Measuring the Stability of Process Outcome Predictions in Online Settings

Suhwan Lee¹  Marco Comuzzi²  Xixi Lu¹  Hajo A. Reijers¹

¹Utrecht University, Utrecht, The Netherlands
²Ulsan National Institute of Science and Technology
Outline

1. Introduction

2. Related works & Research question

3. Meta-measures

4. Experiment results

5. Conclusion & Future works
Predictive process monitoring

- Create loan Application
- Retrieve Application data
- Obtain installments data
- Check Eligibility
  - Approval
  - Rejection

Prediction model

Introduction
Related works
Meta-measures
Experiment
Conclusion
Online predictive process monitoring

- Training
- Training
- Training

Introduction  Related works  Meta-measures  Experiment  Conclusion
Model is updating?

Model performance is also updated
Performance of online predictive model

Accuracy

0.71

0.83

0.67

Average

0.71

0.83

0.73

Prediction model accuracy

Number of finished cases

Introduction

Related works

Meta-measures

Experiment

Conclusion
What is the best model?

Scenario A

Scenario B

Accuracy

Average

Introduction
Related works
Meta-measures
Experiment
Conclusion
Research question

How to assess the **stability** of models for online predictive process monitoring?
Related works

Single aggregated value

- 0.71
- 0.83
- 0.67

Average

Time-series visualization

- 0.73

Do not assess the model’s fluctuation in performance
Motivation - Business scenario

- **Frequency**
  - **High**
    - E.g., Predictive maintenance
  - **Low**
    - Non-critical Scenario

- **Risk**
  - **High**
    - E.g., Diagnosis and treatment in the emergency department
  - **Low**
    - E.g., Diagnosis and treatment of critical diseases

**Adapt** to business environment change

**Stable** performance over time

---

**Introduction**

**Related works**

**Meta-measures**

**Experiment**

**Conclusion**
Continuous performance evaluation

F1-Score

Stable? Drop?

Model update count

Introduction  Related works  Meta-measures  Experiment  Conclusion
Continuous performance evaluation

<table>
<thead>
<tr>
<th></th>
<th>F1-Score ($p_t$)</th>
<th>Moving Average ($ma_t$)</th>
<th>Moving Standard deviation ($\varphi_t$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.550</td>
<td>0.520</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>0.490</td>
<td>0.510</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>0.520</td>
<td>0.520</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>0.520</td>
<td>0.513</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>0.520</td>
<td>0.500</td>
<td></td>
</tr>
</tbody>
</table>

Stable area = ($ma_t - \varphi_t, ma_t + \varphi_t$)

Drop point ($d_t$) = $p_t < (ma_t - \varphi_t)$
Continuous performance evaluation

Model update count

- F1
- Moving average
- Drop points
- Stable area
Meta-measures

1. Frequency of relevant performance drops (F)

Drops: Sequence of consecutive drop points

F: 16.00
2. Volatility of the performance ($V_{perf}$)

$V_{perf}$: The average of the sequence of standard deviations
3. Magnitude of performance drop \( (M_{max,avg}) \)

\[
M_{max} = \max(|p_i - ma_i|) \\
M_{avg} = \text{avg}(|p_i - ma_i|)
\]
Meta-measures

4. Recovery rate ($R_{avg}$)

$R_{avg}$: The number of drop points in drops

\[ \text{Frequency of drops} \]

- $F$: 16.00
- $V_{per}: 0.10$
- $M_{max}: 0.17$
- $M_{avg}: 0.13$
- $R_{avg}: 3.19$
Experiment setting

How to use the meta-measures?

Let’s look at the business scenarios again

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Adapt to changes</td>
</tr>
<tr>
<td>Low</td>
<td>Stable over time</td>
</tr>
</tbody>
</table>
Experiment setting

LOG

Two real-life logs
• BPIC 2015 & BPIC 2017

Two synthetic logs
• Different concept drift

Model

Three algorithms
(Binary outcome prediction)
• Incremental (HATC)
• Sliding window (XGB)
• Train-once (LSTM)

Performance

Four measures
• Accuracy
• Precision, Recall, & F1-Score
## Result

<table>
<thead>
<tr>
<th></th>
<th>BPIC17 Prefix 2 XGB</th>
<th>BPIC17 Prefix 7 XGB</th>
<th>BPIC17 Prefix 14 XGB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average F1-Score</strong></td>
<td>0.54</td>
<td>0.69</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Freq. of Drops</strong></td>
<td>54</td>
<td>67</td>
<td>31</td>
</tr>
<tr>
<td><strong>Volatility of perf.</strong></td>
<td>0.074</td>
<td>0.114</td>
<td>0.037</td>
</tr>
<tr>
<td><strong>Avg. Magnitude</strong></td>
<td>0.115</td>
<td>0.166</td>
<td>0.068</td>
</tr>
<tr>
<td><strong>Recovery rate</strong></td>
<td>8.556</td>
<td>6.776</td>
<td>13.194</td>
</tr>
</tbody>
</table>

![Graphs showing model update count against various metrics for BPIC17 Prefix 2 XGB, BPIC17 Prefix 7 XGB, and BPIC17 Prefix 14 XGB.](image)
### Result

<table>
<thead>
<tr>
<th></th>
<th>IRO5K Prefix: 7 HATC Classifier</th>
<th>IRO5K Prefix: 12 XGB Classifier</th>
<th>IRO5K Prefix: 11 XGB Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Freq. of Drops</strong></td>
<td>44</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td><strong>Volatility of perf.</strong></td>
<td>0.06</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Max. Magnitude</strong></td>
<td>0.29</td>
<td>0.25</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Recovery rate</strong></td>
<td>6.57</td>
<td>11.04</td>
<td>7.69</td>
</tr>
</tbody>
</table>
1) We develop Meta-measures for online process outcome predictive monitoring.
2) We assess the performance stability in various business scenarios.

In-depth analysis with benchmark test
Uncover the causes of the performance drop
Thank you