How to Determine the Optimal Anomaly Detection Method For Your Application

Cynthia Freeman
Research Engineer

Jonathan Merriman
Software Engineer
Background
Time Series

- A time series is a sequence of data points indexed in order of time.
- How are time series used?
  - Stock Market
  - Tracking KPIs
  - Medical Sensors
  - Weather Patterns
Anomalies

An anomaly in a time series is a pattern that does not conform to past patterns of behavior.

Applications:
- Efficient troubleshooting
- Fraud detection
- Ensuring undisrupted business
- Saving lives in system health monitoring
Anomaly Detection is Hard

- What is anomalous?
- Online anomaly detection
- Lack of labeled data
- Data imbalance
- Minimize false positives
- Plethora of anomaly detection methods
Which anomaly detection method should I use?

- Base this decision off of the characteristics the time series possesses
- Evaluate anomaly detection methods on 4 time series characteristics as an example
- Experiment with 2 evaluation criteria
  - Window-based F-score
  - Numenta Anomaly Benchmark (NAB) Score
Signal Processing Flow for Anomaly Detection

1. signal
2. filter
3. residual
4. score
5. detect
Simple Example: Gaussian

- Estimate mean and variance over sliding window
- Compute a score based on the tail probability
  \[ S(y_t) = P(y_t \leq \tau | \mu, \sigma^2) \]
- Use max relative to upper and lower extremes
Simple Example: Gaussian
Time Series Characteristics
Seasonality

- Presence of variations that occur at specific regular intervals
- Real data often exhibits seasonal effects at multiple time scales.
  - Day-of-week
  - Hour-of-day
  - Can be irregular
    - Day-of-month
    - Holidays
- ACF plot is one way to detect seasonality
Concept Drift

The underlying process can change over time.

- Bayesian Online Changepoint Detection
- ecp package in R

https://github.com/hildensia/bayesian_changepoint_detection
Trend

The process mean can change over time.
Missing Time Steps
Time Series Modeling for Anomaly Detection
Nonstationarity: Differencing

- First-order difference to remove trend:

\[ [\Delta y](t) = y(t) - y(t - 1) \]

- Seasonal differencing with period s:

\[ [\Delta_{sy}](t) = y(t) - y(t - s) \]
Nonstationarity: Decomposition STL

Local regression with LOESS

\[ y(t) = S(t) + T(t) + \epsilon(t) \]

- Decompose into season and trend
- LOESS smoothing can interpolate missing data
- Residual should look more stationary
ARMA

A family of Gaussian models with temporal correlation.

\[ y(t) - \sum_{i=1}^{p} \theta_i y(t - i) = \epsilon(t) + \sum_{j=1}^{q} \phi_j \epsilon(t - j) \]

**Autoregressive (AR)**

The value at time \( t \) is a linear combination of \( p \) past values plus current noise signal.

**Moving Average (MA)**

The value at time \( t \) is a linear combination of \( q \) past values of noise.
ARMA for Nonstationary Signals

ARIMA
   ARMA on differenced signal.

SARIMA
   Extend ARIMA to incorporate longer-term seasonal correlation.

SARIMAX
   Add eXogenous variables.
ARMA

- Generative model having Gaussian distribution at each timestep
- Optimal model order selection is not straightforward
- See: Box-Jenkins method
Prophet

Uses an additive model:

\[ y(t) = g(t) + s(t) + h(t) + \epsilon_t \]

- \( g(t) \) is linear/logistic growth trend
- \( s(t) \) is yearly/weekly seasonal component
- \( h(t) \) is user-provided list of holidays

https://github.com/facebook/prophet
Extreme Studentized Deviate Test

How many outliers does the data set contain? ESD test requires an upper bound on the number of outliers. Assuming data is approximately normally distributed,

1. Compute the statistic,

\[ R_i = \frac{\max_j |x_i - \bar{x}|}{s} \]

2. Remove observation that maximizes \( |x_i - \bar{x}| \), and repeat

3. Compare \( R_i \) up to critical value
Twitter Anomaly Detection

- Uses STL but replaces trend with median
  - Anomalies can affect trend estimation
  - Leads to artificial anomalies in the residual
- Apply Extreme Studentized Deviate (ESD) test
  - Need to specify an upper limit on the # of outliers
  - $\bar{x}$ is median and $s$ is Median Absolute Deviation
Recurrent Neural Network

- Given a window of $n_{lag}$ time steps in the past, predict a window of $n_{seq}$ time steps in the future
- Anomaly score is an average of the prediction error
- Adaptive: uses online gradient-based optimizer, built to deal with concept drift
- Choice of $n_{seq}$ can greatly affect false positive rate

Illustration from Saurav et al. ’18
HTM for Anomaly Detection

Hierarchical Temporal Memory Network

- HTM outputs sparse representation of input and next prediction step to determine the prediction error modeled as a rolling normal distribution
- HTM not implemented in a widely accessible way
- Cannot handle missing time steps innately

Illustration from Ahmad et al. ’17
HOT-SAX

Heuristically Ordered Timeseries - Symbolic Aggregated ApproXimation

- Finds Discords: Subsequences of time series that are maximally different from all remaining subsequences
- Transform timeseries into alphabetical symbols and compare the distances between words
- Not built for concept drift detection
- Inefficient for very large time series

Illustrations from Keough et al. 2005
Evaluation Strategies
Anomaly Scores

Anomaly detectors are adapted to output a score between 0 and 1

- HTM: Use provided score
- Twitter AD and HOT-SAX: Use binary determination
- Windowed gaussian: Apply Q function to standardized signal
- STL, SARIMA, Prophet: Apply Q function to standardized residual
Numenta Anomaly Benchmark Scoring

- For every predicted anomaly $y$, its score $\sigma(y)$ is determined by its position relative to its containing window or an immediately preceding window.
- For every ground truth anomaly, construct an anomaly window with the anomaly in the center.

$$\frac{.1 \times \text{length of time series}}{\# \text{ of true anomalies}}$$

Illustration from Lavin & Ahmad '15
Numenta Anomaly Benchmark Scoring (Continued)

- The raw score is computed as:

\[ S_d = \left( \sum_{y \in Y_d} \sigma(y) \right) + A_{FN} f_d \]

- \( A_{FN} \) is cost of false negatives

- Then rescale to get summary score:

\[ 100 \times \frac{S - S_{null}}{S_{perfect} - S_{null}} \]

- Choose threshold that maximizes score
Window-based F-score

- Segment into nonoverlapping windows
- Window is anomalous if it contains an anomaly
- Treat like binary classification and report $F_1$
- Choose threshold that minimizes # of errors
- Prefer detection in case of tie
Results and Conclusions
## Characteristic Corpora

<table>
<thead>
<tr>
<th>Seasonality</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 datasets</td>
<td>10 datasets</td>
</tr>
<tr>
<td>63,336 samples</td>
<td>31,596 samples</td>
</tr>
<tr>
<td>23 ground truth anomalies</td>
<td>17 ground truth anomalies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept Drift</th>
<th>Missing Timesteps</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 datasets</td>
<td>10 datasets</td>
</tr>
<tr>
<td>32,402 samples</td>
<td>33,245 samples</td>
</tr>
<tr>
<td>27 ground truth anomalies</td>
<td>22 ground truth anomalies</td>
</tr>
<tr>
<td></td>
<td>1,254 missing samples</td>
</tr>
</tbody>
</table>

[https://github.com/numenta/NAB](https://github.com/numenta/NAB)
Which methods are promising given a characteristic?

Seasonality and Trend
   STL, SARIMA, Prophet

Concept Drift
   Requires more complex methods such as HTMs

Missing Time Steps
   ▶ Performance varies based on evaluation strategy
   ▶ Area for future work: more methods needed!
Which evaluation strategy should I use?

- F-score scheme is more restrictive
- NAB scores have more wiggle room for false positives due to reward for early detection
- What evaluation metric to use is entirely based on the needs of the user
In Summary

► The existence of an anomaly detection method that is optimal for all domains is a myth
► Determine the characteristics present in the data to narrow down the choices for anomaly detection methods
Questions?

Cynthia Freeman
cynthia.freeman@verint.com

Jonathan Merriman
jonathan.merriman@verint.com

https://github.com/cynthiaw2004/adclasses
Rate today’s session

Cyberconflict: A new era of war, sabotage, and fear

9:15am-10:10am Wednesday, March 27, 2019
Location: Ballroom

We’re living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you’re often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict where we’re uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the New York Times as well as a national security and political contributor for CNN and a frequent guest on CBS This Morning, Face the Nation, and many PBS shows.
Timing

Average time to generate anomaly scores: