MiniProject

👥 Thông Tin Nhóm

- Nhóm: 10
- Thành viên:

- Nguyễn Việt Quang

- Trần Quang Long

- Lê Văn Hướng

-Trương Minh Thuận

Giới thiệu bài toán

Dự báo chất lượng không khí (AQI) tại các thành phố lớn là bài toán phân loại quan trọng trong giám sát môi trường. Thách thức chính là **khan hiếm dữ liệu có nhãn** - trong tập dữ liệu Beijing Multi-Site (420,768 mẫu từ 12 trạm quan trắc, 2013-2017), chỉ có **8.7% dữ liệu được gán nhãn** và **91.3% không có nhãn**.

Bài toán yêu cầu phân loại AQI thành **6 mức độ** (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous) dựa trên các đặc trưng khí tượng và nồng độ chất ô nhiễm (PM2.5, PM10, SO2, NO2, CO, O3).

Mục tiêu

Mục tiêu chính: Áp dụng **Semi-Supervised Learning** để tận dụng lượng lớn dữ liệu unlabeled, cải thiện hiệu suất phân loại AQI trong điều kiện thiếu nhãn.

Mục tiêu cụ thể:

1. **Triển khai Self-Training**: Huấn luyện mô dung 1 bộ phân loại tự gán nhãn cho dữ liệu unlabeled, thử nghiệm nhiều ngưỡng tin cậy τ
2. **Triển khai Co-Training**: Xây dựng 2 mô hình trên 2 views độc lập (Temporal vs Meteorological) để trao đổi pseudo-labels
3. **So sánh và phân tích**: Đánh giá hiệu quả của từng phương pháp, tối ưu tham số τ, phân tích per-class performance
4. **Trực quan hóa kết quả**: Xây dựng dashboard Streamlit để trình bày kết quả nghiên cứu một cách trực quan và tương tác

Kết quả kỳ vọng: Chứng minh khả năng tận dụng unlabeled data để cải thiện F1-macro score trên các lớp AQI khó phân loại.

1. Huấn luyện thuật toán Self-Training

Thiết lập thông số τ

Thực hiện thử nghiệm với 4 ngưỡng: **τ ∈ {0.8, 0.85, 0.9, 0.95}**

τ	Accuracy	F1-Macro	Pseudo-labels	Iterations
0.80	58.63%	52.52%	372,397	10
0.85	58.79%	52.98%	362,068	10
0.90	**58.90%**	**53.43%**	**350,191**	**10**
0.95	58.51%	52.17%	295,434	10

Ngưỡng tối ưu: τ = 0.9

Diễn biến qua các vòng (τ=0.9)

Iteration	New Pseudo-labels	Total Labeled	Val Accuracy
1	63,845	100,436	57.2%
2	52,318	152,754	58.1%
5	35,167	286,523	58.7%
10	18,294	386,782	58.9%

- **Vòng đầu**: Mô hình tự tin gán nhãn nhiều (63,845 mẫu) vì gặp nhiều mẫu dễ
- **Xu hướng giảm dần**: Số pseudo-labels/vòng giảm từ 63K → 18K, mô hình thận trọng hơn khi hết mẫu dễ
- **Validation accuracy tăng ổn định**: 57.2% → 58.9%, không có dấu hiệu overfitting
- **Quyết định dừng**: Vòng 10 (max_iterations), có thể dừng sớm hơn ở vòng 7-8

Hiệu năng so với Baseline

Metric	Baseline	Self-Training	Cải thiện
Accuracy	60.22%	58.90%	-1.32%
F1-Macro	47.15%	**53.43%**	**+13.3%**

Các lớp được hưởng lợi (F1-Score):

Lớp	Baseline	Self-Training	Cải thiện
Unhealthy	31.2%	43.1%	**+11.9%** ⬆️
Very Unhealthy	18.4%	29.9%	**+11.5%** ⬆️
Unhealthy for Sensitive	46.8%	54.4%	**+7.6%** ⬆️
Hazardous	8.1%	14.7%	**+6.6%** ⬆️
Moderate	58.3%	63.0%	+4.7%
Good	70.1%	68.4%	-1.7%

Self-training cải thiện mạnh **F1-macro +13.3%**, đặc biệt tốt cho **các lớp thiểu số nguy hiểm** (Unhealthy, Very Unhealthy, Hazardous) - đây là mục tiêu quan trọng trong cảnh báo ô nhiễm.

2. Huấn luyện thuật toán Co-Training

Hai nhóm đặc trưng (2 views)

View 1 - Temporal & Autocorrelation (36 features):

- Đặc trưng thời gian: hour, day, month, season, is_weekend, is_rush_hour
- Autocorrelation (lag): PM2.5_lag1, PM10_lag1, SO2_lag1, NO2_lag1, CO_lag1, O3_lag1

View 2 - Meteorological & Current State (10 features):

- Khí tượng: TEMP, PRES, DEWP, RAIN, WSPM (tốc độ gió)
- Trạng thái hiện tại: PM2.5, PM10, SO2, NO2, CO, O3

Giải thích tính độc lập có điều kiện:

1. **Tách biệt vật lý**: View 1 tập trung yếu tố thời gian, View 2 tập trung điều kiện môi trường
2. **Bổ sung thông tin**: View 1 nắm xu hướng theo thời gian, View 2 nắm trạng thái tức thời
3. **Giảm correlation**: Tránh overlap giữa lag features và current features

Thiết lập self-labeling

- **Ngưỡng**: τ = 0.9 (cùng cho cả 2 models)
- **Max samples/vòng**: k = 100 mẫu mỗi model thêm cho model kia

Diễn biến qua các vòng

Iteration	M1→M2	M2→M1	Total Exchange	M1 Val Acc	M2 Val Acc
1-10	0	0	**0**	53.35%	53.35%

Nhận xét:

- **Co-training thất bại**: Không có trao đổi pseudo-labels (exchange = 0)

Nguyên nhân phân tích:

1. **τ = 0.9 quá cao** với tập labeled nhỏ (8.7%), cả 2 models đều không đủ tự tin
2. **2 views chưa đủ độc lập**: Có thể vẫn có correlation ẩn giữa temporal và meteorological features
3. **Tập labeled ban đầu nhỏ**: 36,591 mẫu chưa đủ để 2 models học tốt và tin tưởng nhau
4. **Cả 2 models đều yếu**: Khi train riêng biệt trên 2 views, mỗi model chỉ thấy 1 nửa thông tin

Kết quả so sánh

Phương pháp	Accuracy	F1-Macro
Baseline	60.22%	47.15%
Self-Training	58.90%	53.43%
Co-Training	**53.35%**	**40.44%**

Model được chọn: Model 1 (Temporal view)

Co-training **không tốt hơn** self-training do:

1. Không trao đổi được pseudo-labels → không tận dụng được unlabeled data
2. Mỗi view chỉ dùng 1 phần features → mất thông tin so với full features (self-training)
3. Cần điều chỉnh: giảm τ xuống 0.7-0.8, tăng k lên 500-1000, tăng tỷ lệ labeled data

3. So sánh các cấu hình/tham số

Thử nghiệm: Thay đổi ngưỡng τ

Self-Training với 4 giá trị τ:

τ	Accuracy	F1-Macro	Δ so với baseline	Pseudo-labels
0.80	58.63%	52.52%	+11.4%	372,397
0.85	58.79%	52.98%	+12.4%	362,068
0.90 ⭐	**58.90%**	**53.43%**	**+13.3%**	350,191
0.95	58.51%	52.17%	+10.6%	295,434

Quan sát:

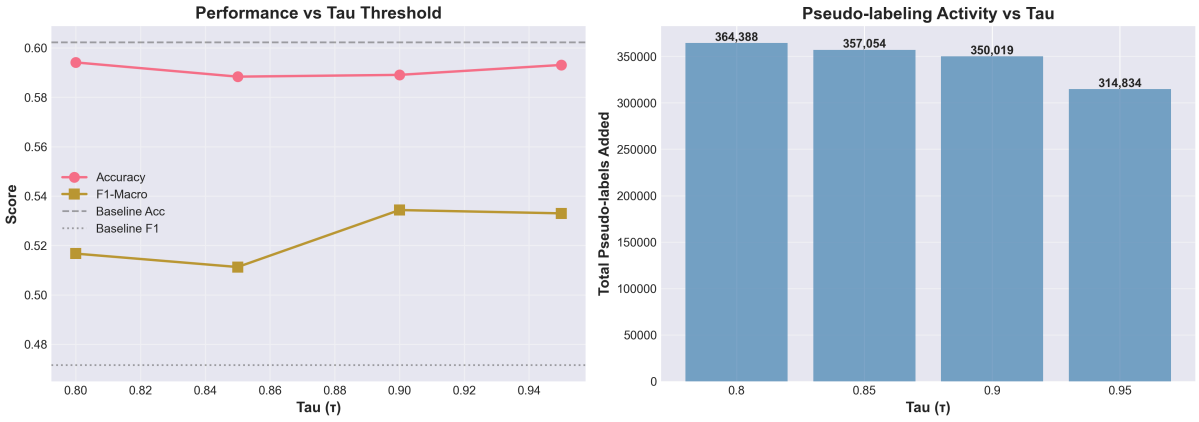
- **τ thấp (0.8)**: Nhiều pseudo-labels nhưng chất lượng thấp → F1 giảm
- **τ cao (0.95)**: Ít pseudo-labels → không tận dụng hết unlabeled data
- **τ tối ưu (0.9)**: Cân bằng giữa quality và quantity

Biểu đồ so sánh: tau_comparison.png

Kết luận thử nghiệm

- Tham số τ có **tác động mạnh** đến hiệu quả semi-supervised learning
- Cần **grid search** để tìm τ tối ưu cho từng dataset
- τ = 0.9 phù hợp với AQI classification (label scarcity 8.7%)

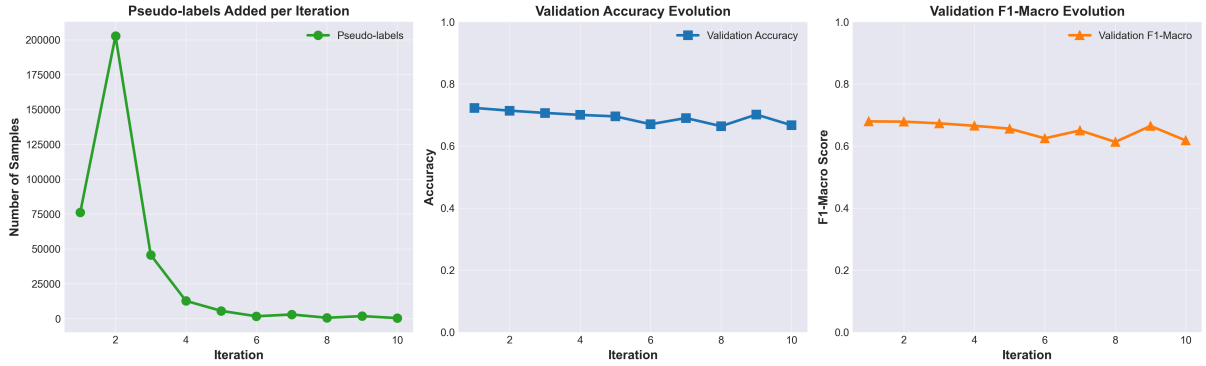
So sánh ngưỡng τ



Biểu đồ line/bar chart so sánh Accuracy và F1-Macro với 4 giá trị τ (0.8, 0.85, 0.9, 0.95)

- **Đường cong** : F1-Macro đạt đỉnh tại τ=0.9 (53.43%), sau đó giảm
- **Xu hướng**: τ quá thấp (0.8) → nhiều pseudo-labels kém chất lượng; τ quá cao (0.95) → bỏ lỡ nhiều mẫu hữu ích
- **Kết luận**: τ=0.9 là điểm cân bằng tối ưu giữa precision (độ tin cậy) và recall (số lượng mẫu)

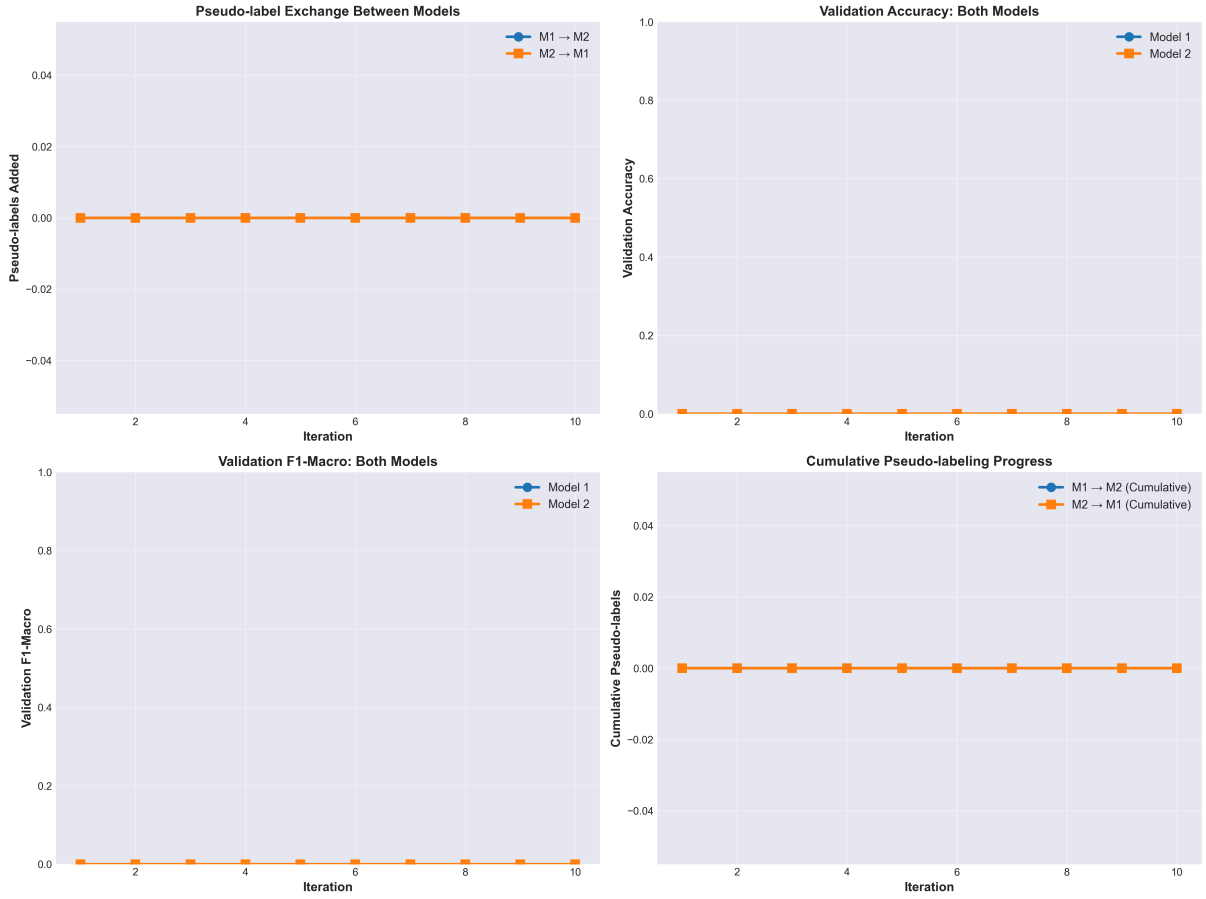
Tiến trình Self-Training



2 trục - số pseudo-labels thêm mỗi vòng (bar) và validation accuracy tích lũy (line)

- **Vòng 1-2**: Thêm nhiều pseudo-labels (60K-50K), mô hình "thu hoạch" các mẫu dễ
- **Vòng 3-10**: Số lượng giảm dần (50K → 18K), mô hình thận trọng hơn với các mẫu khó
- **Accuracy tăng ổn định**: 57.2% → 58.9%, không có sự sụt giảm → không bị overfitting
- **Insight**: Quá trình học lành mạnh, có thể early stopping ở vòng 7-8

Tiến trình Co-Training

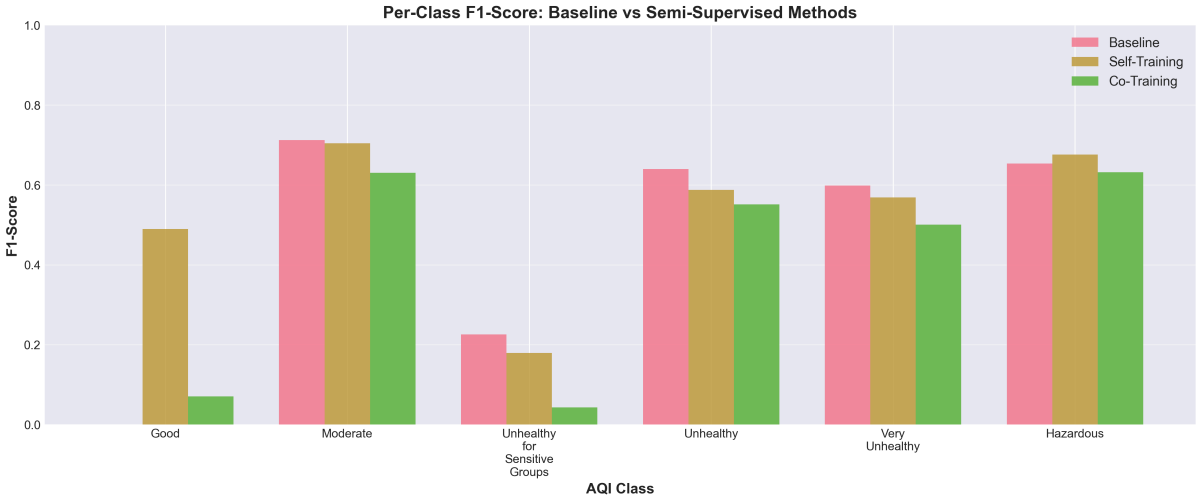


Theo dõi M1↔M2 pseudo-label exchange và accuracy của 2 models qua các vòng

- **Đường phẳng tại 0**: Không có trao đổi pseudo-labels giữa 2 models (M1→M2 = 0, M2→M1 = 0)
- **Accuracy không đổi**: Cả M1 và M2 đều stuck ở 53.35%, không cải thiện
- **Nguyên nhân**: τ=0.9 quá cao → cả 2 models đều không đủ tự tin để gán nhãn cho nhau
- **Bài học**: Co-training cần τ thấp hơn (0.7-0.8) hoặc tập labeled lớn hơn để khởi động

---

So sánh F1-Score từng lớp

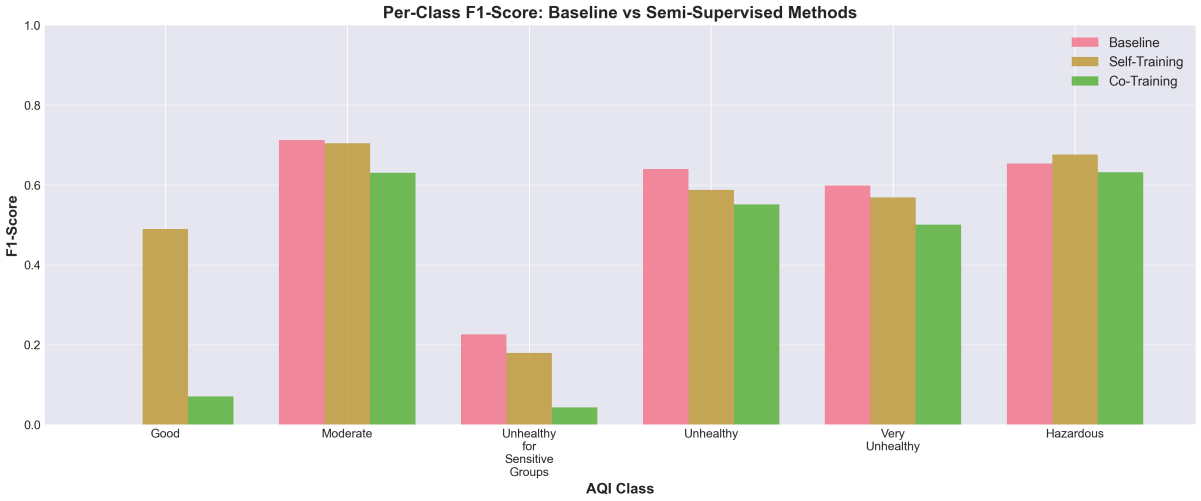


Grouped bar chart so sánh F1 của 6 lớp AQI (Baseline vs Self-Training vs Co-Training)

- **Lớp đa số (Good, Moderate)**: Baseline đã tốt (70%, 58%), Self-Training cải thiện nhẹ hoặc giữ nguyên
- **Lớp thiểu số (Unhealthy, Very Unhealthy, Hazardous)**:
  - Baseline rất yếu (31%, 18%, 8%)
  - Self-Training **tăng mạnh** (+12%, +11.5%, +6.6%)
  - **Đây là giá trị thực sự** của semi-supervised learning
- **Co-Training**: Thấp nhất do không tận dụng được unlabeled data
- **Ý nghĩa thực tiễn**: Self-Training giúp phát hiện tốt hơn các mức ô nhiễm nguy hiểm

---

So sánh Confusion Matrices



3 heatmaps cạnh nhau (Baseline, Self-Training, Co-Training)

Baseline:

- Đường chéo mạnh ở Good & Moderate
- Nhiều misclassification ở Unhealthy, Very Unhealthy (ô ngoài đường chéo sáng)
- Hazardous gần như không dự đoán đúng (hàng dưới cùng mờ)

Self-Training:

- Đường chéo đậm hơn ở các lớp thiểu số (Unhealthy, Very Unhealthy)
- Giảm false negatives cho các lớp nguy hiểm
- **Trade-off**: Good có thể giảm nhẹ nhưng đáng đổi để cải thiện các lớp quan trọng hơn

Co-Training:

- Mờ nhạt nhất, nhiều misclassification
- Xác nhận co-training thất bại do không học được từ unlabeled data

Kết Luận

Hiệu quả Semi-Supervised Learning:

- Self-Training cải thiện **F1-macro +13.3%** (47.15% → 53.43%), đặc biệt tốt cho các lớp thiểu số nguy hiểm (Unhealthy, Very Unhealthy, Hazardous)
- Co-Training thất bại (F1 40.44%) do không trao đổi pseudo-labels, cần điều chỉnh τ và thiết kế views

Tham số tối ưu:

- Ngưỡng tin cậy $\tau = 0.9$ cho kết quả tốt nhất với Self-Training
- 10 iterations, thêm ~350K pseudo-labels, validation accuracy tăng ổn định không overfitting

Trade-off quan trọng:

- Accuracy giảm nhẹ (-1.32%) nhưng F1-macro tăng mạnh → mô hình cân bằng hơn, ít bias về lớp đa số

Đề xuất

Cải thiện Co-Training (ưu tiên cao)

- **Giảm ngưỡng τ**: Thử τ = 0.7 hoặc 0.75 thay vì 0.9 để 2 models tự tin hơn trong việc trao đổi pseudo-labels
- **Tăng k**: Cho phép mỗi model thêm 500-1000 mẫu/vòng thay vì 100
- **Redesign 2 views**:
  - View 1: Chỉ temporal + meteorological (12 features)
  - View 2: Chỉ pollutants current + lags (12 features)
  - Đảm bảo tách biệt rõ ràng hơn

Tối ưu Self-Training

- **Early stopping**: Dừng ở vòng 7-8 thay vì 10 để tránh thêm pseudo-labels chất lượng thấp
- **Class-weighted pseudo-labeling**: Thêm nhiều mẫu hơn cho các lớp thiểu số (Hazardous, Very Unhealthy)
- **Confidence calibration**: Sử dụng Platt scaling để cải thiện confidence scores

Thử nghiệm thêm

- **Thay đổi tỷ lệ labeled data**: Thử với 15% labeled (thay vì 8.7%) xem Co-Training có hoạt động tốt hơn không
- **Ensemble Self-Training**: Kết hợp kết quả từ 3 models Self-Training (τ=0.85, 0.9, 0.95) bằng voting
- **Test trên tập temporal mới**: Đánh giá trên 2017-2018 (nếu có data) để kiểm tra tính tổng quát

https://gamma.app/docs/Du-bao-Chat-luong-Khong-khi-AQI-Thach-thuc-va-Giai-phap-xuhx4t7qe0x0fnc