OPTWIN: DRIFT IDENTIFICATION WITH OPTIMAL SUB-WINDOWS

Dr. Mauro Dalle Lucca Tosi

Research Engineer – Luxembourg Institute of Science and Technology

Prof. Dr. Martin Theobald

Full Professor – University of Luxembourg

13 May 2024

table of contents

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY | LIST

CONTEXT & MOTIVATION

1. CONTEXT & MOTIVATION

- What is Concept Drift?

  - Unforeseeable changes in the statistical properties of the incoming data stream over time.

- Concept Drift Types:
  - Sudden
  - Incremental
  - Gradual
  - Reoccurring

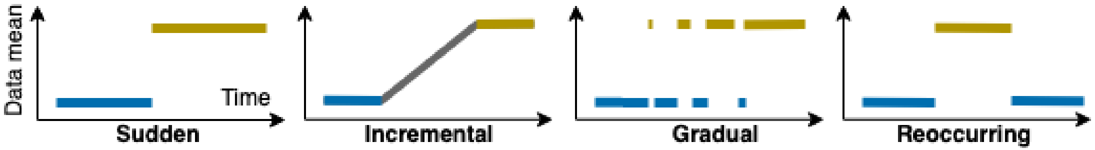
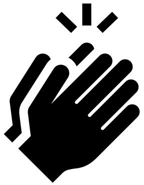


Figure 1: Concept drift types.

- Interesting use cases:

Stock market prediction

Fraud detection



Noise detection

1. CONTEXT & MOTIVATION: REAL-WORLD EXAMPLE

How can a concept change over time?

Figure 2: Tweet containing "Corona" term published before covid-19 pandemic.

Figure 3: Tweet containing "Corona" term published after covid-19 pandemic.

Sentiment Analysis:

Sentiment Analysis:

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY LIST

1. CONTEXT & MOTIVATION: CURRENT CONCEPT DRIFT DETECTORS

Error rate-based drift detector track the error rates produced by an Online Learning learner (e.g. loss of an ANN) in a sliding window W.

Current drift detectors:

- Select sliding window W of the most recent errors from a learner.
- Divide W into historical and new errors. $W_{historical}$ and W_{new}.
- Compare the mean of $W_{historical}$ and W_{new} using a statistical test.
- If the null hypothesis is rejected a concept drift is flagged.

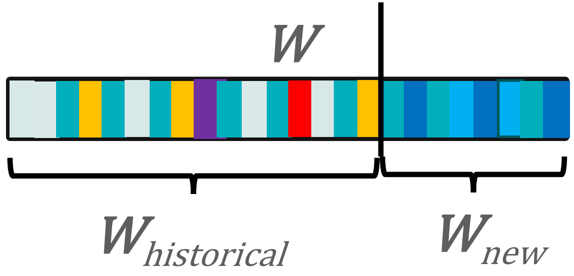


Figure 4: Sliding window illustration.

Problems:

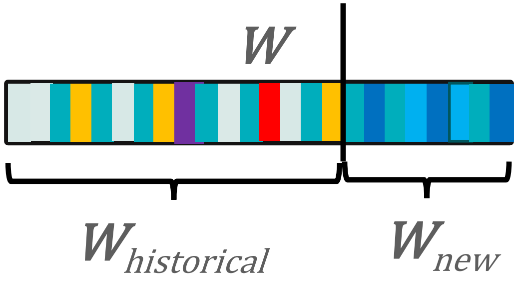
- How to identify the optimal splitting point of W?
  - currently: $\mathcal{O}(\log n)$

- What if the concept drift does not impact the mean of the errors?

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY LIST

OPTWIN - CONCEPT DRIFT DETECTOR

2. OPTWIN - CONCEPT DRIFT DETECTOR: INSIGHT



Standard deviation induced concept drift:

$W_{hist} = \{0.3; 0.7; 0.3; 0.7; 0.3; 0.7; 0.3; 0.7\}$

$W_{new} = \{1.0; 0.0; 1.0; 0.0; 1.0; 0.0; 1.0; 0.0\}$

$$t_{value} = \frac{\mu_{hist} - \mu_{new}}{\sqrt{\frac{\sigma_{hist}^2}{|W_{hist}|} - \frac{\sigma_{new}^2}{|W_{new}|}}}$$

$$t_{value} = \frac{\rho\sigma_{hist}^2}{\sqrt{\frac{\sigma_{hist}^2}{|W_{hist}|} - \frac{\sigma_{new}^2}{|W_{new}|}}}$$

$$t_{value} = \frac{\rho\sigma_{hist}^2}{\sqrt{\frac{\sigma_{hist}^2}{|W_{hist}|} - \frac{f_{value}\,\sigma_{hist}^2}{|W_{new}|}}}$$

$$t_{value} = \frac{\rho}{\sqrt{\frac{1}{|W_{hist}|} - \frac{f_{value}}{|W_{new}|}}}$$

$$\rho = t_{value}\sqrt{\frac{1}{\nu|W|} - \frac{f_{value}}{(1-\nu)|W|}}$$

Considering:

$$\rho\sigma_{hist}^2 = \mu_{hist} - \mu_{new}$$

$$f_{value} = \frac{\sigma_{new}^2}{\sigma_{hist}^2}$$

2. OPTWIN CONCEPT DRIFT DETECTOR: ALGORITHM & THEOREM

OPTWIN calculates the minimum size of W_{new} that guarantees it can identify drifts larger than $\rho\sigma^2_{hist}$

Algorithm 1 OPTWIN

Input parameters:
- δ – confidence level
- ρ – robustness
- w_{max} – max window size

Global variables:
- $W = \langle\,\rangle$ – sliding window
- $w_{min} = 30$ – min window size
- $\eta = 1e^{-5}$ – avoids division by 0

procedure AddElement(x_i)
  $W \leftarrow W \cup x_i$
  if $|W| < w_{min}$ **then**
    return False
  else if $|W| \geq max\_lenght$ **then**
    $W \leftarrow W - W_0$
  $\nu \leftarrow$ OptimalCut$(|W|, \rho, \delta^{\frac{1}{4}})$ — cf. Equation (1)
  $\nu_{split} \leftarrow \lfloor \nu |W| \rfloor$
  $W_{hist} \leftarrow W_{0:\nu_{split}}$
  $W_{new} \leftarrow W_{\nu_{split}:|W|-1}$
  // f-test
  if $\frac{(\sigma_{W_{new}}+\eta)^2}{(\sigma_{W_{hist}}+\eta)^2} > f\_ppf(\delta^{\frac{1}{4}}, \nu|W|-1, (1-\nu)|W|-1)$ **then**
    reset()
    return True
  // t-test – cf. Equations (1) and (2)
  else if $t\_value(W_{hist}, W_{new}) > t\_ppf(\delta^{\frac{1}{4}}, df)$ **then**
    reset()
    return True

$$\rho = t_{value}\sqrt{\frac{1}{\nu|W|} - \frac{f_{value}}{(1-\nu)|W|}}$$

Theorem 4.5.1

- ***False Positive Bound.*** *At every step, if μ_W and σ^2_W remain constant within W, OPTWIN will flag a concept drift at this step with a confidence of at most 1-δ.*
- ***False Negative Bound*** *(for mean drift with large enough W). For any partitioning of W into two sub-windows W_{hist} W_{new}, with $|W| \geq w_{proof}$ and W_{new} containing the most recent elements, if $\mu_{hist} - \mu_{new} > \rho\, \sigma_{hist}$, then, with confidence δ, OPTWIN flags a concept drift in at most $|W| - \nu_{split}$ steps.*
- ***False Negative Bound*** *(for mean drift with small W). For any partitioning of W into two sub-windows W_{hist} W_{new}, with $w_{min} \leq |W| < w_{proof}$ and W_{new} containing the most recent elements, if $\mu_{hist} - \mu_{new} > \rho_{temp}\, \sigma_{hist}$, then, with confidence δ, OPTWIN flags a concept drift in at most $\frac{|W|}{2}$ steps.*
- ***False Negative Bound*** *(for standard-deviation drift with any W). For any partitioning of W into two sub-windows W_{hist} W_{new}, with $|W| \geq w_{min}$ and W_{new} containing the most recent elements, if $\frac{\sigma^2_{new}}{\sigma^2_{hist}} > f\_ppf(\delta', \nu\,|W|-1, (1-\nu)\,|W|-1)$, then, with confidence δ, OPTWIN flags a concept drift in at most ν_{split} steps.*

EXPERIMENTS

3. EXPERIMENTS: DRIFT IDENTIFICATION ON MOA

- Massive Online Analysis (MOA):
  - Binary & Non-binary data
  - Gradual & Sudden Drifts
  - AGRAWL, Random RBF, and STAGGER synthetic datasets
  - Electricity and Covertype real-world datasets

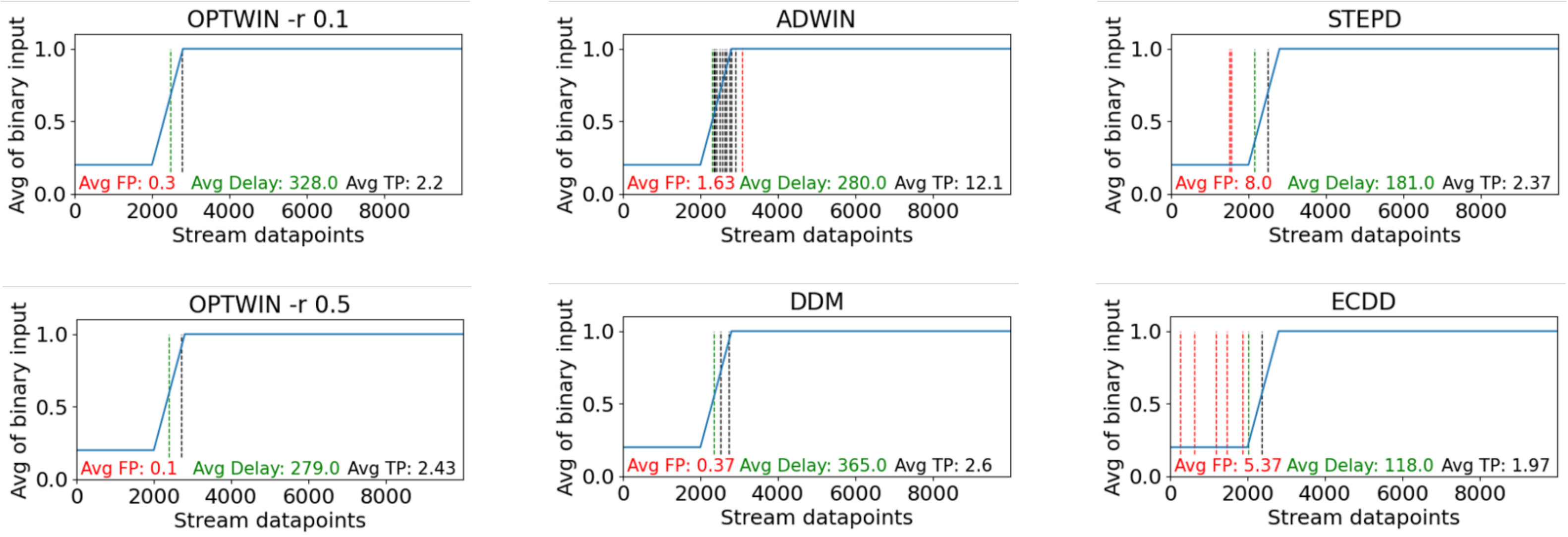
- Regression on ANN models:
  - CNN model
  - CIFAR-10 dataset (simulated concept drifts)

Detector	Avg. Delay	Avg. FP	Avg. F1
ADWIN	132.19	4.39	67%
DDM	569.37	0.73	86%
EDDM	1127.56	8.22	49%
STEPD	196.97	34.93	30%
ECDD	131.84	67.22	37%
$\text{OPTWIN}_{\rho=0.1}$	156.13	**0.17**	**95%**
$\text{OPTWIN}_{\rho=0.5}$	**120.98**	0.19	**95%**
$\text{OPTWIN}_{\rho=1.0}$	414.29	0.32	89%

TABLE I
AVERAGE STATISTICS OF DRIFT IDENTIFICATION ON THE FOLLOWING SYNTHETIC SETTINGS: GRADUAL BINARY DRIFT, GRADUAL NON-BINARY DRIFT, SUDDEN BINARY DRIFT, SUDDEN NON-BINARY DRIFT, SUDDEN STAGGER, SUDDEN RANDOM RBF, AND SUDDEN AGRAWL.

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY LIST

3. EXPERIMENTS: DRIFT IDENTIFICATION ON MOA



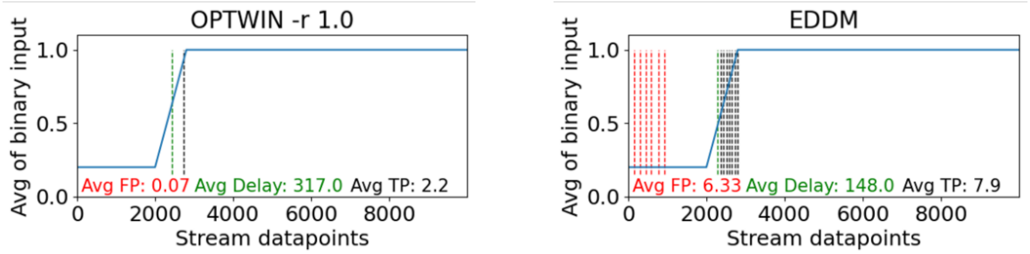


Figure 5: Gradual binary drift detection with average TP and FP rates compared to drift-detection delays.

EXPERIMENTS: DRIFT DETECTION ON ANN MODELS

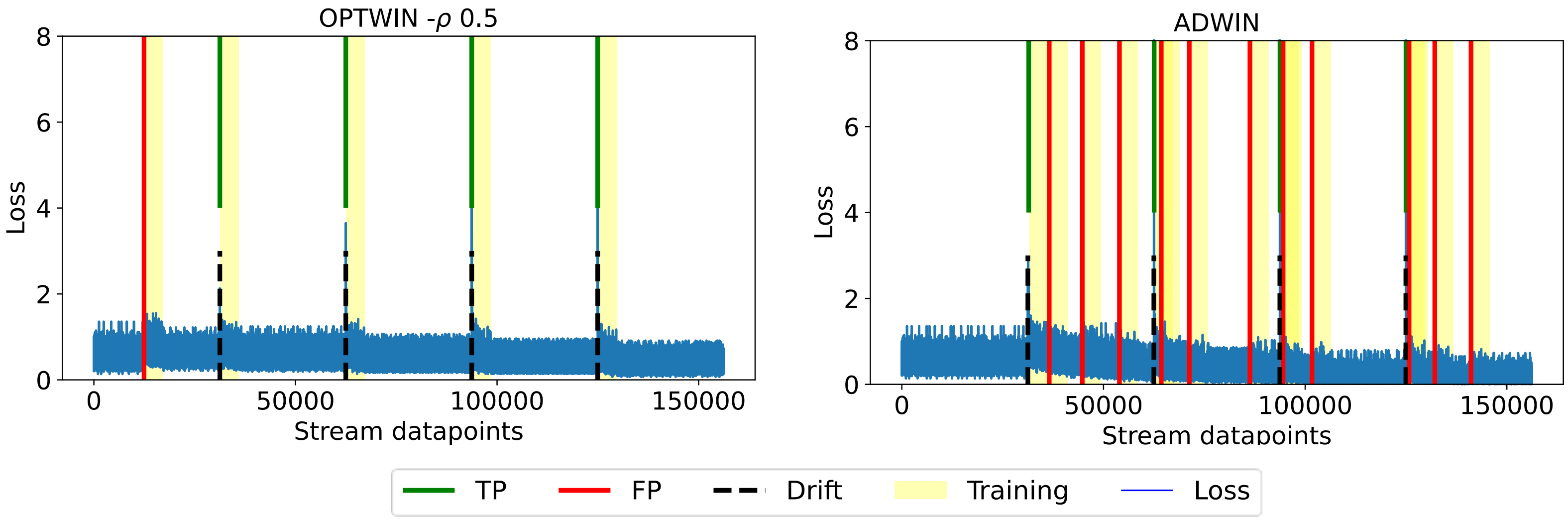


Figure 6: Sudden drift detection over the loss of a CNN.

- **OPTWIN was 21% faster than ADWIN due to a lower number of false positives.**

CONCLUSION

4. CONCLUSION

- Improve optimal splitting point identification from $\mathcal{O}(\log n)$ to $\mathcal{O}(1)$.
- Reduce False Positives (improve F1-score).
- Pre-computable optimal sliding window splitting point.
- Suited for regression and classification problems.
- Available in GitHub in Python, C++, and Java.

http://github.com/maurodlt/optwin

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY LIST

ACKNOLEDGEMENTS

The Doctoral Training Unit Data-driven computational modelling and applications (DRIVEN) is funded by the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781). https://driven.uni.lu

DTU DRIVEN

Fonds National de la Recherche Luxembourg

thank you

github.com/maurodlt/optwin

contact information

For more info, please contact us at

mauro.dalle-lucca-tosi@list.lu