Báo cáo Anomaly detection

Trend, season analysis

Mục lục

I. Giới thiệu chung

I.1. Anomaly detection:

Mục đích:

- Show tất cả các chỉ số có điểm gần nhất hoặc trend gần nhất bị bất thường. Chức năng này chạy đầu ngày trên tất cả các chỉ số. Data được chạy độc lập đầu ngày và site chỉ cần đọc kết quả này để show ra màn hình cho user.

Format data:

Date,metric1,metric2,metric3,...,metric1_anopoint,metric2_anopoint,metric3_anopoint,...,metric1_trend,metric2_trend,metric3_trend,...,metric1_anotrend,metric2_anotrend,metric3_anotrend,...
Các cột data được split bằng dấu ","

Giải thích:

- metric1,metric2,metric3: data gốc
- metric1_anopoint,metric2_anopoint,metric3_anopoint: có giá trị 0/1 (0: không phải anomaly point, 1: là anomaly point. Nếu có giá trị là 1 thì point sẽ được highlight.
- metric1_trend,metric2_trend,metric3_trend: có giá trị là trend
- metric1_anotrend,metric2_anotrend,metric3_anotrend: nhận giá trị 0/1 (0: không phải anomaly trend, 1: là anomaly trend. Nếu có giá trị là 1 thì đường trend nối 3 điểm gần nhất được highlight.

Ví dụ:

date	**msg**	**msgtext**	**msg_anopoint**	**msgtext_anopoint**	**msg_trend**	**msgtext_trend**	**msg_anotrend**	**msgtext_anotrend**
8/16/2016	12345678	2345678	0	0	12345000	2345600	0	0
8/15/2016	12245678	2245678	0	1	12245000	2245600	0	0
8/14/2016	12145678	2145678	1	0	12145000	2145600	1	1

I.2. Trend, season analysis:

Mục đích:

- Show ra chỉ số, trend, season.

Layout:

- User chọn chỉ số, ngày bắt đầu, ngày kết thúc rồi nhấn **Decompose** để phân tích thành 2 phần: trend, season. Mỗi lần chỉ đượcchọn 1 metric.

Format data:

Date,metric1,metric2,metric3,...,metric1_trend,metric2_trend,metric3_trend,...,metric1_season,metric2_season,metric3_season
Giải thích:

- Metric1,metric2,metric3: số liệu gốc tại ngày đó
- Metric1_trend,metric2_trend,metric3_trend: giá trị là trend tại ngày đó
- Metric1_season,metric2_season,metric3_season: giá trị là season tại ngày đó

II. Giải thuật

II.1. Anomaly points detection

- Ý tưởng chung

Decompose timeseries[3] để có thành phần **remainder**. Sau đó, ta tính **quantile 75%** và **quantile 25%** trên tập **remainder** thu được, rồi tìm outliers.

Thông thường, ta sẽ chạy thuật toán decompose cho toàn tập data. Nhưng nhiều trường hợp, data khá lớn, có biến động riêng trong từng khoảng, làm cho outliers detect không chính xác.

Với mỗi observation, ta xét data từ observation đó trở về trước (khoảng 8 chu kì), coi đó là 1 tập **sub_data**, rồi áp dụng thuật toán lên tập **sub_data** đó. Kết quả, ta chỉ thu về 1 điểm giá trị của outlier cho mỗi lần chạy thuật toán.

Với data có n observation, ta cần chạy thuật toán tới O(n) lần.

- Cụ thể hơn về thuật toán:

Để decompose timeseres, ta dùng package **Seasonal**[1] trong python.

Package **Seasonal** hỗ trợ 1 vài phương pháp xác định trend như *line, mean, median, spline*

Ở đây, ta sử dụng **median**[2], áp dụng lên trên từng tập **sub_data** (như đã nói ở trên) để tính **trend**.

Sau khi có **trend** và **seasonal**, ta tính được **remainder**.

Ta định nghĩa khoảng *normal* cho dữ liệu theo **quantile**[4]

low_threshold = q25 - IQR * 1.5
high_threshold = q75 + IQR * 1.5

Những điểm nào nằm ngoài khoảng [low_threshold, high_threshold] là outliers.
Áp dụng công thức đó vào tập **remaider** thu được, ta sẽ tìm ra các outliers trong data gốc.

II.2. Anomaly trend detection.

Ta chạy thuật toán tính **trend** như mục **II.1**, sau đó xét hiệu của k điểm liên tiếp, nếu k hiệu số đó đều âm thì điểm cuối cùng là điểm bất thường.

Ví dụ với trend: 15, 18, 13, 12, 11, 10, 9, 14.

Ta có các điểm bất thường với k = 4 (màu đỏ): 15, 18, 13, 12, 11, 10, 9, 14.

II.3. Tính trend

Trend sẽ được tính trên toàn tập data, thay vì tính riêng trên từng **sub_data** như mục **II.1** và **II.2**

II.4 Tính season[5]

Dựa vào cross-validated residual errors. Test tính ảnh hưởng của seasonal bằng cách dùng R^2 of the leave-one-out cross-validation.

Tài liệu tham khảo:

[1] https://github.com/welch/seasonal

[2] https://en.wikipedia.org/wiki/Median_filter

[3] https://www.otexts.org/fpp/6/1

[4] http://www.statsdirect.com/help/content/nonparametric_methods/quantiles.htm

[5] https://github.com/welch/seasonal/blob/master/seasonal/seasonal.py