How well do clustering similarities-based concept drift detectors identify concept drift in case of synthetic/real-world data?

AUTHOR: Jindřich Pohl, j.pohl@student.tudelft.nl

SUPERVISOR: Lorena Poenaru-Olaru, MSc

RESPONSIBLE PROFESSOR: Dr. J. S. Rellermeyer

1 Introduction

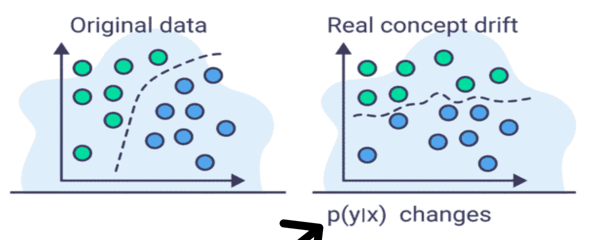


Figure 1: **Concept drift** = change in underlying distribution of streaming data over time. Adapted from (D. Geyshis, 2021)

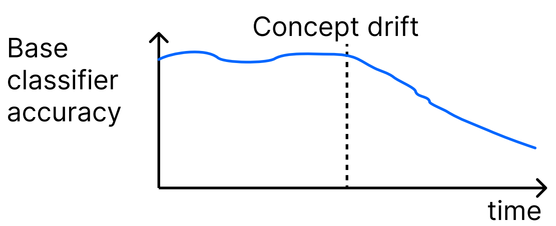


Figure 2: Concept drift leads to **drop in** classifier **accuracy**.

Why this research?

- **Concept drift** (Fig. 1, 2): relevant streaming data problem - examples: fraud detection [1], user modelling [2]
- **Drift detectors**: algorithms detecting concept drift
- → help **prevent drop in accuracy** of deployed classifiers
- **Supervised** drift detectors require data **labels** → **expensive**
- Few available **unsupervised** drift detectors → these only **compare original data to incoming data** → labels not necessary
- Drift detectors **seldom evaluated on real-world** datasets

Contributions?

GitHub

- Two existing unsupervised drift detectors now **publicly available**[1]
- → **reduced limitations in existing** drift detection comparison **research** [3] caused by unavailable implementations
- Evaluation results on both **synthetic and real-world** data

[1]https://github.com/Jindrich455/clustering-drift-detection

2 Related work - chosen clustering similarities-based drift detectors

UCDD [4]: k-means clustering for **class estimates** (Fig. 3), drift detected when **classes shift** away from one another

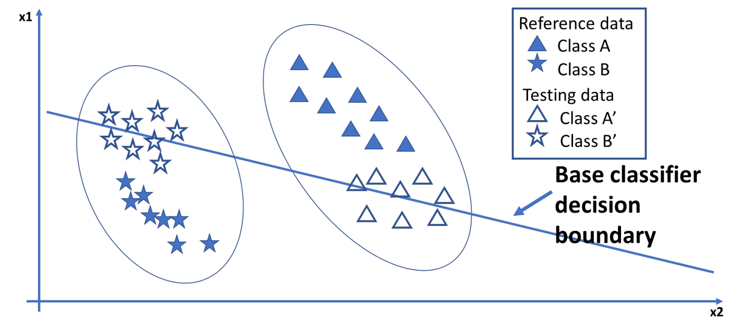


Figure 3: Intuition of unbiased label estimates through clustering in UCDD

MSSW [5]: drift detected when **average total distance** to centroids in weighted k-means clustering **exceeds a threshold** (Fig. 4)

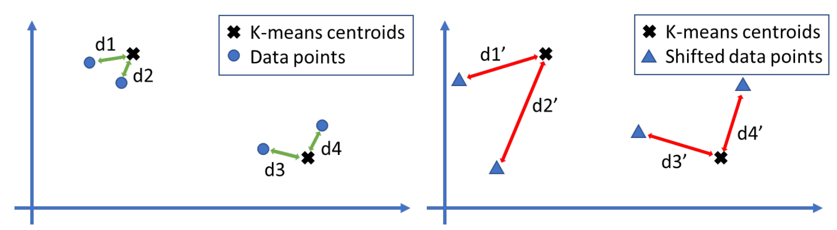


Figure 4: example suddenly too high average distances to centroids in MSSW

3 Methodology

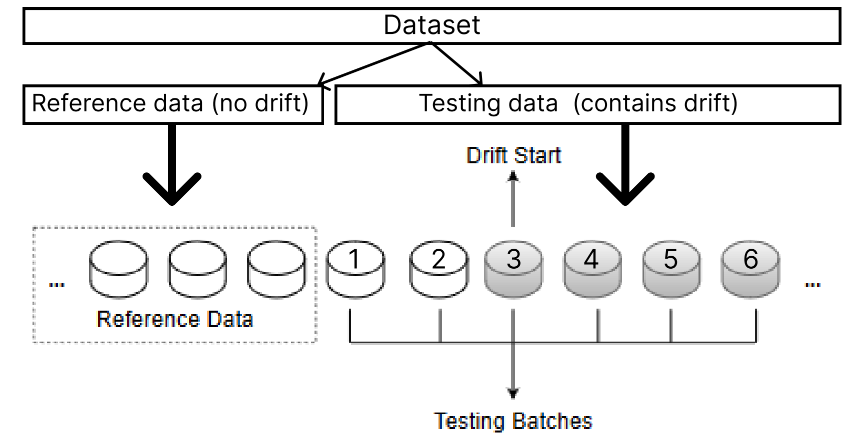


Figure 5: Data setup for evaluation: dataset split to **reference** and **testing** data and then to **equal-sized batches**, of which **some are drifting (grey)**. Adapted from (L. Poenaru-Olaru et al., 2022)

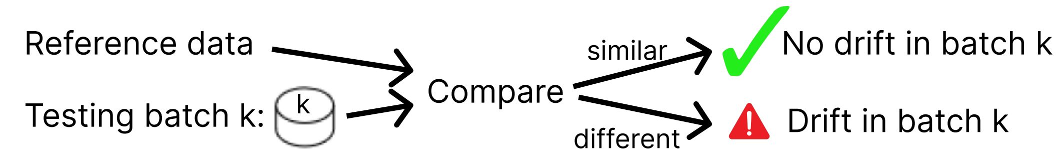


Figure 6: High-level drift detection explanation. Drift detector **goal**: **detect all drifting (grey) batches**.

Data setup: shown in and explained by Fig. 5, 6

Datasets:

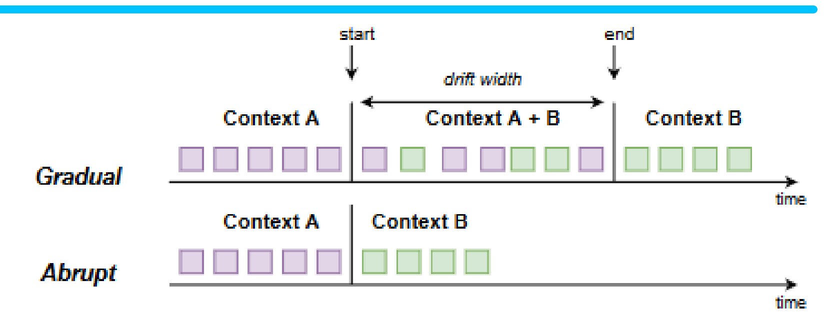
- **Synthetic:**
- → **abrupt** and **gradual** drift (Fig. 7), drift **starts** at a known batch
- **Real-world:**
- → **drift unknown**, estimated for **individual batches** through base classifier accuracy

Evaluation metrics:

- **Synthetic** data: **false-positive rate** (FPR_S) = fraction of false alarms, **latency** (**LTC**) = how late drift was detected
- **Real-world** data: **false-positive rate** (FPR_R) = fraction of false alarms, **detection accuracy** (**ACC**) = fraction of batches where drift was correctly detected

Adaptation: categorical feature handling by excluding them or encoding through one-hot and target encoding

Figure 7: Gradual vs abrupt drift, (L. Poenaru-Olaru et al., 2022)



4 Results with best encoding and parameters

Table 1: Results in synthetic datasets for different algorithms, datasets, and drift widths. The table is color coded to highlight important trends, such as a worsening performance for width >= 10.0.

drift width =		0	0.5	1	5	10	20
algorithm	dataset	(FPR_S, LTC)	(FPR_S, LTC)	(FPR_S, LTC)	(FPR_S, LTC)	(FPR_S, LTC)	(FPR_S, LTC)
UCDD*	SEA	(0.00, 0.17)	(0.00, 0.25)	(0.00, 0.25)	(0.00, 0.00)	(0.00, 0.25)	(0.00, 0.75)
	AGRAW1	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.25)	(0.00, 0.25)
	AGRAW2	(0.00, 0.50)	(0.00, 0.50)	(0.00, 0.50)	(0.00, 0.50)	(0.00, 1.00)	(0.00, 0.00)
MSSW	SEA	(0.00, 0.25)	(0.00, 0.25)	(0.00, 0.25)	(0.00, 0.25)	(0.00, 0.25)	(0.00, 0.25)
	AGRAW1	(0.00, 0.00)	(0.00, 0.17)	(0.00, 0.17)	(0.00, 0.00)	(0.00, 0.25)	(0.00, 0.25)
	AGRAW2	(0.00, 0.20)	(0.00, 0.20)	(0.00, 0.20)	(0.00, 0.20)	(0.00, 0.25)	(0.00, 0.50)

Table 3: Color coding for the metrics. Cells are filled by the worse color in each pair.

	FPR_S, FPR_R, LTC	ACC
0.0 <= x < 0.25	xxxx	xxxx
0.25 <= x < 0.5	xxxx	xxxx
0.5 <= x < 0.75	xxxx	xxxx
0.75 <= x < 1	xxxx	xxxx

* UCDD is dependent on additional parameters, so the results presented here might be biased

Table 2: Results in real-world datasets for different algorithms, datasets and experiments/batch sizes. The table is color coded to show the suboptimal performances of the algorithms in real-world datasets

algorithm	dataset	experiment	(FPR_R, ACC)
UCDD*	Weather	Yearly	(0.73, 0.83)
		Monthly	(0.59, 0.76)
	ELECT2	Yearly	(0.85, 0.96)
	Airlines	One-hot	(0.00, 0.65)
		Target	(0.00, 0.60)
	Spam	All	(1.00, 1.00)
MSSW	Weather	Yearly	(0.27, 0.11)
		Monthly	(0.06, 0.04)
	ELECT2	Yearly	(1.00, 1.00)
	Airlines	One-hot	(0.00, 0.00)
		Target	(1.00, 0.05)
	Spam	All	(0.00, 0.00)

5 Conclusions and further research

- UCDD: very dependent on additional parameters, k-means clustering too simple for label estimates→ detections not guaranteed for more complex datasets
- MSSW: good results in synthetic datasets → LTC capped at 25% for small enough widths
- UCDD and MSSW resilient to drift widths of at most the batch size
- UCDD and MSSW depend on categorical feature handling
- Real-world data: either too little detections (ACC<=65%) or too many detections (FPR>55%)
- → UCDD and MSSW likely not suitable for the real world
- → further research: try other clustering methods in UCDD, try other drift definition strategies to thoroughly confirm that these detectors are not suitable for the real world

[1] Z. Yuan, H. Liu, J. Liu, Y. Liu, Y. Yang, R. Hu, and H. Xiong, Incremental Spatio-Temporal Graph Learning for Online Query-POI Matching. Feb. 2021.
[2] B. Gupta, A. Goyal, C. Sharma, and D. Kumar, "RE-RentFraud: A System to detect Frauds in rent payments for Real-Estate properties," in Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD), CODS-COMAD '23, (New York, NY, USA), pp. 253–257, Association for Computing Machinery, Jan. 2023.
[3] L. Poenaru-Olaru, L. Cruz, A. van Deursen, and J. S. Rellermeyer, "Are concept drift detectors reliable alarming systems? - a comparative study," in 7th Workshop on Real-time Stream Analytics, Stream Mining, CER/CEP Stream Data Management in Big Data, 2022.
[4] D. Shang, G. Zhang, and J. Lu, "Fast concept drift detection using unlabeled data," in Developments of Artificial Intelligence Technologies in Computation and Robotics, vol. Volume 12 of World Scientific Proceedings Series on Computer Engineering and Information Science, pp. 133–140, WORLD SCIENTIFIC, June 2020.
[5] Y. Yuan, Z. Wang, and W. Wang, "Unsupervised concept drift detection based on multi-scale slide windows," Ad Hoc Networks, vol. 111, p. 102325, Feb. 2021.