Anomaly detection for web traffic data

Di He (dh3171), Xinli Gu (xg588), Bo Zhang (bz854)

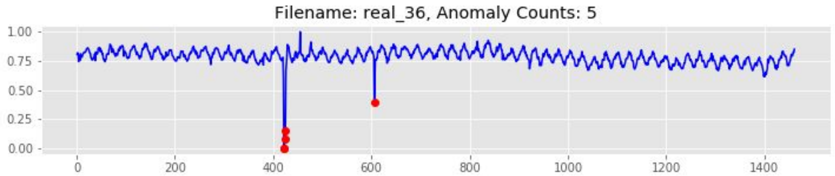
Problem Formulation

Why do we want to detect anomalies?

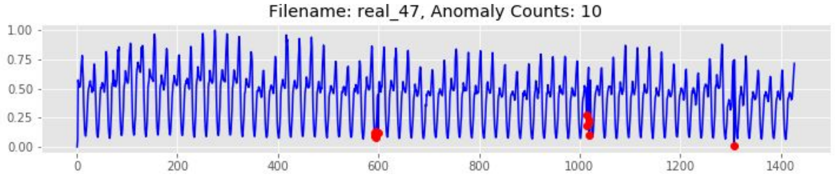
Dataset: Yahoo web traffic data (A1 benchmark)

Three Types of Anomaly

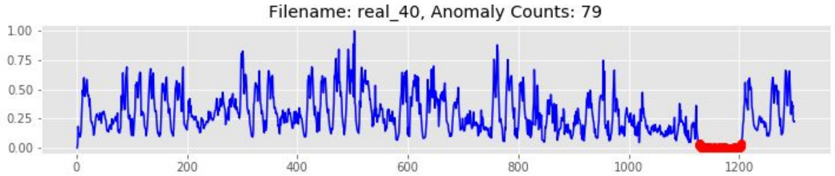
- Point Anomaly



- Contextual Anomaly



- Collective Anomaly



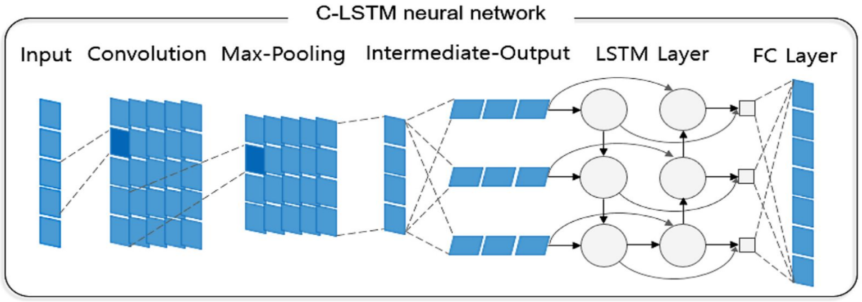
Key idea for the solution

ARIMA: for single dataset, offline

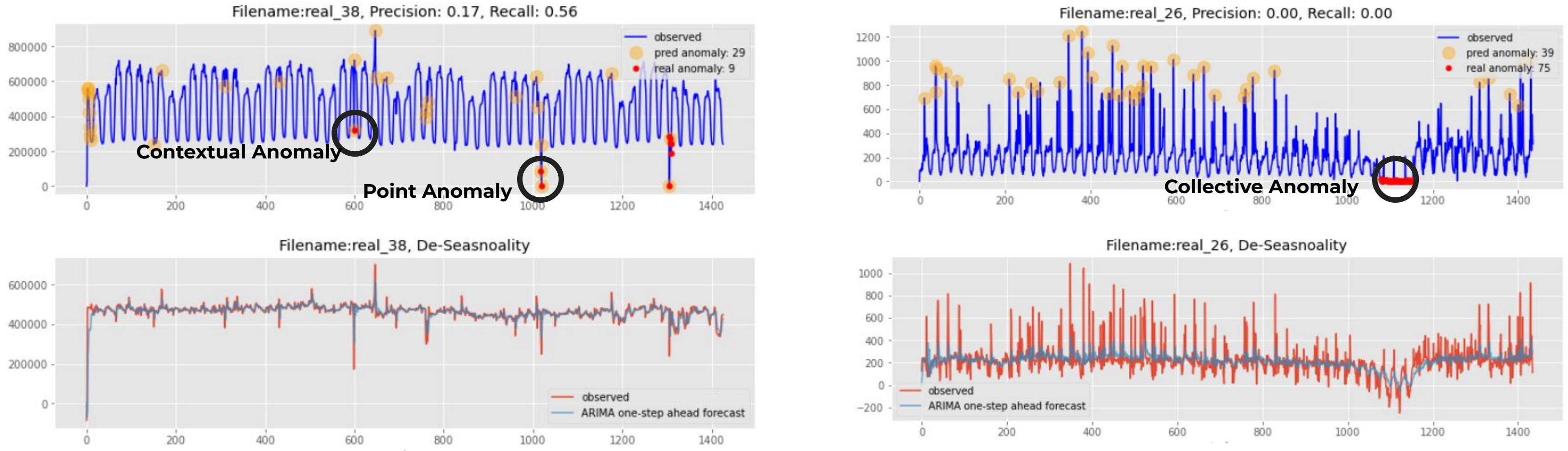
1. Remove the daily and weekly seasonality
2. Fit the ARIMA model on de-seasonality time series
3. Get the prediction error for ARIMA one-step ahead forecast
4. Detect the anomaly timestamps where the prediction error is beyond 3 standard deviation

C-LSTM: for combined dataset, online

1. Spatial features are extracted from time-series data by using a CNN
2. Passing these features through the LSTM to identify how temporal modeling of spatial characteristics in data affects performance



ARIMA



Results

- It can detect some point and contextual anomalies.
- It could hardly detect the collective anomaly.
- Peaks are easily to be falsely detected as anomalies.

C-LSTM

	Accuracy	Recall	Precision	F1
C-LSTM Tanh	91.2%	79.1%	55.9%	65.5%
C-LSTM ReLU	89.6%	72.9%	48.9%	58.5%
CNN	89.2%	71.1%	47.7%	57.1%

Results

- C-LSTM outperforms CNN model
- C-LSTM is able to do online prediction, with low inference time
- C-LSTM results in relatively high recall rate, which is desired for anomaly detection

Conclusions & Future Work

ARIMA: Offline Method

C-LSTM: Online Method

Future work:

- **Solve data imbalance problem (1.76% Anomalies)**
- **Classify Anomalies to Different Categories**

Thanks