CLASS DISTRIBUTION MONITORING FOR CONCEPT DRIFT DETECTION

Diego Stucchi, Luca Frittoli, Giacomo Boracchi

POLITECNICO MILANO 1863

CONCEPT DRIFT DETECTION

Detecting changes in multivariate annotated datastreams:

$$\{(x_t, y_t)\}_t \text{ where } x_t \in \mathbb{R}^d, y_t \in \{1, \dots, M\}$$

Typical approaches:

- Monitor the **distribution** of x_t ignoring class labels
- Monitor the **error rate** of a classifier (e.g., [4]) – ignoring drifts having little impact on classification

Challenges:

- Monitor the data distribution taking into account class labels
- Online monitoring at a **controlled Average Run Length** (ARL_0), i.e., the expected time before a **false alarm**:

$$ARL_0 = \mathbb{E}[t^*], \quad t^* = \textit{detection time}$$

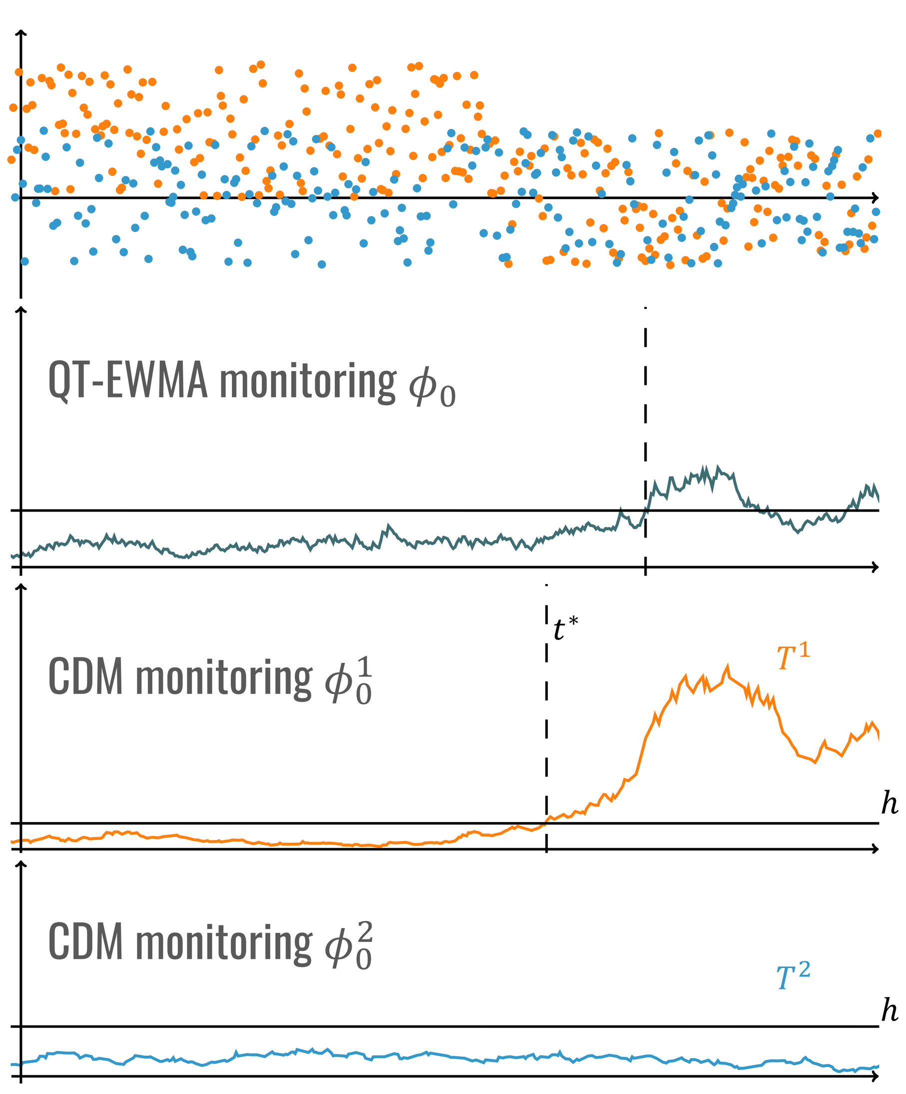
Motivation: a **costly re-training** might be necessary after a concept drift, so **false alarms** must be kept under **control**.

CLASS DISTRIBUTION MONITORING (CDM)

Idea: monitor the class-conditional distributions ϕ_0^m defined by

$$\mathbb{P}_{\phi_0^m}(x_t) = \mathbb{P}_{\phi_0}(x_t \mid y_t = m)$$

- **model** each class-conditional distribution by a **QuantTree histogram** [1]
- **monitor** samples from each class by **QT-EWMA** [2], an **online** and **nonparametric** change-detection algorithm



CDM statistic: $\tilde{T}_t = T_{t_m}^m$ where $m = y_t$ and T^m is the **QT-EWMA statistic** computed on class m.

A **drift** is detected when $T_{t_m}^m > h_{t_m}$ where $\{h_t\}$ are the **QT-EWMA thresholds** guaranteeing the **target** $\textbf{ARL}_0$

Advantages:

- **CDM is faster** than QT-EWMA in detecting the drift, especially when it affects a subset of classes
- CDM indicates **which class** triggered the detection

CONTROL OF THE ARL_0

We demonstrate that the CDM maintains the **same** $\textbf{ARL}_0$ as the **QT-EWMA** monitoring each class-conditional distribution.

Proposition. *Let the CDM statistic be defined by the QT-EWMA statistic* T_t, *and let* $\{h_t\}$ *be thresholds such that:*

$$\mathbb{P}(T_t > h_t \mid T_k \le h_k \;\; \forall k < t) = \alpha = \frac{1}{ARL_0}$$

Then, CDM yields the same ARL_0 *as QT-EWMA.*

EXPERIMENTS

We test our solution on the real-world INSECTS dataset [3]

Empirical ARL_0

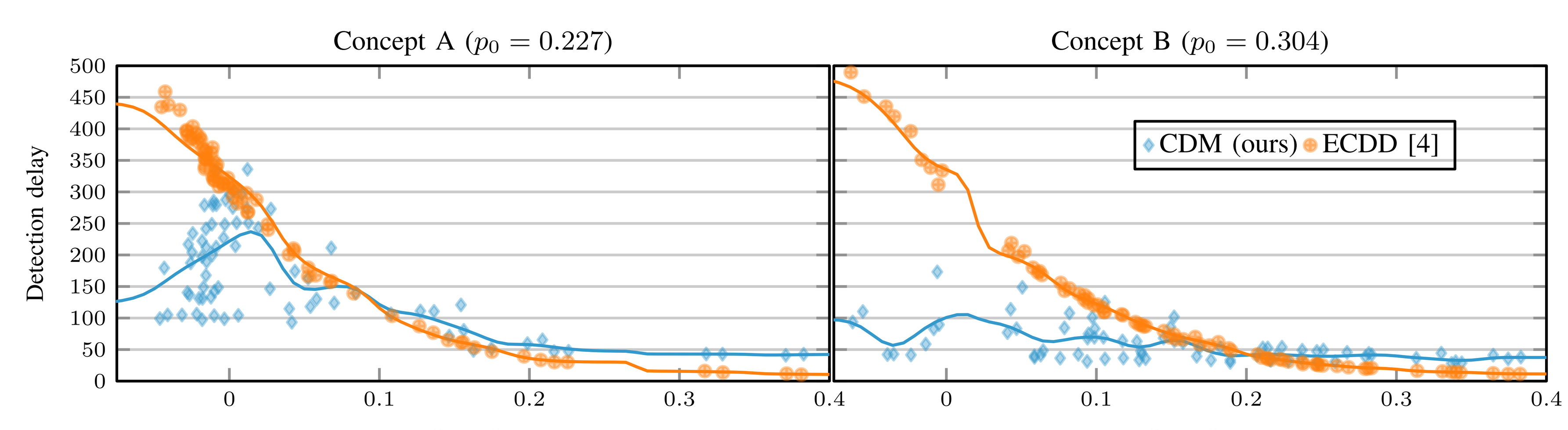
Concept	Methods (target ARL_0)			
	ECDD [4] (400)	Scan-B [5] (300)	QT-EWMA [2] (375)	CDM (ours) (375)
A	376.51	382.08	379.10	375.44
B	371.07	384.56	361.78	374.47
C	373.16	381.65	371.66	365.32
D	374.14	387.17	367.18	369.94
E	371.82	376.28	375.10	374.64
F	377.67	374.22	375.58	371.87

CDM and QT-EWMA control the ARL_0 more accurately than the alternatives

Detection Delay

Drifted Classes	ECDD [4]	Scan-B [5]	QT-EWMA [2]	CDM (ours)
1	207.98	212.53	267.73	**195.45**
2	245.85	162.58	195.44	**124.92**
3	264.27	224.99	278.57	**204.00**
4	224.91	235.87	265.96	**196.74**
1, 2	198.17	131.71	174.80	**114.44**
1, 3	172.62	169.87	223.50	**160.98**
1, 4	165.77	163.63	221.66	**145.82**
2, 3	163.66	126.56	167.55	**112.18**
2, 4	176.53	119.41	154.95	**106.49**
3, 4	210.04	169.88	218.90	**153.51**
1, 2, 3	139.29	115.01	152.91	**103.60**
1, 2, 4	148.03	103.24	141.09	**98.89**
1, 3, 4	144.81	134.83	183.41	**131.38**
2, 3, 4	132.36	**96.92**	136.90	98.57
1, 2, 3, 4	122.38	**88.86**	128.04	91.44
Avg. rank	2.416	2.356	3.524	**1.704**

CDM outperforms the alternatives, particularly when the change affects only a few classes



CDM achieves **excellent detection delays**, outperforming ECDD when the drift has **little impact on classification error** (which happens quite often!)

CONCLUSIONS

- **CDM**: a new approach to concept drift detection by **monitoring class-conditional distributions**
- **Nonparametric** monitoring **controlling the** $\textbf{ARL}_0$ thanks to the properties of **QT-EWMA**
- **State-of-the-art** concept drift detection **performance**

REFERENCES

[1] Boracchi, Carrera, Cervellera, Macciò *"QuantTree: histograms for change detection in multivariate data streams"* ICML 2018

[2] Frittoli, Carrera, Boracchi *"Change detection in multivariate datastreams controlling false alarms"* ECML-PKDD 2021

[3] Souza, dos Reis, Maletzke, Batista *"Challenges in benchmarking stream learning algorithms with real-world data"* Data Mining and Knowledge Discovery 2020

[4] Ross, Adams, Tasoulis, Hand *"Exponentially weighted moving average charts for detecting concept drift"* Pattern Recognition Letters 2012

[5] Li, Xie, Dai, Song *"M-statistic for change-point detection"* NIPS 2015

You can contact us by email at {diego.stucchi, luca.frittoli, giacomo.boracchi}@polimi.it

Our code is available at: