Phân loại nước ngầm phục vụ tưới tiêu

Thành viên nhóm 27:
Trần Đăng Tài - 23001558
Nguyễn Việt Phúc - 23001547
Nguyễn Khắc Huy - 23001525

Giảng viên hướng dẫn: TS. Cao Văn Chung

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - CƠ - TIN HỌC

Ngày 10 tháng 1 năm 2026

Nội dung báo cáo

Chương 1
Giới thiệu chung

Giới thiệu đề tài

- Biến đổi khí hậu và nhu cầu sử dụng nước ngày càng tăng làm cho việc quản lý và khai thác **nước ngầm** trở nên đặc biệt quan trọng trong nông nghiệp.
- Chất lượng nước ngầm ảnh hưởng trực tiếp đến năng suất cây trồng, đồng thời tác động lâu dài đến đất canh tác và môi trường sinh thái.
- **Phân loại nước ngầm** dựa trên các chỉ tiêu hóa lý là một hướng tiếp cận hiệu quả nhằm đánh giá mức độ phù hợp cho tưới tiêu.

Giới thiệu đề tài

- Tiền xử lí dữ liệu.
- Giảm chiều dữ liệu:*Principal Component Analysis* (PCA) và *Linear Discriminant Analysis* (LDA)
- Phân cụm: *Gaussian Mixture Model* (GMM)
- Phân loại: *K-Nearest Neighbors* (KNN), *Multi-Layer Perceptron* (MLP), *Gaussian Naive Bayes* (GNB), *Softmax Regression* và *Support Vector Machine* (SVM).

Chương 2
Dữ liệu và tiền xử lí

Dữ liệu

Bộ dữ liệu

Có 3 tệp được thu thập trong các năm 2018, 2019 và 2020 chứa thông tin chi tiết về chất lượng nước ngầm sau mùa mưa.

Mỗi tập dữ liệu chứa 26 cột như: số sê-ri (sno), quận, mandal, làng, vĩ độ, kinh độ, 16 trường hóa chất (như Ca, Mg, CO3, v.v.), tổng độ cứng của nước, tổng chất rắn hòa tan, RSC, SAR và các biến mục tiêu 'Phân loại' và 'Phân loại 1'. Các cột đặc trưng có thể được sử dụng để dự đoán chất lượng nước, được phân loại thành 9 lớp, bao gồm: C1S1, C2S1, C3S1, C3S2, C3S3, C4S1, C4S2, C4S3, C4S4.

Tiền xử lí

Đặc điểm

Do tính chất dữ liệu được thu thập tại các thời điểm rời rạc (vào mùa mưa) trong 3 năm, nên các quy ước đặt tên trường có phần không thống nhất.

- Thống nhất các giá trị, và tên trường qua các năm.
- Loại bỏ các trường không cần thiết: các trường địa lí và trường có tương quan với trường khác $= 1$.
- Các trường có lượng bản ghi thiếu $< 5\%$ được fill bằng mean.

Fill CO3

Với trường CO3, các giá trị null chiếm tới $\sim 15\%$.

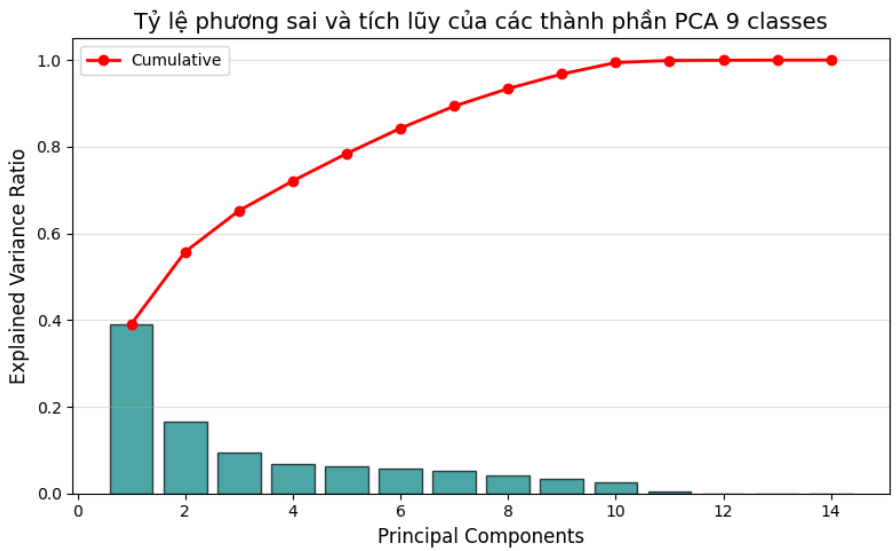
Xây dựng mô hình hồi quy tuyến tính với 4 biến RSC, Ca, Mg, HCO3.

- Log-transform, standard-scale.
- Huấn luyện mô hình tuyến tính đơn biến trong không gian log để giảm sai số do e mũ cho các giá trị > 0.
- MAE tính riêng cho các giá trị > 0 là 4.928.

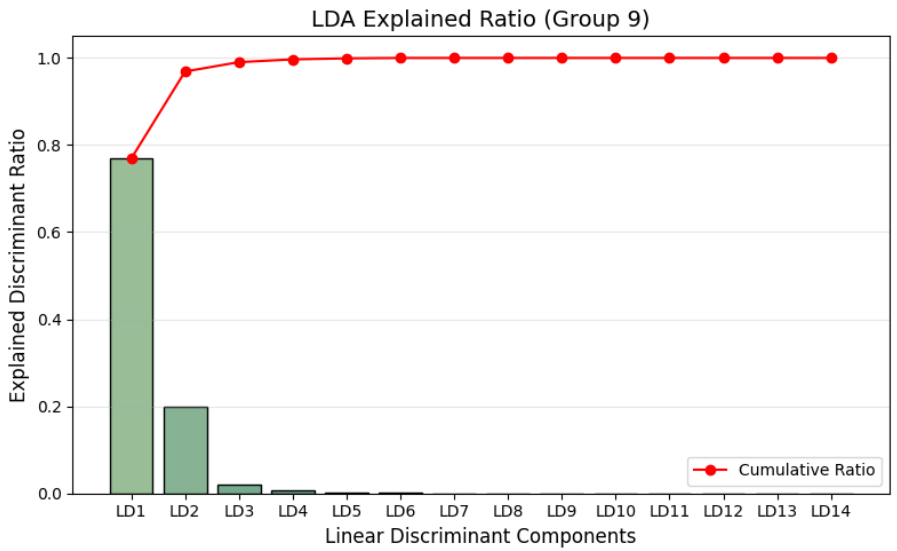
Không thể xóa các bản ghi null do sẽ làm giảm số lượng các lớp hiếm.
Phần lớn ($\sim 70.52\%$) CO3 là giá trị 0.
$\Rightarrow$ Xóa trường CO3 khỏi dữ liệu.

Chương 3
Kết quả thực nghiệm

Giảm chiều dữ liệu



Hình 1: Tỉ lệ giải thích tích lũy của PCA



Hình 2: Tỉ lệ giải thích tích lũy của LDA

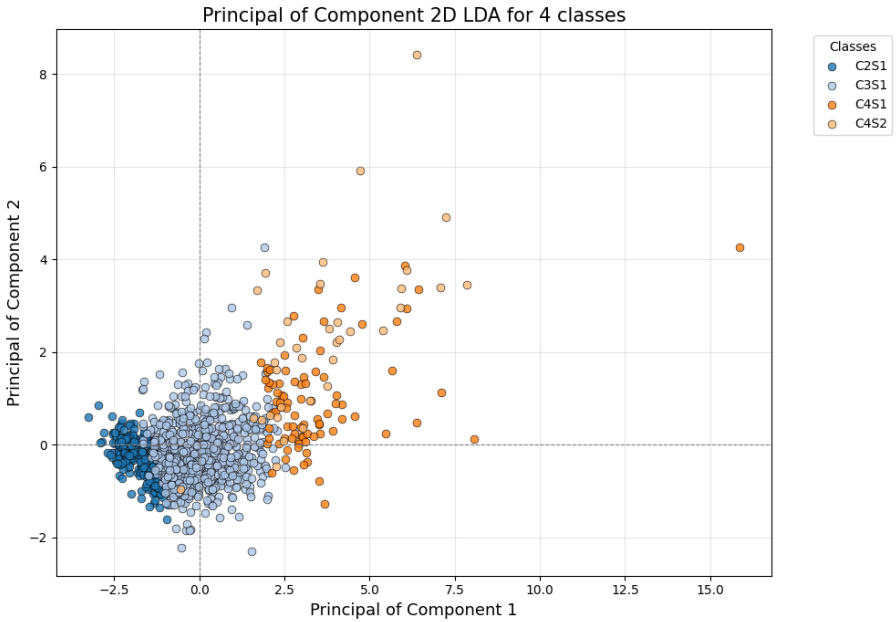
Giảm chiều dữ liệu

Bảng 1: Bảng kết quả đo đạc thời gian chạy của các mô hình

	Dữ liệu gốc	**Dữ liệu LDA 2 chiều**
K-NN	0.79382s	0.39405s
SVM	0.28080s	0.15043s
Softmax	0.30326s	0.17001s

$\Rightarrow$ Nhanh hơn $\sim 50\%$.

Phân cụm



Hình 3: Label các lớp của dữ liệu LDA 4 lớp

Hình 4: Kết quả phân cụm bằng GMM trên dữ liệu LDA 4 lớp

- Silhouette Score: 0.36992
- DBCV Score: -0.58674

Mô hình phân loại

Các kịch bản thử nghiệm

1. Huấn luyện trên bộ dữ liệu 9 class.
2. Huấn luyện trên bộ dữ liệu 4 class (giữ lại các lớp nhiều hơn 35 bản ghi)
3. Huấn luyện trên bộ dữ liệu 4 class áp dụng under/oversampling để tránh tác động từ mất cân bằng dữ liệu.
4. Huấn luyện trên bộ dữ liệu LDA 2 chiều 4 class

Mô hình phân loại – Kịch bản 1

Kịch bản 1: Huấn luyện trên bộ dữ liệu 9 class.

Bảng 2: Bảng kết quả phân loại các mô hình theo kịch bản 1

	Best accuracy	**Macro precision**	**Macro recall**	**Weighted precision**	**Weighted recall**
K-NN	0.94	0.53	0.55	0.94	0.95
SVM	0.93	0.58	0.53	0.94	0.93
GNB	0.85	0.56	0.65	0.87	0.85
MLP	0.92	0.58	0.55	0.91	0.92
Softmax	0.91	0.52	0.50	0.90	0.91

Mô hình phân loại – Kịch bản 2

Kịch bản 2: Huấn luyện trên bộ dữ liệu 4 class (giữ lại các lớp nhiều hơn 35 bản ghi)

Bảng 3: Bảng kết quả phân loại các mô hình theo kịch bản 2

	Best accuracy	**Macro precision**	**Macro recall**	**Weighted precision**	**Weighted recall**
K-NN	0.94	0.92	0.88	0.98	0.98
SVM	0.95	0.84	0.89	0.97	0.95
GNB	0.87	0.73	0.78	0.89	0.87
MLP	0.94	0.82	0.80	0.94	0.94
Softmax	0.93	0.77	0.75	0.92	0.93

Mô hình phân loại – Kịch bản 3

Kịch bản 3: Huấn luyện trên bộ dữ liệu 4 class áp dụng under/oversampling để tránh tác động từ mất cân bằng dữ liệu.

Bảng 4: Bảng kết quả phân loại các mô hình theo kịch bản 3

	Best accuracy	**Macro precision**	**Macro recall**	**Weighted precision**	**Weighted recall**
K-NN	0.96	0.97	0.97	0.97	0.97
SVM	0.95	0.80	0.81	0.96	0.95
GNB	0.86	0.69	0.81	0.89	0.86
MLP	0.9	0.75	0.83	0.91	0.9
Softmax	0.91	0.79	0.81	0.92	0.91

Mô hình phân loại – Kịch bản 4

Kịch bản 4: Huấn luyện trên bộ dữ liệu LDA 2 chiều 4 class

	Best accuracy	**Macro precision**	**Macro recall**	**Weighted precision**	**Weighted recall**
SVM	0.91	0.81	0.84	0.92	0.91
GNB	0.91	0.78	0.77	0.91	0.91
MLP	0.93	0.78	0.77	0.91	0.91
Softmax	0.93	0.90	0.81	0.93	0.93
K-NN	0.93	0.72	0.72	0.92	0.93

Bảng 5: Bảng kết quả phân loại các mô hình theo kịch bản 4

Chuyển về mô hình hồi quy

Lựa chọn một lớp mục tiêu c^* là lớp chiếm đa số trong tập huấn luyện và sử dụng giá trị hàm quyết định tương ứng với lớp c^* làm đầu ra liên tục mới y_{new}:

- **Softmax Regression:** $y_{\text{new}} = P(y = c^* \mid \mathbf{x})$ (xác suất softmax của lớp c^*).
- **MLP:** $y_{\text{new}} = P(y = c^* \mid \mathbf{x})$ (xác suất đầu ra của lớp c^*).
- **SVM:** $y_{\text{new}} = f_{c^*}(\mathbf{x})$ (giá trị *decision score* của lớp c^*).

Chuyển về mô hình hồi quy

Huấn luyện mô hình hồi quy Random Forest Regressor, Linear Regression để học ánh xạ $\mathbf{x} \to y_{\text{new}}$ và đánh giá chất lượng hồi quy trên tập kiểm tra thông qua các chỉ số: RMSE, MAE và R^2.
Nhóm tiến hành thử nghiệm trên 3 kịch bản:

1. Dữ liệu gốc 9 lớp.
2. Dữ liệu gốc 4 lớp.
3. Dữ liệu giảm chiều bằng LDA.

Chuyển về mô hình hồi quy – Kịch bản 1

Kịch bản 1: Dữ liệu gốc 9 lớp.

Bảng 6: Bảng kết quả hồi quy theo kịch bản 1

Mô hình	**RMSE**	**MAE**	R^2
Softmax	0.062924	0.020080	0.973496
MLP	0.071259	0.035548	0.969973
SVM	0.358363	0.159774	0.776685

Chuyển về mô hình hồi quy – Kịch bản 2

Kịch bản 2: Dữ liệu gốc 4 lớp.

Bảng 7: Bảng kết quả hồi quy theo kịch bản 2

Mô hình	**RMSE**	**MAE**	R^2
Softmax	0.062076	0.018708	0.974554
MLP	0.067304	0.032712	0.973246
SVM	0.268025	0.124287	0.837799

Chuyển về mô hình hồi quy – Kịch bản 3

Kịch bản 3: Dữ liệu giảm chiều bằng LDA 2 chiều.

Bảng 8: Bảng kết quả hồi quy theo kịch bản 3

Mô hình	**RMSE**	**MAE**	R^2
Softmax	0.039216	0.011932	0.988905
MLP	0.030365	0.014639	0.993951
SVM	0.171467	0.046370	0.961611

Chương 4
Kết luận

Tổng kết

Một số kết luận:

- Dữ liệu ban đầu bị mất cân bằng nặng và một số lớp có rất ít mẫu, làm hạn chế khả năng đánh giá công bằng trên bộ 9 lớp.
- Các kỹ thuật oversampling dạng nhân bản có thể gây nguy cơ overfitting đối với một số mô hình, đặc biệt khi số mẫu gốc của lớp hiếm quá nhỏ.
- Việc tinh chỉnh siêu tham số và đánh giá ổn định chưa được thực hiện đầy đủ do giới hạn thời gian và tài nguyên.
- Trong phần chuyển đổi sang hồi quy, y_{new} phụ thuộc vào mô hình phân loại; do đó kết quả hồi quy phản ánh khả năng xấp xỉ đầu ra của mô hình phân loại hơn là tối ưu trực tiếp theo nhãn thật.

Hướng phát triển

Trong tương lai, nhóm đề xuất một số hướng cải thiện và mở rộng:

- **Kỹ thuật xử lí mất cân bằng nâng cao:** thử nghiệm SMOTE/Borderline-SMOTE hoặc điều chỉnh trọng số lớp/loss để giảm lệch dự đoán mà không nhân bản quá mức.
- **Tối ưu mô hình và đánh giá:** thực hiện grid-search/bayes-search cho siêu tham số, đánh giá bằng cross-validation, và bổ sung các thước đo phù hợp khi dữ liệu mất cân bằng.
- **Giải thích mô hình:** áp dụng phân tích độ quan trọng đặc trưng để lý giải yếu tố nào ảnh hưởng mạnh đến chất lượng nước.
- **Mở rộng mô hình:** thử nghiệm các mô hình phi tuyến mạnh hơn (Gradient Boosting, XGBoost/LightGBM) và các phương pháp hiệu chuẩn xác suất (calibration) để tăng độ tin cậy của đầu ra xác suất.

Cảm ơn thầy và các bạn
đã lắng nghe!