

Chronos-2

Time Series Forecasting

Dàn ý thuyết trình

1. Giới thiệu & Động lực
2. Tổng quan bộ dữ liệu
3. Thiết lập bài toán
4. Phương pháp & Mô hình
5. Khung đánh giá
6. Kết quả
7. Thảo luận & Phân tích
8. Kết luận & Hướng nghiên cứu trong tương lai

Giới thiệu

Các mô hình nền tảng (foundation models) đã tạo ra cuộc cách mạng trong lĩnh vực học máy trên các bài toán Xử lý Ngôn ngữ Tự nhiên (NLP) và Thị giác Máy tính (Computer Vision).

Chronos-2:

Một mô hình nền tảng mới dành cho bài toán dự báo chuỗi thời gian.

Câu hỏi then chốt:

Liệu Chronos-2 ở chế độ zero-shot có thể sánh ngang với các phương pháp truyền thống hay không?

Mục tiêu nghiên cứu

- So sánh Chronos-2 zero-shot với 3 mô hình nền mạnh làm đường cơ sở
- Đánh giá mức độ hiệu chuẩn (calibration) của dự báo xác suất
 - Kiểm định ý nghĩa thống kê của các khác biệt
- Phân tích các dạng sai số và những trường hợp mô hình hoạt động kém

Tổng quan bộ dữ liệu

Nguồn:

Wikimedia Pageviews
API

Đối tượng dự báo:

Trang Wikipedia về
Bitcoin

Khoảng thời gian:

01/01/2020 – 31/12/2024 (5
năm)

Tần suất:

Quan sát theo ngày

Tổng số bản ghi:

1.827 ngày

Tiền xử lý dữ liệu

- **Giá trị thiếu:** Không có (0 giá trị thiếu)
- **Ngoại lệ (outliers):** 38 (2,08%) được xử lý bằng phương pháp winsorization
- **Giới hạn giá trị:** 3.563 đến 36.912 lượt xem

Chiến lược chia dữ liệu

Chia theo thời gian (Temporal Split)

Huấn luyện: 60% (1.096 ngày)

Tháng 01/2020 – Tháng 12/2022

Xác thực: 20% (365 ngày)

Tháng 01/2023 – Tháng 12/2023

Kiểm tra: 20% (366 ngày)

Tháng 01/2024 – Tháng 12/2024

Nguyên tắc chính

- ✓ Tuân thủ nghiêm ngặt thứ tự thời gian
- ✓ Không rò rỉ dữ liệu
- ✓ Không sử dụng thông tin tương lai
- ✓ Đánh giá sát với kịch bản thực tế

Problem Setup

Chân trời dự báo:

$H = 30$ ngày

Chu kỳ mùa vụ:

$m = 7$ (theo tuần, được xác nhận bằng STL)

Nhiệm vụ:

Dự báo đơn biến
(không sử dụng biến ngoại sinh)

Chỉ số đánh giá chính:

MASE (Mean Absolute Scaled Error
– Sai số tuyệt đối trung bình đã chuẩn hóa)

Vì sao chọn MASE?

- Chỉ số không phụ thuộc vào thang đo
 - Đối xứng và được xác định rõ ràng
- $MASE < 1,0$ cho thấy mô hình tốt hơn đường cơ sở ngây thơ (naive baseline)

Các mô hình được đánh giá

Seasonal Naive

Dự báo = giá trị của
cùng ngày trong tuần
trước

$m = 7$

Gradient Boosting

LightGBM với các đặc
trưng độ trễ (lag) được
thiết kế

100 cây, độ sâu = 5

ETS

Holt–Winters với xu
hướng và mùa vụ dạng
cộng

Tối ưu bằng phương
pháp MLE

Chronos-2

Mô hình T5 encoder–
decoder (200 triệu
tham số), chế độ
zero-shot

20 mẫu → suy ra các
phân vị (quantiles)

Đặc trưng cho Gradient Boosting

Đặc trưng độ trễ (Lag Features):
[1, 7, 14, 28] ngày

Trung bình trượt (Rolling Mean):
cửa sổ [7, 28] ngày

Độ lệch chuẩn trượt (Rolling Std):
cửa sổ [7, 28] ngày

Mã hóa ngày trong tuần (Day-of-Week Encoding):
0–6

Tầm quan trọng của đặc trưng (Top 5)

- lag_7 (tuần trước): 35%
- lag_1 (ngày trước): 22%
- rolling_mean_7: 18%
- rolling_mean_28: 12%
- day_of_week: 8%

Khung đánh giá

Phương pháp:

Backtesting theo điểm gốc trượt (Rolling-Origin) với cửa sổ mở rộng (Expanding Window)

Số lượng fold:

$k = 5$

Chân trời dự báo cho mỗi fold:

$H = 30$ ngày

Các chỉ số được tính toán

- **Dự báo điểm:** MAE, RMSE, sMAPE, MASE

- **Dự báo xác suất:** Pinball loss, độ bao phủ khoảng dự báo (Interval coverage)

- **Thống kê:** Kiểm định Wilcoxon Signed-Rank ($\alpha = 0,05$)

Kết quả: Hiệu năng trên tập xác thực

Cross-Validation 5 fold

Model	MAE	RMSE	sMAPE	MASE
Gradient Boosting ★	873 ± 293	$1,150 \pm 471$	14.3%	0.344
Chronos-2	999 ± 200	$1,375 \pm 428$	16.7%	0.394
Seasonal Naive	$1,007 \pm 243$	$1,348 \pm 385$	16.6%	0.397
ETS	$1,288 \pm 92$	$1,662 \pm 323$	23.0%	0.508

Gradient Boosting đạt MASE thấp hơn Chronos-2 khoảng 14%

Kết quả: Hiệu năng trên tập kiểm tra

Tập holdout cuối cùng (năm 2024)

Model	MAE	RMSE	sMAPE	MASE
Gradient Boosting ★	2,739	4,231	35.8%	1.080
Seasonal Naive	2,745	4,407	35.8%	1.082
Chronos-2	2,836	4,722	37.1%	1.118
ETS	6,984	8,453	149.5%	2.754

Giai đoạn kiểm tra (năm 2024) thể hiện mức độ biến động cao
hơn so với tập xác thực

Ý nghĩa thống kê

Kiểm định Wilcoxon Signed-Rank ($\alpha = 0,05$)

Comparison	Statistic	p-value	Significant?
Seasonal Naïve vs. GB	4,527.0	0.033	✓ Yes
ETS vs. GB	3,093.0	<0.001	✓ Yes
Chronos-2 vs. GB	4,601.5	0.047	✓ Yes

Gradient Boosting vượt trội một cách có ý nghĩa thống kê so với
TẤT CẢ các mô hình, bao gồm cả Chronos-2

Chất lượng dự báo xác suất

Hiệu chuẩn của Chronos-2

Độ bao phủ khoảng 80%: 14,2%
(Kỳ vọng: 80%)

Độ rộng khoảng trung bình:
769

Pinball Loss

- ($\tau = 0,1$): 156
- ($\tau = 0,5$): 499
- ($\tau = 0,9$): 426

Diễn giải

- ⚠ Thiếu bao phủ nghiêm trọng
 - ⚠ Khoảng dự báo quá hẹp
 - ⚠ Dự báo quá tự tin
- Nguyên nhân có khả năng:
- Mức độ biến động cao trong năm 2024
 - Hạn chế về hiệu chuẩn trong chế độ zero-shot

Vì sao Gradient Boosting vượt trội

Mùa vụ theo tuần mạnh:

đặc trưng lag_7 nắm
bắt rõ chu kỳ $m = 7$
(chiếm 35% tầm quan
trọng)

Kỹ thuật tạo đặc trưng hiệu quả:

kết hợp tín hiệu ngắn
hạn và dài hạn

Tinh chỉnh theo miền dữ liệu:

được tối ưu riêng cho
mô hình truy cập
trang Wikipedia

Mã hóa tường minh:

ngày trong tuần và
các thống kê trượt

Nhận định then chốt

- Việc thiết kế đặc trưng nắm bắt rõ các quy luật đã biết giúp mô hình vượt trội hơn các foundation model mang tính tổng quát trong bài toán dự báo đơn chuỗi

Vì sao Chronos-2 hoạt động tốt (nhưng chưa phải tốt nhất)

Học zero-shot:

Không cần huấn luyện
trên dữ liệu Wikipedia

Kiến trúc tổng quát:

Không được tinh chỉnh
riêng cho các mẫu tuần
cụ thể

Hiệu năng cạnh tranh:

Chỉ kém mô hình tốt
nhất khoảng 14%

Điểm mạnh của Chronos-2

- ✓ Không cần thiết kế đặc trưng
 - ✓ Cung cấp dự báo xác suất
 - ✓ Triển khai nhanh cho các danh mục đa dạng
- ✓ Phù hợp hơn với các miền mới thiếu dữ liệu lịch sử

Phân tích sai số

Theo chân trời dự báo:

MAE tăng gấp đôi từ $h = 1$ (~600) lên $h = 30$ (~1.200)

Theo mức độ lưu lượng:

Tất cả các mô hình đều gặp khó khăn ở giai đoạn lưu lượng cao (MAE ~1.500–2.000)

Theo từng fold:

Gradient Boosting ổn định nhất (độ lệch chuẩn = 0,116), Chronos-2 cũng rất ổn định (độ lệch chuẩn = 0,079)

Mẫu hình then chốt

- Các đợt bùng nổ lan truyền và sự kiện tin tức tạo ra mức biến động vốn dĩ khó dự đoán, gây thách thức cho mọi phương pháp dự báo

Hạn chế

- **Miền dữ liệu đơn lẻ:** Chỉ thử nghiệm trên trang Wikipedia về Bitcoin
- **Không tinh chỉnh siêu tham số:** Chronos-2 sử dụng thiết lập mặc định
- **Chân trời cố định:** Chỉ đánh giá $H = 30$
- **Không có biến ngoại sinh:** Không đưa vào các sự kiện tin tức, ngày lễ
- **Yêu cầu GPU:** Chronos-2 cần tối thiểu 4GB VRAM
- **Giai đoạn biến động cao:** Tập kiểm tra năm 2024 có mức biến động bất thường

Hàm ý thực tiễn

- **Đối với chuỗi đơn:** Sử dụng Gradient Boosting với các đặc trưng độ trễ
- **Đối với danh mục đa dạng:** Chronos-2 mang lại sự tiện lợi của zero-shot
- **Đối với môi trường sản xuất:** Cân nhắc mô hình tổ hợp (Gradient Boosting + Chronos-2)
- **Đối với bất định:** Cần hiệu chuẩn hậu kỳ (post-hoc calibration) cho các khoảng dự báo của Chronos-2

Kết luận

Gradient Boosting vượt trội một cách có ý nghĩa thống kê so với Chronos-2.

MASE:
0,344 so với 0,394 ($p = 0,047$)

Việc thiết kế đặc trưng giúp nắm bắt hiệu quả mùa vụ theo tuần.

Chronos-2 có giá trị đối với các danh mục đa dạng, nhưng không tối ưu cho bài toán dự báo một chuỗi đơn lẻ.

Thông điệp then chốt

- Với dự báo mang tính đặc thù miền dữ liệu và có các mẫu hình rõ ràng, nên đầu tư vào thiết kế đặc trưng. Với nhu cầu triển khai nhanh trên nhiều chuỗi, các foundation model zero-shot mang lại giá trị thực tiễn.

Hướng nghiên cứu trong tương lai

Mở rộng sang nhiều trang Wikipedia thuộc các chủ đề khác nhau

Tích hợp các biến ngoại sinh (sự kiện tin tức, ngày lễ)

Tinh chỉnh Chronos-2 trên dữ liệu đặc thù của Wikipedia

Khám phá các phương pháp tổ hợp (Gradient Boosting + Chronos-2)

Thử nghiệm các chân trời dự báo khác nhau

Khả năng tái lập

- Mã nguồn: github.com/Vipproplayerone1/ts-chronos-gpu
 - Seed: 42 | Python 3.10 | PyTorch 2.5.1
 - Thời gian chạy: ~20 phút trên RTX 3050

**Thank
You**