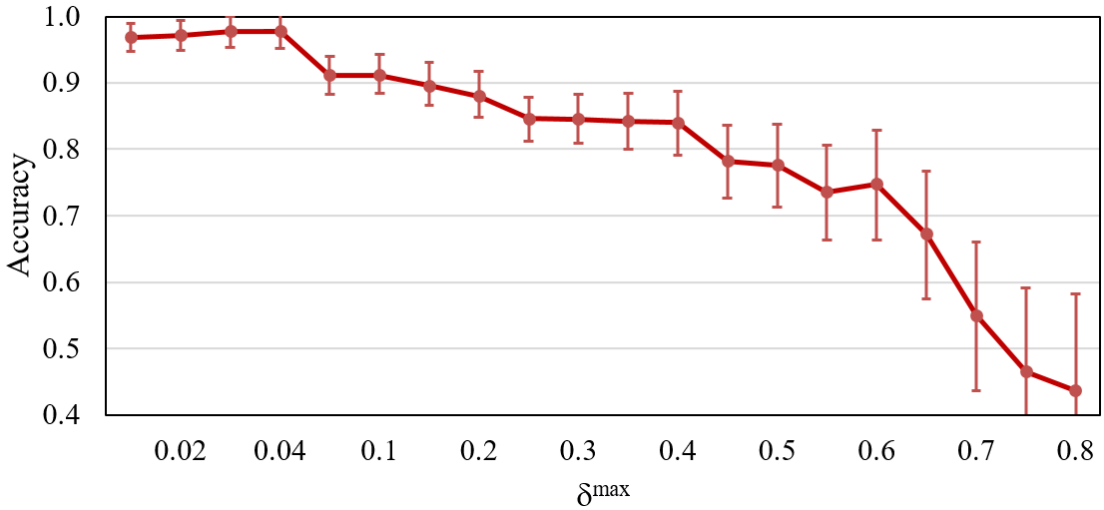
Bảng 2.3 và đồ thị trực quan trên Hình 2.18 cho thấy ảnh hưởng rõ rệt của tham số δ^{max} đến số lượng siêu hộp hình thành trong FMN-SAL. Nhìn chung, khi δ^{max} tăng, số lượng siêu hộp giảm trên tất cả các tập dữ liệu thực nghiệm. Nguyên nhân là do δ^{max} quy định giới hạn mở rộng của siêu hộp; khi giá trị này nhỏ, mỗi



Hình 2.15: *Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Spiral*



Hình 2.16: *Hiệu năng phân cụm trung bình theo chỉ số Accuracy đạt được bởi FMN-SAL trên tập dữ liệu Jain*



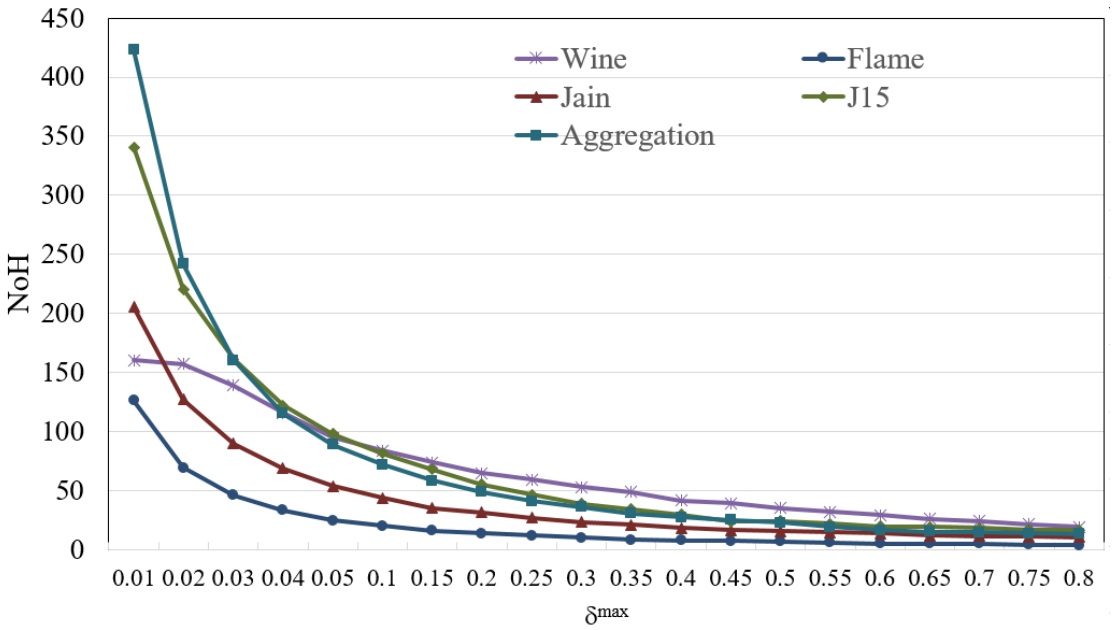
Hình 2.17: *Hiệu năng phân cụm đạt được bởi FMN-SAL trên tập dữ liệu Pathbased*

siêu hộp chỉ có thể bao phủ một vùng dữ liệu hẹp, buộc mô hình phải tạo ra nhiều siêu hộp nhỏ để biểu diễn dữ liệu. Ngược lại, khi δ^{max} lớn hơn, các siêu hộp có khả năng mở rộng và bao phủ nhiều mẫu hơn, từ đó làm giảm số lượng siêu hộp cần thiết. Xu hướng này đặc biệt rõ ở các tập dữ liệu có số lượng mẫu lớn hoặc cấu trúc phân bố phức tạp như *PID*, *Aggregation*, *J15* và *Jain*. Chẳng hạn, trên tập *Aggregation*, số lượng siêu hộp giảm từ 423,1 tại $\delta^{max} = 0{,}01$ xuống còn 13,3 tại $\delta^{max} = 0{,}80$; trên tập *PID*, giá trị này giảm từ 688,0 xuống còn 151.1. Hình 2.18 trực quan hóa xu hướng trên, trong đó các đường biểu diễn đều giảm mạnh ở vùng δ^{max} nhỏ, đặc biệt trong khoảng từ 0,01 đến 0,05, sau đó giảm chậm dần khi δ^{max} tiếp tục tăng. Điều này cho thấy việc tăng nhẹ δ^{max} ở giai đoạn đầu đã giúp mô hình hợp nhất đáng kể các vùng dữ liệu cục bộ, nhưng khi δ^{max} đủ lớn, tốc độ giảm số lượng siêu hộp bị giới hạn bởi cấu trúc phân bố dữ liệu, ranh giới lớp và điều kiện tránh chồng lấn. Do đó, δ^{max} là tham số quan trọng, chi phối trực tiếp mức độ khái quát hóa, độ phức tạp cấu trúc và khả năng biểu diễn dữ liệu của FMN-SAL.

Bảng 2.3: *Ảnh hưởng của δ^{max} đến số lượng siêu hộp hình thành trong FMN-SAL*

δ^{max}	**Iris**	**Thyroid**	**Wine**	**PID**	**Spiral**	**Flame**	**Jain**	**J15**	**Pathbased**	**Aggregation**
0,01	132,6	179,2	160,3	688,0	163,0	126,2	205,5	340,6	174,4	423,1
0,02	132,6	127,5	157,1	666,5	101,4	69,5	127,0	220,5	106,3	241,7
0,03	129,3	102,0	139,0	602,6	76,1	46,0	89,9	161,4	77,1	160,2
0,04	129,3	80,6	115,6	530,7	62,0	33,3	68,8	122,3	61,5	115,2
0,05	126,1	67,3	94,9	458,1	53,0	24,8	54,0	97,9	53,4	88,8
0,10	99,2	57,1	83,7	407,6	46,4	20,3	43,6	81,4	47,8	72,0
0,15	78,8	48,1	74,0	364,6	40,5	16,2	35,2	67,7	41,6	58,4
0,20	64,4	39,4	64,9	329,5	37,0	13,9	31,4	55,0	37,7	48,7
0,25	54,3	35,5	59,4	302,8	32,1	11,8	27,0	46,7	33,3	41,1
0,30	47,7	32,0	52,7	282,9	28,3	10,0	23,5	38,9	28,7	36,2
0,35	40,4	27,7	48,7	263,0	25,3	8,2	21,1	34,1	25,9	30,7
0,40	36,7	25,0	41,6	243,1	22,4	7,9	18,2	29,5	23,5	27,5
0,45	30,7	22,9	39,3	228,6	20,6	7,6	16,5	23,9	21,9	24,9
0,50	27,8	20,9	35,3	212,3	18,2	7,1	15,5	24,1	20,2	22,7
0,55	24,6	18,3	31,8	201,4	18,6	6,1	14,5	22,0	17,5	19,4
0,60	22,6	16,6	29,2	190,0	16,0	5,0	13,8	19,7	16,1	16,5
0,65	20,8	14,6	26,2	177,7	14,6	4,9	11,8	19,3	15,2	15,1
0,70	17,6	14,3	24,0	167,8	13,0	5,0	11,5	18,2	13,4	14,8
0,75	16,2	12,9	21,3	157,8	12,2	4,4	11,1	17,0	13,0	13,8
0,80	14,5	12,9	19,4	151,1	10,7	3,9	10,7	16,6	11,5	13,3

Nhìn chung, các kết quả trên cho thấy $\delta^{\max}$ là tham số quan trọng đối với cả FMN-SAL và FMN-Voting. Giá trị nhỏ hoặc trung bình thấp thường phù hợp hơn vì giúp siêu hộp bám sát cấu trúc phân bố của dữ liệu, hạn chế ảnh hưởng của nhiễu và bảo toàn tốt hơn ranh giới giữa các cụm. Ngược lại, khi $\delta^{\max}$ quá lớn, siêu hộp dễ mất tính đặc trưng cục bộ, dẫn đến suy giảm cả độ chính xác lẫn tính ổn định của phương pháp.



Hình 2.18: *Ảnh hưởng của δ^{max} tới siêu hộp trong mạng FMN-SAL*

Bảng 2.4 cho thấy phương pháp đề xuất đạt hiệu năng rất cạnh tranh trên bốn tập dữ liệu chuẩn. Cụ thể, trên tập *Aggregation*, GCAE đạt giá trị cao nhất là 0,9920, trong khi FMN-SAL đạt 0,9911, cho thấy kết quả của phương pháp đề xuất gần tương đương với mô hình tốt nhất. Trên tập *Jain*, GCAE đạt giá trị tối ưu tuyệt đối là 1,0000, còn FMN-SAL đạt 0,9879, tiếp tục cho thấy khả năng phân cụm rất tốt. Đáng chú ý, trên hai tập dữ liệu *R15* và *Flame*, FMN-SAL lần lượt đạt 0,9984 và 0,9798, là các giá trị cao nhất trong toàn bộ các phương pháp được so sánh. Kết quả này cho thấy phương pháp đề xuất có ưu thế rõ rệt trên các tập dữ liệu có cấu trúc cụm rõ ràng nhưng vẫn tiềm ẩn sự phức tạp về hình dạng hoặc phân bố điểm dữ liệu. Cần lưu ý rằng một số kết quả của các phương pháp đối sánh trong Bảng 2.4 được kế thừa từ tài liệu [101], trong khi kết quả của phương pháp đề xuất được thực nghiệm trong nghiên cứu này. Do các kết quả từ [101] không được thực nghiệm lại với cùng quy trình tiền xử lý và thiết lập thực nghiệm, việc đối sánh có thể chịu ảnh hưởng bởi sự khác biệt về điều kiện thực nghiệm. Vì vậy, các số liệu này được sử dụng chủ yếu như một tham chiếu để đặt kết quả của phương pháp đề xuất trong tương quan với các nghiên cứu liên quan, thay vì khẳng định một so sánh tuyệt đối trong cùng điều kiện.

Xét một cách tổng thể, FMN-SAL duy trì hiệu năng ổn định và nằm trong nhóm tốt nhất trên cả bốn tập dữ liệu, trong khi FMN-Voting cũng cho kết quả khả quan trên *R15* và *Flame*. So với các thuật toán phân cụm cổ điển như K-Means, Hierarchical, Clique hay DPC, các phương pháp dựa trên FMN cho thấy khả năng thích nghi tốt hơn đối với cấu trúc dữ liệu phi tuyến và ranh giới cụm phức tạp. Nhận xét này cũng phù hợp với xu hướng được báo cáo trong tài liệu [28], nơi các phương pháp cải tiến

theo hướng thích nghi và khai thác thông tin lân cận hoặc giả nhãn thường cho hiệu quả cao hơn trên các tập dữ liệu có hình dạng cụm phức tạp, đặc biệt là các bộ dữ liệu tổng hợp như *Pathbased* hoặc *Spiral*. Tài liệu [28] cũng cho thấy các phương pháp cải tiến dựa trên cơ chế bỏ phiếu và tối ưu thích nghi đạt kết quả tốt hơn các thuật toán đối sánh trên phần lớn tập dữ liệu thực nghiệm, đặc biệt với các tập dữ liệu tổng hợp có cấu trúc cụm không lồi hoặc phức tạp.

Nhìn chung, FMN-SAL đạt Accuracy cao và duy trì kết quả tương đối ổn định trên nhiều tập dữ liệu được khảo sát. Các kết quả này cho thấy tiềm năng của phương pháp đề xuất đối với bài toán phân cụm bán giám sát; tuy nhiên, các chênh lệch giữa các phương pháp cần được diễn giải thận trọng do luận án chưa thực hiện kiểm định ý nghĩa thống kê. Bảng 2.5 trình bày giá trị độ đo $Accuracy$ của ba mô hình SCFMN,

Bảng 2.4: *Giá trị ARI trên các tập dữ liệu thực nghiệm [101]*

Mô hình	**Aggregation**	**Jain**	**R15**	**Flame**
GCAE	**0,9920**	**1,0000**	0,9928	0,9501
Clique	0,7623	0,4508	0,4984	0,5143
K-Means	0,7281	0,5528	0,9188	0,4148
Hierarchical	0,7568	0,5146	0,9893	0,2889
DBSCAN	0,9664	0,9382	0,9375	0,9388
DPC	0,7320	0,6438	0,9928	0,9015
HDBSCAN	0,8759	0,9605	0,9572	0,6275
FMN-Voting	-	-	0,9780	0,9622
FMN-SAL	0,9911	0,9879	0,9984	0,9798

MSCFMN và FMN-SAL trên các tập dữ liệu chuẩn với các giá trị ngưỡng mở rộng siêu hộp $\delta^{\max}$ khác nhau. Kết quả cho thấy FMN-SAL đạt giá trị Accuracy cao hơn trên phần lớn các tập dữ liệu được khảo sát. Tuy nhiên, các chênh lệch nhỏ giữa các phương pháp chỉ phản ánh kết quả quan sát được trong thực nghiệm và chưa đủ cơ sở để kết luận về sự khác biệt có ý nghĩa thống kê. Trong đó, độ chính xác của FMN-SAL trên tập *Spiral* đạt tới **99,43%**, cho thấy khả năng khai thác hiệu quả cấu trúc không gian của dữ liệu. Ngược lại, trên các tập dữ liệu *PID* và *Sonar*, mô hình **SCFMN** cho kết quả tốt hơn đôi chút so với FMN-SAL.

Tuy nhiên, FMN-SAL không đạt kết quả cao nhất trên tất cả các tập dữ liệu. Cụ thể, trên PID và Sonar, SCFMN đạt Accuracy lần lượt là 70,57% và 73,91%, cao hơn nhẹ so với 70,04% và 73,58% của FMN-SAL. Mức chênh lệch tương ứng là 0,53% và 0,33%. Kết quả này cho thấy lợi ích của cơ chế bỏ phiếu dựa trên các mẫu mang nhãn phụ thuộc vào đặc điểm của từng tập dữ liệu và chất lượng của thông tin nhãn sẵn có; do đó FMN-SAL không được kỳ vọng cao hơn trong mọi trường hợp.

Trên các tập PID và Sonar, SCFMN đạt Accuracy cao hơn FMN-SAL với mức

chênh lệch nhỏ. Do chưa thực hiện kiểm định thống kê đối với các kết quả này, sự khác biệt quan sát được chưa đủ để kết luận phương pháp nào có ưu thế có ý nghĩa thống kê trên hai tập dữ liệu. Kết quả này cho thấy hiệu quả tương đối của các mô hình có thể thay đổi tùy theo đặc điểm của tập dữ liệu.

Nhìn chung, các kết quả thực nghiệm cho thấy việc khai thác thông tin lân cận trong các mô hình đề xuất có xu hướng cải thiện độ chính xác trên phần lớn các tập dữ liệu được khảo sát. Tuy nhiên, do chưa thực hiện kiểm định thống kê, các kết quả này được xem là bằng chứng thực nghiệm về xu hướng cải thiện và chưa được sử dụng để khẳng định sự khác biệt có ý nghĩa thống kê giữa các phương pháp.

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

Tập dữ liệu	$\delta^{\max}$	**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**

Kết quả thực nghiệm của mô hình FMN-SAL trên tập dữ liệu ILPD được đánh giá thông qua khả năng nhận diện các mẫu bệnh gan và phân loại nhóm bệnh của mô hình. Trong đó, lớp bệnh gan được xác định là lớp dương, còn lớp không bệnh gan là lớp âm.

Bảng 2.6 và Hình 2.19 trình bày kết quả đánh giá khả năng nhận diện lớp bệnh gan của mô hình FMN-SAL trên tập dữ liệu ILPD. Trong Bảng 2.6, lớp bệnh gan được xem là lớp dương (positive class). Do đó, các độ đo Recall, Specificity, Precision và F1-Score được xác định với lớp bệnh gan là lớp quan tâm nhằm đánh giá khả năng nhận diện các trường hợp mắc bệnh của mô hình.

Kết quả trên Bảng 2.6 cho thấy ảnh hưởng của tham số $\delta^{\max}$ đến hiệu quả của mô hình thông qua các độ đo Accuracy, Recall, Specificity, FP Rate, Precision và F1-Score. Khi $\delta^{\max}$ tăng từ 0,01 đến 0,08, hầu hết các độ đo đều được cải thiện và mô hình đạt kết quả tốt nhất tại $\delta^{\max} = 0{,}08$. Tại giá trị này, Accuracy đạt 0,9039, Recall đạt 0,9423, Specificity đạt 0,8084, Precision đạt 0,9245 và F1-Score đạt 0,9333, trong khi FP Rate giảm xuống 0,1916.

Xét theo mục tiêu nhận diện các mẫu bệnh gan, Recall đạt 0,9423 cho thấy mô hình có khả năng phát hiện đúng phần lớn các mẫu thực sự thuộc lớp bệnh gan, trong khi Precision đạt 0,9245 cho thấy phần lớn các mẫu được mô hình nhận diện là bệnh gan thực sự thuộc lớp này. Giá trị F1-Score cao cho thấy sự cân bằng tốt giữa khả năng phát hiện mẫu bệnh gan và độ chính xác của các dự đoán thuộc lớp bệnh.

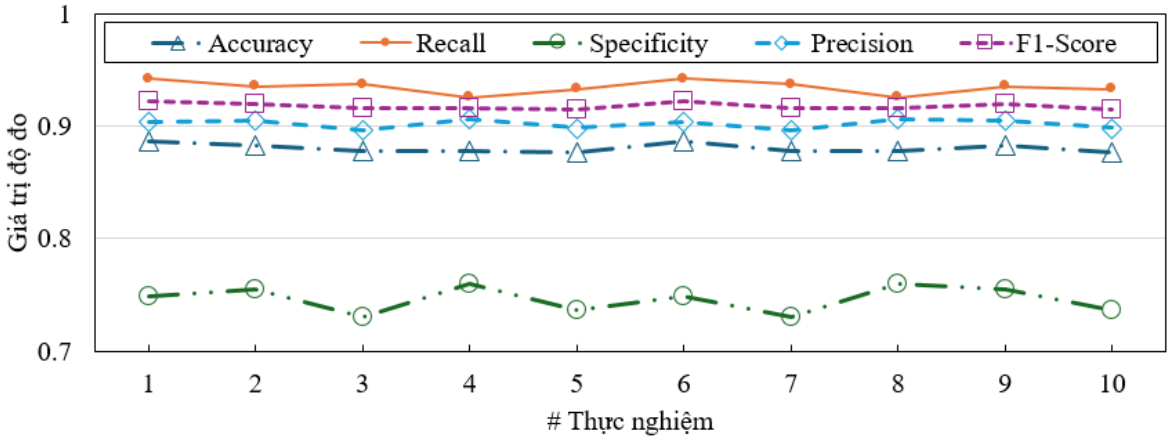
Khi $\delta^{\max} > 0{,}08$, các độ đo Recall, Specificity, Precision và F1-Score có xu hướng giảm, trong khi FP Rate tăng; mức suy giảm trở nên rõ rệt hơn khi $\delta^{\max} \geq 0{,}5$. Điều này cho thấy việc cho phép các siêu hộp mở rộng quá mức có thể làm giảm khả năng phân biệt giữa các mẫu bệnh gan và không bệnh gan. Do đó, $\delta^{\max} = 0{,}08$ được lựa chọn là giá trị phù hợp cho mô hình FMN-SAL trên tập dữ liệu ILPD.

Kết quả minh họa trong Hình 2.19 cho thấy các độ đo có mức dao động tương đối nhỏ giữa các lần thực nghiệm. Recall dao động trong khoảng 0,9255 - 0,9423, Precision trong khoảng 0,8966 - 0,9059 và F1-Score trong khoảng 0,9151 - 0,9224, trong khi Specificity dao động trong khoảng 0,7305 - 0.7605.

Bảng 2.6: *Kết quả đánh giá trên lớp bệnh gan của FMN-SAL*

$\delta^{\max}$	**Accuracy**	**Recall**	**Specificity**	**FP Rate**	**Precision**	**F1-Score**
0,01	0,8199	0,8846	0,6587	0,3413	0,8659	0,8751
0,02	0,8456	0,9014	0,7066	0,2934	0,8844	0,8929
0,04	0,8662	0,9183	0,7365	0,2635	0,8967	0,9074
0,06	0,8902	0,9351	0,7784	0,2216	0,9131	0,9240
0,08	0,9039	0,9423	0,8084	0,1916	0,9245	0,9333
0,10	0,8799	0,9255	0,7665	0,2335	0,9080	0,9167
0,20	0,8593	0,9087	0,7365	0,2635	0,8957	0,9021
0,30	0,8233	0,8702	0,7066	0,2934	0,8808	0,8755
0,40	0,7890	0,8293	0,6886	0,3114	0,8690	0,8487
0,50	0,7204	0,7572	0,6287	0,3713	0,8355	0,7945
0,60	0,6861	0,7163	0,6108	0,3892	0,8209	0,7651
0,70	0,6432	0,6659	0,5868	0,4132	0,8006	0,7270
0,80	0,5695	0,5769	0,5509	0,4491	0,7619	0,6566
0,90	0,5111	0,5288	0,4671	0,5329	0,7120	0,6069

Bên cạnh khả năng nhận diện lớp bệnh gan, chất lượng phân cụm của mô hình FMN-SAL trên tập dữ liệu ILPD cũng được đánh giá. Khác với kết quả trên Bảng 2.6, trong đó các độ đo được xem xét với lớp bệnh gan là lớp dương, Bảng 2.7 đánh giá kết quả trên cả hai lớp bệnh gan và không mắc bệnh. Trước hết, ảnh hưởng của tham số $\delta^{\max}$ được khảo sát thông qua các độ đo *Accuracy, Specificity, Precision* và *F1-Score*,



Hình 2.19: *Đánh giá độ đo trên lớp bệnh gan của FMN-SAL*

đồng thời sử dụng các chỉ số RI, ARI và NMI để đánh giá mức độ tương đồng giữa cấu trúc phân cụm thu được và phân hoạch theo nhãn thực tế.

Kết quả trong Bảng 2.7 cho thấy FMN-SAL phụ thuộc vào giá trị $\delta^{\max}$. Khi $\delta^{\max}$ tăng từ 0,01 đến 0,08, các độ đo Accuracy, Specificity, Precision và F1-Score đều được cải thiện. Tại $\delta^{\max} = 0{,}08$, mô hình đạt kết quả tốt nhất với Accuracy bằng 0,9039, Specificity bằng 0,8467, Precision bằng 0,9029 và F1-Score bằng 0.9032.

Bảng 2.7: *Đánh giá chất lượng phân cụm trung bình trên hai lớp của FMN-SAL khi thay đổi* $\delta^{\max}$

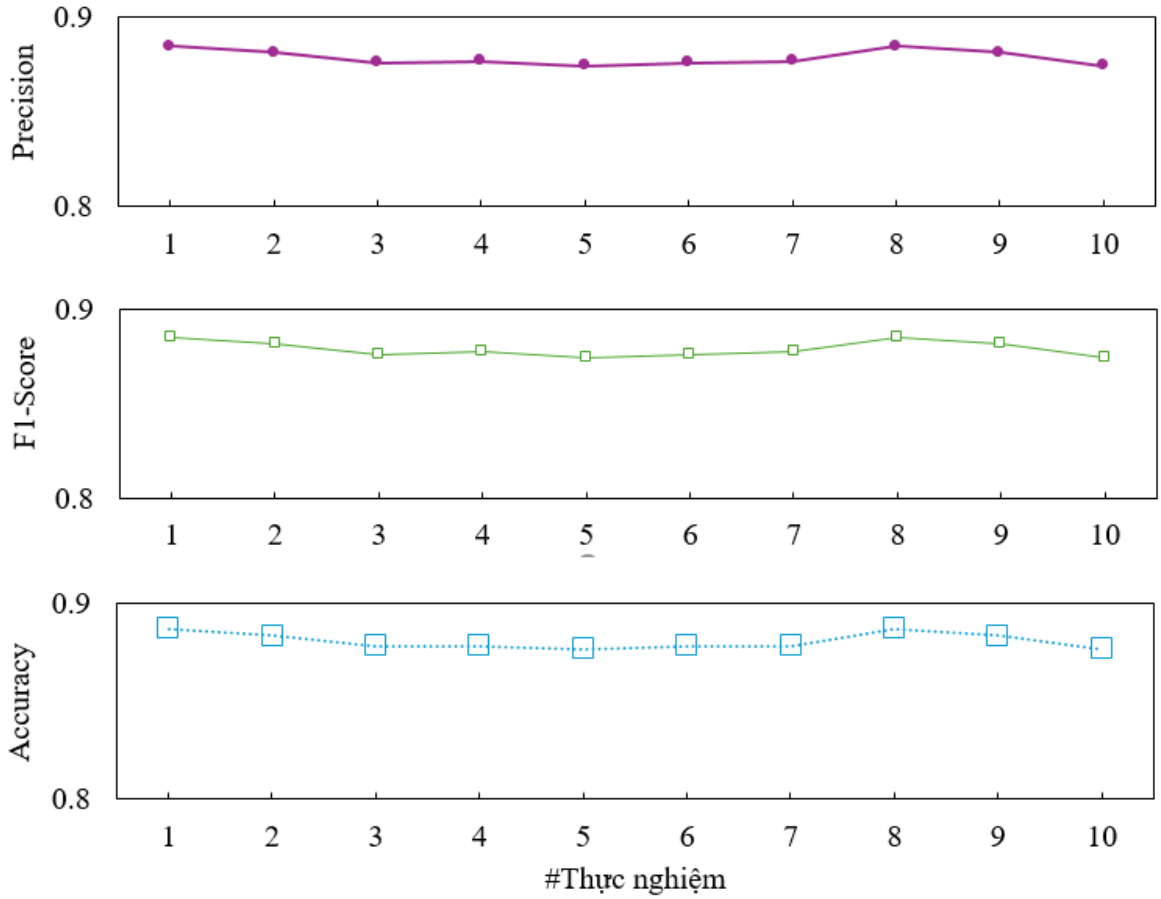
$\delta_{\max}$	**Accuracy**	**Specificity**	**Precision**	**F1-Score**	**RI**	**ARI**	**NMI**
0,01	0,8199	0,7234	0,8173	0,8184	0,7042	0,3851	0,2454
0,02	0,8456	0,7624	0,8437	0,8445	0,7385	0,4567	0,3087
0,04	0,8662	0,7886	0,8643	0,8649	0,7678	0,5172	0,3647
0,06	0,8902	0,8233	0,8888	0,8892	0,8042	0,5928	0,4395
0,08	0,9039	0,8467	0,9029	0,9032	0,8260	0,6387	0,4872
0,10	0,8799	0,8120	0,8785	0,8790	0,7883	0,5603	0,4069
0,20	0,8593	0,7858	0,8580	0,8586	0,7578	0,4976	0,3468
0,30	0,8233	0,7535	0,8250	0,8241	0,7086	0,3990	0,2618
0,40	0,7890	0,7289	0,7972	0,7922	0,6665	0,3170	0,1999
0,50	0,7204	0,6655	0,7422	0,7282	0,5965	0,1804	0,1042
0,60	0,6861	0,6410	0,7186	0,6969	0,5685	0,1278	0,0725
0,70	0,6432	0,6095	0,6897	0,6577	0,5402	0,0749	0,0420
0,80	0,5695	0,5584	0,6420	0,5897	0,5088	0,0168	0,0104
0,90	0,5111	0,4848	0,5896	0,5344	0,4994	-0,0016	0,0000

Đồng thời, các chỉ số RI, ARI và NMI cũng đạt giá trị cao nhất, lần lượt là 0,8260, 0,6387 và 0.4872. Kết quả này cho thấy tại $\delta^{\max} = 0{,}08$, FMN-SAL không chỉ đạt hiệu quả phân loại tổng thể tốt mà cấu trúc phân cụm được hình thành cũng có mức độ tương đồng cao nhất với phân hoạch theo nhãn bệnh gan và không bệnh gan trong

các giá trị tham số được khảo sát.

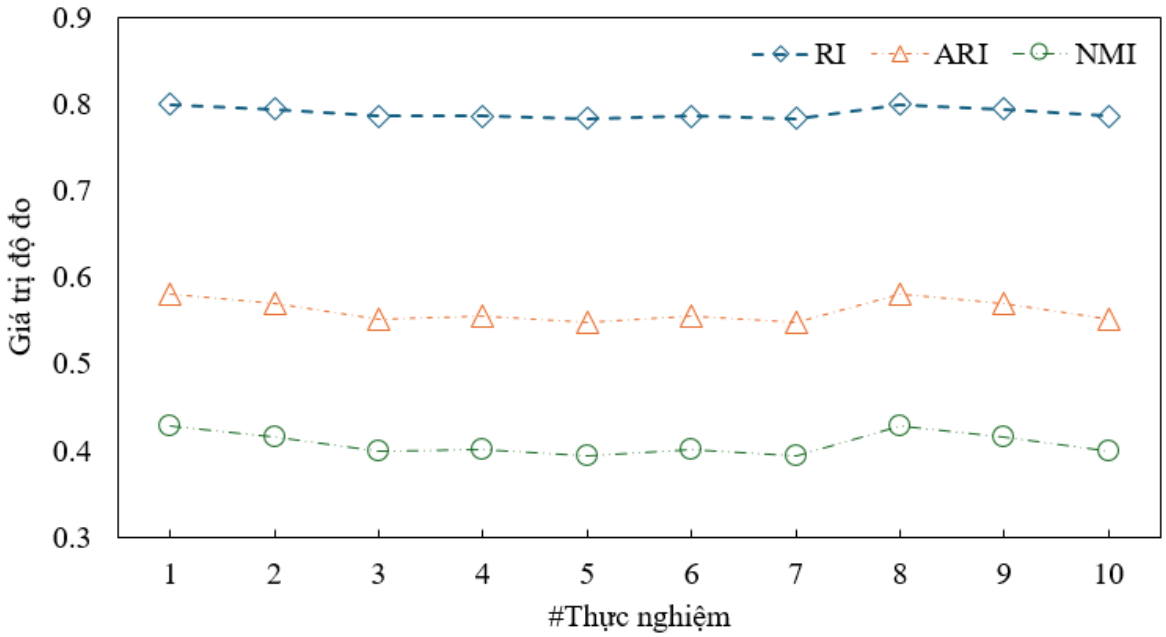
Khi $\delta^{\max} > 0{,}08$, các độ đo phân loại và các chỉ số đánh giá phân cụm đều có xu hướng giảm. Mức suy giảm trở nên rõ rệt hơn tại các giá trị $\delta^{\max}$ lớn, đặc biệt từ 0,5 đến 0.9. Tại $\delta^{\max} = 0{,}90$, Accuracy giảm xuống 0,5111, RI chỉ còn 0,4994, ARI đạt $-0{,}0016$ và NMI xấp xỉ 0. Điều này cho thấy khi kích thước siêu hộp được mở rộng quá mức, khả năng phân biệt các lớp suy giảm và cấu trúc được hình thành gần như không còn tương đồng với phân hoạch theo nhãn thực tế. Vì vậy, $\delta^{\max} = 0{,}08$ được lựa chọn cho mô hình FMN-SAL trên tập dữ liệu ILPD.

Với $\delta^{\max} = 0{,}08$ được lựa chọn, Hình 2.20 và Hình 2.21 cho thấy kết quả của mô hình FMN-SAL có mức biến động tương đối nhỏ qua các lần thực nghiệm. Accuracy dao động trong khoảng 0,8765 - 0,8868, Precision trong khoảng 0,8742 - 0,8848 và F1-Score trong khoảng 0,8746 - 0.8848. RI dao động trong khoảng 0,7831 - 0,7989, ARI trong khoảng 0,5473 - 0,5797 và NMI trong khoảng 0,3943 - 0.4274. Kết quả cho thấy hiệu quả phân loại của mô hình được duy trì ổn định giữa các lần chạy.



Hình 2.20: *Độ đo phân loại của mô hình FMN-SAL trên tập dữ liệu ILPD*

Kết quả trong Bảng 2.8 cho thấy khả năng nhận diện các mẫu bệnh gan của mô hình FMN-Voting chịu ảnh hưởng rõ rệt bởi tham số $\delta^{\max}$. Khi $\delta^{\max}$ tăng từ 0,01 đến 0,08, các độ đo đánh giá nhìn chung được cải thiện. Tại $\delta^{\max} = 0{,}08$, mô hình



Hình 2.21: *Chỉ số đánh giá phân cụm của mô hình FMN-SAL trên tập dữ liệu ILPD*

đạt kết quả tốt nhất với Accuracy bằng 0,8576, Recall bằng 0,9135, Specificity bằng 0,7186, Precision bằng 0,8899 và F1-Score bằng 0,9015; đồng thời FP Rate giảm xuống 0.2814.

Xét theo mục tiêu nhận diện bệnh gan, Recall đạt 0,9135 cho thấy mô hình có khả năng phát hiện đúng phần lớn các mẫu thực sự thuộc lớp bệnh gan. Precision đạt 0,8899 cho thấy phần lớn các mẫu được mô hình nhận diện là bệnh gan thực sự thuộc lớp bệnh. Bên cạnh đó, Specificity đạt 0,7186 cho thấy khả năng nhận diện các mẫu không bệnh ở mức tương đối, trong khi FP Rate bằng 0,2814 phản ánh vẫn còn một tỷ lệ nhất định các mẫu không bệnh bị nhận diện nhầm thành bệnh. F1-Score đạt 0,9015 cho thấy sự cân bằng giữa Recall và Precision trong khả năng nhận diện lớp bệnh gan.

Khi $\delta^{\max}$ tăng vượt quá 0,08, Recall, Specificity, Precision và F1-Score có xu hướng giảm, trong khi FP Rate tăng. Mức suy giảm trở nên rõ hơn khi $\delta^{\max}$ đạt các giá trị lớn. Điều này cho thấy việc mở rộng các siêu hộp quá mức có thể làm giảm khả năng phân biệt giữa hai lớp. Do đó, $\delta^{\max} = 0{,}08$ được lựa chọn là giá trị phù hợp đối với mô hình FMN-Voting trên tập dữ liệu ILPD.

Kết quả trong Bảng 2.9 cho thấy $\delta^{\max}$ có ảnh hưởng đến chất lượng của FMN-Voting. Khi $\delta^{\max}$ tăng từ 0,01 đến 0,08, Accuracy tăng từ 0,7581 lên 0,8576, Specificity tăng từ 0,6628 lên 0,7744, Precision tăng từ 0,7613 lên 0,8554 và F1-Score tăng từ 0,7596 lên 0.8561. RI, ARI và NMI đạt giá trị cao nhất, lần lượt là 0,7554, 0,4910 và 0.3399. Điều này cho thấy tại giá trị $\delta^{\max}$ này, cấu trúc được hình thành bởi FMN-Voting có mức độ tương đồng cao nhất với phân hoạch theo nhãn bệnh gan và không

Bảng 2.8: *Khả năng nhận diện lớp bệnh gan của mô hình FMN-Voting*

$\delta^{\max}$	**Accuracy**	**Recall**	**Specificity**	**FP Rate**	**Precision**	**F1-Score**
0,01	0,7581	0,8221	0,5988	0,4012	0,8362	0,8291
0,02	0,7925	0,8510	0,6467	0,3533	0,8571	0,8540
0,04	0,8130	0,8750	0,6587	0,3413	0,8646	0,8698
0,06	0,8319	0,8894	0,6886	0,3114	0,8768	0,8831
0,08	0,8576	0,9135	0,7186	0,2814	0,8899	0,9015
0,10	0,8233	0,8774	0,6886	0,3114	0,8753	0,8764
0,20	0,7890	0,8534	0,6287	0,3713	0,8513	0,8523
0,40	0,7410	0,8029	0,5868	0,4132	0,8288	0,8156
0,50	0,7136	0,7716	0,5689	0,4311	0,8168	0,7936
0,60	0,6707	0,7212	0,5449	0,4551	0,7979	0,7576
0,70	0,6089	0,6490	0,5090	0,4910	0,7670	0,7031
0,80	0,5540	0,5769	0,4970	0,5030	0,7407	0,6486
0,90	0,5146	0,5288	0,4790	0,5210	0,7166	0,6086

bệnh gan trong số các cấu hình được khảo sát.

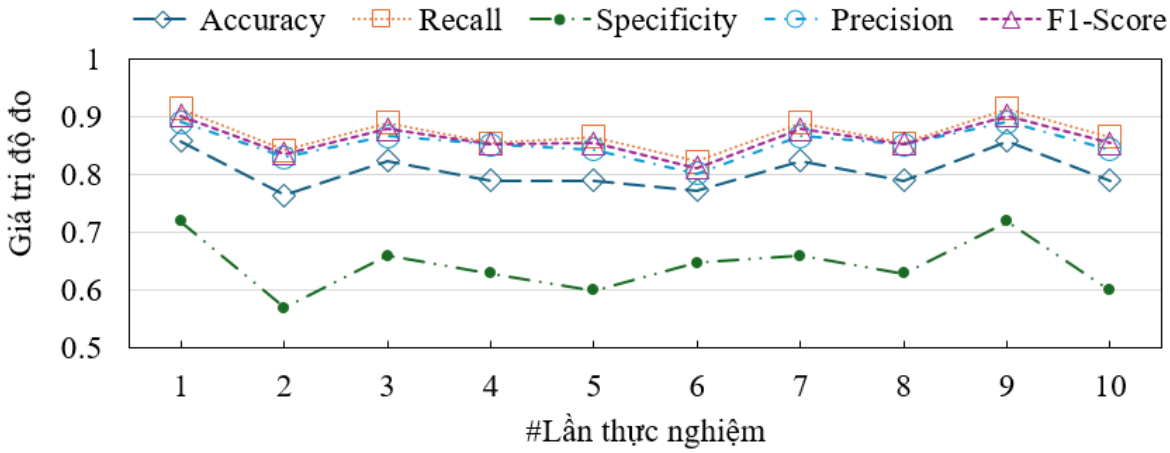
Khi $\delta^{\max} > 0{,}08$, các chỉ số RI, ARI, NMI có xu hướng giảm. Tại $\delta^{\max} = 0{,}90$, Accuracy chỉ còn 0,5146, ARI giảm xuống $-0{,}0011$ và NMI xấp xỉ 0. Kết quả này cho thấy khi kích thước siêu hộp quá lớn, cấu trúc cụm được hình thành gần như không còn tương đồng với phân hoạch theo nhãn thực tế.

Bảng 2.9: *Kết quả các chỉ số đánh giá mô hình FMN-Voting khi thay đổi* $\delta_{\max}$

$\delta^{\max}$	**Accuracy**	**Specificity**	**Precision**	**F1-Score**	**RI**	**ARI**	**NMI**
0,01	0,7581	0,6628	0,7613	0,7596	0,6327	0,2432	0,1366
0,02	0,7925	0,7052	0,7936	0,7930	0,6705	0,3198	0,1948
0,04	0,8130	0,7206	0,8114	0,8122	0,6955	0,3685	0,2323
0,06	0,8319	0,7461	0,8302	0,8310	0,7198	0,4187	0,2752
0,08	0,8576	0,7744	0,8554	0,8561	0,7554	0,4910	0,3399
0,10	0,8233	0,7427	0,8230	0,8232	0,7086	0,3971	0,2578
0,20	0,7890	0,6931	0,7886	0,7888	0,6665	0,3100	0,1856
0,40	0,7410	0,6487	0,7473	0,7438	0,6155	0,2103	0,1145
0,50	0,7136	0,6269	0,7260	0,7187	0,5905	0,1629	0,0846
0,60	0,6707	0,5954	0,6952	0,6800	0,5575	0,1016	0,0493
0,70	0,6089	0,5491	0,6527	0,6241	0,5229	0,0386	0,0166
0,80	0,5540	0,5199	0,6204	0,5745	0,5050	0,0081	0,0035
0,90	0,5146	0,4933	0,5944	0,5377	0,4996	-0,0011	0,0000

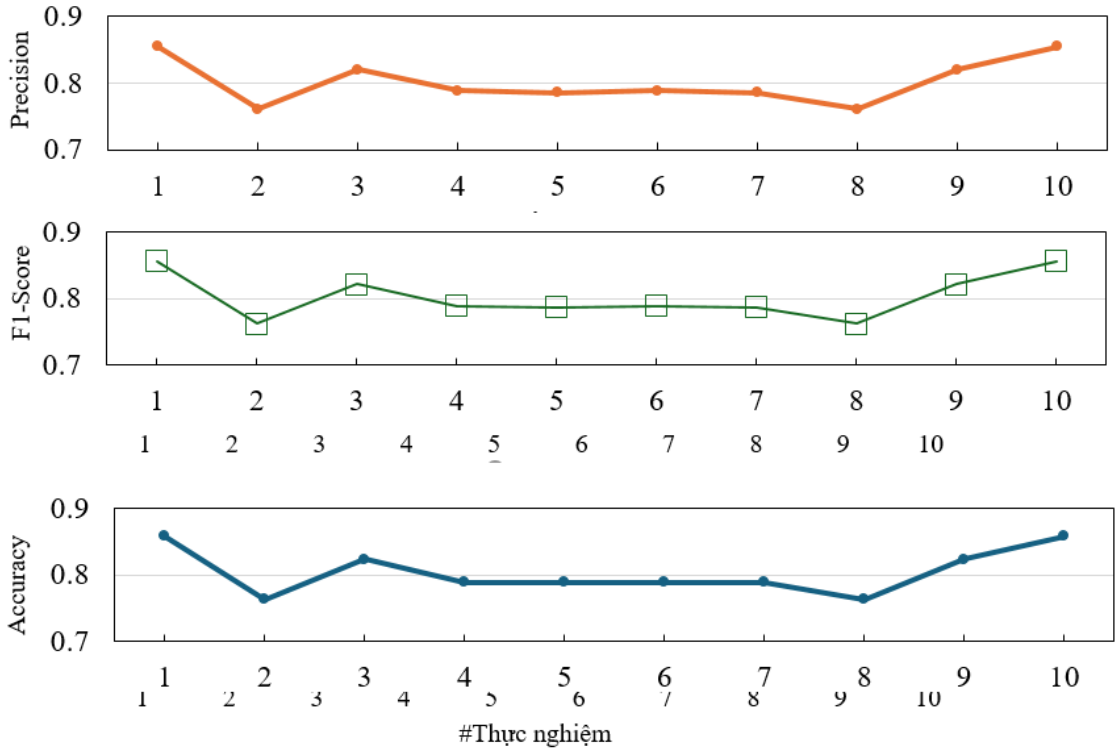
Kết quả trên Hình 2.22 cho thấy khả năng nhận diện lớp bệnh gan của FMN-Voting có sự biến động nhất định giữa các lần thực nghiệm. Recall, Precision và

F1-Score duy trì ở mức cao hơn Specificity. Recall dao động trong khoảng 0,8413 - 0,9135, Precision trong khoảng 0,8294 - 0,8899 và F1-Score trong khoảng 0,8353 - 0.9015. Trong khi đó, Specificity dao động trong khoảng 0,5689 - 0.7186.



Hình 2.22: *Khả năng nhận diện lớp bệnh gan của mô hình FMN-Voting*

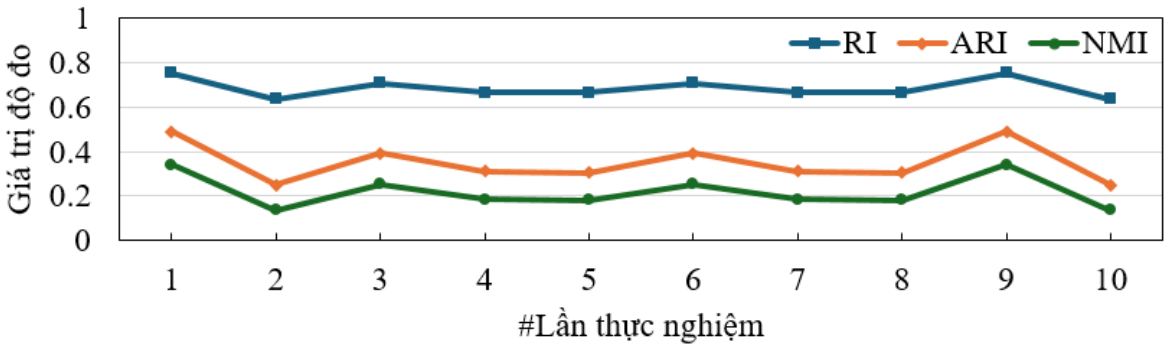
Kết quả trên Hình 2.23 cho thấy các độ đo đánh giá của FMN-Voting có mức biến động tương đối rõ. Accuracy dao động trong khoảng 0,7633 - 0,8576, Precision trong khoảng 0,7608 - 0,8554 và F1-Score trong khoảng 0,7620 - 0.8561.



Hình 2.23: *Sự biến động giá trị độ đo của mô hình FMN-Voting*

Để đánh giá hiệu quả của các mô hình đề xuất, kết quả thực nghiệm của FMN-SAL và FMN-Voting được so sánh với các phương pháp phân cụm K-means++, EM và DBSCAN được thực hiện trên Weka, đồng thời đối sánh với mô hình SCFMN. Việc so sánh được thực hiện trên cùng tập dữ liệu ILPD và xem xét ở ba mức gồm lớp

bệnh gan, lớp không bệnh gan và kết quả trung bình. Bên cạnh các độ đo phân loại, các chỉ số RI, ARI và NMI được sử dụng để đánh giá mức độ tương đồng giữa cấu trúc phân cụm thu được và phân hoạch theo nhãn thực tế.



Hình 2.24: *Giá trị độ đo RI, ARI, NMI của mô hình FMN-Voting*

Kết quả trong Bảng 2.10 và trực quan trên Hình 2.25 cho thấy FMN-SAL đạt hiệu quả tốt nhất trên lớp bệnh gan trong số các phương pháp được so sánh. Mô hình đạt Accuracy bằng $0{,}8806 \pm 0{,}0043$, Recall bằng $0{,}9346 \pm 0{,}0062$, Precision bằng $0{,}9017 \pm 0{,}0041$ và F1-Score bằng $0{,}9178 \pm 0{,}0031$. Đồng thời, Specificity đạt $0{,}7461 \pm 0{,}0124$ và FP Rate ở mức $0{,}2539 \pm 0{,}0124$.

FMN-Voting cũng đạt kết quả cao hơn các phương pháp phân cụm truyền thống trên hầu hết các độ đo, với Accuracy bằng $0{,}8045 \pm 0.0366$, Recall bằng $0{,}8726 \pm 0.0289$, Precision bằng $0{,}8560 \pm 0{,}0232$ và F1-Score bằng $0{,}8642 \pm 0{,}0258$. So với SCFMN, cả FMN-SAL và FMN-Voting đều cải thiện khả năng nhận diện các mẫu bệnh gan.

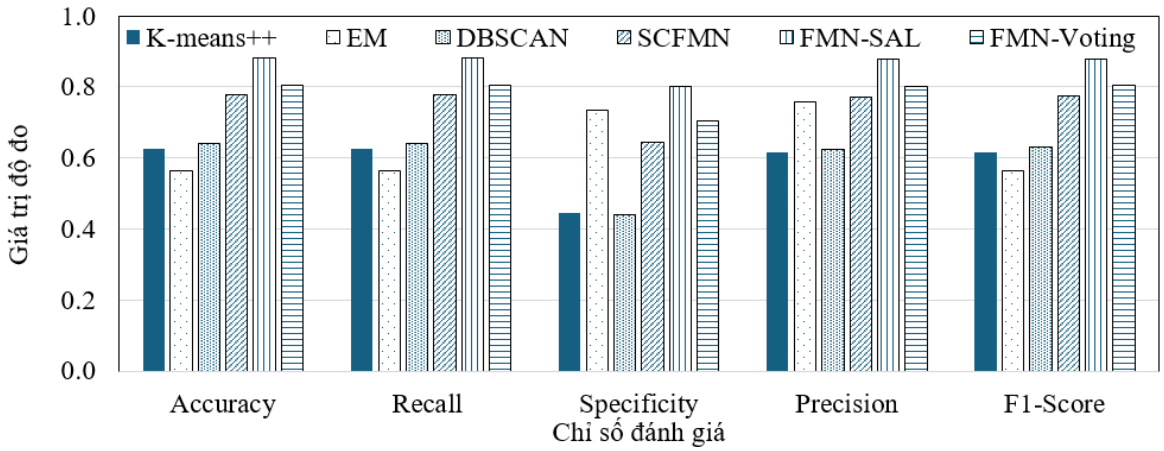
Đối với các phương pháp thực hiện trên Weka, K-means++ và DBSCAN đạt Recall tương đối cao, lần lượt là $0{,}7452$ và $0{,}7831$, nhưng Specificity chỉ đạt $0{,}3263$ và $0{,}2996$, kéo theo FP Rate ở mức cao. Điều này cho thấy hai phương pháp có khả năng nhận diện một tỷ lệ các mẫu bệnh gan nhưng đồng thời dễ nhận diện nhầm các mẫu không bệnh thành bệnh. Ngược lại, EM đạt Specificity và Precision cao, lần lượt là $0{,}8496$ và $0{,}9112$, nhưng Recall chỉ đạt $0{,}4503$, cho thấy nhiều mẫu bệnh gan chưa được nhận diện đúng.

Nhìn chung, FMN-SAL thể hiện sự cân bằng tốt hơn giữa Recall, Specificity, Precision và F1-Score, trong khi FMN-Voting cũng cho kết quả cải thiện rõ rệt so với các phương pháp phân cụm truyền thống.

Bảng 2.11 trình bày khả năng nhận diện lớp không bệnh gan của các phương pháp. FMN-SAL tiếp tục đạt kết quả tốt với Recall bằng $0{,}7461 \pm 0{,}0124$, Specificity bằng $0{,}9346 \pm 0.0062$, Precision bằng $0{,}8210 \pm 0.0130$ và F1-Score bằng $0{,}7817 \pm 0.0077$. Đồng thời, FP Rate của mô hình chỉ ở mức $0{,}0654 \pm 0{,}0062$.

Bảng 2.10: *Kết quả nhận diện lớp bệnh gan của các mô hình đối sánh*

Mô hình	Accuracy	Recall	Specificity	FP Rate	Precision	F1-Score
K-means++	0,6252 ± 0,0222	0,7452 ± 0,0855	0,3263 ± 0,1628	0,6737 ± 0,1628	0,7366 ± 0,0256	0,7376 ± 0,0303
EM	0,5647 ± 0,0351	0,4503 ± 0,1452	0,8496 ± 0,2402	0,1504 ± 0,2402	0,9112 ± 0,0702	0,5860 ± 0,0721
DBSCAN	0,6416 ± 0,0022	0,7831 ± 0,0058	0,2996 ± 0,0006	0,7004 ± 0,0006	0,7299 ± 0,0045	0,7556 ± 0,0018
SCFMN	0,7780 ± 0,0184	0,8678 ± 0,0231	0,5545 ± 0,0478	0,4455 ± 0,0478	0,8293 ± 0,0143	0,8479 ± 0,0134
FMN-SAL	0,8806 ± 0,0043	0,9346 ± 0,0062	0,7461 ± 0,0124	0,2539 ± 0,0124	0,9017 ± 0,0041	0,9178 ± 0,0031
FMN-Voting	0,8045 ± 0,0366	0,8726 ± 0,0289	0,6347 ± 0,0576	0,3653 ± 0,0576	0,8560 ± 0,0232	0,8642 ± 0,0258



Hình 2.25: *So sánh các độ đo đánh giá trên lớp bệnh gan của FMN-SAL, FMN-Voting và các phương pháp đối sánh*

FMN-Voting đạt Recall bằng $0{,}6347 \pm 0.0576$ và F1-Score bằng $0{,}6506 \pm 0.0631$, cao hơn SCFMN và các phương pháp K-means++, EM, DBSCAN về khả năng đánh giá tổng hợp trên lớp không bệnh. SCFMN đạt Specificity tương đối cao nhưng Recall và F1-Score vẫn thấp hơn hai mô hình FMN-SAL và FMN-Voting.

Đáng chú ý, K-means++ và DBSCAN có Recall trên lớp không bệnh khá thấp, lần lượt là $0{,}3263$ và $0{,}2996$. Kết quả này phù hợp với Specificity thấp của hai phương pháp khi lớp bệnh gan được xem là lớp dương, cho thấy khả năng phân biệt hai lớp còn hạn chế. EM có Recall cao trên lớp không bệnh nhưng Precision chỉ đạt $0{,}3794$, do đó F1-Score chỉ đạt $0{,}5122$.

Như vậy, FMN-SAL không chỉ duy trì khả năng phát hiện mẫu bệnh gan tốt mà còn cải thiện khả năng nhận diện các mẫu không bệnh, cho thấy mức độ cân bằng

Bảng 2.11: *Kết quả nhận diện lớp không bệnh gan của các mô hình đối sánh*

Mô hình	Accuracy	Recall	Specificity	FP Rate	Precision	F1-Score
K-means++	0,6252 ± 0,0222	0,3263 ± 0,1628	0,7452 ± 0,0855	0,2548 ± 0,0855	0,3114 ± 0,0945	0,3112 ± 0,1311
EM	0,5647 ± 0,0351	0,8496 ± 0,2402	0,4503 ± 0,1452	0,5497 ± 0,1452	0,3794 ± 0,0151	0,5122 ± 0,0949
DBSCAN	0,6416 ± 0,0022	0,2996 ± 0,0006	0,7831 ± 0,0058	0,2169 ± 0,0058	0,3638 ± 0,0129	0,3285 ± 0,0054
SCFMN	0,7780 ± 0,0184	0,5545 ± 0,0478	0,8678 ± 0,0231	0,1322 ± 0,0231	0,6286 ± 0,0386	0,5882 ± 0,0365
FMN-SAL	0,8806 ± 0,0043	0,7461 ± 0,0124	0,9346 ± 0,0062	0,0654 ± 0,0062	0,8210 ± 0,0130	0,7817 ± 0,0077
FMN-Voting	0,8045 ± 0,0366	0,6347 ± 0,0576	0,8726 ± 0,0289	0,1274 ± 0,0289	0,6676 ± 0,0702	0,6506 ± 0,0631

tốt hơn giữa hai lớp.

Bảng 2.12 trình bày kết quả đánh giá tổng thể trên hai lớp theo trung bình có trọng số, trong đó Precision, Recall và F1-Score được tổng hợp theo trung bình tương ứng với số lượng mẫu của từng lớp; kết quả trên từng lớp được trình bày riêng trong Bảng 2.10 và Bảng 2.11. Đối với bài toán hai lớp đơn nhãn, Recall trung bình có trọng số bằng tỷ lệ tổng số mẫu được gán đúng trên toàn bộ tập dữ liệu, tức là tương đương với Accuracy; do đó hai độ đo này có cùng giá trị trung bình và độ lệch chuẩn ở tất cả các phương pháp trong Bảng 2.12. FMN-SAL đạt Accuracy (cũng là Recall trung bình có trọng số) bằng $0{,}8806 \pm 0{,}0043$, Specificity bằng $0{,}8001 \pm 0{,}0082$, Precision bằng $0{,}8786 \pm 0{,}0044$ và F1-Score bằng $0{,}8788 \pm 0{,}0043$.

FMN-Voting đạt Accuracy bằng $0{,}8045 \pm 0{,}0366$ và F1-Score bằng $0{,}8030 \pm 0{,}0364$, cao hơn SCFMN với Accuracy bằng $0{,}7780 \pm 0.0184$ và F1-Score bằng $0{,}7736 \pm 0{,}0186$. Các phương pháp K-means++, EM và DBSCAN có Accuracy chỉ nằm trong khoảng từ $0{,}5647$ đến $0{,}6416$, thấp hơn các phương pháp dựa trên FMN.

Các kết quả này cho thấy việc khai thác cấu trúc siêu hộp kết hợp với thông tin giám sát trong FMN-SAL và FMN-Voting góp phần cải thiện hiệu quả phân loại trên tập dữ liệu ILPD.

Bên cạnh các độ đo đánh giá khả năng phân loại, mức độ phù hợp giữa cấu trúc phân cụm được hình thành bởi các phương pháp và phân hoạch theo nhãn thực tế cũng được đánh giá thông qua các chỉ số RI, ARI và NMI. Kết quả chi tiết được trình bày trong Bảng 2.13, đồng thời được minh họa trực quan trong Hình 2.26.

Bảng 2.12: *Kết quả đánh giá tổng thể của các mô hình đối sánh trên tập dữ liệu ILPD*

Mô hình	Accuracy	Recall	Specificity	FP Rate	Precision	F1-Score
K-means++	0,6252 ± 0,0222	0,6252 ± 0,0222	0,4463 ± 0,0932	0,5537 ± 0,0932	0,6148 ± 0,0437	0,6154 ± 0,0240
EM	0,5647 ± 0,0351	0,5647 ± 0,0351	0,7353 ± 0,1299	0,2647 ± 0,1299	0,7589 ± 0,0544	0,5648 ± 0,0249
DBSCAN	0,6416 ± 0,0022	0,6416 ± 0,0022	0,4411 ± 0,0048	0,5589 ± 0,0048	0,6229 ± 0,0021	0,6306 ± 0,0019
SCFMN	0,7780 ± 0,0184	0,7780 ± 0,0184	0,6442 ± 0,0330	0,3558 ± 0,0330	0,7718 ± 0,0185	0,7736 ± 0,0186
FMN-SAL	0,8806 ± 0,0043	0,8806 ± 0,0043	0,8001 ± 0,0082	0,1999 ± 0,0082	0,8786 ± 0,0044	0,8788 ± 0,0043
FMN-Voting	0,8045 ± 0,0366	0,8045 ± 0,0366	0,7029 ± 0,0490	0,2971 ± 0,0490	0,8021 ± 0,0365	0,8030 ± 0,0364

Kết quả trong Bảng 2.13 và Hình 2.26 cho thấy sự khác biệt rõ rệt giữa các phương pháp về mức độ tương đồng giữa cấu trúc phân cụm thu được và phân hoạch theo nhãn thực tế. Trong đó, FMN-SAL đạt kết quả cao nhất trên cả ba chỉ số với RI bằng $0{,}7894 \pm 0{,}0066$, ARI bằng $0{,}5606 \pm 0{,}0135$ và NMI bằng $0{,}4077 \pm 0{,}0137$. Bên cạnh các giá trị trung bình cao, độ lệch chuẩn nhỏ cho thấy kết quả phân cụm của FMN-SAL có tính ổn định tốt qua các lần thực nghiệm.

FMN-Voting đạt RI bằng $0{,}6870 \pm 0{,}0458$, ARI bằng $0{,}3499 \pm 0{,}0942$ và NMI bằng $0{,}2189 \pm 0.0793$. Các giá trị này thấp hơn FMN-SAL nhưng vẫn cao hơn SCFMN và các phương pháp phân cụm K-means++, EM và DBSCAN trên phần lớn các chỉ số. SCFMN đạt RI, ARI và NMI lần lượt là $0{,}6546 \pm 0{,}0208$, $0{,}2772 \pm 0{,}0432$ và $0{,}1570 \pm 0{,}0334$, cho thấy việc khai thác cấu trúc siêu hộp giúp tăng mức độ tương đồng giữa kết quả phân cụm và nhãn thực tế so với các phương pháp phân cụm truyền thống.

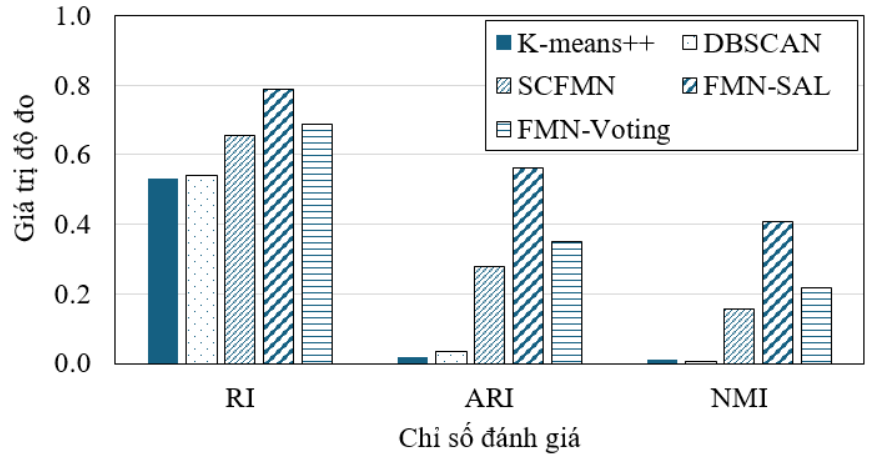
Đối với các phương pháp thực hiện trên Weka, K-means++ và DBSCAN có RI xấp xỉ 0,53 nhưng ARI và NMI đều ở mức rất thấp. Điều này cho thấy mặc dù vẫn tồn tại một tỷ lệ nhất định các cặp mẫu được phân nhóm phù hợp, cấu trúc phân cụm tổng thể của hai phương pháp chưa tương ứng tốt với hai lớp bệnh gan và không bệnh gan. EM đạt NMI bằng $0{,}0926 \pm 0{,}0343$, cao hơn K-means++ và DBSCAN, tuy nhiên ARI có giá trị trung bình âm nhẹ ($-0{,}0130 \pm 0{,}0144$), cho thấy mức độ phù hợp giữa phân hoạch do EM tạo ra và nhãn thực tế xấp xỉ mức kỳ vọng ngẫu nhiên.

Nhìn chung, kết quả của RI, ARI và NMI cho thấy các mô hình dựa trên FMN tạo ra cấu trúc phân cụm phù hợp hơn với phân hoạch bệnh gan và không bệnh gan

so với các phương pháp phân cụm truyền thống. Trong nhóm so sánh, FMN-SAL thể hiện kết quả nổi bật nhất, trong khi FMN-Voting cũng cho thấy sự cải thiện so với SCFMN và các phương pháp được thực hiện trên Weka.

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

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



Hình 2.26: *Kết quả so sánh FMN-SAL, FMN-Voting với các phương pháp khác*

2.6. Kết luận chương 2

Chương 2 đã trình bày hệ thống cơ sở lý thuyết và các mô hình phân cụm bán giám sát được đề xuất dựa trên FMN. Trên cơ sở phân tích những hạn chế của các phương pháp dựa trên FMN truyền thống trong bối cảnh dữ liệu có phân bố phức tạp và số lượng mẫu mang nhãn hạn chế, mô hình FMN-Voting và FMN-SAL đã được xây dựng với các cơ chế học cải tiến nhằm nâng cao chất lượng phân cụm.

Cụ thể, mô hình FMN-Voting khai thác cơ chế bỏ phiếu dựa trên các siêu hộp lân cận, qua đó hạn chế sự phụ thuộc của quyết định gán nhãn vào một siêu hộp đơn lẻ và nâng cao độ ổn định của quá trình suy luận nhãn. Trong khi đó, mô hình FMN-SAL sử dụng chiến lược suy luận nhãn dựa trên các mẫu dữ liệu, cho phép tận dụng thông tin cấu trúc không gian của dữ liệu nhằm cải thiện khả năng lan truyền nhãn và tăng độ tin cậy trong các vùng dữ liệu có mật độ không đồng đều hoặc chồng lấn. Các thuật toán học tương ứng với từng mô hình đã được mô tả chi tiết, kèm theo phân tích độ phức tạp tính toán, cho thấy các phương pháp đề xuất phù hợp với bài toán thực tế.

Bên cạnh đó, chương cũng đã xây dựng thực nghiệm nhằm đánh giá hiệu năng của các mô hình trên nhiều tập dữ liệu với đặc điểm phân bố đa dạng.

Kết quả thực nghiệm cho thấy FMN-Voting và FMN-SAL đạt kết quả cạnh tranh và cải thiện hiệu năng trên nhiều tập dữ liệu được khảo sát, tuy nhiên mức độ cải thiện phụ thuộc vào đặc điểm của từng tập dữ liệu. FMN-Voting khai thác thông tin ở mức siêu hộp và phù hợp hơn khi các siêu hộp lân cận có tính đại diện tốt đối với phân bố cục bộ. FMN-SAL khai thác trực tiếp các mẫu mang nhãn lân cận, qua đó giảm sự phụ thuộc vào tâm hình học và mức độ đại diện của siêu hộp. Ngược lại, hiệu quả của FMN-SAL phụ thuộc nhiều hơn vào số lượng và tính đại diện của các mẫu mang nhãn. Kết quả trên PID và Sonar, nơi SCFMN đạt Accuracy cao hơn nhẹ FMN-SAL, cũng cho thấy không có một cơ chế suy luận nhãn duy nhất chiếm ưu thế trên mọi tập dữ liệu.

Một phần kết quả nghiên cứu trình bày trong Chương 2 đã được công bố trong các công trình [CT2], [CT3], [CT4] và [CT5]. Trong đó, [CT3] liên quan đến mô hình FMN-Voting; [CT2] và [CT5] liên quan đến mô hình FMN-SAL và [CT4] trình bày việc áp dụng FMN-SAL cho bài toán phân vùng khối u phổi. Các công trình [CT2], [CT3] và [CT4] công bố tại kỷ yếu Advances in Information and Communication Technology, thuộc chuỗi Lecture Notes in Networks and Systems của Springer, tương ứng với các hội nghị quốc tế ICTA 2023, ICTA 2024 và ICTA 2025; [CT5] được công bố trong kỷ yếu *Intelligent Systems and Data Science*, thuộc chuỗi *Communications in Computer and Information Science* của Springer.

Những kết quả đạt được trong chương 2 cho thấy tính hợp lý của các mô hình đề xuất về mặt lý thuyết và hiệu quả thông qua thực nghiệm. Đây là cơ sở để triển khai các nghiên cứu mở rộng và đánh giá chuyên sâu hơn trong các chương tiếp theo, đặc biệt là việc so sánh với các phương pháp hiện đại và phân tích chi tiết ảnh hưởng của các tham số học đến hiệu năng của mô hình.

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

Trong Chương 2, luận án đã trình bày hai mô hình phân cụm bán giám sát dựa trên mạng FMN là FMN-Voting và FMN-SAL, với cơ chế suy luận nhãn bằng bỏ phiếu nhằm giảm sự phụ thuộc vào các siêu hộp hoặc mẫu đơn lẻ. Tuy nhiên, hiệu quả của quá trình học vẫn phụ thuộc đáng kể vào chất lượng của các hạt giống ban đầu.

Trong không gian dữ liệu phức tạp, không phải mọi mẫu hoặc siêu hộp mang nhãn đều có mức độ đại diện và độ tin cậy như nhau. Các hạt giống kém chất lượng, đặc biệt ở vùng biên cụm hoặc vùng nhiễu, có thể làm lan truyền lỗi nhãn và làm giảm chất lượng phân cụm. Vì vậy, việc lựa chọn hạt giống một cách phù hợp là yêu cầu quan trọng trong học bán giám sát.

Xuất phát từ đó, chương này đề xuất mô hình FA-Seed, một mô hình học bán giám sát dựa trên mạng nơron FMN với chiến lược lựa chọn hạt giống chủ động. Mục tiêu là lựa chọn các hạt giống có chất lượng cao để dẫn dắt quá trình học, qua đó nâng cao độ ổn định, hạn chế lan truyền lỗi và cải thiện hiệu năng phân cụm trong điều kiện nhãn khan hiếm.

3.1. Giới thiệu về mô hình FA-Seed

Trong luận án, hạt giống (seed) được hiểu là mẫu dữ liệu đại diện được lựa chọn và xác định nhãn để cung cấp thông tin giám sát ban đầu cho quá trình phân cụm bán giám sát. Tập hạt giống là một tập con của dữ liệu huấn luyện, được sử dụng để định hướng quá trình hình thành cụm và lan truyền nhãn. Chất lượng của hạt giống phụ thuộc vào mức độ đại diện của mẫu đối với vùng dữ liệu tương ứng và độ tin cậy của thông tin nhãn được cung cấp.

Trong phạm vi luận án, độ tin cậy của một siêu hộp phản ánh mức độ có thể sử dụng siêu hộp đó làm nguồn thông tin để lựa chọn hạt giống và hỗ trợ quá trình suy luận nhãn. Độ tin cậy được xem xét thông qua các đặc trưng của siêu hộp như mật độ dữ liệu, mức độ sai lệch giữa tâm hình học và tâm dữ liệu, và mức độ thuộc của các mẫu bên trong siêu hộp. Tính đại diện phản ánh mức độ mà một siêu hộp hoặc một mẫu dữ liệu đặc trưng cho phân bố cục bộ của vùng dữ liệu mà nó bao phủ. Siêu hộp có phân bố dữ liệu tập trung và độ lệch nhỏ giữa tâm hình học và tâm dữ liệu được xem là có tính đại diện tốt hơn.

Phân cụm bán giám sát mờ dựa trên mẫu hạt giống là một kỹ thuật phổ biến, trong đó một tập nhỏ các mẫu đã gán nhãn được sử dụng để định hướng quá trình phân cụm. Các hạt giống thường được chọn sao cho có tính đại diện và độ tin cậy cao, chẳng hạn các điểm gần tâm cụm hoặc được chuyên gia xác nhận.

Tuy nhiên, nhiều phương pháp SSC truyền thống vẫn phụ thuộc mạnh vào chất lượng hạt giống ban đầu. Trong nhiều trường hợp, hạt giống được chọn ngẫu nhiên hoặc theo tiêu chí đơn giản, nên không bảo đảm phản ánh đúng cấu trúc tổng thể của dữ liệu. Khi thiếu cơ chế đánh giá độ tin cậy, các hạt giống kém chất lượng vẫn có thể tham gia lan truyền nhãn, từ đó làm giảm chất lượng phân cụm. Bên cạnh đó, nếu quá trình lan truyền nhãn không được kiểm soát chặt chẽ, sai lệch ban đầu dễ bị khuếch đại và gây lan truyền lỗi.

Trong bối cảnh đó, mạng nơron FMN là một mô hình phù hợp cho bài toán học bán giám sát nhờ khả năng biểu diễn dữ liệu bằng các siêu hộp mờ, mô hình hóa tự nhiên các vùng chồng lấn và thích nghi linh hoạt với dữ liệu mới. Trên nền tảng này, mô hình GFMM [15] và RFMN [18] cho phép khai thác thông tin nhãn để giám sát một phần quá trình phân cụm. Tuy nhiên, khi một siêu hộp mới được tạo ra từ dữ liệu không nhãn, siêu hộp đó có thể không được gán nhãn trong suốt quá trình huấn luyện, đặc biệt khi dữ liệu có nhãn khan hiếm hoặc phân bố không đồng đều. Điều này làm giảm khả năng diễn giải và có thể ảnh hưởng đến các bước học tiếp theo.

Mô hình SS-FMM [19] được đề xuất để khắc phục hạn chế này bằng cách xử lý trước các mẫu có nhãn và sử dụng chúng để lan truyền nhãn sang các siêu hộp hình thành từ dữ liệu chưa nhãn. Cách tiếp cận này giúp giảm số lượng siêu hộp không nhãn, nhưng vẫn coi các mẫu có nhãn như các hạt giống mặc định và chưa đánh giá mức độ đại diện hay độ tin cậy của chúng. Tương tự, SSL-FMM [17] khai thác dữ liệu chưa gán nhãn hiệu quả hơn để cải thiện biên quyết định, song cơ chế lựa chọn hạt giống và lan truyền nhãn vẫn mang tính toàn cục, chưa gắn với chất lượng từng siêu hộp.

Khác với các mô hình trên, SCFMN [20], MSCFMN [21], và SSL-FMM [98] tập trung vào lựa chọn hạt giống dựa trên thông tin đặc trưng của dữ liệu. Trong SCFMN và MSCFMN, các điểm nằm ở vùng mật độ cao hoặc có khả năng đại diện tốt cho trung tâm cụm được ưu tiên chọn làm hạt giống, qua đó giúp giảm phụ thuộc vào thứ tự mẫu và tăng độ ổn định của kết quả. MSCFMN còn mở rộng theo hướng đa hạt giống, cho phép mô tả tốt hơn các cụm có hình dạng phức tạp hoặc nhiều vùng mật độ khác nhau.

Mặc dù vậy, các mô hình này vẫn gắn với giả định phải xác định trước số cụm và xem mỗi siêu hộp như đại diện cho một cụm dữ liệu. Giả định này phù hợp với các cụm có hình dạng tương đối đơn giản, nhưng trở nên hạn chế khi dữ liệu có cấu trúc không lồi, mật độ không đồng đều hoặc kích thước cụm khác biệt lớn. Khi đó, việc ép mỗi siêu hộp tương ứng với một cụm có thể làm mất đi cấu trúc hình học thực của dữ liệu và làm giảm chất lượng phân cụm.

Từ những hạn chế trên, nghiên cứu sinh đề xuất mô hình FA-Seed nhằm kết hợp chiến lược lựa chọn hạt giống chủ động với khả năng phân hoạch thích nghi của FMN. Mô hình này cho phép biểu diễn cụm bằng nhiều siêu hộp mờ, không bị ràng buộc bởi giả định cụm hình cầu hay kích thước cân bằng. FA-Seed thực hiện phân vùng dữ liệu thành các siêu hộp, đánh giá độ tin cậy của hạt giống thông qua các phép đo mật độ thành viên và lan truyền nhãn. Những đóng góp chính của FA-Seed gồm: (1) tự động lựa chọn hạt giống ứng viên, (2) mở rộng cụm động không cần huấn luyện lại, (3) phát hiện các vùng có cấu trúc phức tạp dựa trên đặc trưng cụm, và (4) nắm bắt tốt cấu trúc cụm ngay cả khi các cụm khác nhau về mật độ và hình dạng.

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

3.2.1. Mô tả bài toán

Giả sử ta có tập dữ liệu X gồm M mẫu dữ liệu ($X = \{x_r\}_{r=1}^{M} \subset \mathbb{R}^n$), trong đó mỗi $x_r \in X$ có thể được ánh xạ vào C cụm. Tập chỉ số $\{1, \ldots, M\}$ được phân hoạch thành hai tập con: tập có nhãn L và tập chưa có nhãn U, sao cho

$$L \cup U = \{1, \ldots, M\}, \qquad L \cap U = \varnothing, \tag{3.1}$$

trong đó sự phân chia này được hình thành thông qua việc phân hoạch không gian đầu vào thành các siêu hộp mờ. Tập siêu hộp mờ $\mathcal{B}$ được định nghĩa theo Công thức (1.1) ($K = |\mathcal{B}|$), siêu hộp B_j ($B_j \in \mathcal{B}$) được định nghĩa theo Công thức (1.2).

Các tập siêu hộp con có nhãn và chưa có nhãn lần lượt được ký hiệu là $\mathcal{L}$ và $\mathcal{U}$, thỏa mãn

$$\mathcal{B} = \mathcal{L} \uplus \mathcal{U} \iff \begin{cases} \mathcal{L} \cup \mathcal{U} = \mathcal{B}, \\ \mathcal{L} \cap \mathcal{U} = \varnothing. \end{cases} \tag{3.2}$$

Mỗi siêu hộp B_j được gán với một nhãn lớp $c_j \in \{1, \ldots, C\}$ trong quá trình học. Đối với $i \in \mathcal{L}$, nhãn lớp $y_i \in \{1, \ldots, C\}$ là đã biết; trong khi đó, đối với $i \in \mathcal{U}$, nhãn dự đoán được ký hiệu là $\hat{y}_i$.

Gọi x là một mẫu ứng viên ($x \in X$), độ thuộc của x đối với siêu hộp B_j, ký

hiệu là $\mu(x, B_j)$, được tính toán theo Công thức (1.3) ($\mu(x, B_j) \in [0, 1]$).

Với tập $\mathcal{B} = \{B_j\}_{j=1}^{K}$, mẫu x có mức độ phù hợp cao nhất đối với B_j được xác định dựa trên giá trị độ thuộc được xác định theo Công thức (3.3).

$$j^* = \arg \max_{j \in \{1,\ldots,K\}} \mu(x, B_j), \tag{3.3}$$

trong đó:

- j^* là chỉ số của siêu hộp mờ có giá trị độ thuộc lớn nhất đối với mẫu x;
- $\arg\max$ là toán tử xác định giá trị của j tại đó hàm đạt giá trị lớn nhất;
- $j \in \{1, \ldots, K\}$ là tập chỉ số của các siêu hộp hiện có, với K là tổng số siêu hộp;
- $\mu(x, B_j) \in [0, 1]$ là độ thuộc của mẫu x đối với siêu hộp mờ B_j;
- B_j là siêu hộp mờ thứ j trong tập các siêu hộp $\mathcal{B}$;
- x là điểm dữ liệu ứng viên cần được gán nhãn hoặc phân lớp.

Trên cơ sở các định nghĩa trên, bài toán trong mô hình FA-Seed là phân cụm bán giám sát khi dữ liệu huấn luyện ban đầu hầu như chưa có nhãn, trong đó việc xây dựng tập hạt giống giữ vai trò quyết định đối với chất lượng học. Khác với cách giả định tập hạt giống được cho trước hoặc chọn ngẫu nhiên, FA-Seed xem việc xác định và mở rộng hạt giống là một phần của quá trình học.

Cụ thể, từ tập dữ liệu chưa gán nhãn X, FA-Seed phân hoạch không gian dữ liệu thành các siêu hộp để nắm bắt cấu trúc phân bố cục bộ. Trên cơ sở đó, mô hình xác định tập hạt giống mục tiêu $\mathcal{S}_{\text{rep}} \subset X$ có kích thước nhỏ nhưng mang tính đại diện cao, thông qua đánh giá ứng viên và truy vấn chủ động ý kiến chuyên gia.

Từ tập hạt giống mục tiêu đã được xác nhận, FA-Seed tiếp tục mở rộng có kiểm soát để xây dựng tập hạt giống mở rộng $\mathcal{S}_{\text{exp}}$, trong đó chỉ các mẫu có độ thuộc cao đối với siêu hộp tương ứng mới được bổ sung. Tập này sau đó được sử dụng để dẫn dắt quá trình huấn luyện bán giám sát của mạng FMN.

Do đó, bài toán của FA-Seed không chỉ là gán nhãn hoặc phân cụm cho toàn bộ tập dữ liệu X, mà còn 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 trên toàn bộ tập dữ liệu.

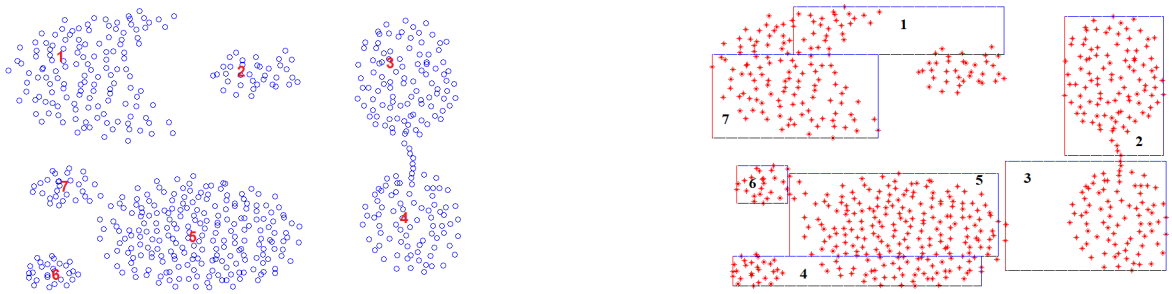
3.2.2. Ý tưởng của FA-Seed

Như đã trình bày, với kỹ thuật học bán giám sát dựa trên hạt giống, hiệu quả của SSC phụ thuộc lớn vào chất lượng hạt giống. Các hạt giống kém chất lượng dễ làm lan truyền sai lệch, trong khi các hạt giống đáng tin cậy giúp giảm chi phí gán nhãn và nâng cao độ chính xác phân cụm. Tuy nhiên, việc lựa chọn hạt giống chất

lượng cao trong thực tế không đơn giản do nhiễu dữ liệu, chồng lấn giữa các cụm và lượng thông tin giám sát hạn chế. Các mô hình bán giám sát dựa trên FMN hiện có thường hoặc chọn hạt giống ngẫu nhiên [15, 17, 19], hoặc xác định hạt giống dựa trên siêu hộp với giả thiết "mỗi siêu hộp đại diện cho một cụm" [20, 21].

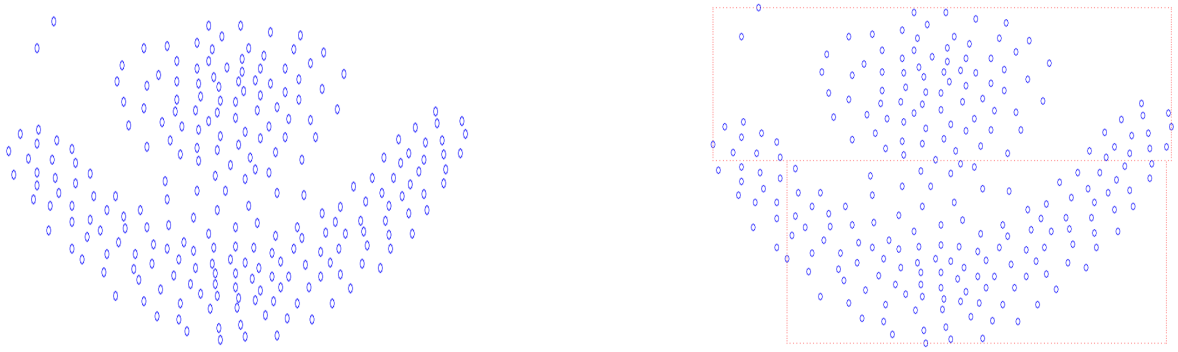
Với chiến lược chọn ngẫu nhiên, hạt giống thường thiếu tính đại diện và độ ổn định: chúng có thể rơi vào vùng biên hoặc vùng chồng lấn, khiến nhãn giám sát ban đầu trở nên mơ hồ và dễ kéo lệch quá trình lan truyền. Đồng thời, cách chọn này không bảo đảm độ phủ cụm, nên một số cụm có thể không có hạt giống hoặc có quá ít hạt giống, làm suy giảm khả năng phát hiện cấu trúc cụm thật.

Các hình từ Hình 3.1 đến Hình 3.3 minh họa rõ hạn chế của giả thiết "mỗi siêu hộp đại diện cho một cụm". Cách tiếp cận này chỉ phù hợp khi các cụm gần cầu và có kích thước tương đối đồng đều. Với các tập dữ liệu có cấu trúc hình học phức tạp như Aggregation, Flame và Pathbased, siêu hộp thường phải mở rộng để bao phủ toàn bộ vùng dữ liệu, dẫn đến chồng lấn nhiều cụm hoặc bao phủ các vùng biên mật độ thấp, làm giảm tính thuần nhất của siêu hộp.



***(a)** Phân bố dữ liệu* ***(b)** Không gian siêu hộp*

***Hình 3.1:** Phân hoạch siêu hộp bằng FMN trên tập dữ liệu Aggregation theo giả thiết mỗi siêu hộp đại diện cho một cụm*

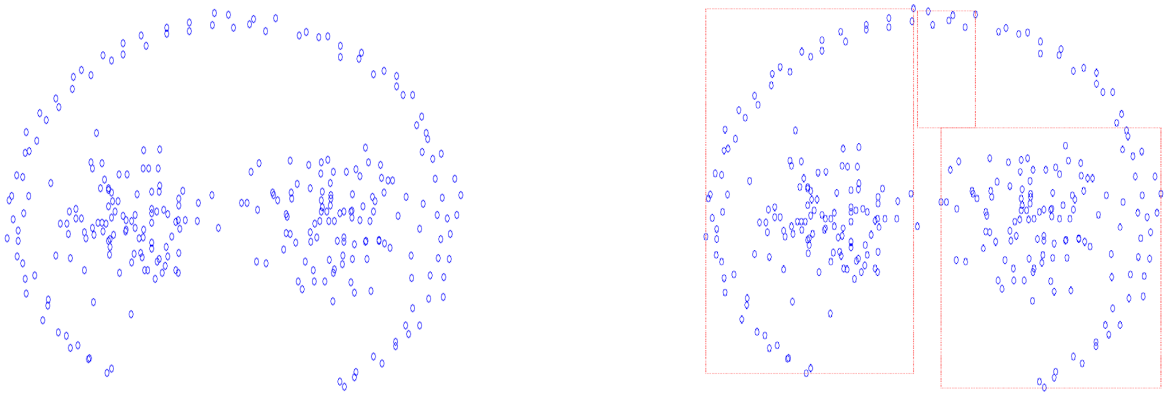


***(a)** Phân bố dữ liệu của tập Flame* ***(b)** Phân hoạch không gian siêu hộp*

***Hình 3.2:** Phân hoạch siêu hộp bằng FMN trên tập dữ liệu Flame theo giả thiết mỗi siêu hộp đại diện cho một cụm*

Cụ thể, trên tập Aggregation, nhiều siêu hộp đồng thời bao phủ các vùng dữ liệu thuộc các cụm khác nhau, nên hạt giống được chọn từ các siêu hộp đó không còn

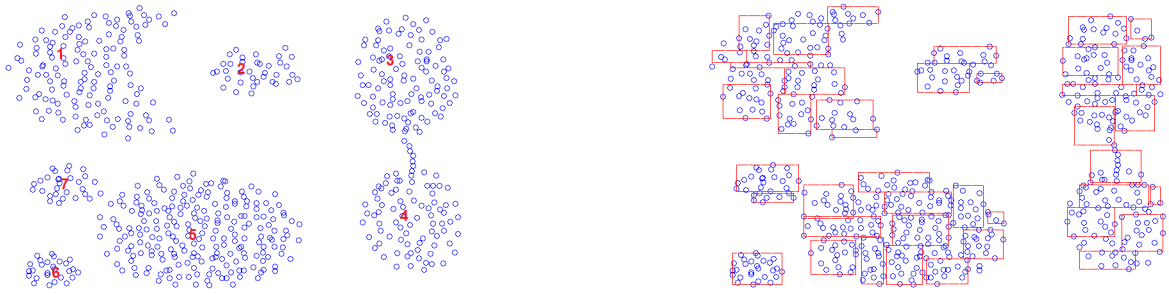
đại diện rõ cho một cụm duy nhất. Tình trạng tương tự xảy ra với tập Flame, nơi các cụm kéo dài và phân bố không đối xứng, khiến một siêu hộp khó mô tả chính xác toàn bộ hình dạng cụm. Với tập Pathbased, các cụm có cấu trúc phi cầu và uốn cong, nên việc ép mỗi cụm vào một siêu hộp làm mất thông tin hình học cục bộ và khiến các hạt giống được chọn kém tin cậy.



(a) *Phân bố dữ liệu* ***(b)*** *Phân hoạch siêu hộp*

Hình 3.3: *Phân hoạch siêu hộp bằng FMN trên tập dữ liệu Pathbased theo giả thiết mỗi siêu hộp đại diện cho một cụm*

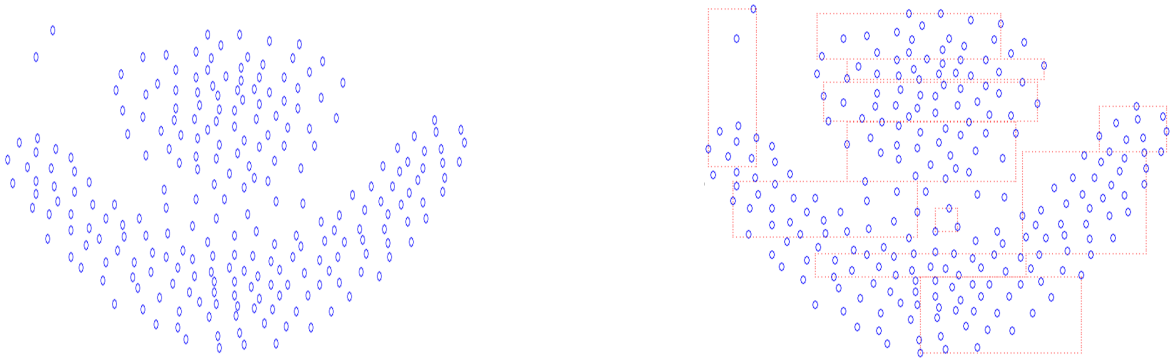
Để khắc phục các hạn chế trên, mô hình FA-Seed [CT6] được đề xuất bằng cách kết hợp phân hoạch siêu hộp, truy vấn chủ động và chiến lược học với số lượng hạt giống hạn chế. Khác với các mô hình trước, ở giai đoạn xác định hạt giống, FA-Seed sử dụng giới hạn kích thước siêu hộp nhỏ hơn, từ đó tạo ra nhiều siêu hộp hơn trong không gian dữ liệu. Các hình từ Hình 3.4 đến Hình 3.6 cho thấy rõ đặc điểm này: trên tập Aggregation mô hình tạo 41 siêu hộp, trên tập Flame tạo 11 siêu hộp, và trên tập Pathbased tạo 19 siêu hộp.



(a) *Phân bố dữ liệu của tập Aggregation* ***(b)*** *Phân hoạch không gian dữ liệu thành 41 siêu hộp*

Hình 3.4: *Phân hoạch siêu hộp trên tập dữ liệu Aggregation với giới hạn kích thước siêu hộp nhỏ của FA-Seed (41 siêu hộp)*

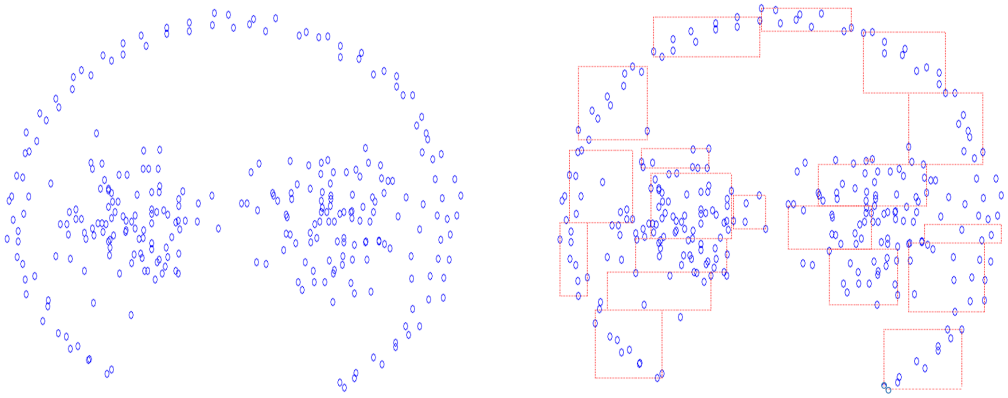
FA-Seed xác định hạt giống trên cơ sở phân vùng siêu hộp sau khi đánh giá các siêu hộp ứng viên theo mật độ và phân bố dữ liệu, từ đó lựa chọn các siêu hộp chứa mẫu mang nhiều thông tin và xác nhận các điểm dữ liệu giàu thông tin với sự hỗ trợ của chuyên gia. Kiến trúc của mô hình gồm hai giai đoạn chính: (i) xây dựng tập



(a) *Phân bố dữ liệu của tập Flame*

(b) *Phân hoạch không gian dữ liệu thành 11 siêu hộp*

Hình 3.5: *Phân hoạch siêu hộp trên tập dữ liệu Flame với giới hạn kích thước siêu hộp nhỏ của FA-Seed (11 siêu hộp)*



(a) *Phân bố dữ liệu của tập Path-based*

(b) *Phân hoạch không gian dữ liệu thành 19 siêu hộp*

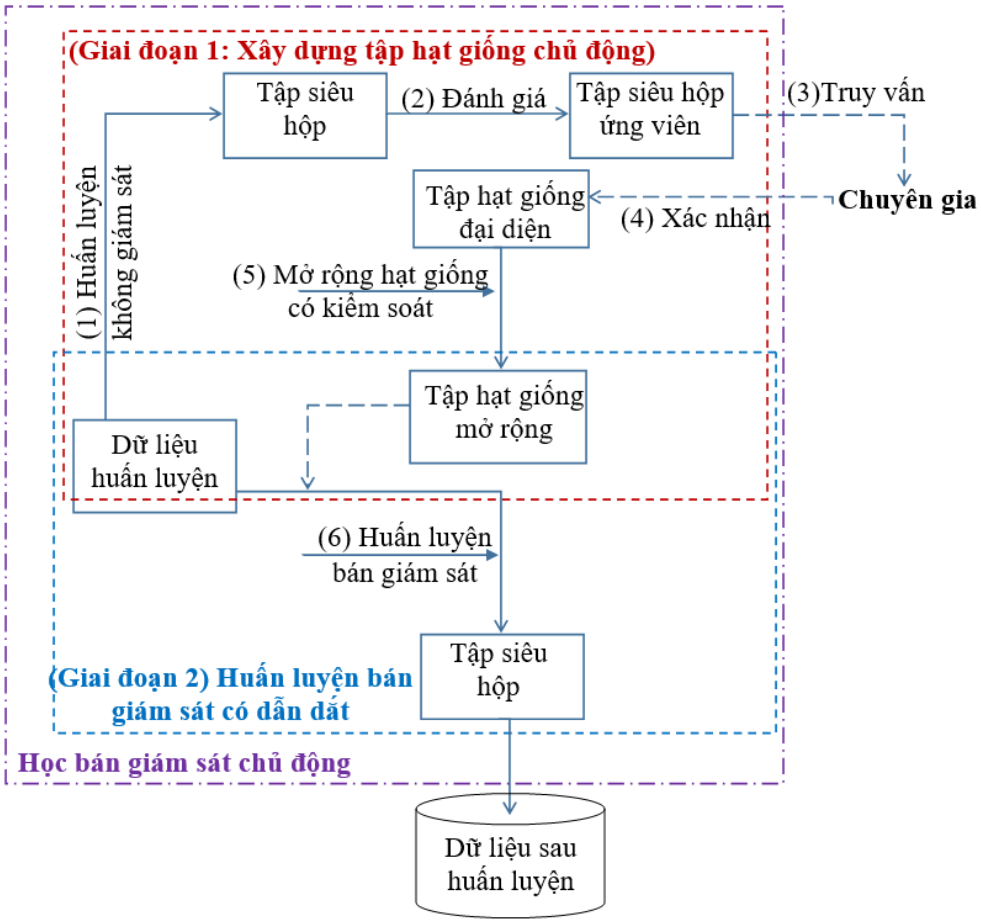
Hình 3.6: *Phân hoạch siêu hộp trên tập dữ liệu Pathbased với giới hạn kích thước siêu hộp nhỏ của FA-Seed (19 siêu hộp)*

hạt giống chủ động và (ii) huấn luyện bán giám sát có dẫn dắt (Hình 3.7).

Kiến trúc minh họa trên Hình 3.7 cho thấy, ở giai đoạn thứ nhất, dữ liệu được đưa vào quá trình huấn luyện không giám sát FMN để hình thành tập siêu hộp biểu diễn cấu trúc phân bố. Từ đó, một tập siêu hộp ứng viên được đánh giá dựa trên đặc trưng mật độ và phân phối dữ liệu, sau đó được truy vấn chuyên gia để xác nhận nhãn. Kết quả là mô hình tạo được một tập hạt giống đại diện, nhỏ nhưng có độ tin cậy cao, rồi tiếp tục mở rộng có kiểm soát để thu được tập hạt giống mở rộng.

Trong giai đoạn thứ hai, tập hạt giống mở rộng được dùng để dẫn dắt quá trình huấn luyện bán giám sát của mạng FMN. Các hạt giống này cung cấp thông tin giám sát đáng tin cậy cho việc mở rộng, điều chỉnh và tinh chỉnh siêu hộp, đồng thời hỗ trợ khai thác hiệu quả các mẫu chưa gán nhãn còn lại. Nhờ đó, mô hình giảm chi phí gán nhãn, hạn chế lan truyền sai lệch và nâng cao chất lượng phân cụm.

Hình 3.8 minh họa quá trình sinh hạt giống với sự hỗ trợ của chuyên gia trên tập



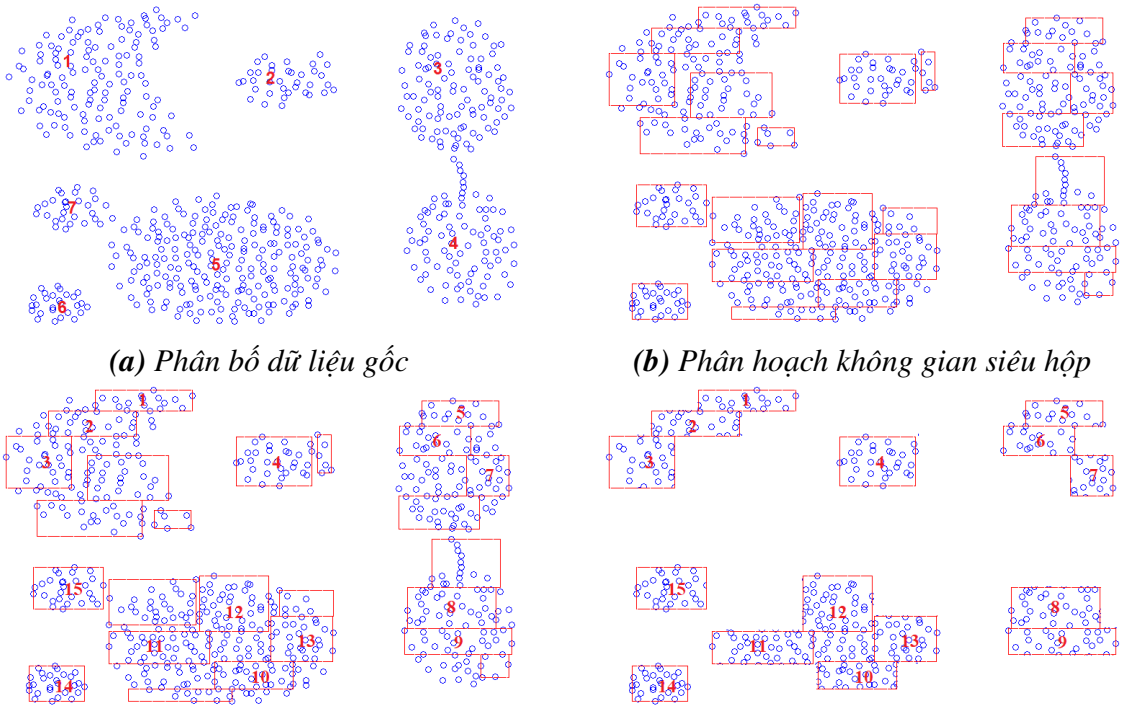
Hình 3.7: *Kiến trúc mô hình phân cụm bán giám sát mờ với học chủ động FA-Seed*

dữ liệu Aggregation. Hình 3.8a thể hiện phân bố dữ liệu gốc gồm 7 cụm; Hình 3.8b minh họa phân hoạch dữ liệu thành 25 siêu hộp; Hình 3.8c cho thấy quá trình đánh giá các siêu hộp ứng viên; và Hình 3.8d minh họa bước lựa chọn siêu hộp ứng viên để sinh hạt giống với sự hỗ trợ của chuyên gia. Trong ví dụ này, 15 trong số 25 siêu hộp được chọn để sinh hạt giống, và các điểm dữ liệu thuộc các siêu hộp đó được gán nhãn để phục vụ giai đoạn học bán giám sát tiếp theo.

Trong đó, việc áp đặt giới hạn kích thước siêu hộp δ^{max} nhỏ hơn giữ vai trò quan trọng trong cả hai giai đoạn của FA-Seed. Ràng buộc này buộc mô hình phân hoạch không gian dữ liệu thành các vùng cục bộ chi tiết hơn, thay vì tạo ra các siêu hộp lớn bao phủ nhiều điểm dữ liệu. Nhờ vậy, các biến động cục bộ về mật độ và cấu trúc phân bố dữ liệu được phản ánh rõ hơn, làm nổi bật các điểm gần ranh giới cụm hoặc tại vùng không chắc chắn, vốn là những mẫu chứa nhiều thông tin để lựa chọn làm hạt giống.

Bên cạnh đó, giới hạn kích thước siêu hộp nhỏ còn giúp cải thiện khả năng nhận diện vùng biên giữa các cụm và phát hiện nhiễu. Những siêu hộp khó mở rộng do ràng buộc kích thước thường tương ứng với các vùng chồng lấn, vùng biên hoặc

điểm bất thường, từ đó cung cấp tín hiệu hữu ích cho việc đánh giá siêu hộp ứng viên và lựa chọn hạt giống. Nhờ cơ chế này, FA-Seed nâng cao chất lượng tập hạt giống, đồng thời hạn chế lan truyền sai lệch trong giai đoạn huấn luyện bán giám sát, qua đó cải thiện độ ổn định và độ chính xác của kết quả phân cụm cuối cùng.



(a) *Phân bố dữ liệu gốc*

(b) *Phân hoạch không gian siêu hộp*

(c) *Đánh giá các siêu hộp ứng viên*

(d) *Lựa chọn các siêu hộp ứng viên*

Hình 3.8: *Sự phân hoạch các siêu hộp, đánh giá và lựa chọn các siêu hộp ứng viên sinh hạt giống trên tập dữ liệu Aggregation*

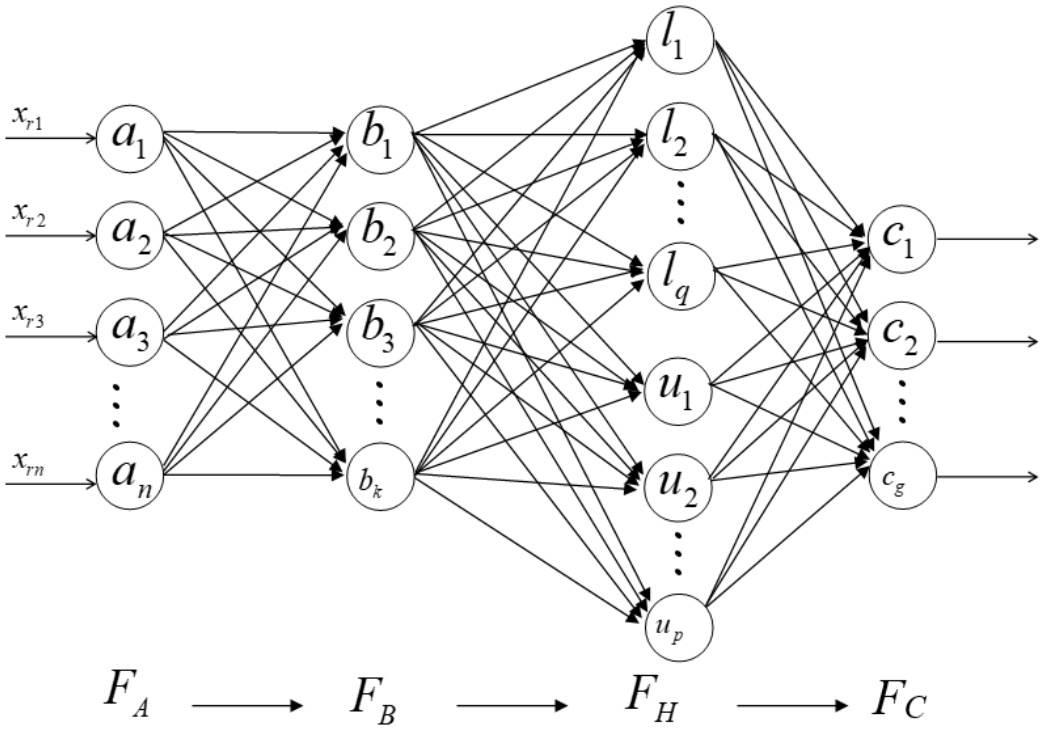
3.2.3. Kiến trúc mạng nơron FA-Seed

Trong đó:

- Lớp đầu vào F_A: Lớp này bao gồm n nút, mỗi nút biểu diễn một thuộc tính đầu vào. Một điểm dữ liệu $x_r = (x_{r1}, x_{r2}, \ldots, x_{rn})$ được ánh xạ vào các siêu hộp tại lớp F_B thông qua các vectơ biên nhỏ nhất (min) v và lớn nhất (max) w.
- Lớp đánh giá siêu hộp F_B: Lớp này bao gồm K nút (tương ứng với K siêu hộp) được tạo ra ở Giai đoạn 1, mỗi nút b_j trong lớp F_B tương ứng với một siêu hộp mờ $B_j = (V_j, W_j) \in \mathcal{B}$. Các siêu hộp này được xây dựng theo cách tăng dần nhằm nắm bắt cấu trúc cục bộ của dữ liệu trong không gian đầu vào. Mức kích hoạt của siêu hộp B_j đối với mẫu x_r được xác định bởi hàm độ thuộc tính toán theo Công thức (3.4):

$$b_j(x_r) = \mu_j(x_r, B_j) \in [0, 1]. \tag{3.4}$$

- Lớp đánh giá và lan truyền nhãn F_H: Các nút siêu hộp trong lớp này được chia thành hai tập: tập siêu hộp đã gán nhãn $\mathcal{B}_{\text{lab}}$ (có kích thước K_{lab}) và tập siêu



Hình 3.9: *Kiến trúc mạng nơron FA-Seed*

hộp chưa gán nhãn $\mathcal{B}_{\text{unl}}$ (có kích thước K_{unl}). Lớp này áp dụng các cơ chế cổng (gates) và quy tắc kế thừa nhãn để chuyển các siêu hộp từ $\mathcal{B}_{\text{unl}}$ sang $\mathcal{B}_{\text{lab}}$. Khi thuật toán kết thúc, $\mathcal{B}_{\text{unl}} = \varnothing$, hay tương đương $K_{\text{unl}} = 0$.

- Lớp đầu ra F_C: Lớp này bao gồm G nút, mỗi nút biểu diễn một cụm đầu ra. Mối quan hệ nhị phân giữa các siêu hộp $L_j \in F_H$ và các lớp $c_g \in F_C$ được biểu diễn thông qua ma trận $U = \{u_{jg}\}$, với:

$$u_{jg} = \begin{cases} 1, & \text{nếu } l_j \in c_g, \\ 0, & \text{ngược lại.} \end{cases} \tag{3.5}$$

Điểm độ thuộc của một mẫu đối với lớp c_g được tính như sau:

$$c_g = \max_{1 \le j \le K} l_j \cdot u_{jg}, \tag{3.6}$$

trong đó K là tổng số siêu hộp, l_j là mức kích hoạt của siêu hộp thứ j. Tập các siêu hộp xác định lớp g được cho bởi:

$$C_g = \bigcup_{j \in G} L_j, \tag{3.7}$$

với G là tập chỉ số của các siêu hộp được gán cho lớp g.

3.2.4. Thuật toán học của mô hình FA-Seed

3.2.4.1. Giai đoạn xây dựng tập hạt giống với huấn luyện chủ động

Huấn luyện chủ động được hiểu là quá trình mô hình chủ động lựa chọn một số mẫu có tính đại diện từ các siêu hộp thỏa mãn tiêu chí lựa chọn để yêu cầu chuyên

gia cung cấp nhãn, thay vì gán nhãn ngẫu nhiên cho các mẫu trong tập dữ liệu. Thông tin nhãn thu được được sử dụng để xây dựng tập hạt giống phục vụ giai đoạn học bán giám sát tiếp theo.

Giai đoạn đầu tiên của thuật toán học xác định các hạt giống dựa trên việc phân hoạch không gian dữ liệu thành các siêu hộp mờ. Các hạt giống ứng viên được đánh giá dựa trên mật độ dữ liệu, độ lệch của trọng tâm và mức độ không chắc chắn (các Công thức (2.2), (1.40), và (3.8)), sau đó các hạt giống đáng tin cậy S sẽ được gán nhãn trong các siêu hộp mờ tương ứng (định nghĩa theo (3.9)). Thuật toán học cho giai đoạn này được trình bày trong Thuật toán 3.1.

$$e(B_j) = \|g(B_j) - d(B_j)\|_2 \tag{3.8}$$

$$S = \bigcup_{j \in J_Q} \{\, x_r \in X \mid \mu(x_r, B_j) \geq \tau \,\}, \qquad \tau \in [0,1] \tag{3.9}$$

Trong đó:

- Tâm dữ liệu $d(B_j)$ của siêu hộp mờ B_j được tính bằng trung bình các mẫu trong tập lõi T_j.
- T_j là tập lõi được sử dụng để tính tâm dữ liệu $d(B_j)$ của siêu hộp mờ B_j, T_j gồm các điểm dữ liệu được định nghĩa theo (3.10):

$$T_j = \{\, r \mid \mu(x_r, B_j) \geq \tau \,\}, \tag{3.10}$$

- $\tau \in [0,1]$ là ngưỡng độ thuộc do người dùng xác định (thường chọn $\tau = 1$ với ý nghĩa x_r có độ thuộc đầy đủ vào B_j).
- Tâm hình học $g(B_j)$ của siêu hộp mờ B_j được xác định dựa trên hai đỉnh của siêu hộp V_j và W_j.
- V_j và W_j lần lượt là các đỉnh nhỏ nhất và lớn nhất của siêu hộp B_j.
- Độ lệch tâm $e(B_j)$ được xác định là khoảng cách Euclid giữa tâm hình học $g(B_j)$ và tâm dữ liệu $d(B_j)$ của siêu hộp B_j.
- ε là tham số ngưỡng do người dùng xác định để đánh giá mức độ "*thuần khiết*" của siêu hộp B_j. Một siêu hộp mờ B_j được xem là thuần khiết nếu độ lệch tâm của nó thỏa mãn điều kiện (2.3):
- J_Q là tập chỉ số các siêu hộp mờ thuần khiết, được xác định bởi:

$$J_Q = \{\, j \mid B_j \in \mathcal{B},\ e(B_j) \leq \varepsilon \,\}. \tag{3.11}$$

- Tập ứng viên hạt giống Q được hình thành từ các tâm dữ liệu tương ứng với các siêu hộp mờ thuần khiết gồm các điểm được định nghĩa bởi (3.12):

$$Q = \{\, d(B_j) \mid j \in J_Q \,\}. \tag{3.12}$$

- S là tập các mẫu được bao phủ hoàn toàn bởi các siêu hộp mờ thuần khiết.

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 (Công thức (3.10)).
    $V_j, W_j \in \mathbb{R}^n$ lần lượt là các đỉnh min và max của B_j;
    Tính *tâm dữ liệu* của B_j theo Công thức (2.2);
    Tính *tâm hình học* của B_j theo Công thức (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 được hình thành từ các tâm dữ liệu tương ứng với các siêu hộp mờ thuần khiết B_j ($B_j \in \mathcal{B}$) gồm các điểm được định nghĩa bởi (3.12) và truy vấn nhãn từ chuyên gia.
Xây dựng tập hạt giống S từ các mẫu dữ liệu nằm hoàn toàn trong ít nhất một siêu hộp thuần khiết B_j ($B_j \in \mathcal{B}$), theo Công thức (3.9);
Xây dựng một hoán vị π của $\{1, \ldots, M\}$ sao cho các phần tử của S được xếp trước, tập S' (các mẫu không thuộc hoàn toàn vào bất kỳ siêu hộp nào) được xếp sau ($S' = X \setminus S$). Hai tập S và S' tạo thành một phân hoạch của tập X;
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)

Mệnh đề 3.1. *Thuật toán 3.1 có độ phức tạp thời gian là*

$$T_1 = \mathcal{O}(nMK), \tag{3.13}$$

trong đó M là tổng số mẫu trong tập dữ liệu huấn luyện, n là số đặc trưng của mỗi mẫu dữ liệu, và K là tổng số siêu hộp mờ được tạo ra trong quá trình huấn luyện.

Chứng minh. Xét Thuật toán 3.1, với $M = |X|$ là số mẫu dữ liệu, n là số chiều đặc trưng và $K = |\mathcal{B}|$ là số siêu hộp mờ được tạo ra.

Bước 1. Thuật toán duyệt toàn bộ tập dữ liệu X để xây dựng tập siêu hộp mờ $\mathcal{B}$. Với mỗi mẫu $x_r \in X$, cần tính độ thuộc đối với tối đa K siêu hộp, mỗi phép tính có độ phức tạp $O(n)$. Do đó, bước này có độ phức tạp

$$O(MKn). \tag{3.14}$$

Bước 2. Thuật toán duyệt qua tập siêu hộp $\mathcal{B}$ để xác định các siêu hộp ứng viên sinh hạt giống. Với mỗi siêu hộp, các phép tính tâm dữ liệu, tâm hình học và độ lệch tâm có độ phức tạp $O(n)$, nên bước này có độ phức tạp

$$O(Kn). \tag{3.15}$$

Bước 3. Thuật toán xây dựng tập hạt giống và sắp xếp lại tập dữ liệu. Quá trình này yêu cầu duyệt qua toàn bộ X, trong đó mỗi mẫu có thể cần kiểm tra với tối đa K siêu hộp, nên có độ phức tạp

$$O(MK). \tag{3.16}$$

Vì vậy, tổng độ phức tạp thời gian của Thuật toán 3.1 là

$$T_1 = O(MKn) + O(Kn) + O(MK). \tag{3.17}$$

Bỏ qua các hạng bậc thấp, suy ra

$$T_1 = O(nMK). \tag{3.18}$$

□

3.2.4.2. Giai đoạn huấn luyện bán giám sát có dẫn dắt

Huấn luyện bán giám sát có dẫn dắt được hiểu là quá trình học bán giám sát trong đó tập hạt giống đã được lựa chọn và xác định nhãn ở giai đoạn huấn luyện chủ động được sử dụng làm thông tin giám sát ban đầu để định hướng quá trình hình thành siêu hộp và lan truyền nhãn cho các mẫu chưa có nhãn. Khác với huấn luyện chủ động, giai đoạn này không tập trung vào việc lựa chọn mẫu để truy vấn nhãn, mà khai thác các nhãn đã thu được nhằm dẫn dắt quá trình học trên phần dữ liệu chưa có nhãn.

Giai đoạn thứ hai của thuật toán học mô tả quá trình lan truyền nhãn từ tập hạt giống đã được xác thực sang các mẫu dữ liệu chưa có nhãn, nhằm đảm bảo tính nhất quán và độ chính xác giữa các cụm. Bằng cách khai thác cấu trúc siêu hộp mờ, giai đoạn này giảm thiểu việc lan truyền nhãn sai thông qua ràng buộc kích thước siêu hộp và ngưỡng độ thuộc β khi gán nhãn cho các siêu hộp chưa có nhãn. Thuật toán cho giai đoạn này được trình bày trong Thuật toán 3.2.

$$j^* = \arg\max_{j \in \mathcal{J}} \mu(x_r, B_j), \quad B_j \in \mathcal{L} \cup \mathcal{U}, \tag{3.19}$$

với

$$\mathcal{J} = \{j \in \{1, \ldots, K\} \mid \delta(x_r, B_j) \leq \delta_2^{max}\}, \quad (K = |\mathcal{L}| + |\mathcal{U}|). \tag{3.20}$$

trong đó, $\mathcal{L}$ và $\mathcal{U}$ lần lượt là tập các siêu hộp mờ đã có nhãn và chưa có nhãn, thỏa

Thuật toán 3.2 Huấn luyện bán giám sát có dẫn dắt

Input: X'; β; δ_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 chỉ số j^* theo (3.19);
    Mở rộng siêu hộp B_{j^*} theo Công thức (3.21);
    if (B_{j^*} có chồng lấn với B_k) với $\ell(B_{j^*}) \neq \ell(B_k)$ **then**
      Điều chỉnh B_{j^*} và B_k theo các Công thức (1.15) đến (1.26);
    end if
  else
    Tạo B_{new}, khởi tạo các điểm V_{new}, W_{new} bởi x_r;
    if ($\ell_{x_r} \neq \varnothing$) **then**
      $\ell(B_{new}) = \ell_{x_r}$; $\mathcal{L} = \mathcal{L} \cup \{B_{new}\}$;
    else
      $\mathcal{U} = \mathcal{U} \cup \{B_{new}\}$;
    end if
  end if
end for
repeat
  for mỗi $B_j \in \mathcal{U}$ **do**
    Tính trọng tâm $g(B_j)$ theo Công thức (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**
  for mỗi $B_j \in \mathcal{U}$ **do**
    Tính trọng tâm $g(B_j)$ theo Công thức (1.40);
    Đặt $k^* \leftarrow \arg\max_k \mu(g(B_j), B_k)$ với $B_k \in \mathcal{L}$; $\ell(B_j) = \ell(B_{k^*})$;
  end for
end if
$\mathcal{B} = \mathcal{L} \cup \mathcal{U}$
return $\mathcal{B}$

mãn (3.2):

$$\mathcal{B} = \mathcal{L} \uplus \mathcal{U}, \qquad |\mathcal{B}| = K.$$

$$B'_{j^*} = \big(\min(V_{j^*}, x),\ \max(W_{j^*}, x)\big). \tag{3.21}$$

$$k^* = \arg\max_{k \in \mathcal{J}_\mathcal{L}} \ \mu(g(B_j), B_k), \quad B_k \in \mathcal{L} \tag{3.22}$$

với

$$\mathcal{J}_\mathcal{L} = \{k \in \{1, \ldots, |\mathcal{L}|\} \mid \mu(g(B_j), B_k) \geq \beta\}. \tag{3.23}$$

Mệnh đề 3.2. *Thuật toán 3.2 có độ phức tạp thời gian là*

$$T_2 = \mathcal{O}(nMK), \tag{3.24}$$

trong đó M là tổng số mẫu trong tập dữ liệu huấn luyện, n là số đặc trưng của mỗi mẫu dữ liệu, và K là tổng số siêu hộp mờ được tạo ra trong quá trình huấn luyện.

Chứng minh. Xét Thuật toán 3.2, với $M = |X|$ là số mẫu dữ liệu, n là số chiều đặc trưng và K là số siêu hộp mờ được tạo ra trong quá trình huấn luyện bán giám sát.

Giai đoạn 1. Thuật toán duyệt qua toàn bộ tập dữ liệu X'. Với mỗi mẫu $x_r \in X'$, cần xác định chỉ số j^* theo Công thức (3.19), tức là tìm siêu hộp có độ thuộc lớn nhất thỏa mãn ràng buộc kích thước. Trong trường hợp xấu nhất, mỗi mẫu phải so sánh với tối đa K siêu hộp, mỗi phép tính có độ phức tạp $O(n)$. Do đó, chi phí của giai đoạn này là

$$O(MKn). \tag{3.25}$$

Giai đoạn 2. Sau khi xác định j^*, thuật toán thực hiện mở rộng siêu hộp hoặc khởi tạo siêu hộp mới. Các phép cập nhật biên, kiểm tra ràng buộc kích thước và điều kiện chồng lấn đều có độ phức tạp $O(n)$ cho mỗi siêu hộp, nên không làm thay đổi bậc độ phức tạp tổng thể.

Giai đoạn 3. Thuật toán tiếp tục gán nhãn cho các siêu hộp chưa nhãn. Trong mỗi vòng lặp, với mỗi siêu hộp trong $\mathcal{U}$ (tối đa K siêu hộp), thuật toán tính trọng tâm và đánh giá độ thuộc với tối đa K siêu hộp trong $\mathcal{L}$, nên chi phí của một vòng lặp là $O(K^2 n)$. Gọi R_β là số vòng lặp của bước này; do β giảm sau mỗi vòng lặp cho đến ngưỡng tin cậy, R_β bị chặn bởi một hằng số không phụ thuộc vào M. Do đó, chi phí của giai đoạn này là $O(R_\beta K^2 n)$; vì $K \leq M$ nên hạng này không vượt quá bậc $O(nMK)$.

Vì vậy, độ phức tạp thời gian của Thuật toán 3.2 là

$$T_2 = O(MKn) + O(R_\beta K^2 n) = O(nMK). \tag{3.26}$$

□

3.2.4.3. Độ phức tạp tính toán của FA-Seed

Thuật toán học trong FA-Seed gồm hai giai đoạn chính.

Giai đoạn 1 xây dựng tập hạt giống thông qua cơ chế học chủ động. Với mỗi mẫu dữ liệu, mô hình đánh giá mức độ đại diện và độ không chắc chắn dựa trên các siêu hộp hiện có. Chi phí tính toán của giai đoạn này tỷ lệ với số mẫu n, số chiều đặc trưng M và số siêu hộp K:

$$T_1 = O(nMK).$$

Ngoài ra, FA-Seed phát sinh chi phí truy vấn nhãn cho q mẫu được chọn trong học chủ động:

$$C_q = O(q).$$

Do $q \ll n$, chi phí này không làm thay đổi bậc độ phức tạp tổng thể.

Giai đoạn 2 thực hiện huấn luyện bán giám sát có dẫn dắt dựa trên tập hạt giống đã xác định. Quá trình này bao gồm mở rộng, kiểm tra chồng lấn và điều chỉnh các siêu hộp, với chi phí tính toán:

$$T_2 = O(nMK).$$

Vì vậy, tổng chi phí tính toán của thuật toán FA-Seed là

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

Trong thực tế, số siêu hộp K thường nhỏ hơn nhiều so với số mẫu, đồng thời số truy vấn nhãn q được giữ ở mức thấp. Do đó, FA-Seed không chỉ giảm chi phí gán nhãn mà còn có tiềm năng mở rộng về mặt độ phức tạp tính toán; việc kiểm chứng trên các tập dữ liệu có quy mô lớn hơn là hướng nghiên cứu tiếp theo.

3.3. Thực nghiệm và đánh giá

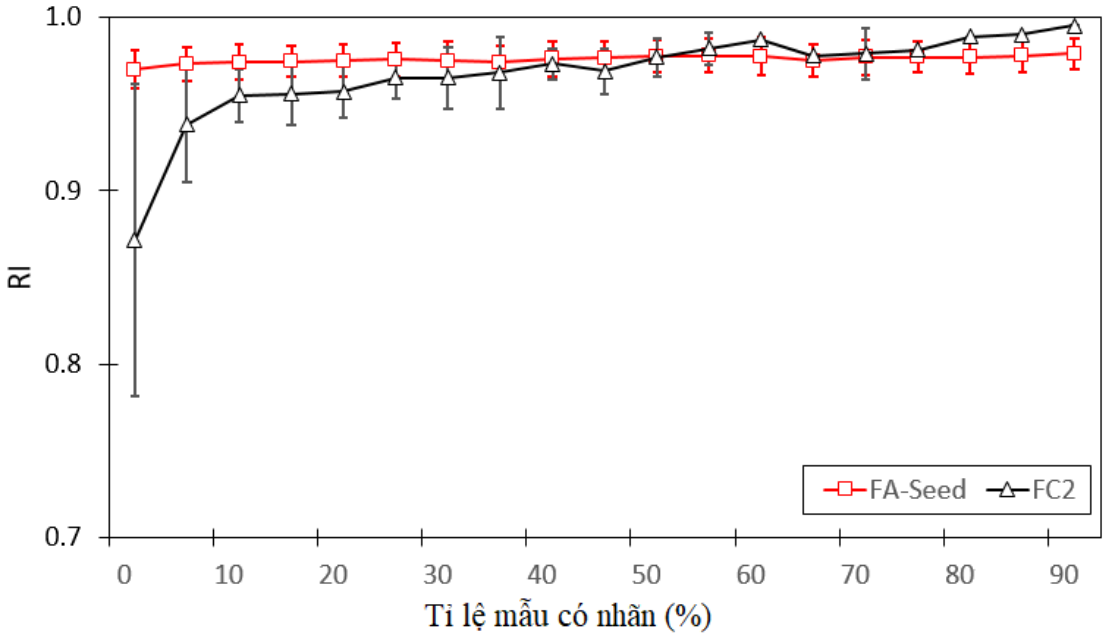
3.3.1. Thiết lập thực nghiệm

Các kết quả thực nghiệm được trình bày trong chương này đều được thực hiện trên cơ sở thiết lập thực nghiệm đã mô tả chi tiết ở Chương 2. Cụ thể, toàn bộ các thí nghiệm tuân theo cùng một quy trình về dữ liệu, phương pháp đánh giá, môi trường thực thi, cơ chế kiểm tra chéo, cách thiết lập tham số và các bước tiền xử lý. Cách tiếp cận này nhằm bảo đảm tính nhất quán trong thực nghiệm, đồng thời tạo cơ sở cho việc đánh giá khách quan và so sánh công bằng hiệu năng giữa các mô hình đề xuất và các phương pháp đối sánh. Ngoài các tập dữ liệu ở Bảng 2.1, chương này sử dụng

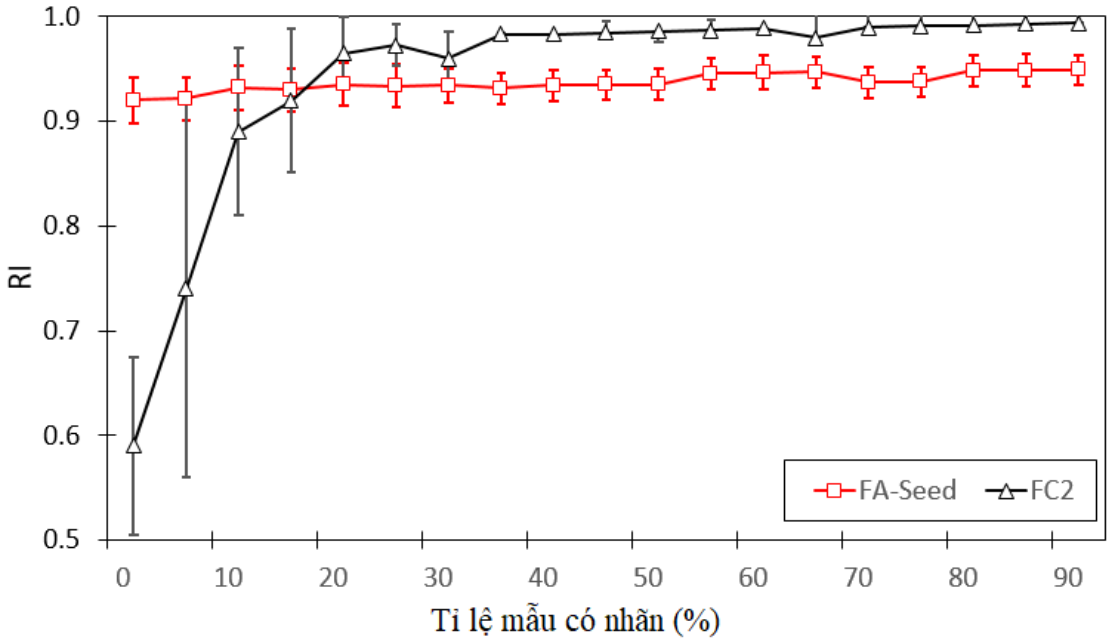
thêm các tập dữ liệu *Ecoli*, *Soybean* và *Zoo* từ kho dữ liệu UCI để đánh giá số lượng hạt giống và chất lượng phân cụm của FA-Seed.

3.3.2. Kết quả thực nghiệm và đánh giá

Đồ thị trong Hình 3.10 và Hình 3.11 minh họa sự so sánh hiệu năng phân cụm giữa phương pháp đề xuất FA-Seed và phương pháp FC2 [102] dựa trên chỉ số RI khi tỷ lệ dữ liệu được gán nhãn thay đổi.



Hình 3.10: *Hiệu năng phân cụm theo chỉ số RI đạt được bởi phương pháp đề xuất FA-Seed và FC2 trên tập dữ liệu Iris*



Hình 3.11: *Hiệu năng phân cụm theo chỉ số RI đạt được bởi phương pháp đề xuất FA-Seed và FC2 trên tập dữ liệu Wine*

Trên cả hai tập dữ liệu Iris và Wine, FA-Seed thể hiện ưu thế trong các thiết lập

có tỷ lệ dữ liệu gán nhãn thấp. Cụ thể, tại các mức nhãn ban đầu thấp (dưới khoảng 20%), FA-Seed đạt giá trị RI trung bình cao hơn đáng kể so với FC2, đồng thời có biên độ dao động nhỏ hơn. Ngược lại, FC2 cho thấy sự biến động lớn của RI giữa các lần chạy, đặc biệt ở các mức nhãn rất thấp, được phản ánh qua các error bar rộng. Điều này cho thấy chiến lược lựa chọn hạt giống đại diện của FA-Seed giúp khai thác hiệu quả thông tin giám sát hạn chế, đồng thời giảm độ nhạy của mô hình đối với khởi tạo ngẫu nhiên.

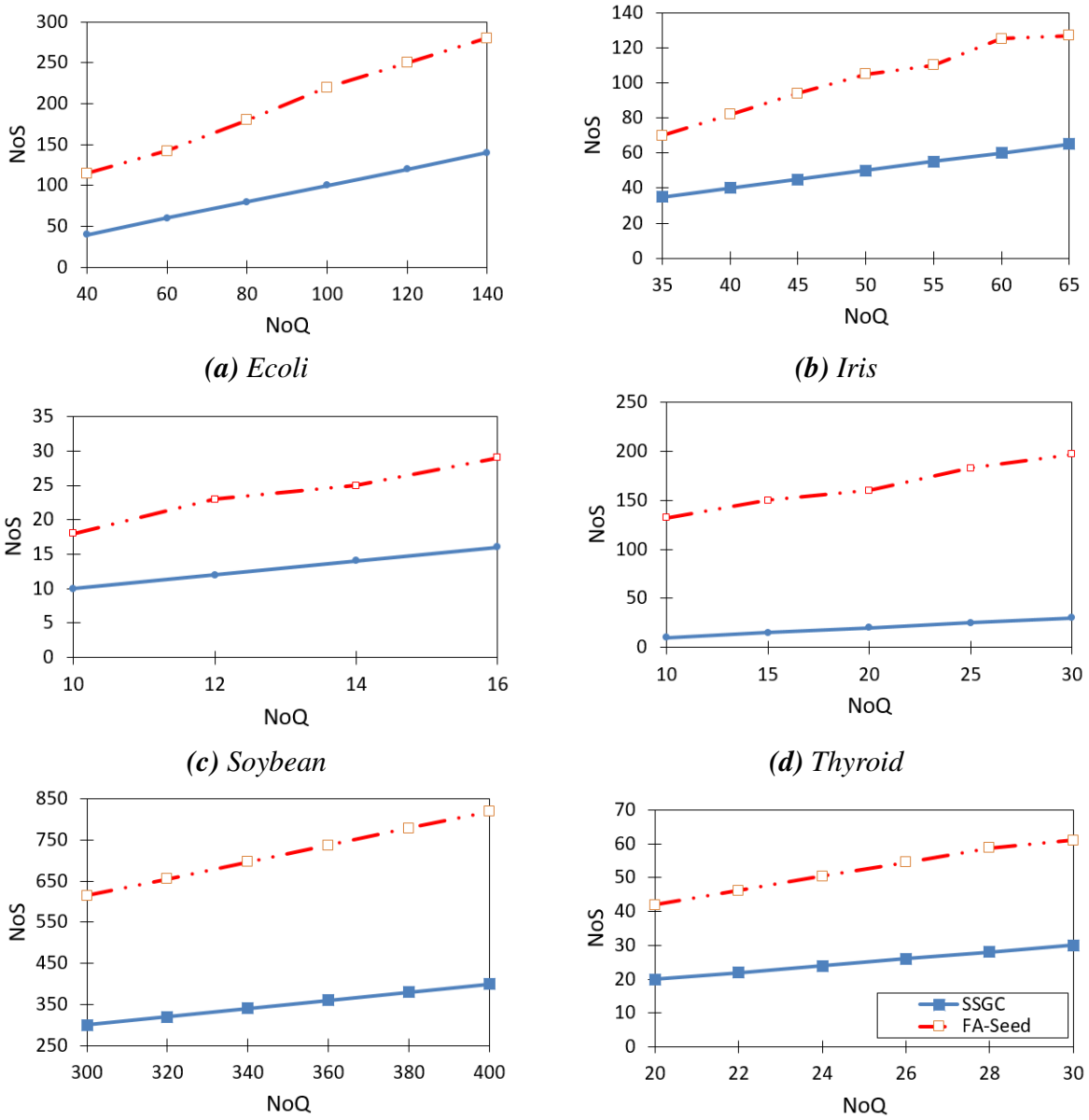
Khi tỷ lệ dữ liệu được gán nhãn tăng lên, giá trị RI trung bình của cả hai phương pháp đều được cải thiện và mức độ biến động giảm dần. Trên tập Iris, FA-Seed tiếp tục duy trì hiệu năng RI trung bình tương đương hoặc nhỉnh hơn FC2 trong hầu hết các mức tỷ lệ nhãn, cùng với mức dao động thấp và ổn định. Trong khi đó, trên tập Wine, FC2 có xu hướng đạt giá trị RI trung bình cao hơn ở các mức tỷ lệ nhãn trung bình và cao; tuy nhiên, sự khác biệt này đi kèm với mức biến động nhỏ, cho thấy khi thông tin giám sát tăng, lợi thế của chiến lược lựa chọn hạt giống dần suy giảm và các phương pháp có xu hướng hội tụ về hiệu năng tương đương.

Các kết quả thực nghiệm cho thấy FA-Seed không chỉ cải thiện chỉ số RI với dữ liệu thiếu nhãn, mà còn làm giảm đáng kể phương sai hiệu năng giữa các lần chạy, qua đó nâng cao độ tin cậy và tính bền vững của mô hình phân cụm bán giám sát.

Hình 3.12 minh họa trực quan sự so sánh về số lượng hạt giống thu được giữa các phương pháp khi sử dụng cùng một số lượng truy vấn. Kết quả cho thấy, tại cùng một mức NoQ, FA-Seed luôn thu được số lượng hạt giống cao hơn so với MSCFMN [21]. Đáng chú ý, FA-Seed vẫn đạt được tỷ lệ hạt giống cao ngay cả khi số lượng truy vấn thấp. Đây là một ưu điểm quan trọng, vì nó giúp giảm đáng kể chi phí gán nhãn. Bên cạnh đó, tỷ lệ hạt giống cao còn trực tiếp góp phần nâng cao hiệu năng của mô hình. Nguyên nhân chính khiến FA-Seed đạt được số lượng hạt giống lớn hơn là do cơ chế phân hoạch siêu hộp tinh hơn so với MSCFMN, cho phép mô hình tiếp cận ranh giới cụm một cách sát hơn, từ đó khai thác hiệu quả hơn các điểm dữ liệu tiềm năng cho việc gán nhãn.

Hình 3.13 trình bày kết quả so sánh giữa FA-Seed và các phương pháp SSGC, SSDBSCAN, SSK-Means, K-Means và MSCFMN [21, 103], trong đó chỉ số RI được sử dụng để đánh giá chất lượng phân cụm. Trong mỗi lần chạy, các hạt giống được lựa chọn ngẫu nhiên và quá trình được lặp lại 10 lần, với giá trị RI trung bình được ghi nhận để phục vụ so sánh.

Từ kết quả trực quan có thể nhận thấy rằng FA-Seed nhìn chung đạt hiệu năng



(a) Ecoli *(b) Iris*

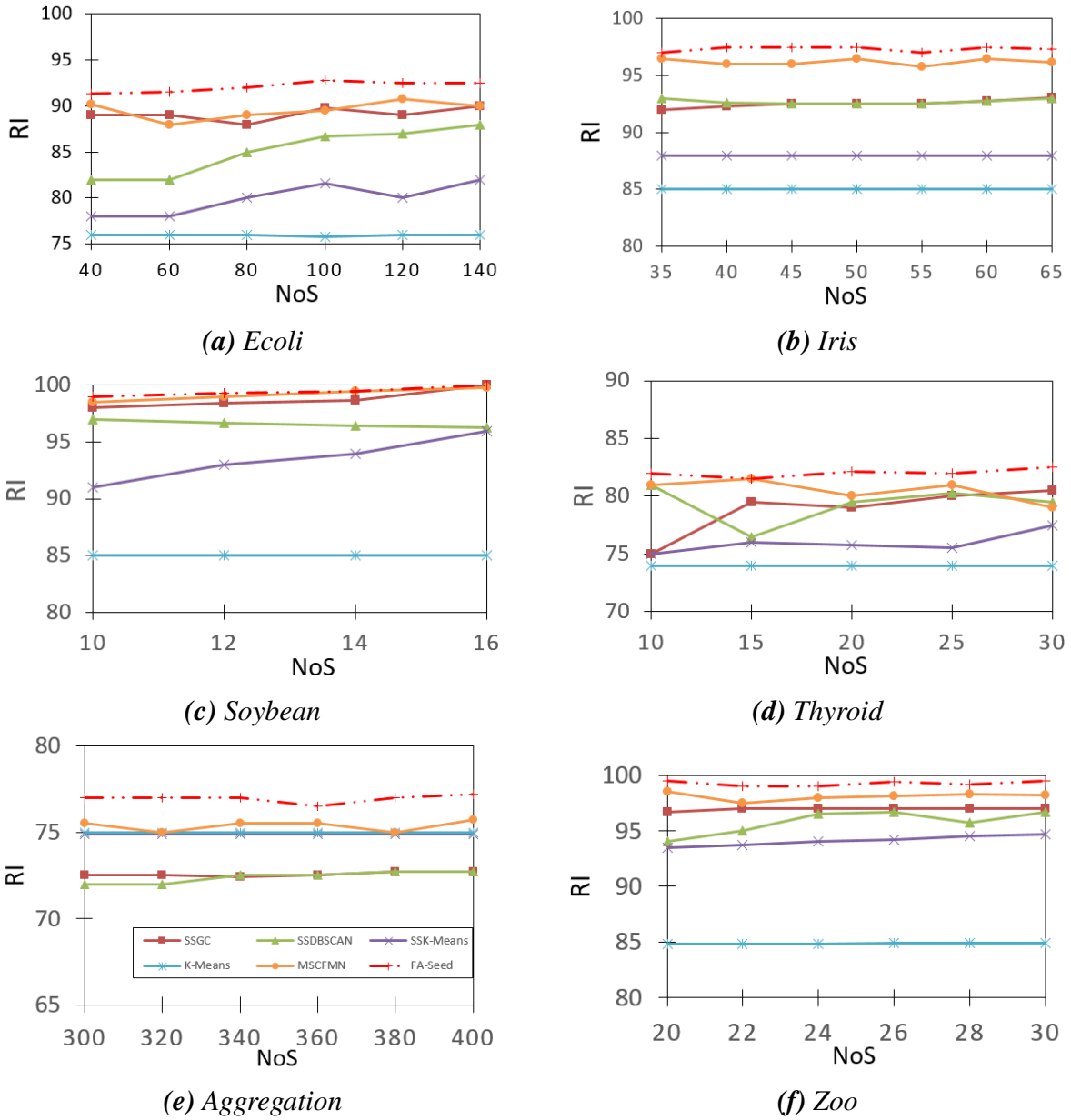
(c) Soybean *(d) Thyroid*

(e) Aggregation *(f) Zoo*

Hình 3.12: *So sánh giá trị NoS giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu: (a) Ecoli, (b) Iris, (c) Soybean, (d) Thyroid, (e) Aggregation, (f) Zoo*

tốt hơn hoặc ít nhất là tương đương với các phương pháp còn lại trên hầu hết các tập dữ liệu. Đối với các tập dữ liệu có cấu trúc cụm rõ ràng như Soybean, Zoo và Iris, FA-Seed đạt được các giá trị RI gần như hoàn hảo và thể hiện độ ổn định cao hơn so với các phương pháp khác. Điều này cho thấy khả năng khai thác hiệu quả thông tin hạt giống của FA-Seed nhằm gán nhãn chính xác ngay từ giai đoạn ban đầu.

Hình 3.14 minh họa sự so sánh chỉ số Accuracy giữa FA-Seed và MSCFMN. Kết quả cho thấy FA-Seed tương đương hoặc cao hơn MSCFMN trên hầu hết các tập dữ liệu, với những cải thiện đáng kể trên các tập dữ liệu có phân bố lớp mất cân bằng (như tập Thyroid, Pathbased, Zoo). Điều này chứng tỏ rằng cơ chế lựa chọn hạt giống chất lượng cao của FA-Seed có khả năng nâng cao hiệu năng phân cụm ngay cả trong



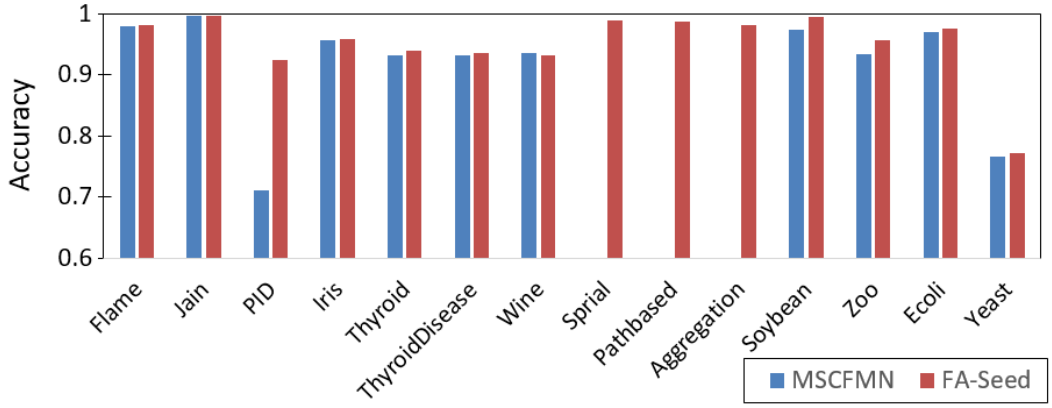
(a) Ecoli *(b) Iris*

(c) Soybean *(d) Thyroid*

(e) Aggregation *(f) Zoo*

Hình 3.13: *So sánh giá trị RI giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu: (a) Ecoli, (b) Iris, (c) Soybean, (d) Thyroid, (e) Aggregation, (f) Zoo*

các điều kiện dữ liệu khó khăn.

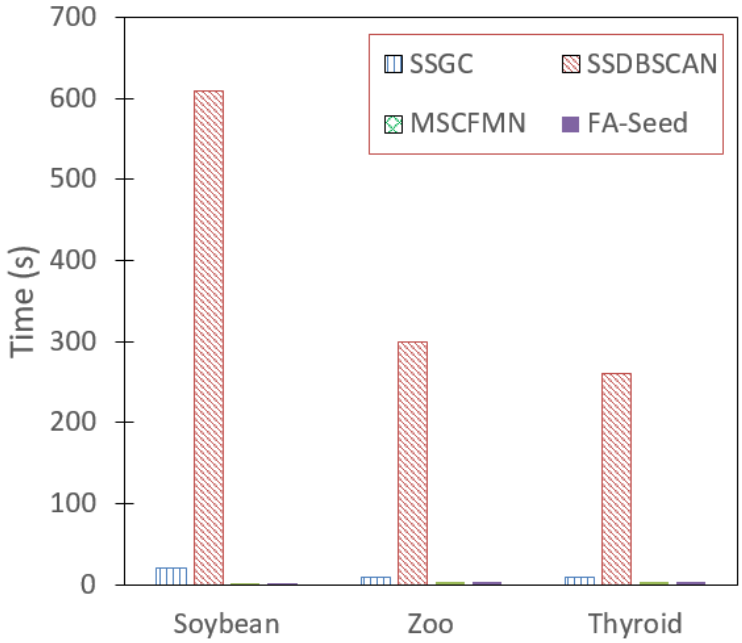


Hình 3.14: *Biểu đồ so sánh trực quan chỉ số Accuracy giữa FA-Seed và MSCFMN*

Hình 3.15 và Hình 3.16 so sánh thời gian chạy giữa FA-Seed và các thuật toán

cơ sở trên các tập dữ liệu chuẩn. Như có thể quan sát, FA-Seed chiếm lợi thế so với các phương pháp còn lại. Xét về độ phức tạp tính toán, FA-Seed có chi phí tính toán là $O(nMK)$ (Công thức (3.27)), trong đó $M = |X|$ là số lượng mẫu dữ liệu, n là số thuộc tính (số chiều dữ liệu), và $K = |B|$ là số lượng siêu hộp (HXs). Trong thực tế, K thường nhỏ hơn đáng kể so với M, do đó chi phí tính toán của FA-Seed tăng gần tuyến tính theo số mẫu.

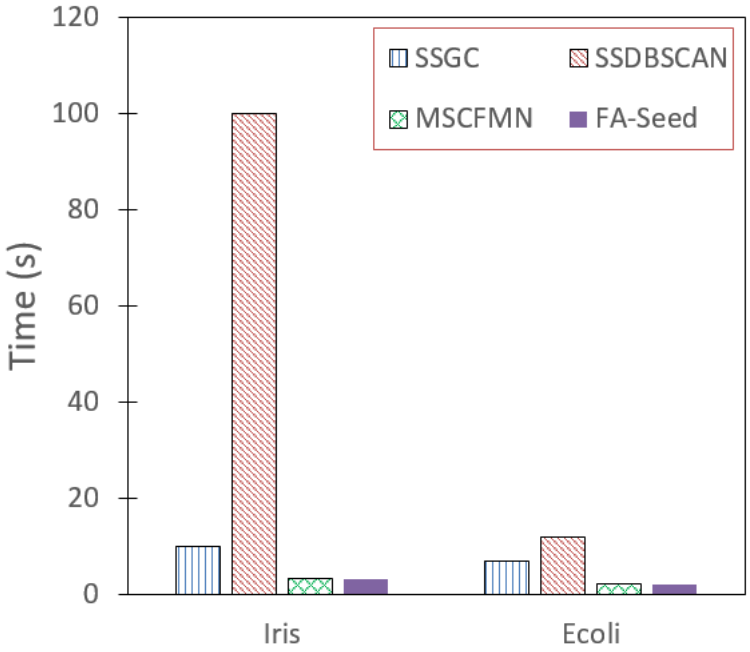
Ngược lại, SSGC có độ phức tạp $O(n^2)$ do yêu cầu xây dựng đồ thị k-láng giềng gần nhất và thực hiện lan truyền nhãn, khiến chi phí tính toán trở nên lớn đối với các tập dữ liệu quy mô lớn [103]. SSDBSCAN đạt độ phức tạp trung bình $O(n \log n)$ khi sử dụng các cấu trúc dữ liệu tối ưu, nhưng trong trường hợp xấu nhất vẫn có thể đạt tới $O(n^2)$.



Hình 3.15: *So sánh thời gian chạy (s) giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu thực nghiệm (Soybean, Zoo, Thyroid)*

Kết quả thực nghiệm của mô hình FA-Seed trên tập dữ liệu ILPD được đánh giá nhằm xem xét khả năng nhận diện các mẫu bệnh gan, hiệu quả tổng thể và chất lượng phân cụm của mô hình. Trước hết, ảnh hưởng của tham số $\delta^{\max}$ đến khả năng nhận diện lớp bệnh gan được khảo sát thông qua các độ đo Accuracy, Recall, Specificity, FP Rate, Precision và F1-Score. Kết quả được trình bày trong Bảng 3.1.

Kết quả trong Bảng 3.1 cho thấy khả năng nhận diện các mẫu bệnh gan của mô hình FA-Seed chịu ảnh hưởng rõ rệt bởi tham số $\delta^{\max}$. Tại $\delta^{\max} = 0{,}08$, mô hình đạt Accuracy bằng 0,9211, Recall bằng 0,9567 và F1-Score bằng 0,9454, là các giá trị cao nhất tương ứng trong các cấu hình được khảo sát. Precision đạt 0,9343 và Specificity đạt 0,8323, trong khi FP Rate ở mức 0,1677.



Hình 3.16: *So sánh thời gian chạy (s) giữa FA-Seed và các thuật toán cơ sở trên các tập dữ liệu thực nghiệm (Iris, Ecoli)*

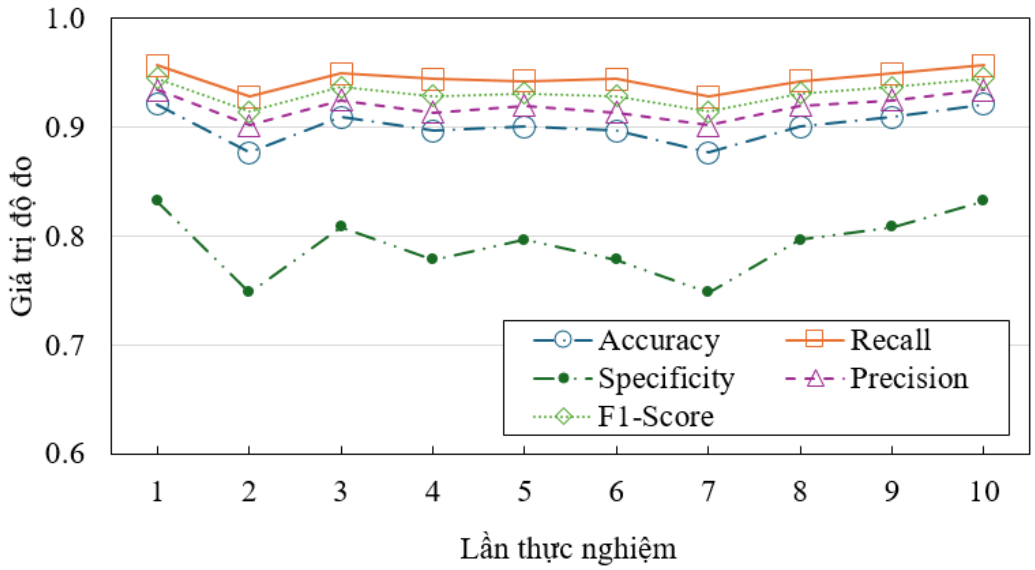
Xét theo mục tiêu nhận diện bệnh gan, Recall đạt 0,9567 cho thấy FA-Seed có khả năng phát hiện đúng phần lớn các mẫu thực sự thuộc lớp bệnh gan. Precision đạt 0,9343 cho thấy phần lớn các mẫu được mô hình nhận diện là bệnh gan thực sự thuộc lớp này. Đồng thời, Specificity đạt 0,8323 cho thấy mô hình vẫn duy trì khả năng nhận diện các mẫu không bệnh ở mức tương đối tốt. Tại $\delta^{\max} = 0{,}06$, Specificity và Precision đạt giá trị cao hơn một chút, lần lượt là 0,8383 và 0,9360, đồng thời FP Rate thấp nhất ở mức 0,1617. Tuy nhiên, $\delta^{\max} = 0{,}08$ đạt Accuracy, Recall và F1-Score cao hơn, cho thấy sự cân bằng tổng thể tốt hơn đối với mục tiêu nhận diện lớp bệnh gan.

Khi $\delta^{\max}$ tăng vượt quá 0,08, các độ đo Accuracy, Recall, Specificity, Precision và F1-Score nhìn chung có xu hướng giảm, trong khi FP Rate tăng. Sự suy giảm trở nên rõ rệt hơn tại các giá trị $\delta^{\max}$ lớn, đặc biệt từ 0,5 đến 0,9. Tại $\delta^{\max} = 0{,}90$, Accuracy chỉ còn 0,5317, Recall đạt 0,5529 và F1-Score đạt 0,6276. Kết quả này cho thấy việc mở rộng kích thước siêu hộp quá mức làm giảm khả năng phân biệt giữa các mẫu bệnh gan và không bệnh gan. Từ kết quả khảo sát, $\delta^{\max} = 0{,}08$ được lựa chọn để tiếp tục đánh giá mô hình FA-Seed trên tập dữ liệu ILPD.

Hình 3.17 cho thấy các độ đo của FA-Seed duy trì tương đối ổn định. Recall luôn đạt giá trị cao trong nhóm các độ đo, trong khi Precision và F1-Score cũng được duy trì ở mức cao và có xu hướng biến động tương đồng. Accuracy chỉ dao động trong một khoảng tương đối hẹp giữa các lần chạy. Specificity có giá trị thấp hơn các độ đo còn lại và thể hiện mức biến động lớn hơn, cho thấy khả năng nhận diện các mẫu không bệnh chưa ổn định bằng khả năng phát hiện các mẫu bệnh gan.

Bảng 3.1: *Đánh giá trên lớp bệnh gan của mô hình FA-Seed khi thay đổi* $\delta^{\max}$

$\delta^{\max}$	**Accuracy**	**Recall**	**Specificity**	**FP Rate**	**Precision**	**F1-Score**
0,01	0,8491	0,9014	0,7186	0,2814	0,8886	0,8950
0,02	0,8679	0,9135	0,7545	0,2455	0,9026	0,9080
0,04	0,8937	0,9399	0,7784	0,2216	0,9136	0,9265
0,06	0,9177	0,9495	0,8383	0,1617	0,9360	0,9427
0,08	0,9211	0,9567	0,8323	0,1677	0,9343	0,9454
0,10	0,8885	0,9327	0,7784	0,2216	0,9129	0,9227
0,20	0,8679	0,9135	0,7545	0,2455	0,9026	0,9080
0,30	0,8370	0,8846	0,7186	0,2814	0,8867	0,8857
0,40	0,7976	0,8413	0,6886	0,3114	0,8706	0,8557
0,50	0,7650	0,7933	0,6946	0,3054	0,8661	0,8281
0,60	0,7118	0,7452	0,6287	0,3713	0,8333	0,7868
0,70	0,6690	0,6971	0,5988	0,4012	0,8123	0,7503
0,80	0,5918	0,6010	0,5689	0,4311	0,7764	0,6775
0,90	0,5317	0,5529	0,4790	0,5210	0,7256	0,6276



Hình 3.17: *Đánh giá khả năng nhận diện lớp bệnh gan của FA-Seed*

Sau khi đánh giá khả năng nhận diện riêng đối với lớp bệnh gan, hiệu quả tổng thể của mô hình FA-Seed trên tập dữ liệu ILPD tiếp tục được xem xét thông qua các độ đo Accuracy, Specificity, Precision và F1-Score. Trong đó, Specificity, Precision và F1-Score được tính theo trung bình trên hai lớp, còn Accuracy phản ánh tỷ lệ dự đoán đúng trên toàn bộ tập dữ liệu. Ảnh hưởng của tham số $\delta^{\max}$ đến các độ đo này được trình bày trong Bảng 3.2 và được minh họa trực quan trong các Hình 3.18 - 3.21.

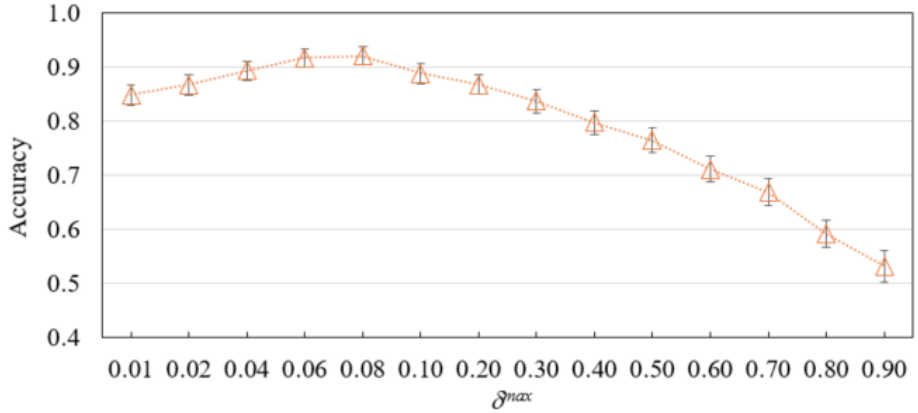
Kết quả trong Bảng 3.2 và các Hình 3.18 - 3.21 cho thấy hiệu quả của mô hình FA-Seed phụ thuộc vào $\delta^{\max}$. Khi $\delta^{\max}$ tăng từ 0,01 đến khoảng 0,06 - 0,08, các độ

đo Accuracy, Specificity, Precision và F1-Score nhìn chung đều được cải thiện. Điều này cho thấy việc tăng giới hạn kích thước siêu hộp trong khoảng phù hợp giúp mô hình biểu diễn tốt hơn cấu trúc của dữ liệu và nâng cao khả năng phân biệt giữa hai lớp bệnh gan và không bệnh gan.

Tại $\delta^{\max} = 0{,}08$, mô hình đạt Accuracy cao nhất bằng $0{,}9211 \pm 0.0170$, Precision trung bình có trọng số cao nhất bằng $0{,}9203 \pm 0{,}0173$ và F1-Score trung bình có trọng số cao nhất bằng $0{,}9203 \pm 0.0169$. Specificity đạt $0{,}8680 \pm 0.0182$, thấp hơn không đáng kể so với giá trị cao nhất $0{,}8702 \pm 0.0182$ tại $\delta^{\max} = 0{,}06$. Như vậy, mặc dù $\delta^{\max} = 0{,}06$ cho Specificity cao hơn một mức nhỏ, $\delta^{\max} = 0{,}08$ tạo ra sự cân bằng tốt hơn về hiệu quả tổng thể khi đồng thời đạt các giá trị cao nhất về Accuracy, Precision và F1-Score.

Khi $\delta^{\max}$ tăng vượt quá 0,08, các độ đo bắt đầu giảm. Mức suy giảm trở nên rõ hơn khi $\delta^{\max}$ tăng từ 0,40 đến 0,90. Xu hướng thể hiện trong các Hình 3.18 - 3.21 cho thấy các độ đo đều tăng đến vùng cực đại tại $\delta^{\max} = 0{,}06$ hoặc $0{,}08$, sau đó giảm dần khi tham số tiếp tục tăng.

Kết quả này cho thấy nếu $\delta^{\max}$ quá nhỏ, các siêu hộp có thể chưa bao phủ đầy đủ cấu trúc phân bố của dữ liệu. Ngược lại, khi $\delta^{\max}$ quá lớn, các siêu hộp có thể mở rộng quá mức và bao phủ các mẫu thuộc những lớp khác nhau, làm giảm khả năng phân biệt giữa lớp bệnh gan và lớp không bệnh gan. Do đó, $\delta^{\max} = 0{,}08$ được lựa chọn là giá trị phù hợp để tiếp tục đánh giá mô hình FA-Seed trên tập dữ liệu ILPD.



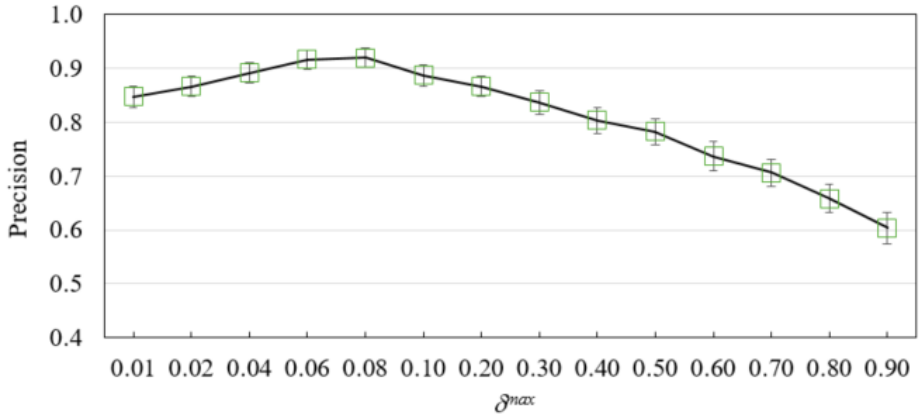
Hình 3.18: *Sự biến động Accuracy của mô hình FA-Seed khi thay đổi giá trị* δ^{max}

Hình 3.22 cho thấy các độ đo của mô hình FA-Seed duy trì ở mức tương đối cao qua các lần thực nghiệm. Accuracy, Precision và F1-Score có xu hướng biến động tương đồng, các đường biểu diễn nằm tương đối gần nhau và chỉ dao động trong một khoảng hẹp. Kết quả này cho thấy sự nhất quán giữa tỷ lệ dự đoán đúng tổng thể, độ chính xác của các dự đoán và sự cân bằng giữa Precision với Recall của mô hình.

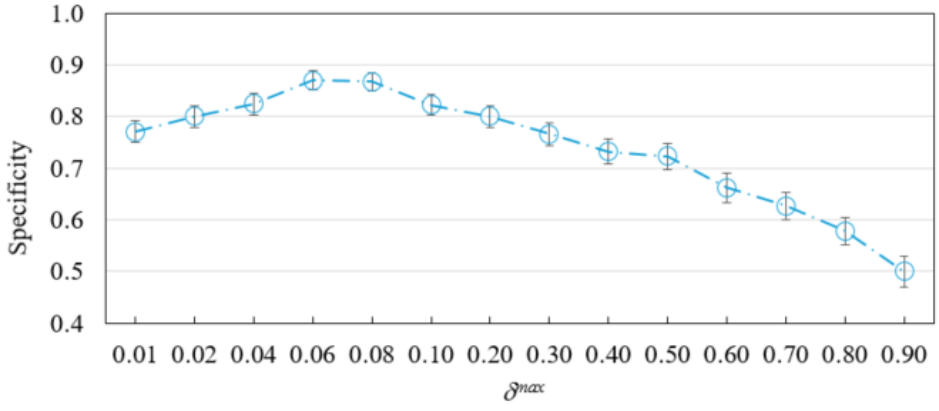
Specificity có giá trị thấp hơn Accuracy, Precision và F1-Score, đồng thời thể

Bảng 3.2: *Kết quả đánh giá mô hình FA-Seed khi thay đổi $\delta^{\max}$ trên tập dữ liệu ILPD*

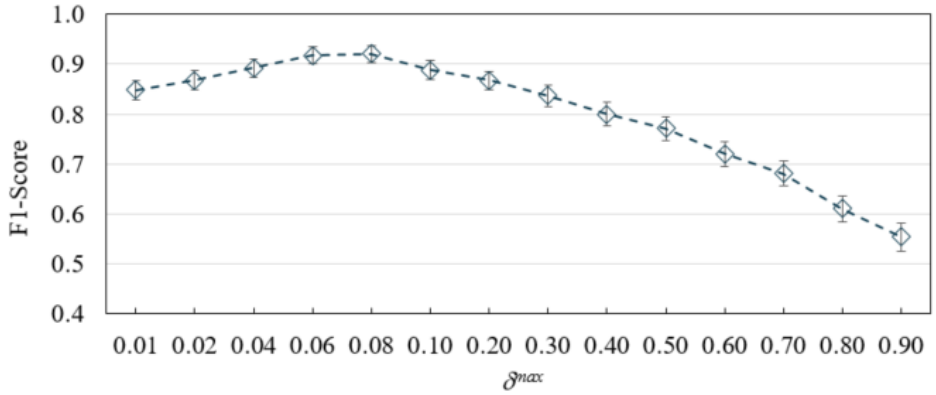
$\delta^{\max}$	**Accuracy**	**Specificity**	**Precision**	**F1-Score**
0,01	0,8491 ± 0,0190	0,7709 ± 0,0204	0,8476 ± 0,0194	0,8482 ± 0,0190
0,02	0,8679 ± 0,0190	0,8000 ± 0,0200	0,8669 ± 0,0195	0,8673 ± 0,0192
0,04	0,8937 ± 0,0180	0,8247 ± 0,0210	0,8921 ± 0,0193	0,8924 ± 0,0184
0,06	0,9177 ± 0,0170	0,8702 ± 0,0182	0,9170 ± 0,0176	0,9172 ± 0,0173
0,08	0,9211 ± 0,0170	0,8680 ± 0,0182	0,9203 ± 0,0173	0,9203 ± 0,0169
0,10	0,8885 ± 0,0190	0,8226 ± 0,0198	0,8871 ± 0,0194	0,8876 ± 0,0191
0,20	0,8679 ± 0,0180	0,8000 ± 0,0211	0,8669 ± 0,0194	0,8673 ± 0,0185
0,30	0,8370 ± 0,0210	0,7661 ± 0,0228	0,8373 ± 0,0220	0,8372 ± 0,0217
0,40	0,7976 ± 0,0230	0,7324 ± 0,0248	0,8032 ± 0,0236	0,7999 ± 0,0231
0,50	0,7650 ± 0,0230	0,7229 ± 0,0246	0,7825 ± 0,0240	0,7710 ± 0,0237
0,60	0,7118 ± 0,0240	0,6621 ± 0,0287	0,7372 ± 0,0263	0,7206 ± 0,0251
0,70	0,6690 ± 0,0250	0,6270 ± 0,0262	0,7064 ± 0,0254	0,6812 ± 0,0249
0,80	0,5918 ± 0,0250	0,5781 ± 0,0274	0,6583 ± 0,0261	0,6106 ± 0,0255
0,90	0,5317 ± 0,0290	0,5002 ± 0,0299	0,6039 ± 0,0292	0,5536 ± 0,0289



Hình 3.19: *Sự biến động Precision của mô hình FA-Seed khi thay đổi giá trị δ^{max}*

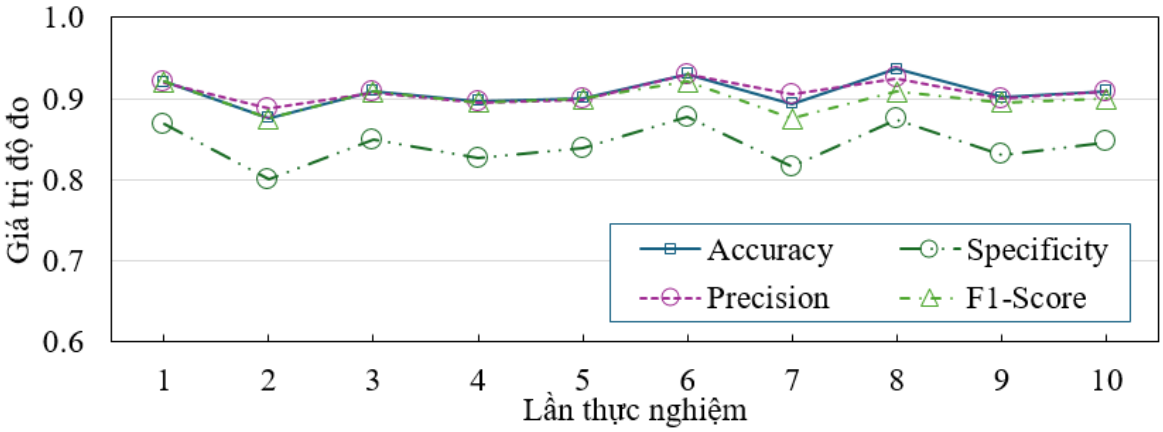


Hình 3.20: *Sự biến động Specificity của mô hình FA-Seed khi thay đổi giá trị δ^{max}*



Hình 3.21: *Sự biến động F1-Score của mô hình FA-Seed khi thay đổi giá trị* δ^{max}

hiện mức dao động rõ hơn giữa các lần thực nghiệm. Điều này cho thấy khả năng nhận diện các mẫu không bệnh gan của FA-Seed chưa ổn định bằng hiệu quả tổng thể và khả năng nhận diện lớp bệnh gan. Tuy nhiên, Specificity vẫn được duy trì ở mức tương đối cao trong phần lớn các lần chạy, cho thấy mô hình không chỉ tập trung vào lớp bệnh gan mà vẫn có khả năng phân biệt các mẫu không bệnh.



Hình 3.22: *Sự biến động của Accuracy, Precision và F1-Score của mô hình FA-Seed*

Bên cạnh các độ đo đánh giá khả năng nhận diện lớp và chất lượng phân cụm của mô hình FA-Seed trên tập dữ liệu ILPD được đánh giá thông qua các chỉ số RI, ARI và NMI. Ảnh hưởng của tham số $\delta^{\max}$ đến các chỉ số này được trình bày chi tiết trong Bảng 3.3 và được minh họa trực quan trong Hình 3.23. Sau khi lựa chọn giá trị $\delta^{\max}$ phù hợp, sự biến động của các chỉ số qua các lần thực nghiệm (Hình 3.24).

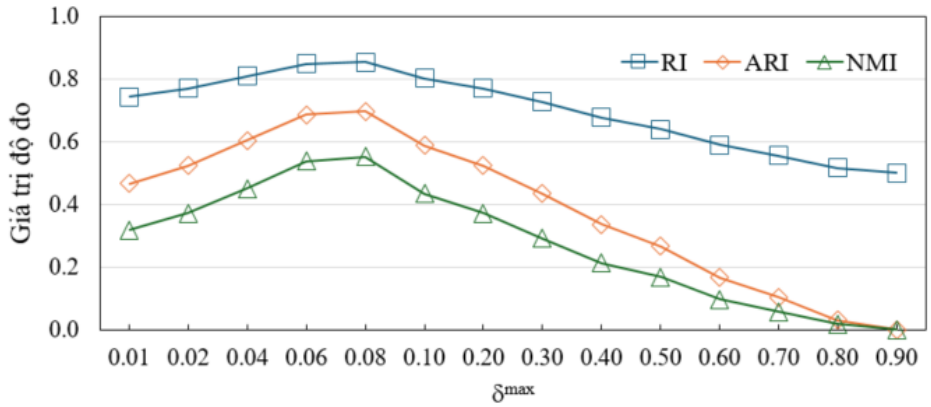
Kết quả trong Bảng 3.3 và Hình 3.23 cho thấy chất lượng phân cụm của mô hình FA-Seed phụ thuộc rõ rệt vào giá trị $\delta^{\max}$. Khi $\delta^{\max}$ tăng từ 0,01 đến 0,08, cả ba chỉ số RI, ARI và NMI đều có xu hướng tăng. Tại $\delta^{\max} = 0{,}08$, mô hình đạt kết quả cao nhất với RI bằng $0{,}8544 \pm 0{,}0263$, ARI bằng $0{,}6972 \pm 0{,}0549$ và NMI bằng $0{,}5516 \pm 0{,}0577$. Các kết quả này cho thấy cấu trúc phân cụm do FA-Seed hình thành tại giá trị tham số này có mức độ tương đồng tương đối cao với phân hoạch theo nhãn bệnh gan và không bệnh gan.

Bảng 3.3: *Kết quả đánh giá phân cụm của mô hình FA-Seed khi thay đổi* $\delta^{\max}$

$\delta^{\max}$	**RI**	**ARI**	**NMI**
0,01	$0{,}7432 \pm 0{,}0280$	$0{,}4673 \pm 0{,}0560$	$0{,}3187 \pm 0{,}0591$
0,02	$0{,}7703 \pm 0{,}0270$	$0{,}5238 \pm 0{,}0567$	$0{,}3718 \pm 0{,}0609$
0,04	$0{,}8096 \pm 0{,}0260$	$0{,}6036 \pm 0{,}0494$	$0{,}4507 \pm 0{,}0626$
0,06	$0{,}8486 \pm 0{,}0265$	$0{,}6859 \pm 0{,}0477$	$0{,}5384 \pm 0{,}0644$
0,08	$0{,}8544 \pm 0{,}0263$	$0{,}6972 \pm 0{,}0549$	$0{,}5516 \pm 0{,}0577$
0,10	$0{,}8015 \pm 0{,}0216$	$0{,}5875 \pm 0{,}0432$	$0{,}4341 \pm 0{,}0432$
0,20	$0{,}7703 \pm 0{,}0243$	$0{,}5238 \pm 0{,}0510$	$0{,}3718 \pm 0{,}0510$
0,30	$0{,}7267 \pm 0{,}0275$	$0{,}4353 \pm 0{,}0591$	$0{,}2919 \pm 0{,}0591$
0,40	$0{,}6766 \pm 0{,}0277$	$0{,}3361 \pm 0{,}0609$	$0{,}2129 \pm 0{,}0609$
0,50	$0{,}6398 \pm 0{,}0278$	$0{,}2674 \pm 0{,}0626$	$0{,}1694 \pm 0{,}0626$
0,60	$0{,}5890 \pm 0{,}0280$	$0{,}1668 \pm 0{,}0644$	$0{,}0967 \pm 0{,}0644$
0,70	$0{,}5563 \pm 0{,}0284$	$0{,}1047 \pm 0{,}0667$	$0{,}0587 \pm 0{,}0567$
0,80	$0{,}5160 \pm 0{,}0287$	$0{,}0304 \pm 0{,}0689$	$0{,}0185 \pm 0{,}0567$
0,90	$0{,}5012 \pm 0{,}0285$	$0{,}0012 \pm 0{,}0693$	$0{,}0007 \pm 0{,}0494$

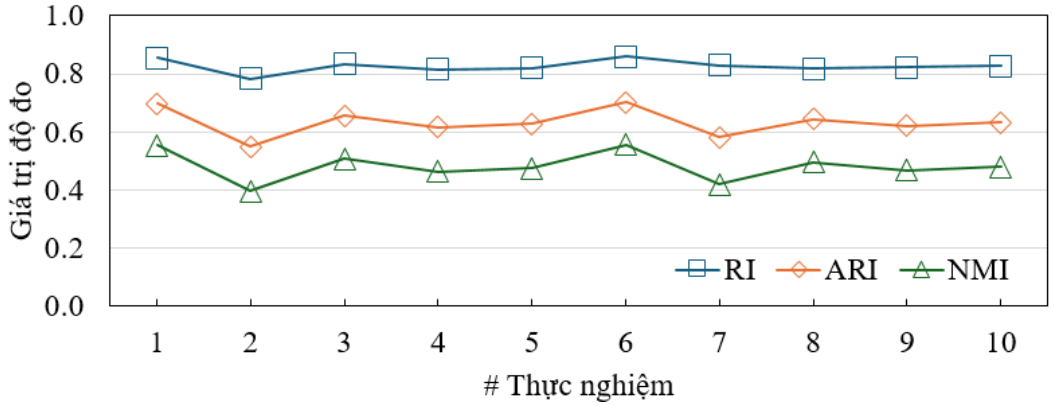
Tại $\delta^{\max} = 0{,}06$, các chỉ số RI, ARI và NMI cũng đạt giá trị cao, lần lượt là $0{,}8486 \pm 0{,}0265$, $0{,}6859 \pm 0{,}0477$ và $0{,}5384 \pm 0{,}0644$. Tuy nhiên, khi tăng $\delta^{\max}$ lên 0,08, cả ba chỉ số tiếp tục được cải thiện. Kết quả này phù hợp với các độ đo phân loại, qua đó củng cố việc lựa chọn $\delta^{\max} = 0{,}08$ cho mô hình FA-Seed trên tập dữ liệu ILPD.

Khi $\delta^{\max}$ tăng vượt quá 0,08, RI, ARI và NMI đều giảm dần. Mức suy giảm trở nên rõ rệt hơn tại các giá trị $\delta^{\max}$ lớn. Tại $\delta^{\max} = 0{,}90$, RI chỉ còn $0{,}5012 \pm 0{,}0285$, ARI bằng $0{,}0012 \pm 0{,}0693$ và NMI bằng $0{,}0007 \pm 0{,}0494$. Giá trị ARI và NMI xấp xỉ 0 cho thấy cấu trúc phân cụm được hình thành gần như không còn tương đồng đáng kể với phân hoạch theo nhãn thực tế. Nguyên nhân có thể là khi $\delta^{\max}$ quá lớn, các siêu hộp được mở rộng quá mức và có khả năng bao phủ các mẫu thuộc những lớp khác nhau, từ đó làm giảm khả năng phân tách giữa lớp bệnh gan và lớp không bệnh gan.



Hình 3.23: *Sự biến động của các độ đo mô hình FA-Seed khi thay đổi giá trị* δ^{max}

Hình 3.24 cho thấy các chỉ số RI, ARI và NMI có sự biến động nhất định trong các lần thực nghiệm. RI dao động trong khoảng 0,7831 - 0,8544, ARI trong khoảng 0,5485 - 0,6972 và NMI trong khoảng 0,3950 - 0,5516. Trong đó, RI duy trì ở mức cao nhất, còn ARI và NMI có mức dao động rõ hơn do hai chỉ số này đánh giá chặt chẽ hơn mức độ tương đồng giữa kết quả phân cụm và phân hoạch theo nhãn thực tế.



Hình 3.24: *Sự biến động độ đo mô hình FA-Seed qua các lần thực nghiệm*

Để đánh giá hiệu quả của mô hình FA-Seed trên tập dữ liệu ILPD, kết quả thực nghiệm của mô hình được so sánh với MSCFMN và các phương pháp phân cụm K-means++, EM, DBSCAN được thực hiện trên Weka. Việc so sánh được tiến hành trên lớp bệnh gan, lớp không bệnh gan và kết quả trung bình có trọng số trên hai lớp. Bên cạnh các độ đo dựa trên ma trận nhầm lẫn, các chỉ số RI, ARI và NMI cũng được sử dụng để đánh giá mức độ tương đồng giữa cấu trúc phân cụm thu được và phân hoạch theo nhãn thực tế.

Bảng 3.4: *So sánh nhận diện lớp bệnh gan của FA-Seed với các phương pháp khác*

Mô hình	Accuracy	Recall	Specificity	FP Rate	Precision	F1-Score
K-means++	0,6252 ± 0,0222	0,7452 ± 0,0855	0,3263 ± 0,1628	0,6737 ± 0,1628	0,7366 ± 0,0256	0,7376 ± 0,0303
EM	0,5647 ± 0,0351	0,4503 ± 0,1452	0,8496 ± 0,2402	0,1504 ± 0,2402	0,9112 ± 0,0702	0,5860 ± 0,0721
DBSCAN	0,6416 ± 0,0022	0,7831 ± 0,0058	0,2996 ± 0,0006	0,7004 ± 0,0006	0,7299 ± 0,0045	0,7556 ± 0,0018
MSCFMN	0,7976 ± 0,0332	0,8798 ± 0,0334	0,5928 ± 0,0444	0,4072 ± 0,0444	0,8432 ± 0,0179	0,8610 ± 0,0240
FA-Seed	0,9009 ± 0,0165	0,9442 ± 0,0107	0,7928 ± 0,0316	0,2072 ± 0,0316	0,9191 ± 0,0121	0,9315 ± 0,0113

Kết quả trong Bảng 3.4 và Hình 3.25 cho thấy FA-Seed đạt hiệu quả tốt trong nhận diện lớp bệnh gan xét trên phần lớn các độ đo. Cụ thể, mô hình đạt Accuracy bằng $0{,}9009 \pm 0{,}0165$, Recall bằng $0{,}9442 \pm 0{,}0107$, Precision bằng $0{,}9191 \pm 0{,}0121$ và

F1-Score bằng $0{,}9315 \pm 0{,}0113$. Các giá trị này đều cao hơn MSCFMN và các phương pháp phân cụm được thực hiện trên Weka.

Recall đạt 0,9442 cho thấy FA-Seed có khả năng phát hiện đúng phần lớn các mẫu thực sự thuộc lớp bệnh gan. Precision đạt 0,9191 cho thấy phần lớn các mẫu được mô hình nhận diện là bệnh gan thực sự thuộc lớp này. F1-Score đạt 0,9315 thể hiện sự cân bằng tốt giữa khả năng phát hiện mẫu bệnh gan và độ chính xác của các dự đoán thuộc lớp bệnh.

So với FA-Seed, MSCFMN đạt Recall tương đối cao, bằng $0{,}8798 \pm 0{,}0334$, nhưng Specificity chỉ đạt $0{,}5928 \pm 0{,}0444$, dẫn đến FP Rate ở mức $0{,}4072 \pm 0{,}0444$. Điều này cho thấy MSCFMN vẫn còn nhận diện nhầm một tỷ lệ đáng kể các mẫu không bệnh thành bệnh. FA-Seed cải thiện đồng thời Recall, Specificity, Precision và F1-Score, qua đó tạo ra sự cân bằng tốt hơn giữa khả năng nhận diện hai lớp.

Đối với các phương pháp trên Weka, K-means++ và DBSCAN đạt Recall lần lượt là $0{,}7452 \pm 0{,}0855$ và $0{,}7831 \pm 0{,}0058$, nhưng Specificity chỉ đạt $0{,}3263 \pm 0{,}1628$ và $0{,}2996 \pm 0{,}0006$. FP Rate của hai phương pháp tương ứng lên tới 0,6737 và 0,7004, cho thấy chúng có xu hướng gán nhiều mẫu vào lớp bệnh gan và nhận diện nhầm phần lớn các mẫu không bệnh.

Ngược lại, EM đạt Specificity cao nhất, bằng $0{,}8496 \pm 0{,}2402$, và FP Rate thấp nhất, bằng $0{,}1504 \pm 0{,}2402$. Tuy nhiên, Recall của EM chỉ đạt $0{,}4503 \pm 0{,}1452$, cho thấy phương pháp bỏ sót nhiều mẫu bệnh gan. Vì vậy, mặc dù EM đạt Precision cao, F1-Score chỉ bằng $0{,}5860 \pm 0.0721$. Nhìn chung, FA-Seed thể hiện sự cân bằng tốt hơn giữa Recall và Specificity so với các phương pháp trên Weka.

Kết quả trong Bảng 3.5 cho thấy FA-Seed tiếp tục đạt hiệu quả tốt trong nhận diện lớp không bệnh gan. Mô hình đạt Recall bằng $0{,}7928 \pm 0{,}0316$, Specificity bằng $0{,}9442 \pm 0{,}0107$, Precision bằng $0{,}8508 \pm 0{,}0291$ và F1-Score bằng $0{,}8208 \pm 0{,}0302$. Đồng thời, FP Rate chỉ ở mức $0{,}0558 \pm 0{,}0107$, thấp nhất trong số các phương pháp được so sánh.

So với MSCFMN, FA-Seed cải thiện Recall từ $0{,}5928 \pm 0{,}0444$ lên $0{,}7928 \pm 0{,}0316$, Precision từ $0{,}6676 \pm 0.0735$ lên $0{,}8508 \pm 0.0291$ và F1-Score từ $0{,}6273 \pm 0.0540$ lên $0{,}8208 \pm 0{,}0302$. Kết quả này cho thấy FA-Seed không chỉ phát hiện tốt các mẫu bệnh gan mà còn nâng cao khả năng nhận diện đúng các mẫu không bệnh gan.

K-means++ và DBSCAN có Recall trên lớp không bệnh thấp, lần lượt là $0{,}3263 \pm 0{,}1628$ và $0{,}2996 \pm 0{,}0006$. Điều này phù hợp với kết quả trên lớp bệnh gan, khi hai

phương pháp có xu hướng gán nhiều mẫu vào lớp bệnh. EM đạt Recall cao nhất trên lớp không bệnh, bằng $0{,}8496 \pm 0{,}2402$, nhưng Precision chỉ đạt $0{,}3794 \pm 0{,}0151$. Do đó, F1-Score của EM chỉ đạt $0{,}5122 \pm 0{,}0949$, thấp hơn đáng kể so với FA-Seed.

Như vậy, FA-Seed duy trì hiệu quả tương đối cân bằng trên cả lớp bệnh gan và lớp không bệnh gan, trong khi các phương pháp trên Weka thường thể hiện sự chênh lệch rõ rệt giữa khả năng nhận diện hai lớp.

Bảng 3.5: *So sánh nhận diện lớp không bệnh của FA-Seed với các phương pháp khác*

Mô hình	**Accuracy**	**Recall**	**Specificity**	**FP Rate**	**Precision**	**F1-Score**
K-means++	0,6252 ± 0,0222	0,3263 ± 0,1628	0,7452 ± 0,0855	0,2548 ± 0,0855	0,3114 ± 0,0945	0,3112 ± 0,1311
EM	0,5647 ± 0,0351	0,8496 ± 0,2402	0,4503 ± 0,1452	0,5497 ± 0,1452	0,3794 ± 0,0151	0,5122 ± 0,0949
DBSCAN	0,6416 ± 0,0022	0,2996 ± 0,0006	0,7831 ± 0,0058	0,2169 ± 0,0058	0,3638 ± 0,0129	0,3285 ± 0,0054
MSCFMN	0,7976 ± 0,0332	0,5928 ± 0,0444	0,8798 ± 0,0334	0,1202 ± 0,0334	0,6676 ± 0,0735	0,6273 ± 0,0540
FA-Seed	0,9009 ± 0,0165	0,7928 ± 0,0316	0,9442 ± 0,0107	0,0558 ± 0,0107	0,8508 ± 0,0291	0,8208 ± 0,0302

Kết quả trong Bảng 3.6 cho thấy FA-Seed đạt hiệu quả tốt nhóm phương pháp được so sánh. Mô hình đạt Accuracy bằng $0{,}9009 \pm 0{,}0165$, Specificity bằng $0{,}8362 \pm 0{,}0255$, Precision $0{,}8995 \pm 0{,}0169$ và F1-Score bằng $0{,}8998 \pm 0{,}0167$.

So với MSCFMN, FA-Seed cải thiện Accuracy từ $0{,}7976 \pm 0{,}0332$ lên $0{,}9009 \pm 0{,}0165$ và F1-Score từ $0{,}7941 \pm 0{,}0323$ lên $0{,}8998 \pm 0{,}0167$. Đồng thời, FP Rate giảm từ $0{,}3250 \pm 0{,}0383$ xuống $0{,}1638 \pm 0{,}0255$. Các kết quả này cho thấy cơ chế lựa chọn tập hạt giống của FA-Seed góp phần nâng cao khả năng hướng dẫn quá trình phân cụm và hạn chế việc lan truyền thông tin nhãn không phù hợp.

Các phương pháp K-means++, EM và DBSCAN đạt Accuracy lần lượt là 0,6252, 0,5647 và 0,6416, thấp hơn đáng kể so với FA-Seed. Mặc dù EM đạt Specificity và Precision tương đối cao, F1-Score chỉ đạt $0{,}5648 \pm 0.0249$ do sự mất cân bằng lớn giữa kết quả trên hai lớp. K-means++ và DBSCAN cũng có Specificity thấp và FP Rate cao, cho thấy khả năng phân biệt tổng thể giữa lớp bệnh gan và không bệnh gan còn hạn chế.

Nhìn chung, kết quả trung bình có trọng số khẳng định FA-Seed đạt sự cân bằng tốt hơn giữa hai lớp và có hiệu quả tổng thể cao hơn MSCFMN cũng như các phương pháp phân cụm trên Weka.

Bảng 3.7 và Hình 3.26 trình bày kết quả so sánh mức độ tương đồng giữa cấu trúc phân cụm thu được và phân hoạch theo nhãn thực tế. FA-Seed đạt kết quả cao nhất trên cả ba chỉ số với RI bằng $0{,}8215 \pm 0{,}0263$, ARI bằng $0{,}6285 \pm 0{,}0549$ và NMI bằng $0{,}4779 \pm 0{,}0577$. Kết quả này cho thấy cấu trúc phân cụm do FA-Seed hình thành có mức độ phù hợp tương đối cao với hai lớp bệnh gan và không bệnh gan.

Bảng 3.6: *So sánh FA-Seed với MSCFMN và các phương pháp phân cụm trên Weka: các độ đo phân loại*

Method	Accuracy	Recall	Specificity	FP Rate	Precision	F1-Score
K-means++	0,6252 ± 0,0222	0,6252 ± 0,0222	0,4463 ± 0,0932	0,5537 ± 0,0932	0,6148 ± 0,0437	0,6154 ± 0,0240
EM	0,5647 ± 0,0351	0,5647 ± 0,0351	0,7353 ± 0,1299	0,2647 ± 0,1299	0,7589 ± 0,0544	0,5648 ± 0,0249
DBSCAN	0,6416 ± 0,0022	0,6416 ± 0,0022	0,4411 ± 0,0048	0,5589 ± 0,0048	0,6229 ± 0,0021	0,6306 ± 0,0019
MSCFMN	0,7976 ± 0,0332	0,7976 ± 0,0332	0,6750 ± 0,0383	0,3250 ± 0,0383	0,7929 ± 0,0329	0,7941 ± 0,0323
FA-Seed	0,9009 ± 0,0165	0,9009 ± 0,0165	0,8362 ± 0,0255	0,1638 ± 0,0255	0,8995 ± 0,0169	0,8998 ± 0,0167

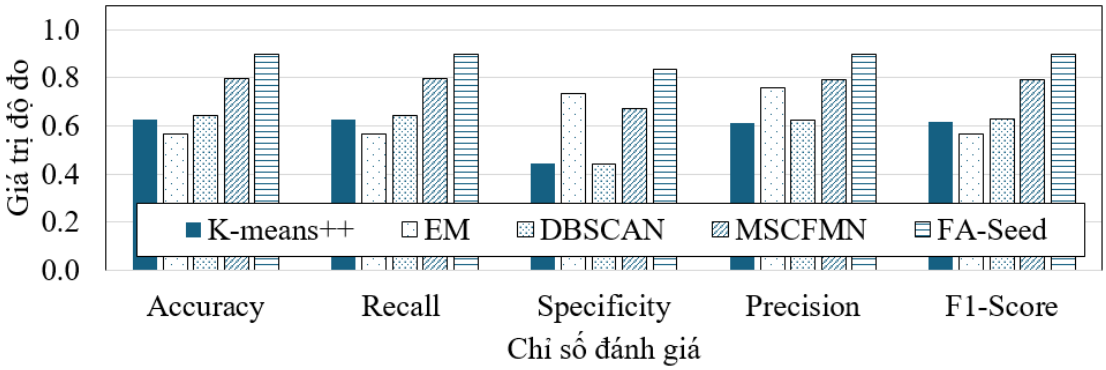
MSCFMN đạt RI bằng $0{,}6783 \pm 0{,}0402$, ARI bằng $0{,}3279 \pm 0{,}0812$ và NMI bằng $0{,}1993 \pm 0{,}0665$. Mặc dù cao hơn các phương pháp trên Weka, các chỉ số này vẫn thấp hơn đáng kể so với FA-Seed. Điều đó cho thấy việc xây dựng tập hạt giống phù hợp đã hỗ trợ quá trình phân cụm bán giám sát đạt được phân hoạch gần với nhãn thực tế.

Đối với các phương pháp trên Weka, K-means++ và DBSCAN có RI xấp xỉ 0,53, nhưng ARI và NMI đều ở mức rất thấp. K-means++ chỉ đạt ARI bằng $0{,}0189 \pm 0{,}0336$ và NMI bằng $0{,}0107 \pm 0.0110$, trong khi DBSCAN đạt ARI bằng $0{,}0347 \pm 0.0030$ và NMI bằng $0{,}0066 \pm 0{,}0011$. Điều này cho thấy cấu trúc cụm do hai phương pháp tạo ra chưa tương ứng tốt với phân hoạch theo nhãn bệnh gan và không bệnh gan.

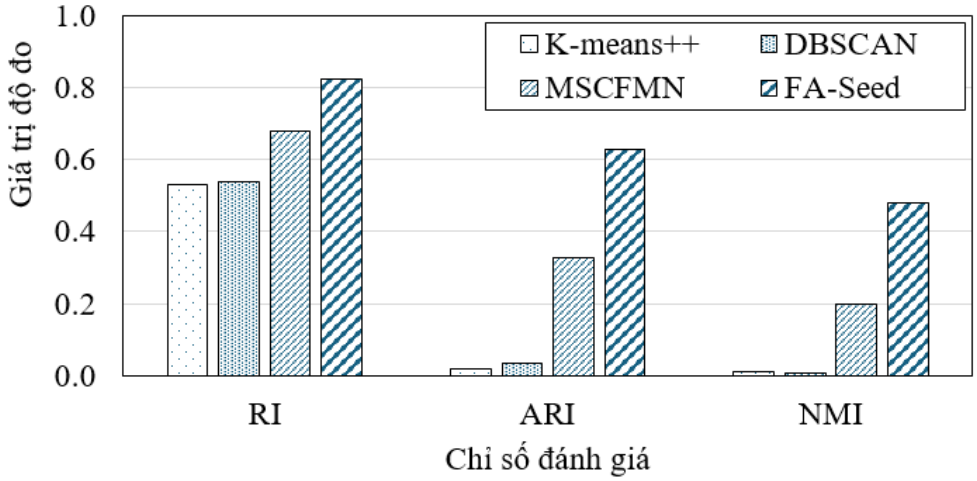
EM đạt RI bằng $0{,}5097 \pm 0{,}0146$ và NMI bằng $0{,}0926 \pm 0{,}0343$, trong khi ARI có giá trị âm nhẹ, bằng $-0{,}0130 \pm 0{,}0144$. Giá trị ARI gần 0 và âm cho thấy mức độ tương đồng đã hiệu chỉnh giữa phân hoạch do EM tạo ra và nhãn thực tế không cao hơn mức kỳ vọng do ngẫu nhiên.

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

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



Hình 3.25: *So sánh các độ đo đánh giá khả năng nhận diện lớp bệnh gan của FA-Seed với MSCFMN và các phương pháp phân cụm trên Weka*



Hình 3.26: *So sánh các chỉ số RI, ARI và NMI của FA-Seed với MSCFMN và các phương pháp phân cụm trên Weka*

3.4. Kết luận chương 3

Chương 3 đã trình bày mô hình FA-Seed cho phân cụm bán giám sát, trong đó trọng tâm là chiến lược lựa chọn hạt giống chủ động nhằm nâng cao chất lượng và độ ổn định của quá trình học. Mô hình FA-Seed không giả định rằng mọi hạt giống đều có vai trò tương đương, mà xem xét đồng thời mức độ đại diện và độ tin cậy của từng hạt giống trong không gian dữ liệu. Việc lựa chọn hạt giống theo hướng thích nghi

giúp giảm nguy cơ lan truyền lỗi nhãn, đặc biệt trong các vùng dữ liệu chồng lấn, nhiễu hoặc phân bố không đồng đều. Bên cạnh đó, phân tích độ phức tạp tính toán cũng cho thấy FA-Seed có chi phí tính toán phù hợp trên các tập dữ liệu thực nghiệm có kích thước và mức độ phức tạp khác nhau. Cách tiếp cận này cho phép mô hình ưu tiên các hạt giống có khả năng dẫn dắt quá trình lan truyền nhãn tốt hơn, đồng thời hạn chế ảnh hưởng của các mẫu hoặc siêu hộp không ổn định.

Các thực nghiệm giúp kiểm chứng cơ chế hoạt động của FA-Seed. Các kết quả thu được cho thấy chiến lược lựa chọn hạt giống chủ động có tiềm năng cải thiện chất lượng phân cụm, đồng thời củng cố cơ sở thực nghiệm cho các phân tích lý thuyết đã trình bày. Các kết quả trên cho thấy FA-Seed là một hướng tiếp cận phù hợp để nâng cao hiệu năng phân cụm bán giám sát trong mạng nơron Min-Max mờ.

Một phần kết quả nghiên cứu trong Chương 3 đã được công bố trong các công trình [CT6], [CT7] và [CT8]. Trong đó, [CT6] được công bố trên tạp chí *Information*; [CT7] được công bố trong kỷ yếu *Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông*, Hội thảo Quốc gia lần thứ XXIV, Nhà xuất bản Khoa học và Kỹ thuật; [CT8] trình bày việc tích hợp học chủ động với lan truyền nhãn có kiểm soát trong mạng nơron min-max mờ, đã gửi *Journal of Computer Science and Cybernetics* và đang chờ quyết định cuối cùng sau phản biện vòng 2.

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 (FMN) 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. Trên cơ sở tổng quan các mô hình FMN bán giám sát (GFMM, RFMN, SS-FMM, SSL-FMM, SCFMN, MSCFMN), luận án chỉ ra rằng các mô hình hiện có thường xem các siêu hộp và các mẫu mang nhãn có vai trò tương đương, trong khi tâm hình học của siêu hộp có thể lệch so với phân bố thực của dữ liệu và hạt giống thường được chọn ngẫu nhiên hoặc theo giả thiết mỗi siêu hộp đại diện cho một cụm.

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 $0{,}8806 \pm 0{,}0043$ và các chỉ số RI, ARI, NMI cao nhất trong nhóm so sánh (Bảng 2.10, Bảng 2.13). 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 $0{,}9009 \pm 0{,}0165$ và RI $0{,}8215 \pm 0{,}0263$, so với $0{,}7976 \pm 0{,}0332$ và $0{,}6783 \pm 0{,}0402$ của MSCFMN (Bảng 3.6, Bảng 3.7). Kết quả đã được công bố trong các công trình [CT6], [CT7] và [CT8].

3. 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 (độ lệch tâm, chỉ số sử dụng, độ thuộc) vào quá trình suy luận nhãn và lựa chọn hạt giống của FMN bán giám sát, đồng thời phân tích độ phức tạp của các thuật toán đề xuất. Về thực tiễn, các mô hình được kiểm chứng trên bộ dữ liệu xét nghiệm bệnh gan ILPD, cho thấy khả năng áp dụng FMN bán giám sát trong bài toán

hỗ trợ chẩn đoán khi dữ liệu có nhãn hạn chế; cấu trúc siêu hộp cho phép kết xuất các luật IF–THEN có thể diễn giải.

4. Hạn chế

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 một số tập dữ liệu (PID, 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 $\delta^{\max}$; (iv) thực nghiệm mới được 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:

Trong các hướng nghiên cứu tiếp theo, luận án định hướng mở rộng các mô hình FMN bán giám sát theo các hướng: (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 các cơ chế suy luận nhãn thích nghi hơn dựa trên độ tin cậy và động học của siêu hộp; (iii) nghiên cứu các chiến lược song song và phân tán nhằm nâng cao khả năng mở rộng của mô hình trên các tập dữ liệu lớn; (iv) tiếp tục khai thác khả năng kết xuất luật mờ để tăng cường tính diễn giải và khả năng ứng dụng của mạng nơron Min-Max mờ trong các hệ thống hỗ trợ ra quyết định thực tế; và (v) kiểm định ý nghĩa thống kê các kết quả so sánh và đánh giá các mô hình trên các tập dữ liệu có quy mô lớn hơn.

Luận án đã đóng góp một số kết quả mới cả về phương pháp và ứng dụng cho bài toán phân cụm bán giám sát trong mạng nơron FMN, góp phần mở rộng hướng nghiên cứu FMN theo hướng ổn định hơn, tin cậy hơn và phù hợp hơn với các bài toán dữ liệu thực tế trong điều kiện nhãn hạn chế.

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.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Sugato Basu, Arindam Banerjee, and Raymond J Mooney, "Semi-supervised clustering by seeding", in: *Proceedings of the nineteenth international conference on machine learning*, 2002, pp. 27–34.

[2] Jianghui Cai et al. (2023), "A review on semi-supervised clustering", *Information Sciences*, 632, pp. 164–200.

[3] Kamal Taha (2023), "Semi-supervised and un-supervised clustering: A review and experimental evaluation", *Information Systems*, 114, p. 102178.

[4] Asma Chebli, Akila Djebbar, and Hayet Farida Marouani, "Semi-supervised learning for medical application: A survey", in: *2018 international conference on applied smart systems (ICASS)*, IEEE, 2018, pp. 1–9.

[5] Kiri Wagstaff et al., "Constrained k-means clustering with background knowledge", in: *ICML*, vol. 1, 2001, pp. 577–584.

[6] Mikhail Bilenko, Sugato Basu, and Raymond J Mooney, "Integrating constraints and metric learning in semi-supervised clustering", in: *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 11.

[7] Sugato Basu, Arindam Banerjee, and Raymond J Mooney, "Active semi-supervision for pairwise constrained clustering", in: *Proceedings of the 2004 SIAM international conference on data mining*, SIAM, 2004, pp. 333–344.

[8] Jin Wang et al. (2015), "Patient admission prediction using a pruned fuzzy min–max neural network with rule extraction", *Neural Computing and Applications*, 26 (2), pp. 277–289.

[9] Hong-Yu Wang, Jie-Sheng Wang, and Guan Wang (2022), "A survey of fuzzy clustering validity evaluation methods", *Information Sciences*, 618, pp. 270–297.

[10] Seyed Emadedin Hashemi, Fatemeh Gholian-Jouybari, and Mostafa Hajiaghaei-Keshteli (2023), "A fuzzy C-means algorithm for optimizing data clustering", *Expert Systems with Applications*, 227, p. 120377.

[11] Imran Shafi et al. (2024), "A review of approaches for rapid data clustering: Challenges, opportunities, and future directions", *IEEE Access*, 12, pp. 138086–138120.

[12] N Yogeesh et al. (2025), "From Crisp to Fuzzy: A Comparative Review of Statistical and Fuzzy Approaches to Problem Solving", *Appl Math Inf Sci*, 19 (3), pp. 647–658.

[13] Witold Pedrycz and James Waletzky (1997), "Fuzzy clustering with partial supervision", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27 (5), pp. 787–795.

[14] Amin Golzari Oskouei, Negin Samadi, and Jafar Tanha (2024), "Feature-weight and cluster-weight learning in fuzzy c-means method for semi-supervised clustering", *Applied Soft Computing*, 161, p. 111712.

[15] Bogdan Gabrys and Andrzej Bargiela (2000), "General fuzzy min-max neural network for clustering and classification", *IEEE Transactions on Neural Networks*, 11 (3), pp. 769–783.

[16] Patrick K Simpson (1992), "Fuzzy min-max neural networks. I. Classification", *IEEE Transactions on Neural Networks*, 3 (5), pp. 776–786.

[17] Jinhai Liu et al. (2020), "Semi-supervised fuzzy min–max neural network for data classification", *Neural Processing Letters*, 51 (2), pp. 1445–1464.

[18] AV Nandedkar and PK Biswas (2008), "Reflex fuzzy min max neural network for semi-supervised learning", *Journal of Intelligent Systems*, 17 (1-3), pp. 5–18.

[19] Dinh Minh Vu, Viet Hai Nguyen, and Ba Dung Le, "Semi-supervised clustering in fuzzy min-max neural network", in: *International Conference on Advances in Information and Communication Technology*, Springer, 2016, pp. 541–550.

[20] Thi Ngan Tran et al. (2019), "The combination of fuzzy min–max neural network and semi-supervised learning in solving liver disease diagnosis support problem", *Arabian Journal for Science and Engineering*, 44 (4), pp. 2933–2944.

[21] VD Minh et al. (2022), "An improvement in integrating clustering method and neural network to extract rules and application in diagnosis support", *Iranian Journal of Fuzzy Systems*, 19 (5), p. 147.

[22] David W. Aha, *UC Irvine Machine Learning Repository*, `https://archive.ics.uci.edu/`, Accessed: 2 August 2025, 2025.

[23] Joni-Kristian Kämäräinen, Olli-Pekka Kauppi, and Pasi Fränti, *Clustering Datasets*, `https://cs.joensuu.fi/sipu/datasets/`, Accessed: 2 August 2025, 2025.

[24] Gustavo EAPA Batista and Maria Carolina Monard (2003), "An analysis of four missing data treatment methods for supervised learning", *Applied Artificial Intelligence*, 17 (5-6), pp. 519–533.

[25] Satish Chander and P Vijaya, "Unsupervised learning methods for data clustering", in: *Artificial intelligence in data mining*, Elsevier, 2021, pp. 41–64.

[26] Bhanu Chander and Kumaravelan Gopalakrishnan, “Data clustering using unsupervised machine learning”, in: *Statistical Modeling in Machine Learning*, Elsevier, 2023, pp. 179–204.

[27] Kanishka Tyagi et al., “Unsupervised learning”, in: *Artificial intelligence and machine learning for edge computing*, Elsevier, 2022, pp. 33–52.

[28] Haixun Wang et al., “Clustering by pattern similarity in large data sets”, in: *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, 2002, pp. 394–405.

[29] Ziling Ma et al. (2025), “FCPCA: Fuzzy clustering of high-dimensional time series based on common principal component analysis”, *International Journal of Approximate Reasoning*, p. 109552.

[30] Selen Avcı Azkeskin and Zerrin Aladağ (2025), “Evaluating regional sustainable energy potential through hierarchical clustering and machine learning”, *Environmental Research Communications*, 7 (1), p. 015002.

[31] Songul Cinaroglu (2025), “Clustering public hospitals based on crisp and fuzzy clustering techniques and probabilistic fuzzy efficiency estimates”, *Informatics for Health and Social Care*, 50 (3-4), pp. 114–132.

[32] José Nataniel A de Sá, Marcelo RP Ferreira, and Francisco de AT de Carvalho (2025), “Fuzzy and Crisp Gaussian Kernel-Based Co-clustering with Automatic Width Computation”, *IEEE Transactions on Fuzzy Systems*.

[33] Fachao Li, Panxiang Yue, and Lianqing Su, “Research on the convergence of fuzzy genetic algorithm based on rough classification”, in: *International Conference on Natural Computation*, Springer, 2006, pp. 792–795.

[34] Yue Yang et al. (2024), “Integrating fuzzy clustering and graph convolution network to accurately identify clusters from attributed graph”, *IEEE Transactions on Network Science and Engineering*, 12 (2), pp. 1112–1125.

[35] Zhe Liu and Sukumar Letchmunan (2024), “Enhanced fuzzy clustering for incomplete instance with evidence combination”, *ACM Transactions on Knowledge Discovery from Data*, 18 (3), pp. 1–20.

[36] Philipp Baumann and Dorit S Hochbaum (2025), “An algorithm for clustering with confidence-based must-link and cannot-link constraints”, *INFORMS Journal on Computing*, 37 (4), pp. 1044–1068.

[37] Chaoqi Jia et al. (2023), “Efficient Algorithm for the k-Means Problem with Must-Link and Cannot-Link Constraints”, *Tsinghua Science and Technology*, 28 (6), pp. 1050–1062.

[38] Bahram Jafrasteh et al. (2024), “Enhanced spatial Fuzzy C-Means algorithm for brain tissue segmentation in T1 images”, *Neuroinformatics*, 22 (4), pp. 407–420.

[39] Tran Manh Truong et al. (2025), “A novel distributed semi-supervised fuzzy clustering method applied on dental X-ray images”, *Vietnam Journal of Science and Technology*, 63 (1), pp. 149–160.

[40] Zuoxun Wang et al. (2024), “A Control Method for Improved Fuzzy PID of GSOM for Fish Scale Evolution”, *IEEE Access*, 12, pp. 55007–55018.

[41] Hengdong Zhu et al. (2025), “A new semi-supervised fuzzy clustering method based on latent representation learning and information fusion”, *Applied Soft Computing*, 170, p. 112717.

[42] Tran Manh Tuan et al. (2022), “An improvement of trusted safe semi-supervised fuzzy clustering method with multiple fuzzifiers”, *Journal of Computer Science and Cybernetics*, 38 (1), pp. 47–61.

[43] Tien Dung Duong et al. (2025), “An Innovative Semi-Supervised Fuzzy Clustering Technique Using Cluster Boundaries”, *Computers, Materials, & Continua*, 85 (3), p. 5341.

[44] Benfei Zhang, Yizhang Jiang, and Kaijian Xia (2024), “Interclass Balance Factor-Based Membership Fusion Semi-Supervised Fuzzy Clustering Algorithm for Lesion Segmentation in Cerebral Infarction Images”, *IEEE Access*, 12, pp. 107077–107088.

[45] Benjamin M Knisely and Holly H Pavliscsak (2023), “Research proposal content extraction using natural language processing and semi-supervised clustering: A demonstration and comparative analysis”, *Scientometrics*, 128 (5), pp. 3197–3224.

[46] Mohd Saqib, Benjamin CM Fung, and Philippe Charland (2025), “PAC-X: Fuzzy Explainable AI for Multi-Class Malware Detection”, *IEEE Transactions on Fuzzy Systems*.

[47] Silvana Trindade and Nelson LS Da Fonseca (2026), “Multi-Criteria Scoring for Cluster and Client Selection in Heterogeneous Hierarchical Federated Learning”, *IEEE Internet of Things Journal*.

[48] Liang Ding, Wenzhang Yang, and Yinxing Xue, “Function Clustering-Based Fuzzing Termination: Toward Smarter Early Stopping”, in: *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, IEEE, 2025, pp. 1731–1743.

[49] Amine M Bensaid et al. (1996), “Partially supervised clustering for image segmentation”, *Pattern Recognition*, 29 (5), pp. 859–871.

[50] Wen-Zhao Liu and Min Li (2024), "An Approximation Algorithm Based on Seeding Algorithm for Fuzzy k-Means Problem with Penalties", *Journal of the Operations Research Society of China*, 12 (2), pp. 387–409.

[51] Jing Lin et al. (2022), "A novel defect segmentation for pulse infrared images based on improved fuzzy C-means algorithm with weighted distance", *IEEE Transactions on Magnetics*, 59 (2), pp. 1–13.

[52] Amin Golzari Oskouei et al. (2024), "SSFCM-FWCW: Semi-supervised fuzzy c-means method based on feature-weight and cluster-weight learning", *Software Impacts*, 21, p. 100678.

[53] Benfei Zhang et al. (2024), "Semi-supervised fuzzy C means based on membership integration mechanism and its application in brain infarction lesion segmentation in DWI images", *Journal of Intelligent & Fuzzy Systems*, 46 (1), pp. 2713–2726.

[54] Hao-Ran Chen et al. (2024), "Adaptive semi-supervised fuzzy C-means method with local spatial information and pre-clustering for image segmentation", *IEEE Access*, 12, pp. 196328–196346.

[55] Kamil Kmita, Katarzyna Kaczmarek-Majer, and Olgierd Hryniewicz (2024), "Explainable impact of partial supervision in semi-supervised fuzzy clustering", *IEEE Transactions on Fuzzy Systems*, 32 (5), pp. 3189–3198.

[56] Jun Yan and Xiangfeng Wang (2022), "Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology", *The Plant Journal*, 111 (6), pp. 1527–1538.

[57] Huy Thong Pham et al. (2026), "A Novel Semi-Supervised Multi-View Picture Fuzzy Clustering Approach for Enhanced Satellite Image Segmentation", *Computers, Materials, & Continua*, 86 (3).

[58] Arezou Najafi Moghaddam et al. (2025), "A robust possibilistic semi-supervised fuzzy clustering algorithm with neighborhood-aware feature weighting", *International Journal of Machine Learning and Cybernetics*, 16 (11), pp. 8803–8838.

[59] Diogo PP Branco and Francisco de AT de Carvalho (2023), "Medoid based semi-supervised fuzzy clustering algorithms for multi-view relational data", *Fuzzy Sets and Systems*, 469, p. 108630.

[60] Zhaorui Zhu and Quanxue Gao (2022), "Semi-supervised clustering via cannot link relationship for multiview data", *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (12), pp. 8744–8755.

[61] Zhanhu Zhang et al. (2023), "Knowledge augmentation-based soft constraints for semi-supervised clustering", *Applied Soft Computing*, 144, p. 110484.

[62] Sina Shaham et al. (2025), “Privacy and fairness in machine learning: A survey”, *IEEE Transactions on Artificial Intelligence*, 6 (7), pp. 1706–1726.

[63] Xiaowei Wang et al. (2022), “Fuzzy-clustering and fuzzy network based interpretable fuzzy model for prediction”, *Scientific Reports*, 12 (1), p. 16279.

[64] Ma Lina (2025), “Evaluation system of college English teaching quality based on fuzzy information of artificial intelligence”, *Discover Artificial Intelligence*, 5 (1), p. 267.

[65] Zihe Lv, Zhengda Wu, and Jinghua Zhu (2025), “Clustering-guided contrastive prototype learning: Towards semi-supervised medical image segmentation”, *Pattern Recognition*, p. 112321.

[66] Zhiyu Zheng, Liang Lv, and Lefei Zhang (2024), “Enhancing the semi-supervised semantic segmentation with prototype-based supervision for remote sensing images”, *IEEE Geoscience and Remote Sensing Letters*, 21, pp. 1–5.

[67] Mahnoor Chaudhry et al. (2023), “A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective”, *Symmetry*, 15 (9), p. 1679.

[68] Cindy Espinoza and Jesús Carretero (2025), “Profiling and archetyping of higher education applicants using intelligent data analysis techniques”, *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*.

[69] Lukas Grützner and Michael H Breitner (2025), “Bridging the implementation gap in MCABC inventory management: from a taxonomy to practical archetypes”, *International Transactions in Operational Research*.

[70] Guojie Li et al. (2024), “Exploring feature selection with limited labels: A comprehensive survey of semi-supervised and unsupervised approaches”, *IEEE Transactions on Knowledge and Data Engineering*, 36 (11), pp. 6124–6144.

[71] Wei Shen et al. (2023), “A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (8), pp. 9284–9305.

[72] Zhiyu Pan et al. (2024), “Pseudo label fusion with uncertainty estimation for semi-supervised cropping box regression”, *IEEE Transactions on Multimedia*, 26, pp. 8157–8171.

[73] Dymitr Ruta, Ernesto Damiani, and Bogdan Gabryś, “Scalable Hyperbox Clustering for Geospatial Data”, in: *2025 IEEE International Conference on Fuzzy Systems (FUZZ)*, IEEE, 2025, pp. 1–4.

[74] Thien M Tran and My-Ha Le, “Simultaneous Low-Light Image Enhancement and Deblurring for Surveillance Systems Using Intelligent”, in: *Intelligent*

Systems and Data Science: Third International Conference, ISDS 2025, Can Tho City, Vietnam, October 18-19, 2025, Proceedings, Part II. Vol. 2191, Springer Nature, 2024, p. 246.

[75] Noam Shental et al. (2008), "Gaussian mixture models with equivalence constraints", *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, pp. 33–58.

[76] Patrick K Simpson (1993), "Fuzzy min-max neural networks—Part 2: Clustering", *IEEE Transactions on Fuzzy Systems*, 1 (1), pp. 32–45.

[77] Ömer Nedim Kenger and Eren Özceylan (2023), "Fuzzy min–max neural networks: a bibliometric and social network analysis", *Neural Computing and Applications*, 35 (7), pp. 5081–5111.

[78] P Simpson, "Fuzzy adaptive resonance theory", in: *Proceedings of the Neuroengineering Workshop*, Southern Illinois University at Carbondale, IL, 1990.

[79] Gail A Carpenter and Stephen Grossberg (1987), "ART 2: Self-organization of stable category recognition codes for analog input patterns", *Applied Optics*, 26 (23), pp. 4919–4930.

[80] Hon Keung Kwan and Yaling Cai (1994), "A fuzzy neural network and its application to pattern recognition", *IEEE Transactions on Fuzzy Systems*, 2 (3), pp. 185–193.

[81] David Martínez-Rego, Oscar Fontenla-Romero, and Amparo Alonso-Betanzos (2012), "Nonlinear single layer neural network training algorithm for incremental, nonstationary and distributed learning scenarios", *Pattern Recognition*, 45 (12), pp. 4536–4546.

[82] Chuan Luo et al. (2013), "Incremental approaches for updating approximations in set-valued ordered information systems", *Knowledge-Based Systems*, 50, pp. 218–233.

[83] Reza Davtalab, Mir Hossein Dezfoulian, and Muharram Mansoorizadeh (2013), "Multi-level fuzzy min-max neural network classifier", *IEEE Transactions on Neural Networks and Learning Systems*, 25 (3), pp. 470–482.

[84] Abhijeet V Nandedkar and Prabir K Biswas, "A general fuzzy min max neural network with compensatory neuron architecture", in: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, 2005, pp. 1160–1167.

[85] Huaguang Zhang et al. (2011), "Data-core-based fuzzy min–max neural network for pattern classification", *IEEE Transactions on Neural Networks*, 22 (12), pp. 2339–2352.

[86] Dazhong Ma, Jinhai Liu, and Zhanshan Wang, “The pattern classification based on fuzzy min-max neural network with new algorithm”, in: *International Symposium on Neural Networks*, Springer, 2012, pp. 1–9.

[87] Mohammed Falah Mohammed and Chee Peng Lim (2014), “An enhanced fuzzy min–max neural network for pattern classification”, *IEEE Transactions on Neural Networks and Learning Systems*, 26 (3), pp. 417–429.

[88] Swati Shinde et al. (2014), “Diabetes diagnosis using fuzzy min-max neural network with rule extraction and Apriori algorithm”, *The International Journal of Science and Technoledge*, 2 (4), p. 369.

[89] Anas Quteishat and Chee Peng Lim (2008), “A modified fuzzy min–max neural network with rule extraction and its application to fault detection and classification”, *Applied Soft Computing*, 8 (2), pp. 985–995.

[90] Mohammed Falah Mohammed and Chee Peng Lim (2017), “Improving the fuzzy min-max neural network with a k-nearest hyperbox expansion rule for pattern classification”, *Applied Soft Computing*, 52, pp. 135–145.

[91] Mohammed Falah Mohammed and Chee Peng Lim (2017), “A new hyperbox selection rule and a pruning strategy for the enhanced fuzzy min–max neural network”, *Neural Networks*, 86, pp. 69–79.

[92] Peixin Hou et al. (2018), “Contribution-factor based fuzzy min-max neural network: order-dependent clustering for fuzzy system identification”, *International Journal of Computational Intelligence Systems*, 11 (1), pp. 737–756.

[93] Jinhai Liu et al. (2017), “A modified fuzzy min–max neural network for data clustering and its application on pipeline internal inspection data”, *Neurocomputing*, 238, pp. 56–66.

[94] Preetee M Sonule and Balaji S Shetty (2017), “An enhanced fuzzy min–max neural network with ant colony optimization based-rule-extractor for decision making”, *Neurocomputing*, 239, pp. 204–213.

[95] Manjeevan Seera, Kuldeep Randhawa, and Chee Peng Lim (2018), “Improving the fuzzy min–max neural network performance with an ensemble of clustering trees”, *Neurocomputing*, 275, pp. 1744–1751.

[96] Ashish Kumar Dehariya and Pragya Shukla (2020), “Medical data classification using fuzzy min max neural network preceded by feature selection through moth flame optimization”, *International Journal of Advanced Computer Science and Applications*, 11 (12).

[97] Tao Hou et al., “Quality-Aware Fuzzy Min-Max Neural Networks for Dynamic Brain Network Analysis”, in: *International Conference on Medical Im-*

age Computing and Computer-Assisted Intervention, Springer, 2024, pp. 356–366.

[98] Farhad Pourpanah et al. (2021), "A semisupervised learning model based on fuzzy min–max neural networks for data classification", *Applied Soft Computing*, 112, p. 107856.

[99] Anas Quteishat and Chee Peng Lim, "Application of the fuzzy min-max neural networks to medical diagnosis", in: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, 2008, pp. 548–555.

[100] Thomas Cover and Peter Hart (1967), "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, 13 (1), pp. 21–27.

[101] Honglei He (2025), "A clustering algorithm based on grids for core data and adjacency relationships for edge data", *Scientific Reports*, 15 (1), p. 18390.

[102] Xin Sun (2025), "Semi-Supervised Clustering via Constraints Self-Learning", *Mathematics*, 13 (9), p. 1535.

[103] Viet-Vu Vu (2018), "An efficient semi-supervised graph based clustering", *Intelligent Data Analysis*, 22 (2), pp. 297–307.