The 14th International Conference on Learning Representations
April 23 - April 27, 2026 | Rio de Janeiro, Brazil

Pitfalls of Unlabeled Disagreement-Based Drift Detection in Streaming Tree Ensembles

Lara Sá Neves[1], Afonso Lourenço[1], Lizy K. John[2], Goreti Marreiros[1]
1 GECAD, ISEP, Polytechnic of Porto, 4249-015, Portugal
2 The University of Texas at Austin, 78712, United States

Motivation

- **Problem: model deterioration monitoring goals**
  - R1: Detect using unlabeled deployment data
  - R2: Avoid false alarms from benign shifts with few samples
  - **Raw data & posterior distribution detectors do R1 but fail R2**
- **Solution: batch-specific model disagreement framework**
  - Transductive reasoning: model-specific **impact of conflicting info, instead of accumulated uncertainty**
  - Successful with ensembles of neural networks in large batches, but not studied **on incremental decision trees**

Can disagreeing critics work with IDTs?

Theory

- **Labeled update (Lemma 1)**

$$\varepsilon_{D_t}(h) = \varepsilon_{D_t}(h, h_{\theta_{t-1}}) + \varepsilon_{D_t}(h_{\theta_{t-1}})$$

one-hot disagreement; history model

- **Drift-based update (Lemma 2)**

$$\varepsilon_{D_t}(h) \leqslant \varepsilon_{D_t}(h, h_{\theta_{t-1}}) + \varepsilon_{D_{t-1}}(h_{\theta_{t-1}}) + \frac{1}{2}\Delta(h_{\theta_{t-1}})$$

disagreeing critic: $h^* = \arg\max_{h' \in H'} \Delta(h_{\theta_{t-1}}, h^*)$, H' per $h_{\theta_{t-1}}$

  - conservative splits & parent hyper-rectangles of $h_{\theta_{t-1}}$ regularize h
  - bias is minimized only if $h_{\theta_{t-1}}$ is localized around D_t
  - useful **drift detectors must account for both data and model complexity**: if D_{t-1}/D_t are similar, bound is small & $h_{\theta_{t-1}}$ can be reused; otherwise, h is updated via pruning, or ensemble modification. **Fig. 1** illustrates:
    - Left: **loss-based false negative in over-regularized model**
    - Center: **data-based false positive for proper regularized model**
    - Right: **both methods succeed in overly complex model**

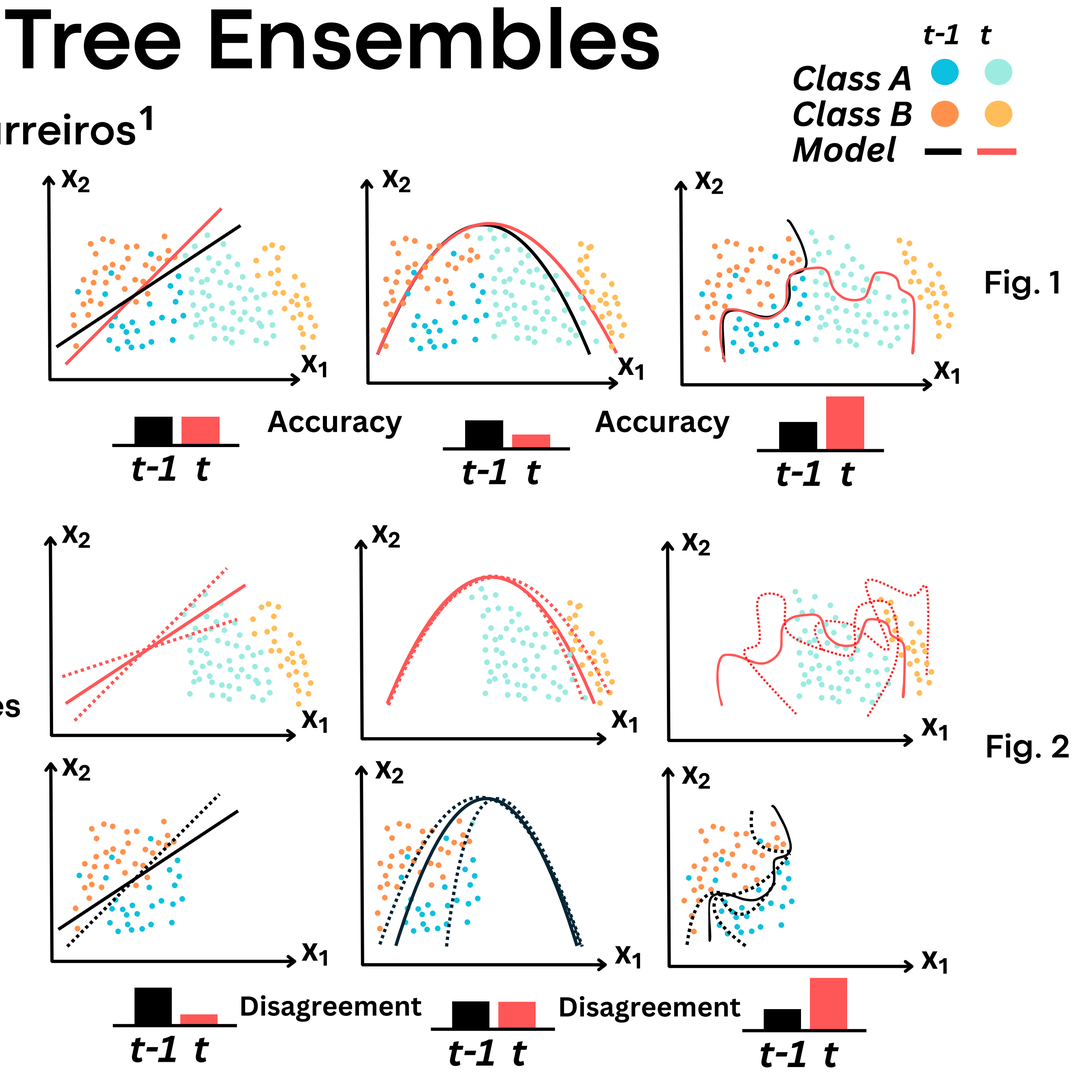


Fig. 1

Fig. 2

- **Disagreement-based update (Lemma 3)**

$$\varepsilon_{D_t}(h) \leqslant \varepsilon_{D_t}(h, h_{\theta_{t-1}}) + \varepsilon_{D_{t-1}}(h_{\theta_{t-1}}) + \frac{1}{2}\Delta(h_{\theta_{t-1}}, h^*)$$

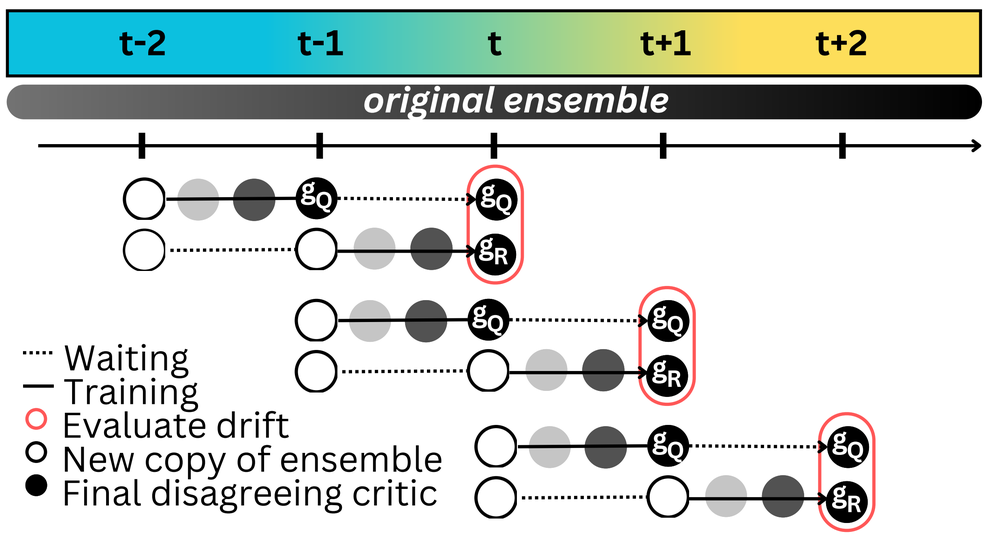
  - **Maximize $\Delta(h_{\theta_{t-1}}, h^*)$ to identify regions most affected** by drift
  - In binary ensembles, simple label flipping captures effects
  - **Fig. 2** shows disagreement-based drift across complexities:
    - Left: hardly induced in far input space
    - Center: evenly induced across input space
    - Right: easily induced in under-regularized far input space

Methodology

- Split batch into two sub-windows (Q, R), training 2 ensemble copies on flipped pseudo-labeled data
- A Kolmogorov–Smirnov (KS) test detects **predictive consistency across disagreement distributions**
- Adaptive resampling increases instance exploitation under high error, accelerating convergence without large windows

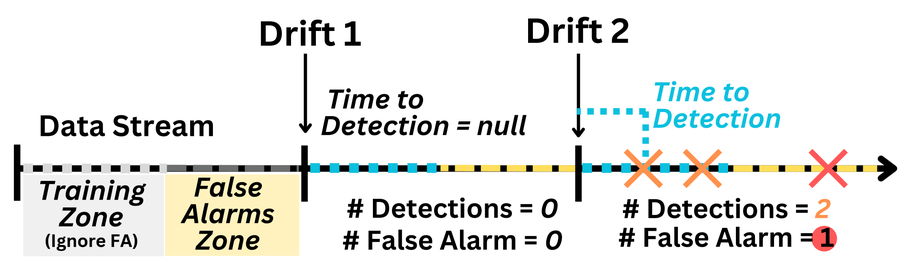
Algorithm Disagreement framework

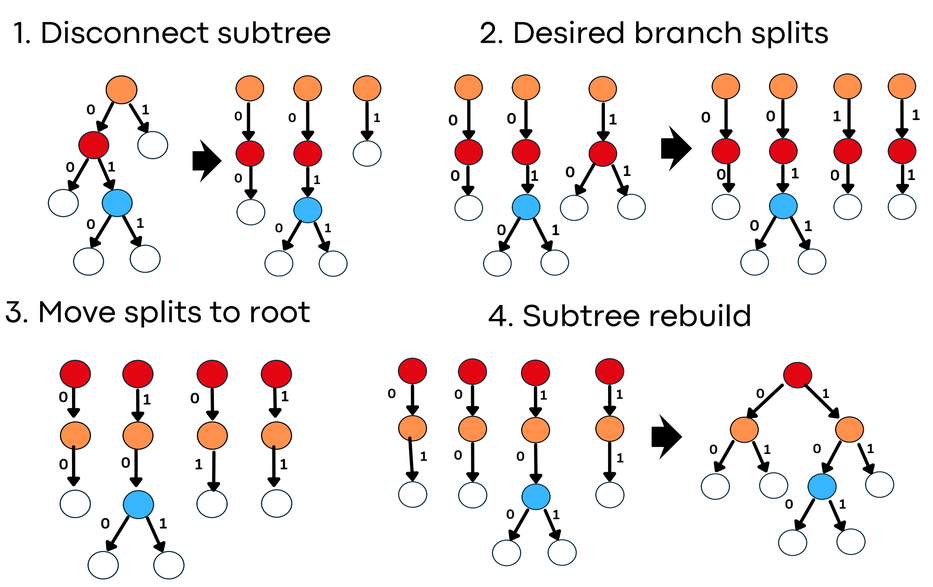
Initialize ensemble g on past data P
while stream has new batch do
  Q′, R′ ← pseudo-label and flip in Q, R
  g_Q, g_R ← copies of g
  Train g_Q on Q′, g_R on R′
  For each ensemble g_X in (g_Q, g_R) do
    For each pairs of models g_a, g_b in g_X do
      $d_{a,b} = \frac{1}{K}\sum_{i=1}^{K} I[g_a(x_i) \neq g_b(x_i)]$
    D_X ← collection of all $d_{a,b}$
  If KS_test(D_Q, D_R) rejects H_0 then
    Drift detected



Results

- Evaluate IDT & MLP ensembles with 6 loss- & 5 data-based detectors, tuned to 0.5×Detection Accuracy+0.3×(1−False Alarms(FA))+0.2×(1−Mean Time to Detect(MTD))
- MTD(FA) for gradual (G) and abrupt (A) drifts, in Disagreement- (Dis.), Data-, & Loss-based detectors show **disagreeing critic works for NNs but performs poorly for IDTs**
- Likely because **IDTs lack the plasticity to generate useful signals. Solution lies in restructuring** with their intrinsic, non-overlapping **rules** that fully partition the space





		T	RBF	RBF2	SEA0	SEA1	SEA2	SineA	Sine4	SineL	Hyp0	Hyp1
Dis.	NNs	G	1137(4)	1383(4)	843(3)	1475(1)	2427(1)	643(3)	2863(1)	1747(4)	1573(2)	2187(1)
		A	820(6)	910(4)	365(1)	980(0)	1620(1)	410(1)	1980(0)	810(0)	685(1)	490(0)
	IDTs	G	2267(6)	1333(15)	1167(4)	3700(2)	3300(3)	2100(2)	6600(0)	3900(3)	1533(10)	2467(2)
		A	3133(14)	2840(12)	2025(0)	1775(6)	2275(5)	1600(13)	1367(0)	1200(0)	1400(17)	2000(18)
Data	D3	G	1662(10)	434(16)	505(14)	486(17)	231(14)	436(5)	978(8)	421(4)	639(14)	452(12)
		A	123(5)	121(3)	728(44)	583(50)	547(40)	129(0)	129(0)	129(0)	1036(53)	1005(56)
	IBDD	G	92(60)	139(52)	254(39)	379(37)	168(39)	343(11)	158(20)	101(11)	232(31)	232(31)
		A	60(154)	60(147)	86(113)	57(100)	61(90)	59(18)	231(3)	212(0)	64(108)	64(108)
Loss	ADWIN	G	1333(5)	3733(1)	4017(1)	1700(0)	5117(0)	2633(1)	4650(0)	7325(0)	2083(2)	3600(2)
		A	1870(7)	1533(5)	320(2)	1680(0)	1750(2)	275(0)	500(0)	470(0)	190(6)	360(3)
	DDM	G	5752(0)	1336(0)	4052(0)	4292(0)	5160(0)	585(1)	4417(0)	3129(0)	3091(0)	3858(0)
		A	1839(4)	-(0)	486(0)	731(0)	1321(0)	317(0)	280(0)	269(0)	336(0)	1209(0)
	PH	G	1363(7)	1474(3)	2060(1)	1937(0)	2884(0)	6466(1)	2062(0)	1225(0)	1434(4)	1769(3)
		A	1364(7)	1562(6)	249(1)	144(0)	184(1)	1916(0)	114(0)	103(0)	152(7)	170(8)

gecad · P.PORTO · TEXAS The University of Texas at Austin · UTAustin Portugal · fct Fundação para a Ciência e a Tecnologia