

DỰ ĐOÁN HẠN HÀM DỰA TRÊN DỮ LIỆU THỜI TIẾT VÀ ĐẤT

CS313.P23 - Khai thác dữ liệu và ứng dụng

Group 5

Trường Đại học Công nghệ Thông tin - ĐHQG TP.HCM

Ngày 16 tháng 5 năm 2025

Mục Lục

- 1 Giới Thiệu
- 2 Phân Tích Khám Phá Dữ Liệu
- 3 Tiền xử lý dữ liệu
- 4 Mô hình
- 5 Thực nghiệm
- 6 Demo

Giới Thiệu về Hạn Hán

Hạn Hán là gì

- Hạn Hán là hiện tượng thời tiết khắc nghiệt do thiếu mưa kéo dài gây ảnh hưởng nghiêm trọng đến:
 - Nông nghiệp
 - Đời sống sinh hoạt
 - Môi trường



Lý Do Thực Hiện

Tầm quan trọng của việc dự đoán Hạn Hán

Dự báo chính xác mức độ nghiêm trọng và tác động của hạn hán đối với nông nghiệp và môi trường đóng vai trò quan trọng trong:

- Phòng ngừa cháy rừng
- Hỗ trợ người nông dân lên kế hoạch mùa vụ



Tập Dữ Liệu

- **Nguồn:** Bộ *Predict Droughts using Weather and Soil Data* của Christoph Minixhofer.
- **Dữ liệu thời tiết:**
 - Hơn 23 triệu bản ghi thời tiết và chỉ số hạn hán (DSCI).
 - Thu thập từ 2000–2020 trên 3.108 khu vực tại Hoa Kỳ.
- **Dữ liệu đất:**
 - 3.109 bản ghi, tương ứng với các mã FIPS.
 - Gồm thông tin độ cao, độ dốc, hướng, độ mặn, v.v.
- **Chia dữ liệu:**
 - Huấn luyện: 2000–2017, Kiểm thử: 2018–2020.

Phân tích khám phá dữ liệu: EDA

Time Data: EDA for weather data

Static Data: EDA for soil data

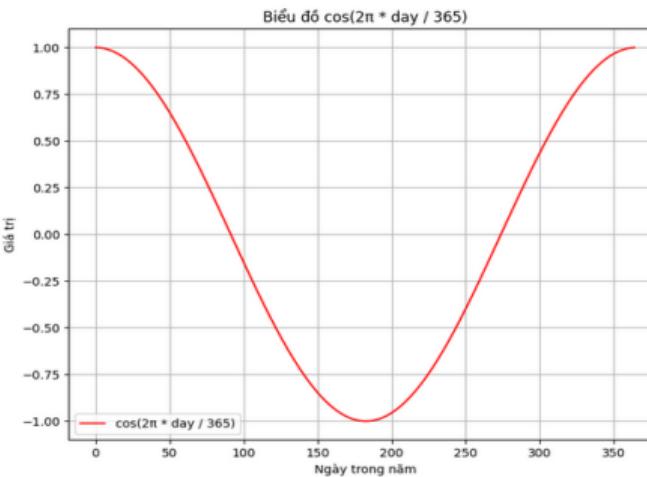
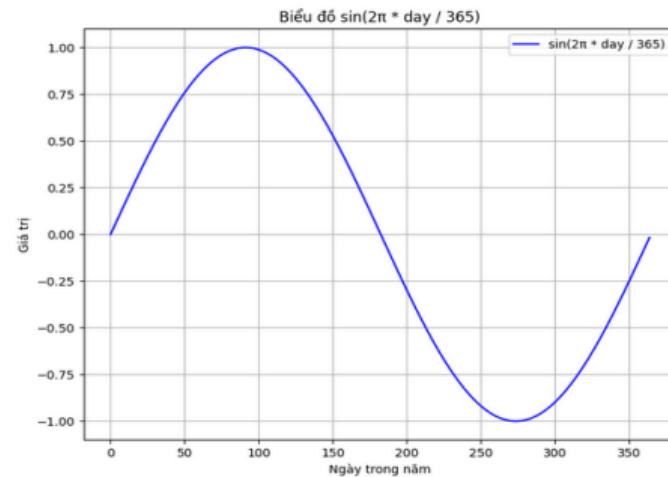
Tiền xử lý: Chuẩn hóa bằng RobustScaler

- **RobustScaler** là một phương pháp chuẩn hóa dữ liệu giúp giảm ảnh hưởng của các giá trị ngoại lai.
 - Thay vì sử dụng giá trị trung bình và độ lệch chuẩn, phương pháp này sử dụng:
 - **Trung vị (Median)**
 - **IQR (Khoảng tứ phân vị) = Q3 - Q1**
 - Công thức chuẩn hóa:
- $$x' = \frac{x - \text{median}}{\text{IQR}}$$
- **Ứng dụng:** Phù hợp với dữ liệu có nhiều nhiễu.

Tiền xử lý: Mã hóa tính chu kỳ thời gian

Dữ liệu có chu kỳ hàng năm (365 ngày), do đó cần mã hóa thời gian để mô hình nhận biết được chu kỳ này.

- $\sin\left(2\pi \frac{\text{ngày}}{365}\right)$
- $\cos\left(2\pi \frac{\text{ngày}}{365}\right)$



Tiền xử lý: Giá trị thiếu

Nội suy/ngoại suy tuyến tính

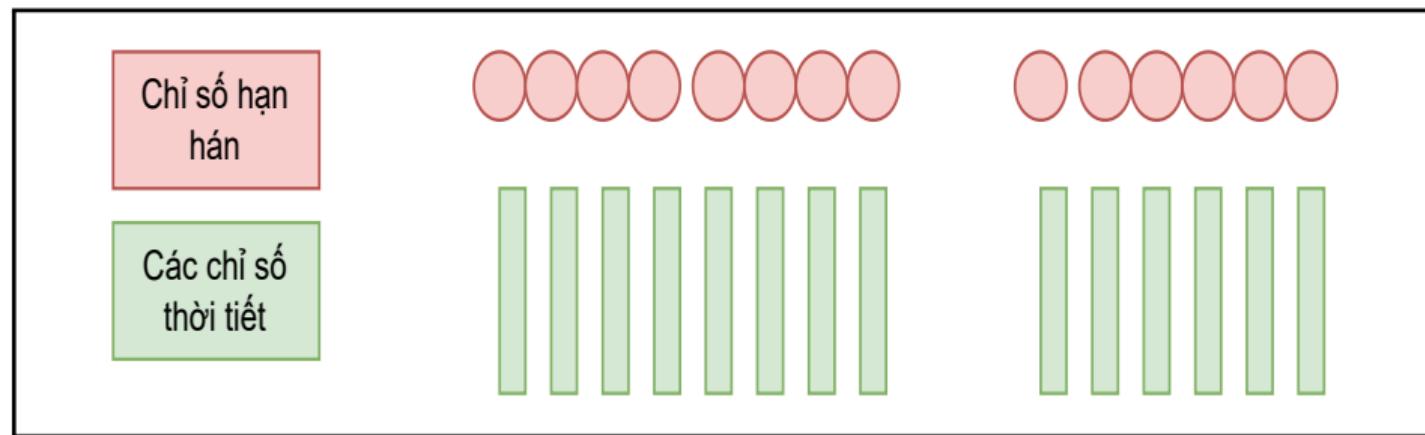
Phương pháp nội suy hoặc ngoại suy tuyến tính được áp dụng cho biến **score**. Lưu ý rằng các giá trị được điền thêm không được sử dụng làm nhãn (y), mà chỉ phục vụ cho các mục đích khác.

1	nan	nan	nan	5
---	-----	-----	-----	---

1	2	3	4	5
---	---	---	---	---

Ý tưởng chính

- Trong dự báo hạn hán, các dự đoán quá gần thường không mang lại nhiều giá trị, trong khi các dự đoán quá xa lại thiếu chính xác.
- Do đó, mục tiêu của nhóm là phát triển một mô hình có thể dự đoán mức độ và tác động của hạn hán trong 6 tuần tới, cân bằng giữa tính hữu ích và độ chính xác.



Hình 1: Hình minh họa ý tưởng chính của đề tài

Mô hình: Yêu cầu

Mô hình cần đáp ứng các yêu cầu sau:

- Là mô hình hồi quy, vì đầu ra là một giá trị thực trong khoảng [0, 5].
- Có khả năng xử lý dữ liệu chuỗi thời gian.
- Có thể tích hợp các đặc trưng tĩnh, ví dụ như thông tin đất đai.
- Hỗ trợ dự đoán nhiều đầu ra, tức là dự đoán chỉ số hạn hán cho 6 tuần tiếp theo.

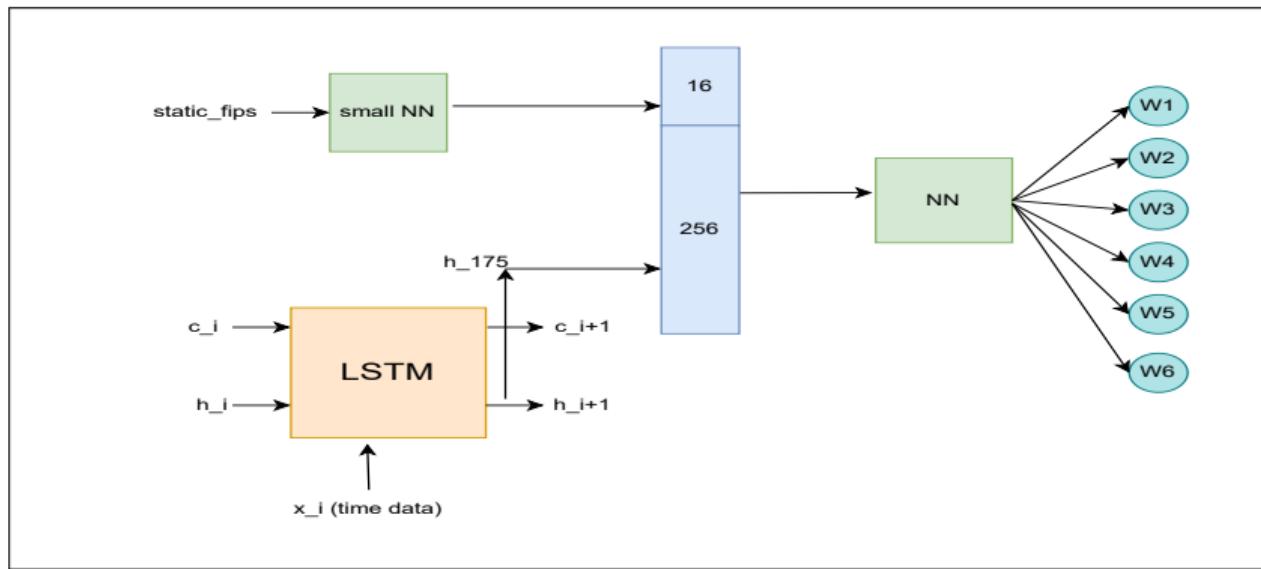
Kiến trúc mô hình

Đầu vào:

- Dữ liệu chuỗi thời gian: $\mathbf{X}_{\text{time}} \in \mathbb{R}^{175 \times 21}$, Đặc trưng tĩnh: $\mathbf{X}_{\text{static}} \in \mathbb{R}^{29}$

Đầu ra:

- Vector dự đoán: $\mathbf{y} \in \mathbb{R}^6$



Chuẩn bị dữ liệu đầu vào

Để huấn luyện mô hình, ta cần bộ ba:

$$(X_{\text{time}}, X_{\text{static}}, y)$$

- Xét riêng từng **fips** trong tập dữ liệu thời gian theo ngày (18 năm).
- Lấy X_{static} tương ứng từ dữ liệu đất (soil data).
- **Bắt đầu từ thứ Ba đầu tiên của năm 2000.**
- Sau đó:
 - Lấy 175 ngày liên tiếp để tạo X_{time} (kết thúc vào một ngày thứ Hai).
 - Dùng nội/ngoại suy tuyến tính để điền giá trị thiếu của biến **score** trong X_{time} .
 - Gán y là giá trị **score** vào:
 - Ngày thứ Ba kế tiếp (ngày 176).
 - Ngày thứ Ba của tuần thứ 6 (tức 5 tuần sau).
- Thực hiện **stride** với bội số của 7.
- Tiếp tục cho đến hết 18 năm, sau đó chuyển sang fips khác.

Kết quả thực nghiệm

Thiết lập:

- Tập đánh giá: dữ liệu từ năm 2018 đến cuối năm 2020.
- Tham số bước nhảy cố định = 7 cho tập đánh giá.
- Hàm mất mát MSELoss, thuật toán tối ưu AdamW, OneCycleLR schedule, 15 epochs.

Sai số MAE cho dự báo từng tuần:

Mô hình	Tuần 1	Tuần 2	Tuần 3	Tuần 4	Tuần 5	Tuần 6	Trung bình
GRU	0.0918	0.1456	0.1906	0.2357	0.2730	0.3102	0.2078
LSTM	0.0922	0.1432	0.1881	0.2338	0.2686	0.3075	0.2056
LSTM + Attention	0.1016	0.1495	0.1935	0.2343	0.2697	0.3003	0.2082
Transformer	0.1045	0.1484	0.1902	0.2324	0.2603	0.2910	0.2045

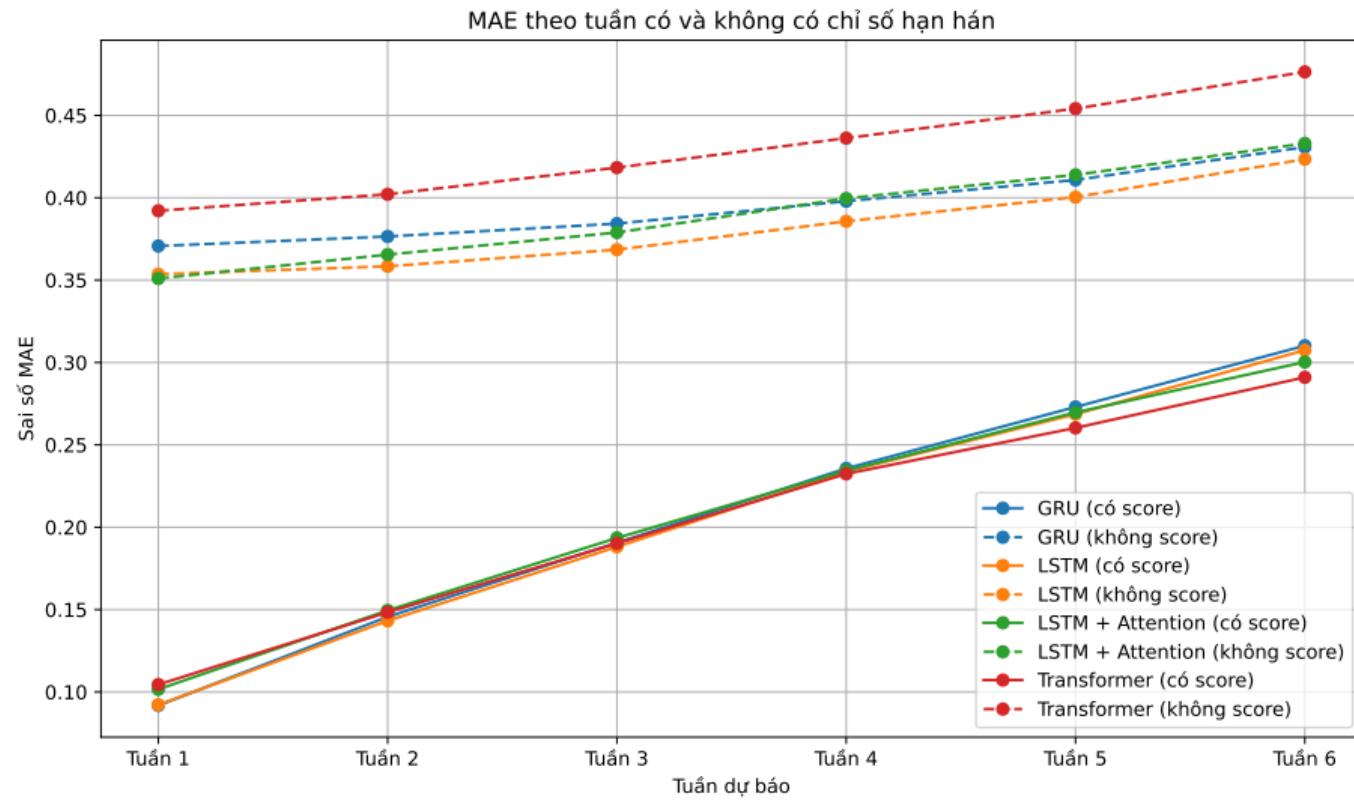
Kết quả thực nghiệm

- Chỉ số hạn hán score được lấy từ U.S. Drought Monitor và chỉ được thống kê trên nước Mỹ. Khó áp dụng cho những nơi khác.
- Do đó, loại bỏ score trong đầu vào để xây dựng một mô hình chỉ dựa vào thời tiết và đất để dự đoán mức độ hạn hán. Lưu ý rằng dữ liệu thời tiết và đất được lấy từ NASA POWER và Harmonized World Soil Database có trên khắp thế giới.

Sai số MAE cho dự báo từng tuần (không sử dụng score):

Mô hình	Tuần 1	Tuần 2	Tuần 3	Tuần 4	Tuần 5	Tuần 6	Trung bình
GRU	0.3707	0.3765	0.3843	0.3979	0.4108	0.4307	0.3951
LSTM	0.3536	0.3584	0.3685	0.3857	0.4004	0.4234	0.3817
LSTM + Attention	0.3510	0.3655	0.3789	0.3996	0.4139	0.4330	0.3903
Transformer	0.3921	0.4021	0.4183	0.4362	0.4541	0.4764	0.4298

Kết quả thực nghiệm



Ứng dụng dự đoán hạn hán

- **Mục tiêu:** Dự đoán chỉ số hạn hán 6 tuần tới từ:

- Dữ liệu thời tiết 175 ngày gần nhất.
- Thông tin đất tại khu vực phân tích

- **Dữ liệu thời tiết:**

- Nguồn: NASA POWER (tổn cầu, cập nhật hàng ngày, trễ 2–3 ngày)
- Truy xuất qua API theo tọa độ và thời gian
- Thông số: T2M, T2M_MAX, v.v.

- **Dữ liệu đất:**

- Nguồn: Harmonized World Soil Database v1.2
- Thông tin: loại đất, độ sâu, khả năng giữ nước theo tọa độ

Demo

Bản Đồ

Tim kiếm địa điểm...

Leaflet | © MapTiler © OpenStreetMap contributors

Thông Tin Vị Trí

Vĩ Độ

Nhập vĩ độ hoặc chọn trên bản đồ

Kinh Độ

Nhập kinh độ hoặc chọn trên bản đồ

Ngày Bắt Đầu Ngày Kết Thúc

11/13/2024 [Calendar]

05/12/2025 [Calendar]

Khu Vực

Toàn Thế Giới

Mô Hình

LSTM

Lấy Dữ Liệu

<https://github.com/l1aF-2027/Drought-Prediction.git>

Phân công công việc

Họ tên	MSSV	Nhiệm vụ	Mức độ đóng góp
Lê Văn Hoàng	22520465	Nghiên cứu, cài đặt mô hình, làm slide, thuyết trình	16.67%
Hà Huy Hoàng	22520460	Xây dựng ứng dụng (frontend, backend)	16.67%
Nguyễn Duy Hoàng	22520467	Nghiên cứu, viết báo cáo	16.67%
Đặng Vĩnh Hội	22520490	Nghiên cứu, backend	16.67%
Phạm Đức Huy Hoàng	22520474	Nghiên cứu, EDA	16.67%
Thạch Minh Hoàng	22520477	Nghiên cứu, EDA, thuyết trình	16.67%