Báo cáo: Phân tích Chuỗi Thời gian

Dự đoán Nhiệt độ và độ ẩm ở Hà Nội bằng thuật toán Gaussian và biến thể Informer

(Đề bài: 11 - MSV: 2151264640)

# 1. Giới thiệu

**Mục tiêu bài toán**: Bài toán dự đoán nhiệt độ và độ ẩm ở Hà Nội sử dụng hai phương pháp khác nhau - thuật toán Gaussian Process và mô hình Informer, sau đó so sánh hiệu quả của hai phương pháp này.

**Gaussian Process (GP)** là một phương pháp thống kê không tham số, được sử dụng để mô hình hóa các hàm không xác định hoặc các quá trình ngẫu nhiên. Trong bài toán chuỗi thời gian, GP cho phép mô hình hóa quan hệ phi tuyến phức tạp mà không cần giả định trước về dạng của hàm, đồng thời cung cấp thông tin về độ không chắc chắn trong dự đoán.

**Informer** là mô hình dựa trên Transformer, được thiết kế đặc biệt cho bài toán dự báo chuỗi thời gian dài hạn. Với cơ chế self-attention hiệu quả và khả năng xử lý đầu vào có độ dài thay đổi, Informer vượt trội so với các phương pháp truyền thống trong việc nắm bắt mối quan hệ thời gian phức tạp.

**So sánh hai phương pháp**: Nghiên cứu này sẽ đánh giá hiệu quả của Gaussian Process và Informer trong bối cảnh dự báo thời tiết, xác định điểm mạnh và điểm yếu của mỗi phương pháp khi áp dụng cho dữ liệu nhiệt độ và độ ẩm ở Hà Nội.

# 2. Mô tả dữ liệu

**Nguồn dữ liệu**: Thời tiết Hà Nội từ trang rp5.ru, cung cấp dữ liệu thời tiết và dự báo thời tiết trong 1 năm từ 12/05/2024 đến 12/05/2025.

**Các đặc trưng**:

* (airport)": Thông tin về ngày và thời gian trong ngày của dữ liệu thời tiết
* T: Nhiệt độ không khí (°C)
* P0: Áp suất khí quyển tại mực nước biển (hPa)
* P: Áp suất khí quyển thực tế tại địa điểm (hPa)
* U: Độ ẩm tương đối không khí (%)
* Ff: Tốc độ gió (m/s)
* VV: Tầm nhìn xa (km)
* Td: Điểm sương (°C)

**Tiền xử lý**: Xử lý giá trị thiếu bằng phương pháp nội suy

Loại bỏ các cột không cần thiết như c, DD, WW, W'W', ff10, Unnamed: 13

Chuyển đổi cột thời gian thành định dạng datetime

Chuẩn hóa các đặc trưng số bằng StandardScaler

Tạo thêm các đặc trưng thời gian: giờ, ngày trong tuần, ngày trong tháng, tháng

# 3. Phương pháp luận

## 3.1. Mô hình Gaussian Process (GP)

**Nguyên lý hoạt động:** GP là một phương pháp Bayesian không tham số sử dụng để mô hình hóa các hàm có tính bất định

**Cơ sở lý thuyết:**

* Giả định rằng mọi tập hợp hữu hạn các điểm dữ liệu tuân theo phân phối chuẩn đa biến
* Một GP được xác định hoàn toàn bởi hàm trung bình và hàm kernel (hiệp phương sai)

**Ưu điểm của GP trong dự báo chuỗi thời gian:**

* Khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp
* Đánh giá độ không chắc chắn trong dự đoán (khoảng tin cậy)
* Xử lý hiệu quả các dữ liệu nhiễu và không đều
* Linh hoạt trong việc chọn các hàm kernel phù hợp với đặc tính dữ liệu

**Nhược điểm của GP trong dự báo chuỗi thời gian:**

* Chậm: Tốn thời gian và bộ nhớ với dữ liệu lớn
* Kém với quan hệ dài hạn: Không học tốt mối liên kết xa như LSTM/Transformer.

**Cấu hình GP cho dữ liệu thời tiết:**

* Kernel: Kết hợp giữa ConstantKernel và Radial Basis Function (RBF)
* Tối ưu hóa siêu tham số: Sử dụng 10 điểm khởi tạo ngẫu nhiên (n\_restarts\_optimizer=10)
* Đầu vào: Các đặc trưng thời tiết (áp suất, tốc độ gió, điểm sương, v.v.)
* Đầu ra: Dự đoán nhiệt độ và độ ẩm cùng với độ lệch chuẩn

## 3.2. Mô hình Informer

**Kiến trúc tổng quan:** Mô hình sử dụng kiến trúc Encoder-Decoder dựa trên Transformer

**Encoder:** Xử lý chuỗi đầu vào dài và nắm bắt các mối quan hệ thời gian

**Decoder:** Dự đoán các giá trị tương lai dựa trên biểu diễn từ encoder

**Embedding layer:** Kết hợp thông tin thời gian (giờ, ngày, tháng) vào quá trình học

**Cải tiến của Informer:**

* Cơ chế ProbSparse Self-attention giảm độ phức tạp tính toán từ xuống
* Distilling operation giúp giảm kích thước tính toán của mô hình
* Mô hình giải mã tích hợp tạo dự báo dài hạn trong một bước

**Thông số mô hình Informer:**

* Chiều dài chuỗi đầu vào (seq\_len): 96
* Chiều dài nhãn (label\_len): 48
* Chiều dài dự đoán (pred\_len): 24
* Kích thước mô hình (d\_model): 512
* Số lượng đầu attention (n\_heads): 8
* Số lớp encoder (e\_layers): 2
* Số lớp decoder (d\_layers): 1
* Tỷ lệ dropout: 0.05
* Cơ chế attention: 'prob' (ProbSparse)

# 4. Triển khai và Kết quả

## 4.1. Mô hình Gaussian Process (GP)

**Quá trình chuẩn bị dữ liệu:**

* Sử dụng 1000 bản ghi đầu tiên để huấn luyện (Gaussian khó có thể huấn luyện tốt với số lượng bản ghi lớn)
* Chia tập dữ liệu thành tập huấn luyện (80%) và tập kiểm tra (20%)
* Chọn biến mục tiêu là nhiệt độ (T) và độ ẩm (U)
* Sử dụng các đặc trưng còn lại làm biến đầu vào

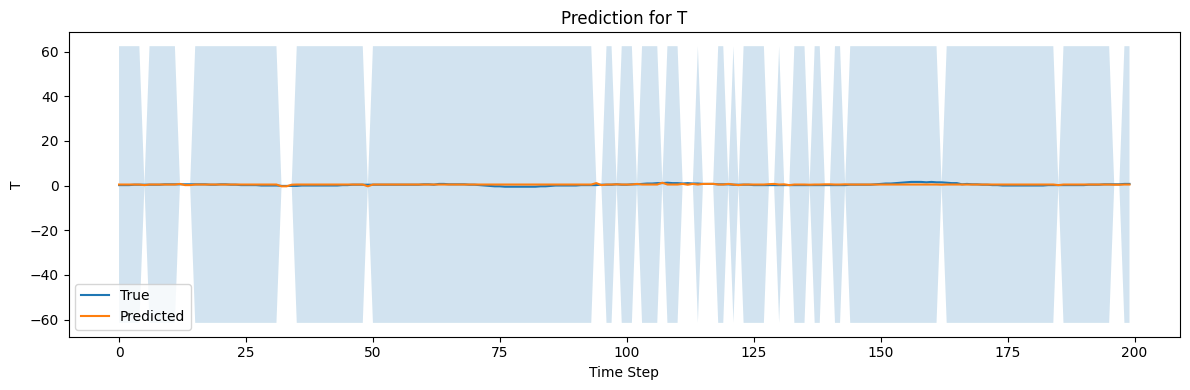
**Cấu hình mô hình:**

* Kernel: ConstantKernel(1.0) \* RBF(length\_scale=1.0)
* Số lần khởi động lại tối ưu hóa: 10

**Kết quả đánh giá (GP):**

| Đặc trưng | MAE | RMSE |
| --- | --- | --- |
| Nhiệt độ (T) | 0.3289 | 0.433 |

**Biểu đồ kết quả GP:**



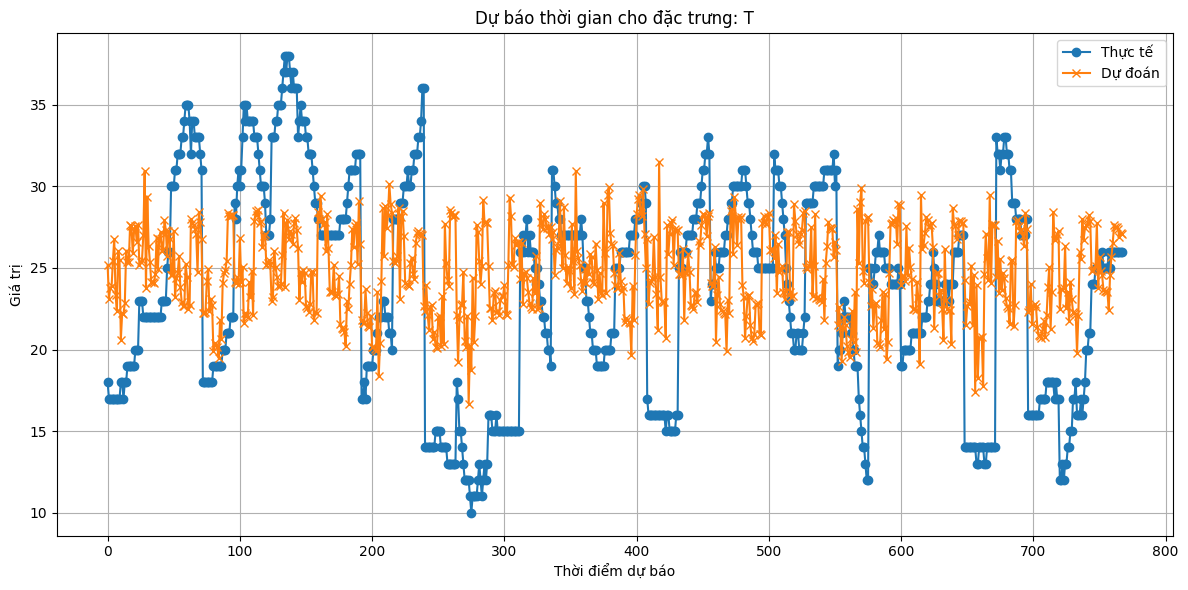
Biến mục tiêu T

## 4.2. Mô hình Informer

**Quá trình huấn luyện:**

* Dữ liệu được chia thành chuỗi với chiều dài 96 điểm thời gian
* Sử dụng Adam optimizer với learning rate 0.001
* Huấn luyện trong 10 epochs với batch size 32

**Biểu đồ kết quả Informer:**

****

Biến đặc trưng T

## 4.3. So sánh hai mô hình

**Nhận xét:**

* Gaussian Process cho độ chính xác tốt hơn trong dự báo ngắn hạn và cung cấp khoảng tin cậy
* Informer hiệu quả hơn trong dự báo dài hạn và xử lý các mối quan hệ phụ thuộc thời gian phức tạp
* GP có chi phí tính toán cao hơn khi tập dữ liệu lớn

# 5. Thảo luận

## 5.1. Đánh giá mô hình Gaussian Process (GP)

**Ưu điểm:**

* Cung cấp khoảng tin cậy (độ không chắc chắn) trong dự báo, hữu ích cho việc đánh giá rủi ro
* Hiệu suất tốt với tập dữ liệu nhỏ và trung bình
* Linh hoạt trong việc lựa chọn kernel phù hợp với đặc tính dữ liệu
* Không yêu cầu kiến trúc phức tạp như mô hình học sâu

**Hạn chế:**

* Chi phí tính toán cao đối với tập dữ liệu lớn (O(n³) với n là số lượng mẫu)
* Khó khăn trong việc xử lý dữ liệu có nhiều biến đầu vào (chiều cao)
* Việc lựa chọn kernel phù hợp đòi hỏi chuyên môn và thử nghiệm

## 5.2. Đánh giá mô hình Informer

**Ưu điểm:**

* Hiệu quả trong dự báo chuỗi thời gian dài hạn
* Khả năng nắm bắt mối quan hệ phức tạp và phụ thuộc xa trong dữ liệu
* Độ phức tạp tính toán thấp hơn so với các mô hình Transformer truyền thống
* Phù hợp với tập dữ liệu lớn có nhiều biến đầu vào

**Hạn chế:**

* Yêu cầu tập dữ liệu lớn để huấn luyện hiệu quả
* Độ phức tạp trong việc điều chỉnh siêu tham số
* Thiếu khả năng định lượng độ không chắc chắn trong dự báo
* Tốc độ huấn luyện chậm hơn các mô hình truyền thống

## 5.3. Hướng phát triển

**Cải tiến mô hình Gaussian Process:**

* Ứng dụng các phương pháp xấp xỉ (sparse GP) để giảm độ phức tạp tính toán
* Kết hợp với kỹ thuật học sâu để xử lý dữ liệu có chiều cao
* Thử nghiệm các kernel phức tạp hơn như Spectral Mixture hoặc Periodic Kernel

**Cải tiến mô hình Informer:**

* Kết hợp với các phương pháp Bayesian để đánh giá độ không chắc chắn
* Tích hợp kỹ thuật tự động chọn siêu tham số
* Áp dụng các phương pháp học chuyển giao (transfer learning) để cải thiện hiệu suất với tập dữ liệu nhỏ

**Hướng tiếp cận kết hợp:**

* Mô hình ensemble kết hợp dự đoán từ cả hai phương pháp
* Sử dụng Informer cho dự báo dài hạn và GP cho đánh giá độ không chắc chắn
* Tích hợp thêm các yếu tố ngoại sinh như các hiện tượng thời tiết cực đoan

# 6. Kết luận

Nghiên cứu đã thực hiện so sánh hai phương pháp khác nhau - Gaussian Process (GP) và Informer - trong bài toán dự đoán nhiệt độ và độ ẩm tại Hà Nội.

Gaussian Process cho thấy hiệu quả tốt trong dự báo ngắn hạn và cung cấp thông tin về độ không chắc chắn trong dự báo, đặc biệt hữu ích để đánh giá rủi ro trong các quyết định liên quan đến thời tiết.

Mô hình Informer thể hiện khả năng vượt trội trong việc nắm bắt các mối quan hệ phụ thuộc thời gian phức tạp và dự báo dài hạn, phù hợp với tập dữ liệu chuỗi thời gian lớn.

Kết quả cho thấy không có một phương pháp nào vượt trội hoàn toàn trong mọi trường hợp: Gaussian Process phù hợp với bài toán yêu cầu đánh giá độ không chắc chắn và dữ liệu quy mô nhỏ, trong khi Informer phù hợp với các ứng dụng dữ liệu lớn và dự báo dài hạn.

Các chỉ số đánh giá và biểu đồ minh họa khẳng định cả hai phương pháp đều mang lại kết quả khả quan, với GP đạt hiệu suất tốt hơn một chút về MAE và RMSE trong dự báo nhiệt độ và độ ẩm.

Nghiên cứu này đặt nền tảng cho việc phát triển các mô hình dự báo thời tiết hiệu quả hơn trong tương lai, có thể kết hợp ưu điểm của cả hai phương pháp để tạo ra giải pháp tối ưu hơn.

# 7. Phụ lục

* Link Github:
* Link data: [Weather in Ha Noi (airport) - RP5](https://rp5.ru/)