PRECISION AND RECALL FOR TIME SERIES

Nesime Tatbul
Mejbah Alam
Justin Gottschlich
Tae Jun Lee
Stan Zdonik

intel

BROWN

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada
Motivation: Time Series Anomaly Detection

- **Anomaly:** Patterns that do not conform to expected behavior.
- Anomalies can have critical impact: loss of life, property damage, monetary loss, ...
- Applications of anomaly detection (AD) are numerous and diverse.

**Autonomous Driving**

- Six levels of autonomy:
  - L0: No automation
  - L1: Driver assistance
  - L2: Partial automation
  - L3: Conditional automation
  - L4: High automation
  - L5: Full automation

L3+ autonomy requires robust AD systems.

**Cancer Detection**

- Anomalies often occur over a period of time.

Source: Society of Automotive Engineers (SAE), National Highway and Traffic Safety Administration (NHTSA)

Motivation: Range-based Anomalies

- Time series anomalies are **range based**, i.e., they occur over a period of time.

- There are **domain-specific application preferences**.
  - Cancer detection, Real-time systems:
    - Early response; Avoid false negatives!
  - Robotic defense systems:
    - Delayed response; Avoid false positives!
  - Emergency braking in self-driving cars:
    - Neither too early nor too late; Avoid false negatives!

Problem: How to Measure Accuracy?

Point-based Anomalies

- False Negatives (FN)
- True Positives (TP)
- False Positives (FP)

Real

Predicted

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

Range-based Anomalies

- Partial overlap at back-end of R₁ & front-end of P₁
- Two real anomaly ranges R₂ and R₃ overlapping with one predicted anomaly range P₂

- Must express partial detection
- Must support flexible time bias
State of the Art

- Classical Precision and Recall
  - Point-based anomalies
  - Precision penalizes FP, Recall penalizes FN
  - $F_\beta$-Score to combine and weight them

- Numenta Anomaly Benchmark (NAB) [2]
  - Point-based anomalies
  - Focuses specifically on early detection use cases
  - Difficult to use in practice (irregularities, ambiguities, magic numbers) [3]

- Activity recognition metrics
  - No support for flexible time bias

$F_\beta = (1 + \beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}$

$\beta$: relative importance of Recall to Precision
- $\beta = 1$: evenly weighted (harmonic mean)
- $\beta = 2$: weights Recall higher (i.e., no FN!)
- $\beta = 0.5$: weights Precision higher (i.e., no FP!)

We extend classical Precision and Recall to measure ranges.

Our model is:
- expressive
- flexible
- extensible
Customization Examples

**Overlap Size \( \omega() \)**

```python
function \( \omega(AnomalyRange, OverlapSet, \delta) \)
    MyValue ← 0
    MaxValue ← 0
    AnomalyLength ← length(AnomalyRange)
    for \( i \) ← 1, AnomalyLength do
        Bias ← \( \delta(i, AnomalyLength) \)
        MaxValue ← MaxValue + Bias
        if AnomalyRange[i] in OverlapSet then
            MyValue ← MyValue + Bias
    return MyValue/MaxValue
```

**Positional Bias \( \delta() \)**

```python
function \( \delta(i, AnomalyLength) \)
    return 1  # Flat bias
function \( \delta(i, AnomalyLength) \)
    return AnomalyLength - i + 1  # Front-end bias
function \( \delta(i, AnomalyLength) \)
    return i  # Back-end bias
function \( \delta(i, AnomalyLength) \)
    if \( i \leq \frac{AnomalyLength}{2} \) then
        return i
    else
        return AnomalyLength - i + 1  # Middle bias
```

**Cancer Detection:**
- Set \( \delta() = \text{Front-end}, \beta = 2 \)

**Robotic Defense:**
- Set \( \delta() = \text{Back-end}, \beta = 0.5 \)

**Emergency Braking:**
- Set \( \delta() = \text{Middle}, \beta = 1.5 \)

Our model **subsumes the classical point-based model**, when:
- all ranges are represented as unit-size ranges, and
- \( \alpha = 0, \gamma() = 1, \omega() \) is as above, and \( \delta() = \text{Flat} \)
Selected Experimental Results

Please see our paper for details of this experimental study and additional results.

Comparison to Classical model
- Our model
  - subsumes the classical model
  - is sensitive to positional bias

Comparison to Numenta model
- Our model can
  - mimic the Numenta model
  - catch additional intricacies

Multiple Anomaly Detectors
- Our model is more effective in
  - evaluating multiple detectors
  - capturing subtleties in data
Key Takeaways

- This work extends the classical Precision and Recall model to time series data.
- We provide tunable parameters to capture domain-specific application preferences.
- Experiments with diverse datasets and anomaly detectors prove the benefits of our approach.
- Future work includes:
  - designing new training strategies for range-based anomaly detection
  - exploring use in other time series classification tasks and applications
More Information

Watch our short video:

https://www.youtube.com/watch?v=K5f-dUBiQP4

Read our paper:


Download our tool:

https://github.com/IntelLabs/TSAD-Evaluator/

Visit our poster session at NeurIPS’18:

Today at 5:00 - 7:00 PM in Room 210 & 230 AB #116

Thanks to Intel and NSF for funding this research.