Prequential AUC for Classifier Evaluation in Evolving Data Streams

Dariusz Brzeziński

Poznan University of Technology

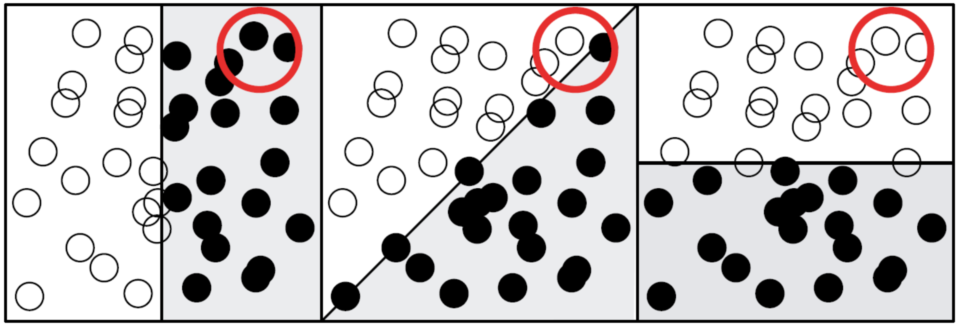
Outline

- Data stream mining
- Evaluation measures
- Prequential AUC
  - Algorithm
  - Properties
    - Visualization
    - Consistency and Discriminancy
    - Processing speed
  - Model selection and drift detection
- Conclusions

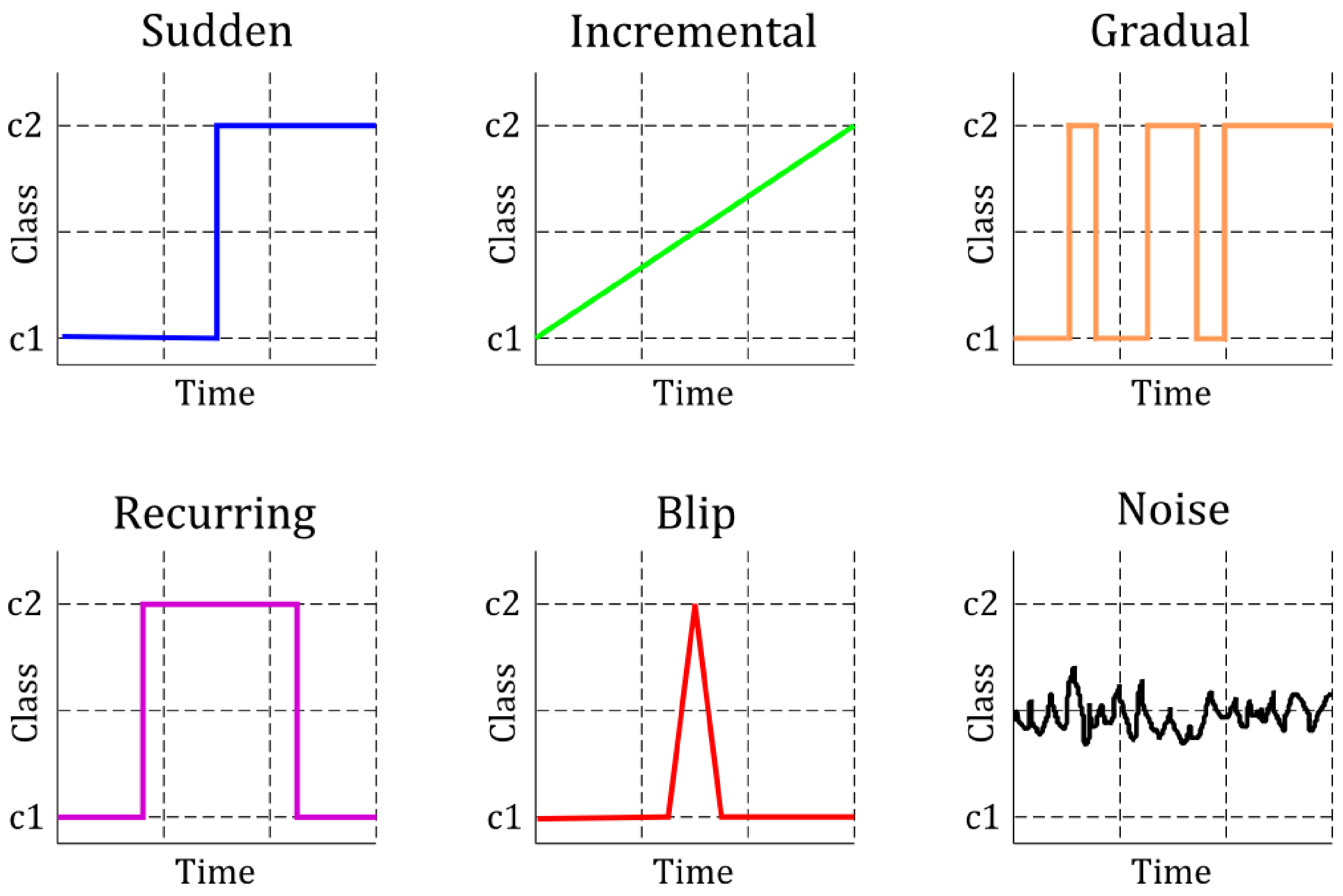
Data stream mining

New challenges for data mining algorithms!

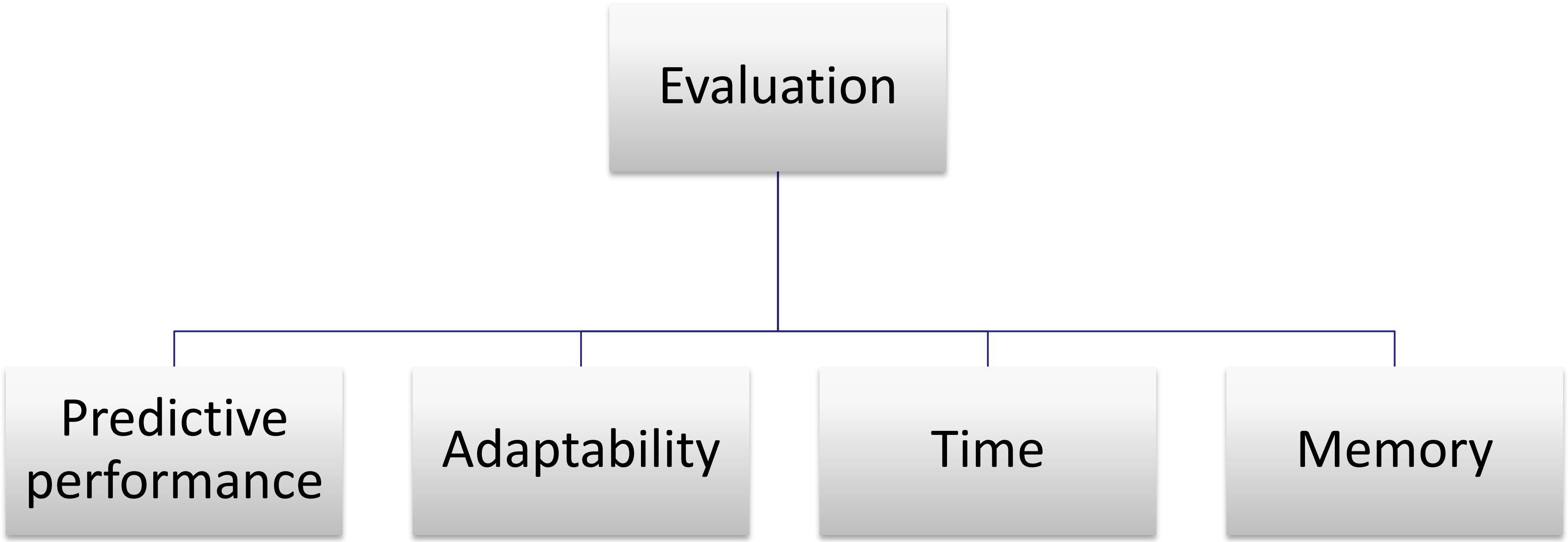
- **Limited time**
  - examples arrive rapidly
  - each example can be processed only once
- **Limited memory**
  - streams are often too large to be processed as a whole
- **Concept drift**
  - data streams can evolve over time



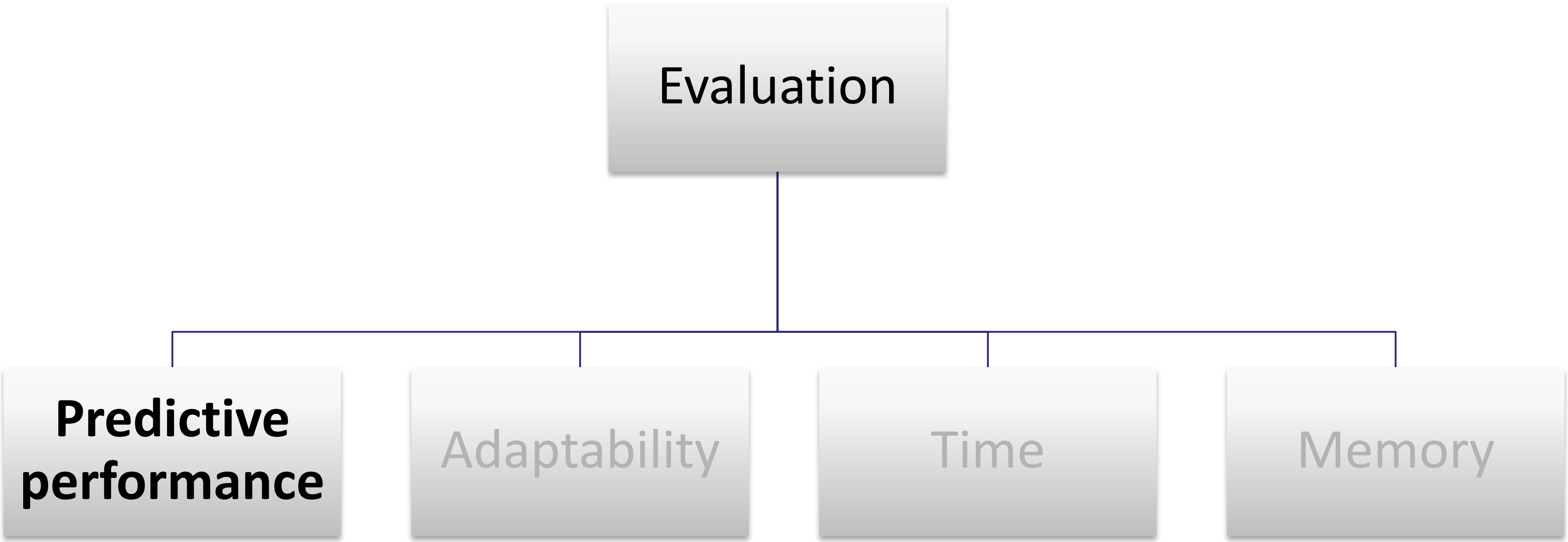
Concept drift



Stream classifier evaluation



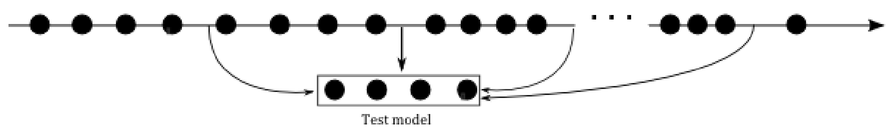
Stream classifier evaluation



Existing evaluation methods

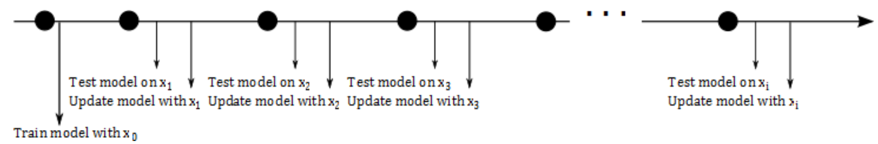
- Holdout

  [e.g., Kirkby 2007]



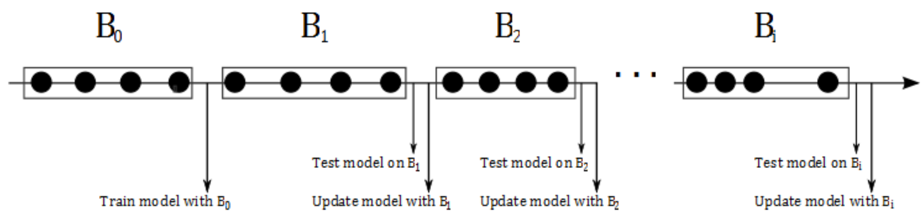
- Test-then-train

  [e.g., Kirkby 2007]



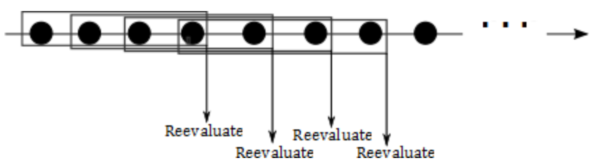
- Block-based evaluation

  [e.g., Brzezinski & Stefanowski 2010]



- Prequential accuracy

  [Gama et al. 2013]



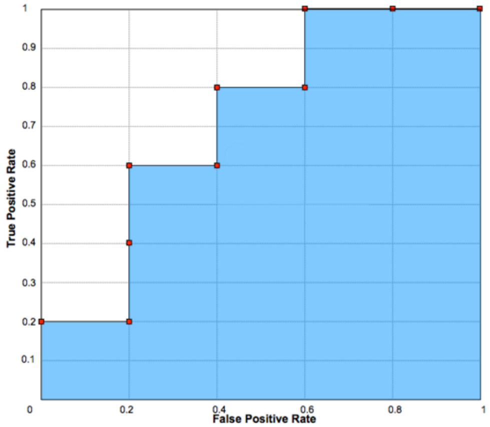
- Kappa statistic

  [Bifet & Frank 2010, Zliobaite et al. 2014]

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

Area Under the ROC Curve

- Quantification of ROC analysis
- Performance of a ranking classifier averaged over all cost ratios
- Several assets
  - Probabilistic interpretation (WMW)
  - More discriminant than accuracy [Huang & Ling al. 2005]
  - Suitable for imbalanced datasets
- Data streams
  - Only on entire streams, holdout sets, or incrementally



Prequential AUC

Main idea:

Calculate AUC prequentially after each example with a sliding window forgetting mechanism

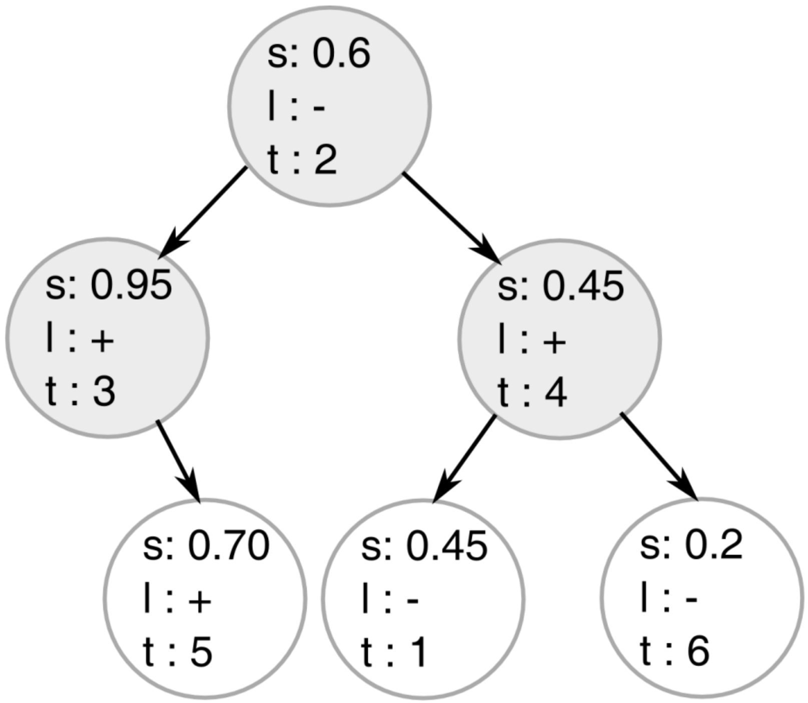
- Keep instance scores in a sorted structure
- Remove oldest instance before adding a new one
- Scan through the sorted structure to calculate AUC

$$AUC = \frac{1}{mn} \sum_{j=1}^{n} \sum_{t=1}^{s_j - j} 1$$

[Wu, Flach, Ferri 2007]

Prequential AUC

Red-black tree as a sorted structure



w:

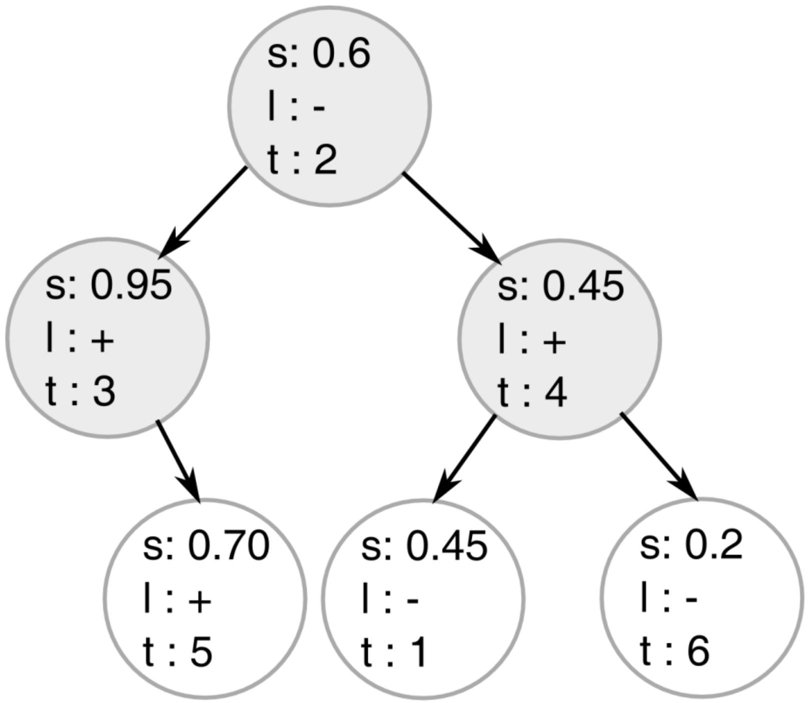
s: 0.45	s: 0.60	s: 0.95	s: 0.45	s: 0.70	s: 0.20
l : -	l : -	l : +	l : +	l : +	l : -
t : 1	t : 2	t : 3	t : 4	t : 5	t : 6

Add O(log(*d*))
Remove O(log(*d*))
Iterate O(*d*)
Space O(*d*)

d: window size

Prequential AUC

Red-black tree as a sorted structure



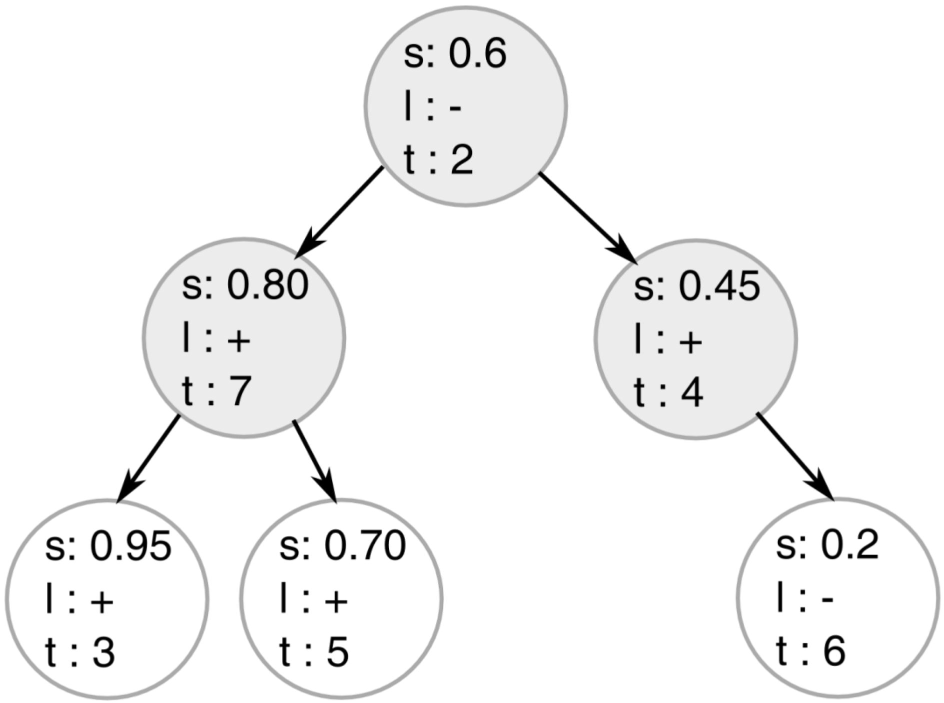
Add O(1)
Remove O(1)
Iterate O(1)
Space O(1)

w:

s: 0.45 l : - t : 1	s: 0.60 l : - t : 2	s: 0.95 l : + t : 3	s: 0.45 l : + t : 4	s: 0.70 l : + t : 5	s: 0.20 l : - t : 6

Prequential AUC

Red-black tree as a sorted structure



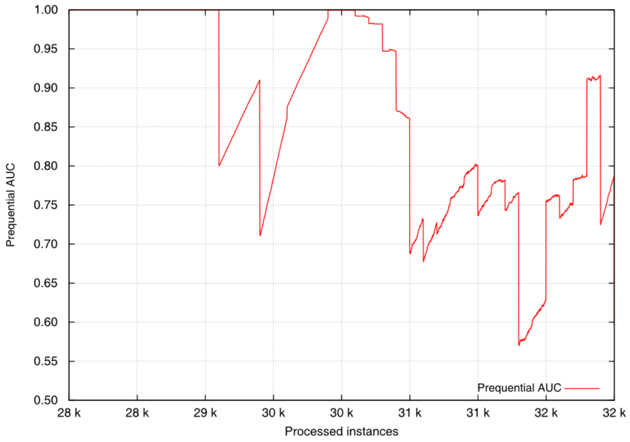
Add O(1)
Remove O(1)
Iterate O(1)
Space O(1)

w:

s: 0.60	s: 0.95	s: 0.45	s: 0.70	s: 0.20	s: 0.80
l : -	l : +	l : +	l : +	l : -	l : +
t : 2	t : 3	t : 4	t : 5	t : 6	t : 7

Prequential AUC

- Calculated after each example
- Constant time and memory
- Suitable for imbalanced data
- Forgetting mechanism
- Can be used with drift detectors (PH Test)



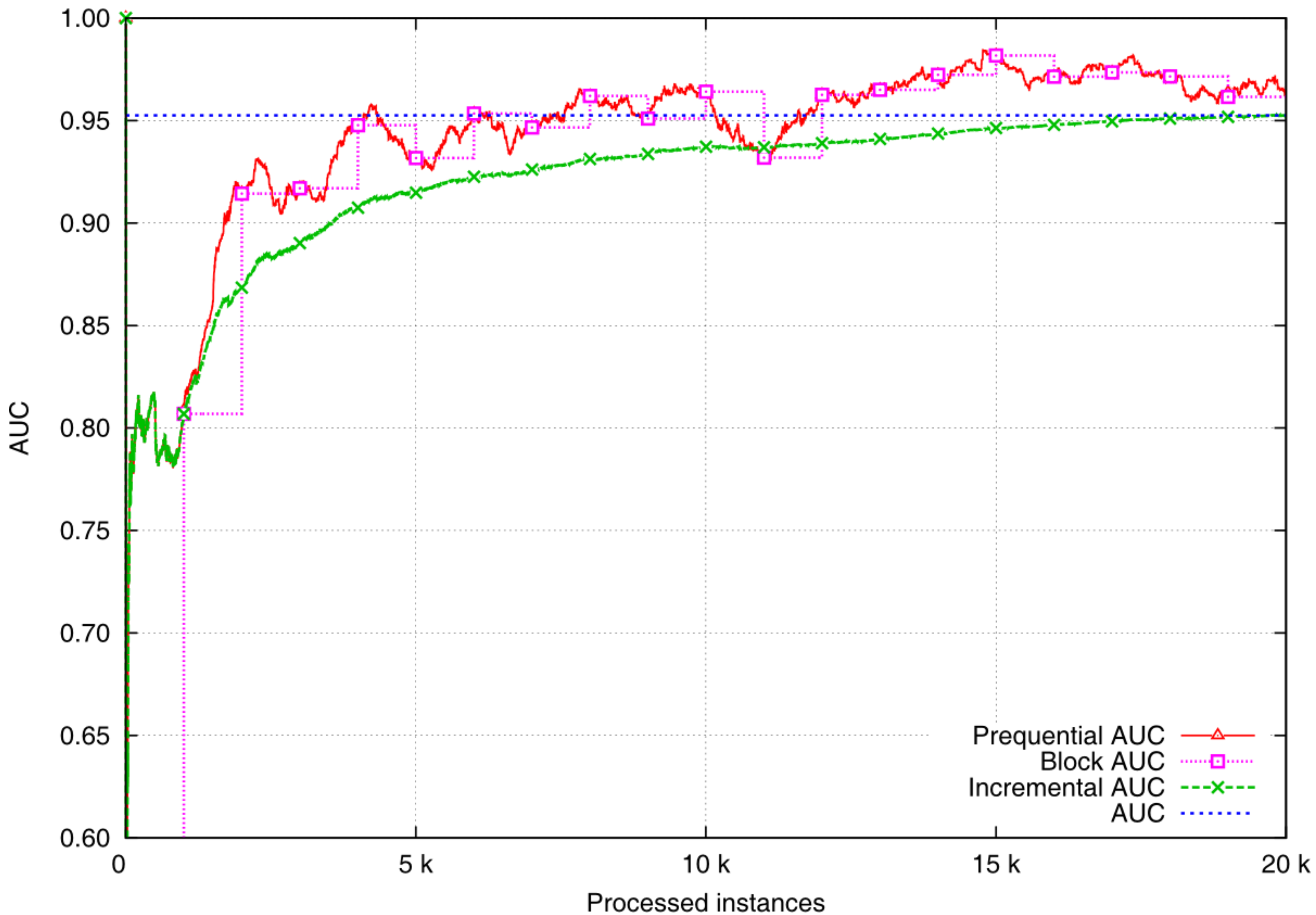
Outline

- Data stream mining
- Evaluation measures
- Prequential AUC
  - Algorithm
  - **Properties**
    - Visualization
    - Consistency and Discriminancy
    - Processing speed
  - Model selection and drift detection
- Conclusions

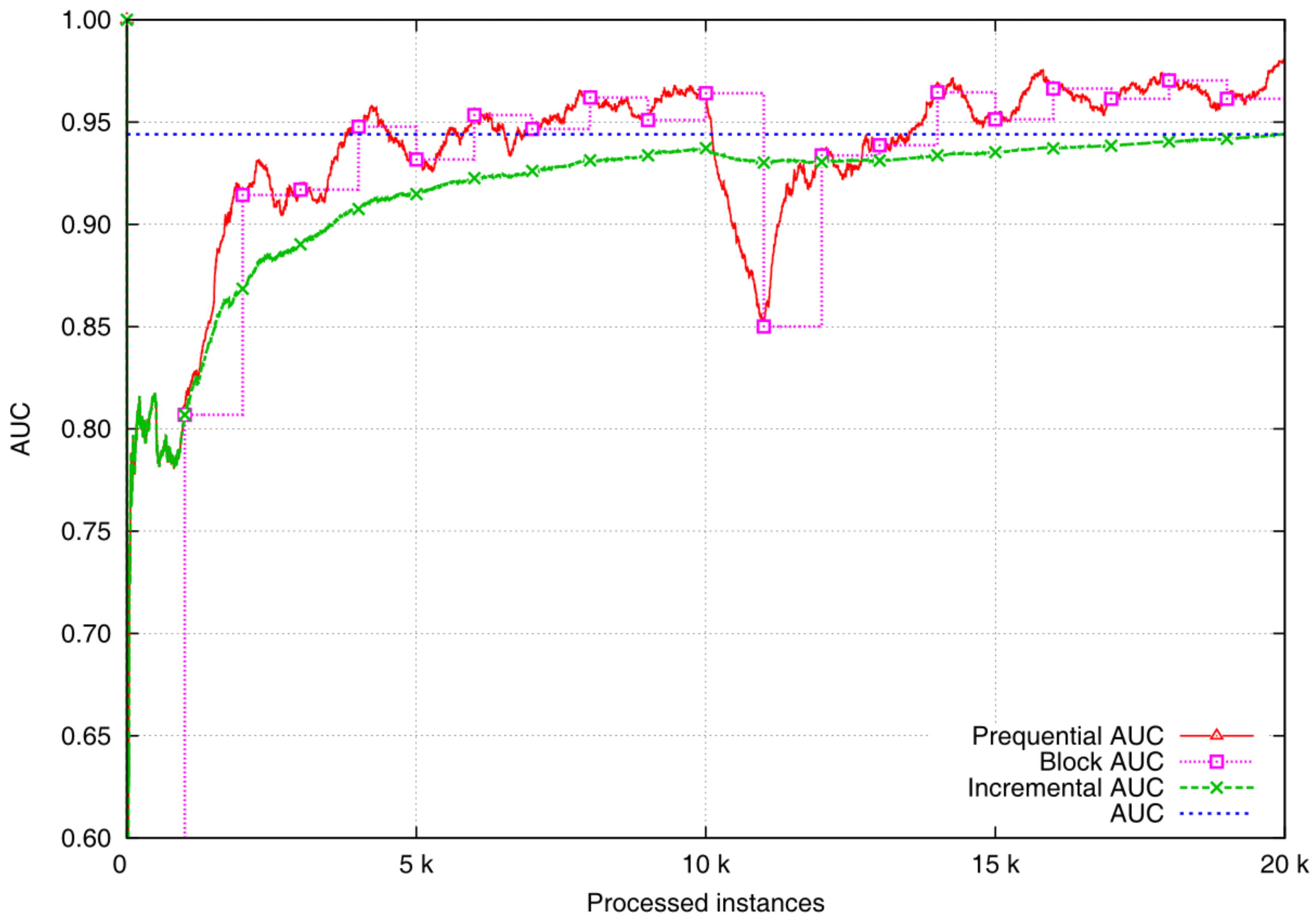
Visualization over time

- Four alternative ways of computing AUC:
  - Traditional (batch)
  - Incremental
  - In blocks
  - Prequential
- Block and prequential AUC are less pessimistic
- Prequential AUC offers the best visualization in the presence of concept drift
- Important for drift detection

Visualization over time – no drift



Visualization over time – sudden drift



Prequential AUC averaged over entire streams

- Often a single numeric metric is more useful than an entire plot
- Prequential AUC can be averaged over time
- Better than batch AUC for streams with drift
- What about stationary streams?
  - Is it identical to batch AUC?
  - If not, is it consistent? To what degree?
  - What is the range of possible absolute errors?

Consistency and discriminancy

Is averaged Prequential AUC equal to batch AUC?

Not always. It depends on the order of examples.

Table 6.1: An example in which two classifiers have the same batch AUC but different prequential (and block) AUC (for a window of $d = 2$ examples)

Classifier 1	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Classifier 2	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
t	1	3	5	7	9	11	13	15	2	4	6	8	10	12	14	16

Table 6.2: An example in which one classifiers has higher batch AUC but lower prequential (and block) AUC (for a window of $d = 2$ examples)

Classifier 3	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Classifier 4	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Consistency and discriminancy

Is averaged Prequential AUC consistent with batch AUC?

Statistically, but not strictly.

Huang and Ling proposed the notion of **strict** and **statistical consistency**, **discriminancy**, and **indifference** [Huang & Ling al. 2005].

We have experimentally shown that Prequential AUC is statistically consistent and comparably discriminant compared to batch AUC. We have also compared Prequential AUC to AUC calculated on blocks, which is less consistent and less discriminant. All measures have a very low degree of indifference.

Consistency and discriminancy

- Exhaustive rankings from 4 to 10
- All possible orderings of all possible pairs of rankings for all possible window sizes
- For a dataset with n_p minority class examples and n examples in total this gives

$$\binom{\binom{n}{n_p} + 2 - 1}{2} \cdot n! \text{ pairs}$$

- Three different class ratios (50%, 34%, 14% minority)

Consistency and discriminancy

Table 6.3: Statistical consistency compared to batch-calculated AUC for balanced datasets (50% both classes)

(a) Prequential AUC

n	d	$AUC(a) < AUC(b)\&$ $pAUC(a) < pAUC(b)$	$AUC(a) < AUC(b)\&$ $pAUC(a) > pAUC(b)$	**C**
4	2	226	30	0.883
4	3	276	0	1.000
6	2	78,558	14,206	0.847
6	3	95,086	11,688	0.891
6	4	106,566	10,098	0.913
6	5	109,152	3,840	0.966
8	2	53,158,628	11,185,976	0.826
8	3	63,240,034	10,425,237	0.858
8	4	69,182,304	9,972,770	0.874
8	5	70,761,192	7,693,012	0.902
8	6	73,479,168	6,411,408	0.920
8	7	75,503,520	2,403,360	0.969
10	2	60,634,866,784	14,063,999,524	0.812
10	3	71,130,163,463	13,290,485,752	0.843
10	4	74,037,644,404	14,163,943,582	0.839
10	5	78,920,700,364	11,176,341,290	0.876
10	6	73,168,730,742	13,467,350,816	0.845
10	7	79,076,275,296	9,293,058,744	0.895
10	8	83,944,244,880	6,701,320,800	0.926
10	9	86,539,824,000	2,926,627,200	0.967

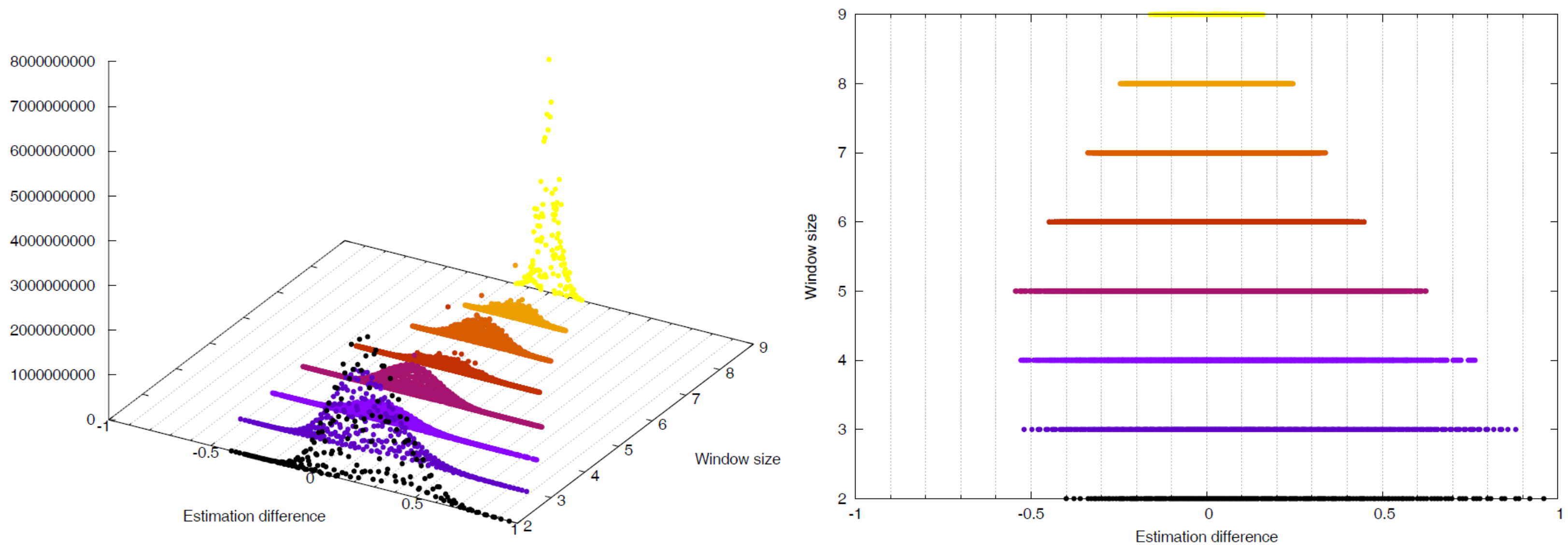
(b) Block AUC

n	d	$AUC(a) < AUC(b)\&$ $bAUC(a) < bAUC(b)$	$AUC(a) < AUC(b)\&$ $bAUC(a) > bAUC(b)$	**C**
4	2	208	56	0.788
4	3	236	28	0.894
6	2	67,230	17,428	0.794
6	3	87,150	14,164	0.860
6	4	92,930	19,356	0.828
6	5	101,520	7,608	0.930
8	2	45,405,586	13,358,810	0.773
8	3	53,850,202	14,391,887	0.789
8	4	66,567,040	8,624,167	0.885
8	5	63,572,560	12,086,730	0.840
8	6	64,611,792	14,232,192	0.819
8	7	72,136,080	4,448,880	0.942
10	2	51,487,887,104	17,000,441,762	0.752
10	3	59,490,757,935	16,883,741,696	0.779
10	4	64,512,419,513	20,509,708,299	0.759
10	5	76,714,147,735	9,078,493,566	0.894
10	6	71,855,703,700	13,934,726,872	0.838
10	7	70,312,608,720	16,552,837,632	0.809
10	8	74,378,685,600	15,590,530,080	0.827
10	9	83,365,107,840	4,817,836,800	0.945

Consistency and discriminancy

What is the range of absolute errors?

Decreases with window size.

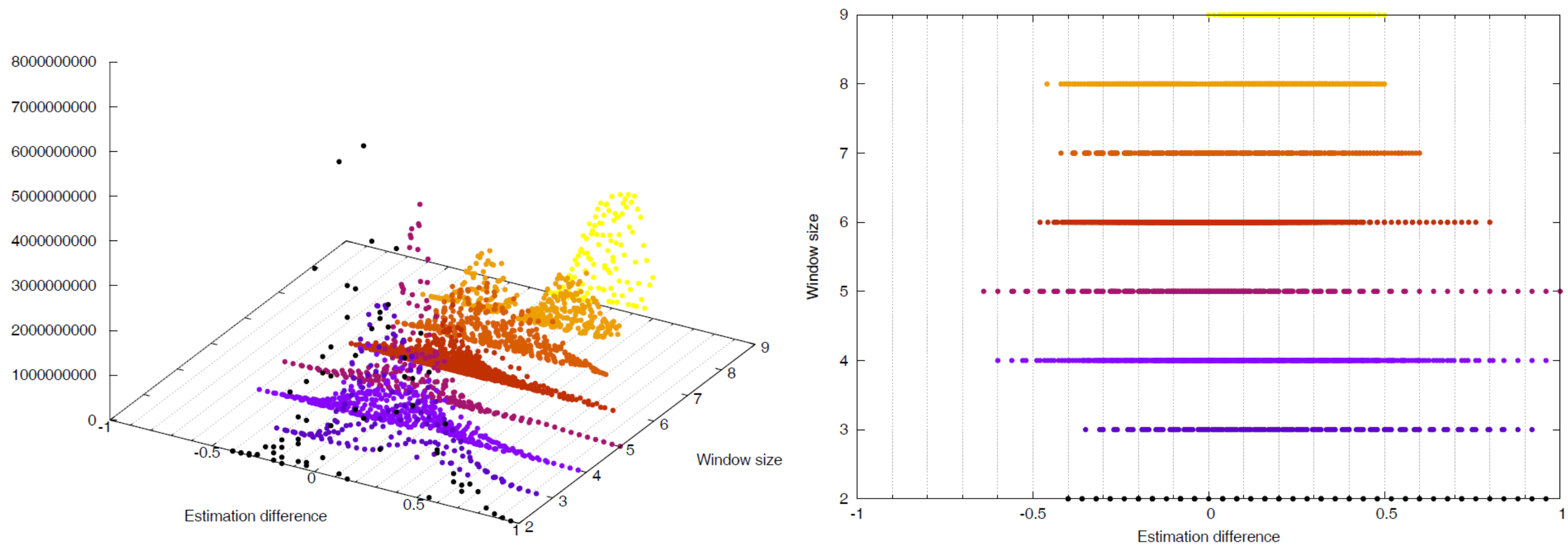


(a) 3-dimensional plot (b) 2-dimensional view

Figure 6.5: Differences between prequential and batch AUC for different window sizes on the largest balanced dataset (50% examples of both classes)

Consistency and discriminancy

Block AUC has larger errors than Prequential AUC and is on average overly optimistic.



(a) 3-dimensional plot (b) 2-dimensional view

Figure 6.6: Differences between block and batch AUC for different window sizes on the largest balanced dataset (50% examples of both classes)

Experiments with classifiers

- Verify processing speed, consistency, discriminancy, and interpretability on larger datasets
- 7 algorithms:
  - NB, VFDT, DWM, ACE, Online Bagging, Lev, OAUE
- 14 datasets
  - 2 real and 12 synthetic
  - different class ratios (1:1, 1:10, 1:100)
- Various types of drift
  - gradual, sudden, virtual, combined real and virtual, no drift
- Seperate tests for drift detection using AUC and Acc.

Processing speed

- Checked for a stationary and driftng stream
- For a window of *d* = 1000 examples prequential AUC was only between 0.50% and 0.90% slower than prequential Kappa

Table 6.10: Evaluation time per example [ms] using prequential AUC and prequential κ on a window of $d = 1000$ examples (averaged over 10 runs $\pm$ standard deviation)

	Prequential AUC	Prequential κ
RBF_{20k}	7.230 ± 0.158	7.189 ± 0.246
RBF_{20kSD}	7.344 ± 0.519	7.287 ± 0.243

Drift detection

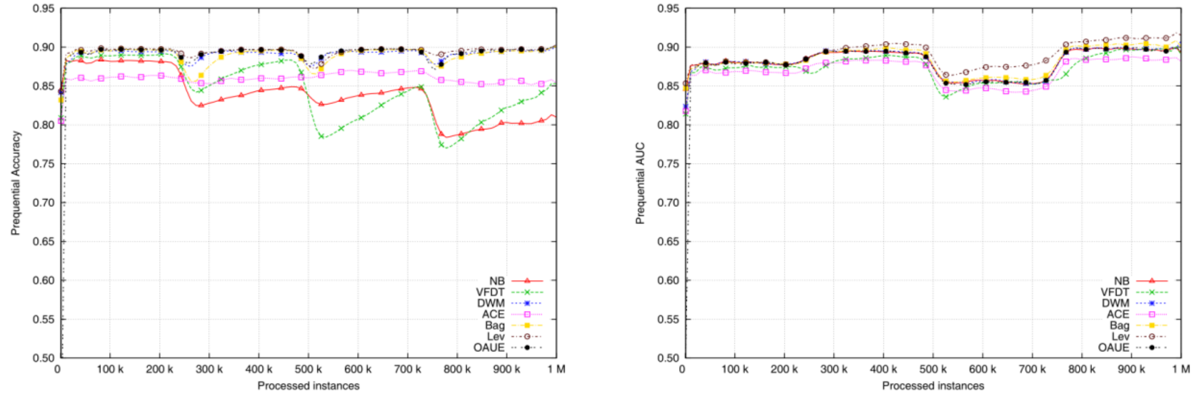
Table 6.11: Number of missed and false detections (in the format missed:false) obtained using the PH test with prequential accuracy (Acc) and prequential AUC (AUC). Average delays of correct detections are given in parenthesis, where (-) means that the detector was not triggered or the dataset did not contain any change. Subscripts in column names indicate the number of examples used for estimating errors.

	Acc_{1k}	Acc_{2k}	Acc_{3k}	Acc_{4k}	Acc_{5k}
$SEA_{NoDrift}$	0:0 (-)	0:0 (-)	0:0 (-)	0:0 (-)	0:0 (-)
Agr_1	0:2 (1040)	0:1 (1859)	0:0 (2843)	1:0 (4033)	5:0 (4603)
Agr_{10}	0:9 (1202)	0:3 (1228)	0:2 (1679)	0:2 (2190)	0:2 (2817)
Agr_{100}	2:12 (1610)	2:17 (2913)	2:10 (3136)	3:12 (3903)	3:10 (4612)
RT	6:0 (1843)	7:0 (2621)	8:0 (2933)	8:0 (3754)	8:0 (4695)
SEA_{Ratio}	10:0 (-)	10:0 (-)	10:0 (-)	10:0 (-)	10:0 (-)
RBF_{Blips}	0:2 (-)	0:1 (-)	0:0 (-)	0:0 (-)	0:0 (-)
	AUC_{1k}	AUC_{2k}	AUC_{3k}	AUC_{4k}	AUC_{5k}
$SEA_{NoDrift}$	0:0 (-)	0:0 (-)	0:0 (-)	0:0 (-)	0:0 (-)
Agr_1	2:2 (1042)	3:1 (1760)	4:1 (2726)	4:0 (3773)	7:0 (4640)
Agr_{10}	0:5 (868)	0:5 (1539)	0:1 (1506)	0:1 (1778)	1:1 (2197)
Agr_{100}	0:19 (1548)	0:18 (2461)	1:9 (2664)	1:11 (3563)	2:9 (4835)
RT	3:0 (1815)	5:0 (2407)	6:0 (3105)	6:0 (4121)	7:0 (4725)
SEA_{Ratio}	0:0 (1339)	0:0 (2249)	0:0 (3152)	0:0 (4057)	0:0 (4959)
RBF_{Blips}	0:3 (-)	0:1 (-)	0:0 (-)	0:0 (-)	0:0 (-)

Model selection

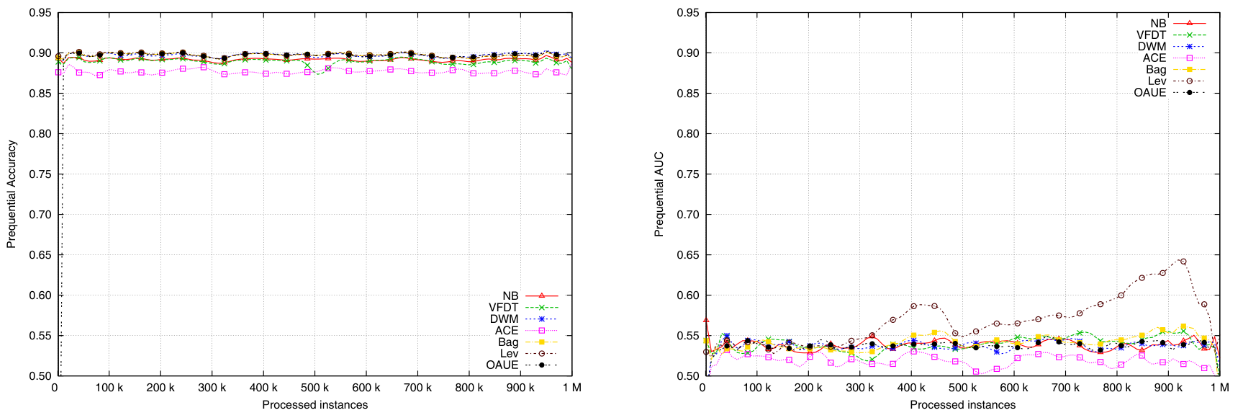
Table 6.12: Average prequential accuracy (Acc.) and AUC (AUC)

	NB		VFDT		DWM		ACE		Bag		Lev		OAUE	
	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
$\texttt{SEA}_{ND}$	0.86	0.90	0.89	0.89	0.88	0.90	0.86	0.89	0.89	0.90	0.90	0.90	0.89	0.90
$\texttt{SEA}_{1}$	0.84	0.88	0.85	0.87	0.89	0.88	0.86	0.87	0.89	0.88	0.89	0.89	0.89	0.88
$\texttt{SEA}_{10}$	0.84	0.74	0.87	0.73	0.89	0.74	0.87	0.72	0.89	0.74	0.89	0.75	0.89	0.74
$\texttt{SEA}_{100}$	0.89	0.54	0.89	0.54	0.90	0.54	0.88	0.52	0.90	0.54	0.90	0.57	0.90	0.54
$\texttt{Hyp}_{1}$	0.78	0.85	0.81	0.87	0.88	0.92	0.78	0.83	0.88	0.93	0.86	0.92	0.88	0.93
$\texttt{Hyp}_{10}$	0.94	0.57	0.93	0.53	0.94	0.52	0.89	0.50	0.94	0.56	0.94	0.55	0.94	0.55
$\texttt{Hyp}_{100}$	0.88	0.80	0.89	0.75	0.91	0.76	0.88	0.71	0.91	0.81	0.91	0.80	0.91	0.82
RBF	0.74	0.83	0.97	0.99	0.98	1.00	0.87	0.89	0.99	1.00	0.99	1.00	0.99	1.00
$\texttt{SEA}_{RC}$	0.86	0.77	0.89	0.77	0.89	0.77	0.87	0.75	0.90	0.77	0.90	0.78	0.90	0.77
$\texttt{SEA}_{RC+D}$	0.82	0.77	0.85	0.76	0.89	0.77	0.86	0.75	0.89	0.77	0.89	0.77	0.89	0.77
$\texttt{Hyp}_{RC}$	0.93	0.67	0.93	0.63	0.93	0.61	0.89	0.55	0.94	0.66	0.93	0.66	0.93	0.66
$\texttt{Hyp}_{RC+D}$	0.92	0.64	0.92	0.61	0.93	0.63	0.88	0.59	0.93	0.65	0.93	0.64	0.93	0.65
Air	0.65	0.66	0.64	0.66	0.65	0.65	0.65	0.61	0.64	0.65	0.62	0.60	0.67	0.68
PAKKD	0.56	0.64	0.73	0.57	0.80	0.50	-	-	0.80	0.63	0.80	0.62	0.80	0.62



(a) Prequential accuracy (b) Prequential AUC

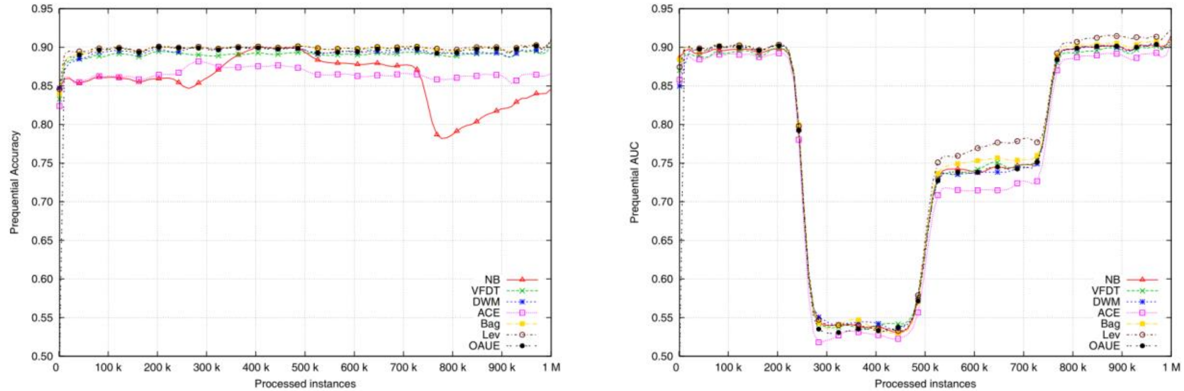
Figure 6.11: Comparison of prequential accuracy and AUC on a data stream with sudden drifts and a balanced class ratio ($\mathtt{SEA}_1$)



(a) Prequential accuracy (b) Prequential AUC

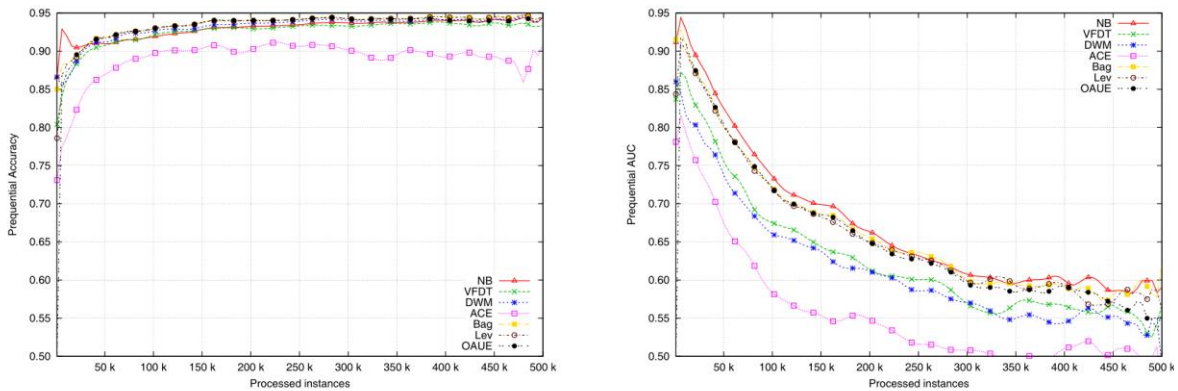
Figure 6.12: Comparison of prequential accuracy and AUC on a data stream with sudden drifts and a 1:100 class imbalance ratio ($\mathtt{SEA}_{100}$)

Experiments



(a) Prequential accuracy (b) Prequential AUC

Figure 6.15: Comparison of prequential accuracy and AUC on a data stream with sudden class ratio changes ($\texttt{SEA}_{RC}$)



(a) Prequential accuracy (b) Prequential AUC

Figure 6.16: Comparison of prequential accuracy and AUC on a data stream with a gradual class ratio change ($\texttt{Hyp}_{RC}$)

Conclusions

- **Data stream classification**
- **Prequential AUC**
  - Suitable for imbalanced data
  - Statistically consistent with traditional AUC for stationary streams
  - More discriminant than accuracy
  - Constant time and memory
  - Forgetting mechanism
  - Virtual drift detection
- **Future work**
  - Different drift detector
  - Scored AUC

Thank you!

Consistency 50% minority

Table 6.3: Statistical consistency compared to batch-calculated AUC for balanced datasets (50% both classes)

(a) Prequential AUC

n	d	$AUC(a) < AUC(b)\&$ $pAUC(a) < pAUC(b)$	$AUC(a) < AUC(b)\&$ $pAUC(a) > pAUC(b)$	**C**
4	2	226	30	0.883
4	3	276	0	1.000
6	2	78,558	14,206	0.847
6	3	95,086	11,688	0.891
6	4	106,566	10,098	0.913
6	5	109,152	3,840	0.966
8	2	53,158,628	11,185,976	0.826
8	3	63,240,034	10,425,237	0.858
8	4	69,182,304	9,972,770	0.874
8	5	70,761,192	7,693,012	0.902
8	6	73,479,168	6,411,408	0.920
8	7	75,503,520	2,403,360	0.969
10	2	60,634,866,784	14,063,999,524	0.812
10	3	71,130,163,463	13,290,485,752	0.843
10	4	74,037,644,404	14,163,943,582	0.839
10	5	78,920,700,364	11,176,341,290	0.876
10	6	73,168,730,742	13,467,350,816	0.845
10	7	79,076,275,296	9,293,058,744	0.895
10	8	83,944,244,880	6,701,320,800	0.926
10	9	86,539,824,000	2,926,627,200	0.967

(b) Block AUC

n	d	$AUC(a) < AUC(b)\&$ $bAUC(a) < bAUC(b)$	$AUC(a) < AUC(b)\&$ $bAUC(a) > bAUC(b)$	**C**
4	2	208	56	0.788
4	3	236	28	0.894
6	2	67,230	17,428	0.794
6	3	87,150	14,164	0.860
6	4	92,930	19,356	0.828
6	5	101,520	7,608	0.930
8	2	45,405,586	13,358,810	0.773
8	3	53,850,202	14,391,887	0.789
8	4	66,567,040	8,624,167	0.885
8	5	63,572,560	12,086,730	0.840
8	6	64,611,792	14,232,192	0.819
8	7	72,136,080	4,448,880	0.942
10	2	51,487,887,104	17,000,441,762	0.752
10	3	59,490,757,935	16,883,741,696	0.779
10	4	64,512,419,513	20,509,708,299	0.759
10	5	76,714,147,735	9,078,493,566	0.894
10	6	71,855,703,700	13,934,726,872	0.838
10	7	70,312,608,720	16,552,837,632	0.809
10	8	74,378,685,600	15,590,530,080	0.827
10	9	83,365,107,840	4,817,836,800	0.945

Discriminancy 50% minority

Table 6.4: Statistical discriminancy compared to batch-calculated AUC for balanced datasets (50% both classes)

(a) Prequential AUC

n	d	$AUC(a) < AUC(b) \&$ $pAUC(a) = pAUC(b)$	$AUC(a) = AUC(b) \&$ $pAUC(a) < pAUC(b)$	D
4	2	80	8	10.00
4	3	60	8	7.50
6	2	26,756	3,500	7.64
6	3	12,746	4,044	3.15
6	4	2,856	4,724	0.60
6	5	6,528	4,032	1.62
8	2	16,214,756	2,353,182	6.89
8	3	6,894,089	2,738,339	2.52
8	4	1,304,298	2,996,764	0.44
8	5	2,104,792	2,870,944	0.73
8	6	356,712	2,891,136	0.12
8	7	2,652,480	2,349,360	1.13
10	2	16,474,070,244	2,325,714,636	7.08
10	3	6,751,652,794	2,692,735,085	2.51
10	4	1,135,333,025	2,782,776,394	0.41
10	5	1,523,949,099	2,693,165,272	0.57
10	6	205,901,522	2,464,416,666	0.08
10	7	837,730,392	2,688,388,944	0.31
10	8	163,838,880	2,513,982,960	0.07
10	9	1,707,148,800	2,113,695,360	0.81

(b) Block AUC

n	d	$AUC(a) < AUC(b) \&$ $bAUC(a) = bAUC(b)$	$AUC(a) = AUC(b) \&$ $bAUC(a) < bAUC(b)$	D
4	2	72	0	∞
4	3	72	4	18.00
6	2	34,862	2,010	17.34
6	3	18,206	2,802	6.50
6	4	7,234	4,122	1.75
6	5	10,392	2,712	3.83
8	2	21,794,964	1,827,868	11.92
8	3	12,317,271	2,152,120	5.72
8	4	5,268,165	2,258,632	2.33
8	5	4,899,706	2,262,634	2.17
8	6	1,403,304	2,475,192	0.57
8	7	3,974,400	1,751,040	2.27
10	2	22,684,607,686	1,764,461,488	12.86
10	3	14,797,802,378	2,108,350,183	7.02
10	4	4,314,793,199	2,648,700,613	1.63
10	5	5,828,349,452	2,133,672,486	2.73
10	6	1,051,552,508	2,577,780,700	0.41
10	7	2,341,618,080	2,381,208,000	0.98
10	8	840,188,880	2,230,413,840	0.38
10	9	2,990,655,360	1,607,114,880	1.86

Consistency 34% minority

Table 6.5: Statistical consistency compared to batch-calculated AUC for imbalanced datasets (34% minority class)

(a) Prequential AUC

n	d	$AUC(a) < AUC(b)$& $pAUC(a) < pAUC(b)$	$AUC(a) < AUC(b)$& $pAUC(a) > pAUC(b)$	**C**
4	2	96	8	0.923
4	3	112	8	0.933
6	2	44,328	7,430	0.856
6	3	53,318	7,312	0.879
6	4	58,774	6,904	0.895
6	5	62,064	3,312	0.949
8	2	34,250,730	6,980,006	0.831
8	3	40,699,497	6,793,142	0.857
8	4	44,193,331	6,982,899	0.864
8	5	45,463,956	5,426,998	0.893
8	6	47,516,640	4,326,864	0.917
8	7	49,281,120	2,127,600	0.959
10	2	14,056,095,136	2,991,723,204	0.825
10	3	16,506,930,025	3,058,609,668	0.844
10	4	17,397,851,565	3,458,536,944	0.834
10	5	17,857,902,489	3,077,583,820	0.853
10	6	17,368,267,282	3,208,774,634	0.844
10	7	18,541,658,208	2,198,965,392	0.894
10	8	19,469,178,720	1,653,883,920	0.922
10	9	20,338,174,080	834,825,600	0.961

(b) Block AUC

n	d	$AUC(a) < AUC(b)$& $bAUC(a) < bAUC(b)$	$AUC(a) < AUC(b)$& $bAUC(a) > bAUC(b)$	**C**
4	2	92	8	0.920
4	3	102	18	0.850
6	2	39,222	9,582	0.804
6	3	49,482	7,590	0.867
6	4	52,302	11,468	0.820
6	5	59,496	5,352	0.917
8	2	28,945,260	8,463,258	0.774
8	3	34,682,961	8,761,157	0.798
8	4	43,361,987	6,740,645	0.865
8	5	41,705,650	7,623,636	0.845
8	6	42,382,608	8,985,336	0.825
8	7	47,743,920	3,418,560	0.933
10	2	12,195,033,276	3,413,021,306	0.781
10	3	13,792,398,281	3,675,831,232	0.790
10	4	15,903,298,910	4,113,647,610	0.794
10	5	17,306,025,784	3,102,200,565	0.848
10	6	17,469,134,532	2,947,901,726	0.856
10	7	16,864,120,896	3,466,868,304	0.829
10	8	17,293,451,760	3,656,527,920	0.825
10	9	19,822,360,320	1,216,010,880	0.942

Discriminancy 34% minority

Table 6.6: Statistical discriminancy compared to batch-calculated AUC for imbalanced datasets (34% minority class)

(a) Prequential AUC

n	d	$AUC(a) < AUC(b)$& $pAUC(a) = pAUC(b)$	$AUC(a) = AUC(b)$& $pAUC(a) < pAUC(b)$	D
4	2	40	0	∞
4	3	24	0	∞
6	2	15,922	1,676	9.50
6	3	7,050	2,080	3.39
6	4	1,988	2,408	0.83
6	5	2,304	2,040	1.13
8	2	10,822,384	1,445,690	7.49
8	3	4,560,481	1,698,162	2.69
8	4	904,810	1,861,822	0.49
8	5	1,156,828	1,864,496	0.62
8	6	191,496	1,831,680	0.10
8	7	646,560	1,517,760	0.43
10	2	4,231,464,860	512,002,316	8.26
10	3	1,713,743,507	605,486,578	2.83
10	4	295,397,178	680,067,479	0.43
10	5	297,697,500	660,113,562	0.45
10	6	43,162,624	665,026,736	0.06
10	7	104,005,512	765,190,632	0.14
10	8	79,602,480	614,255,040	0.13
10	9	114,589,440	486,864,000	0.24

(b) Block AUC

n	d	$AUC(a) < AUC(b)$& $bAUC(a) = bAUC(b)$	$AUC(a) = AUC(b)$& $bAUC(a) < bAUC(b)$	D
4	2	44	0	∞
4	3	24	0	∞
6	2	18,876	1,050	17.98
6	3	10,608	1,042	10.18
6	4	3,896	2,156	1.81
6	5	2,832	1,416	2.00
8	2	14,644,602	1,061,404	13.80
8	3	8,609,002	1,206,625	7.13
8	4	1,978,408	1,835,340	1.08
8	5	2,718,496	1,620,150	1.68
8	6	667,056	1,522,224	0.44
8	7	892,800	1,100,880	0.81
10	2	5,671,228,618	384,816,444	14.74
10	3	3,811,053,687	451,977,687	8.43
10	4	1,134,839,167	602,144,260	1.88
10	5	824,957,460	625,114,249	1.32
10	6	203,168,282	669,160,838	0.30
10	7	513,639,912	557,695,848	0.92
10	8	252,685,440	474,176,160	0.53
10	9	249,217,920	341,107,200	0.73

Consistency 14% minority

Table 6.7: Statistical consistency compared to batch-calculated AUC for imbalanced datasets (14% minority class)

(a) Prequential AUC

n	d	$AUC(a) < AUC(b)\&$ $pAUC(a) < pAUC(b)$	$AUC(a) < AUC(b)\&$ $pAUC(a) > pAUC(b)$	**C**
4	2	96	8	0.923
4	3	112	8	0.933
6	2	6,888	840	0.891
6	3	8,068	1,240	0.867
6	4	8,568	1,220	0.875
6	5	9,264	816	0.919
8	2	701,568	96,768	0.879
8	3	841,168	145,632	0.852
8	4	868,680	164,584	0.841
8	5	891,960	164,384	0.844
8	6	936,144	139,056	0.871
8	7	1,006,560	82,080	0.925
10	2	99,792,000	14,636,160	0.872
10	3	121,113,792	21,683,232	0.848
10	4	125,448,672	25,481,808	0.831
10	5	127,000,512	27,198,096	0.824
10	6	128,755,248	27,134,520	0.826
10	7	132,687,744	24,559,776	0.844
10	8	139,400,640	19,555,920	0.877
10	9	148,861,440	10,805,760	0.932

(b) Block AUC

n	d	$AUC(a) < AUC(b)\&$ $bAUC(a) < bAUC(b)$	$AUC(a) < AUC(b)\&$ $bAUC(a) > bAUC(b)$	**C**
4	2	92	8	0.920
4	3	102	18	0.850
6	2	6,096	784	0.886
6	3	8,152	644	0.927
6	4	8,276	934	0.899
6	5	8,880	1,200	0.881
8	2	571,680	71,040	0.889
8	3	796,032	90,528	0.898
8	4	914,944	64,888	0.934
8	5	923,656	77,832	0.922
8	6	940,368	99,312	0.904
8	7	982,800	105,840	0.903
10	2	81,250,560	11,632,320	0.875
10	3	105,151,848	16,303,848	0.866
10	4	126,349,656	12,907,656	0.907
10	5	139,374,336	8,251,152	0.944
10	6	142,754,112	12,808,368	0.918
10	7	141,324,576	10,824,720	0.929
10	8	142,572,240	12,827,520	0.917
10	9	146,603,520	13,063,680	0.918

Discriminancy 14% minority

Table 6.8: Statistical discriminancy compared to batch-calculated AUC for imbalanced datasets (14% minority class)

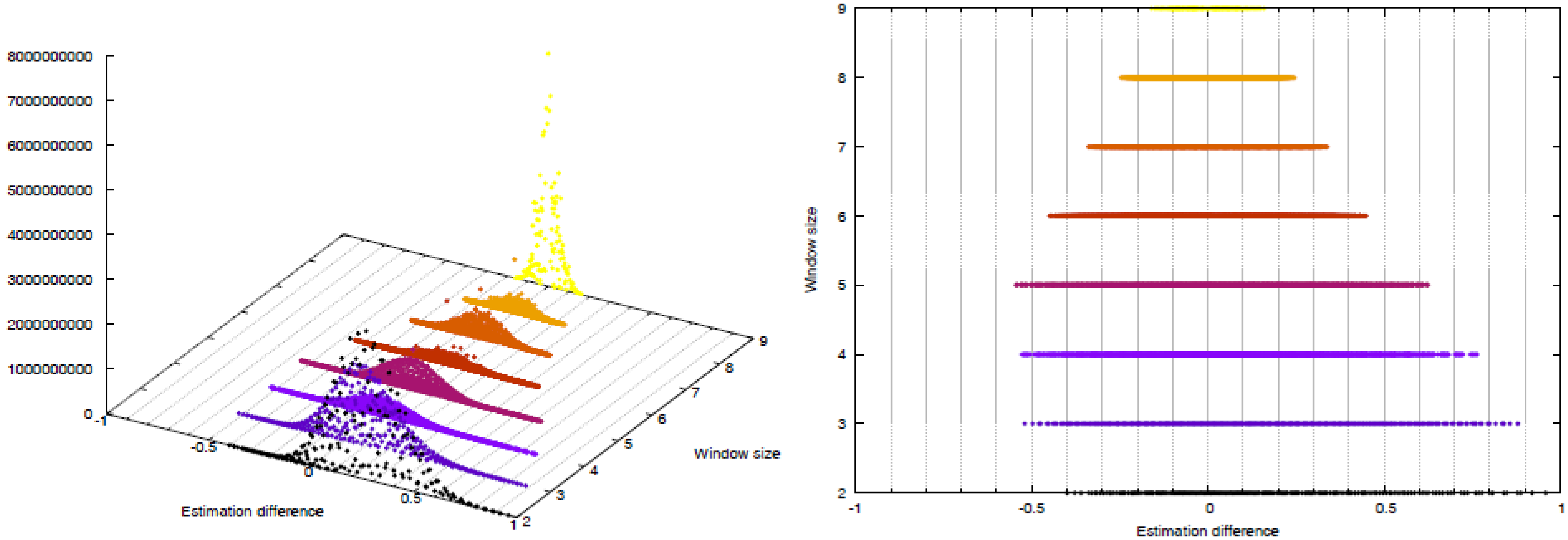
(a) Prequential AUC

n	d	$AUC(a) < AUC(b)$& $pAUC(a) = pAUC(b)$	$AUC(a) = AUC(b)$& $pAUC(a) < pAUC(b)$	**D**
4	2	40	0	∞
4	3	24	0	∞
6	2	3,072	0	∞
6	3	1,492	0	∞
6	4	1,012	0	∞
6	5	720	0	∞
8	2	330,624	0	∞
8	3	142,160	0	∞
8	4	95,696	0	∞
8	5	72,616	0	∞
8	6	53,760	0	∞
8	7	40,320	0	∞
10	2	48,867,840	0	∞
10	3	20,498,976	0	∞
10	4	12,365,520	0	∞
10	5	9,097,392	0	∞
10	6	7,406,232	0	∞
10	7	6,048,480	0	∞
10	8	4,339,440	0	∞
10	9	3,628,800	0	∞

(b) Block AUC

n	d	$AUC(a) < AUC(b)$& $bAUC(a) = bAUC(b)$	$AUC(a) = AUC(b)$& $bAUC(a) < bAUC(b)$	**D**
4	2	44	0	∞
4	3	24	0	∞
6	2	3,920	0	∞
6	3	2,004	0	∞
6	4	1,590	0	∞
6	5	720	0	∞
8	2	486,240	0	∞
8	3	242,400	0	∞
8	4	149,128	0	∞
8	5	127,472	0	∞
8	6	89,280	0	∞
8	7	40,320	0	∞
10	2	70,413,120	0	∞
10	3	41,840,304	0	∞
10	4	24,038,688	0	∞
10	5	15,670,512	0	∞
10	6	7,733,520	0	∞
10	7	11,146,704	0	∞
10	8	7,896,240	0	∞
10	9	3,628,800	0	∞

Prequential error 50% minority

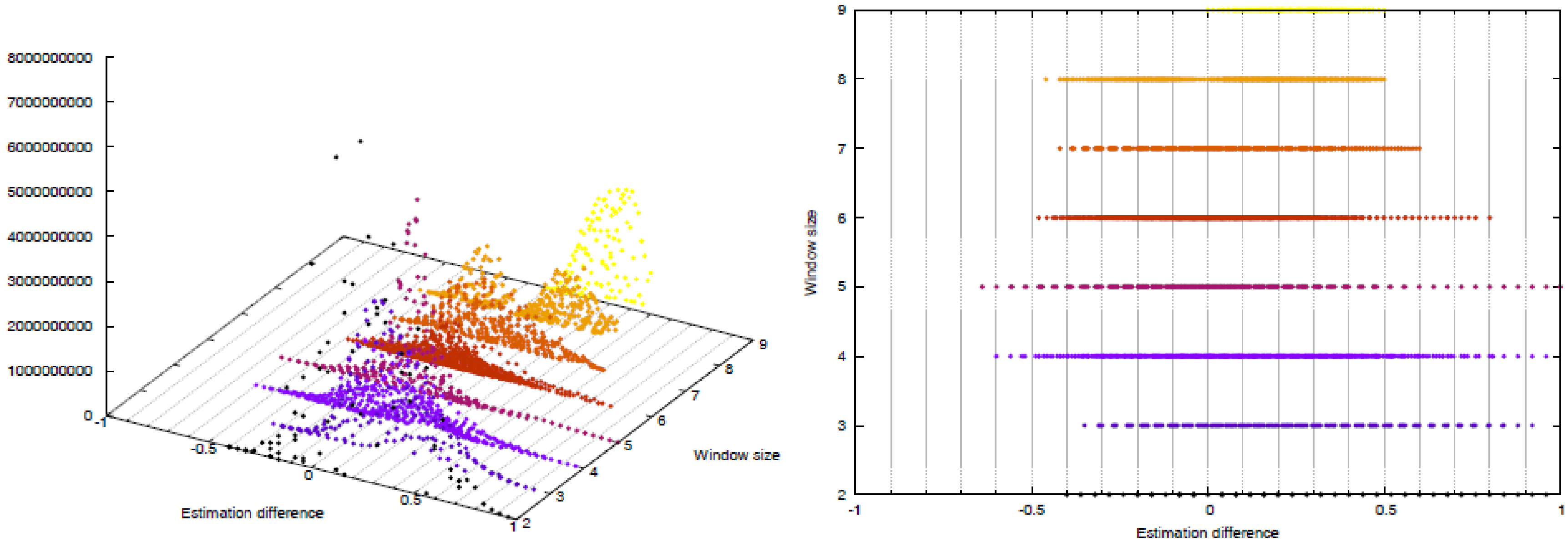


(a) 3-dimensional plot

(b) 2-dimensional view

Figure 6.5: Differences between prequential and batch AUC for different window sizes on the largest balanced dataset (50% examples of both classes)

Block error 50% minority

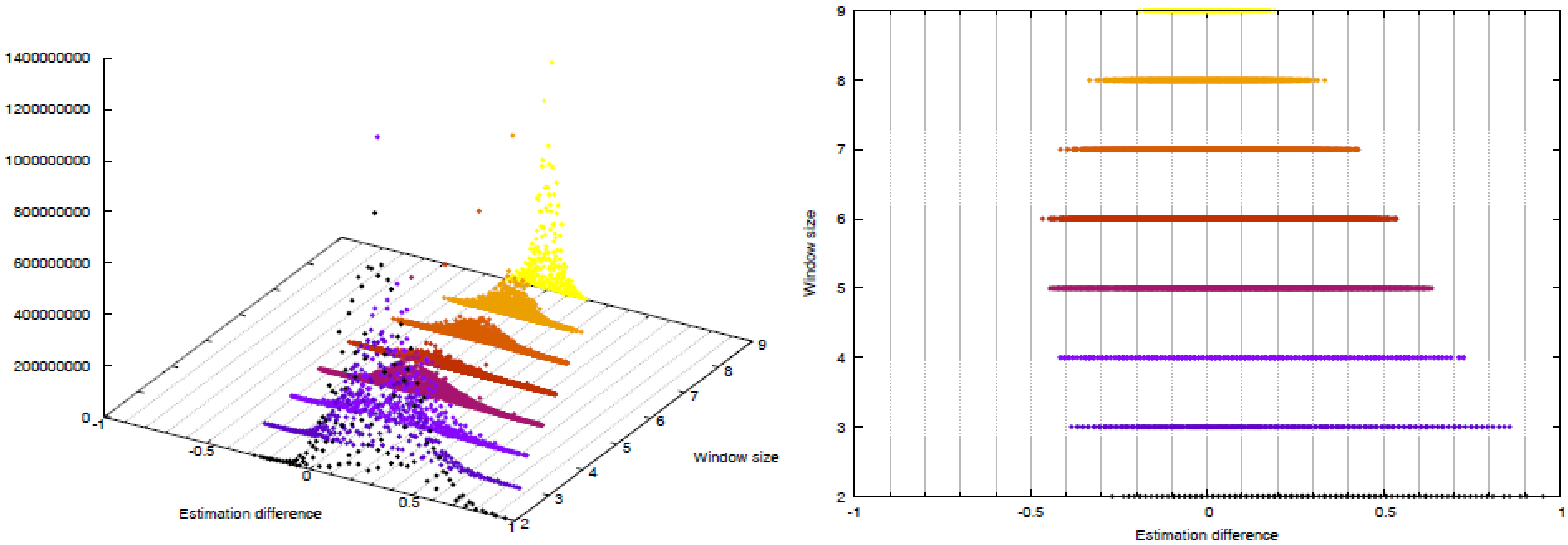


(a) 3-dimensional plot

(b) 2-dimensional view

Figure 6.6: Differences between block and batch AUC for different window sizes on the largest balanced dataset (50% examples of both classes)

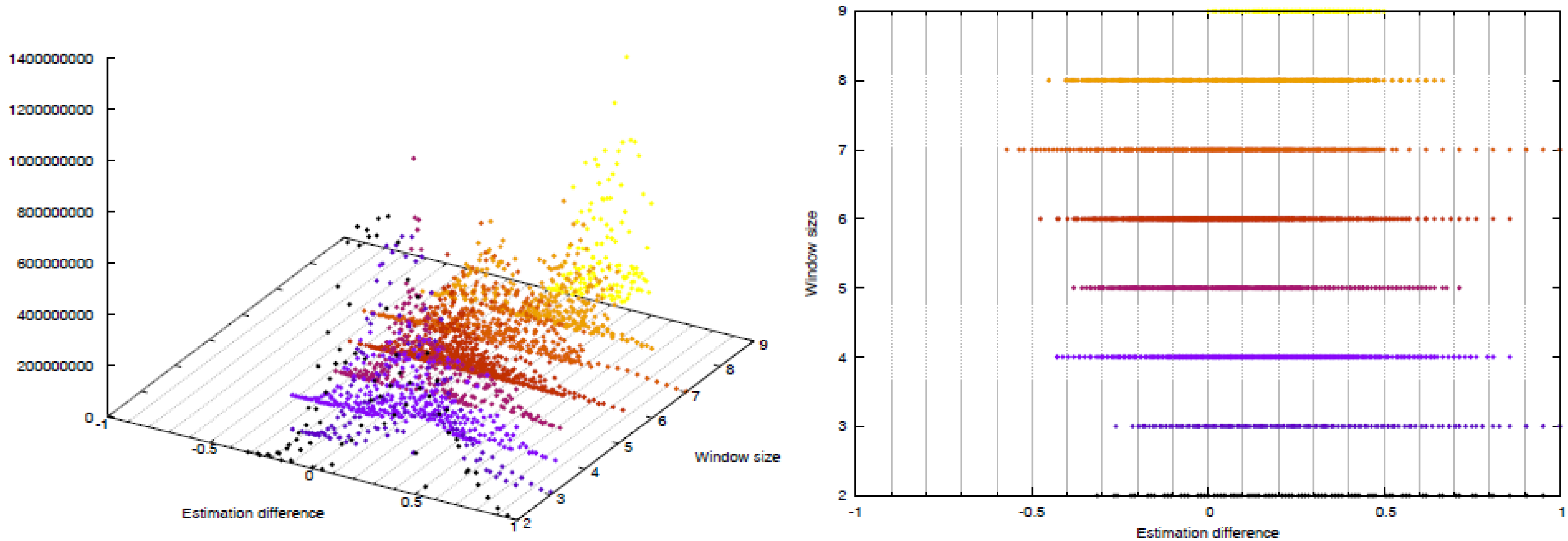
Prequential error 34% minority



(a) 3-dimensional plot (b) 2-dimensional view

Figure 6.7: Differences between prequential and batch AUC for different window sizes on the largest dataset with medium class imbalance (34% minority class examples)

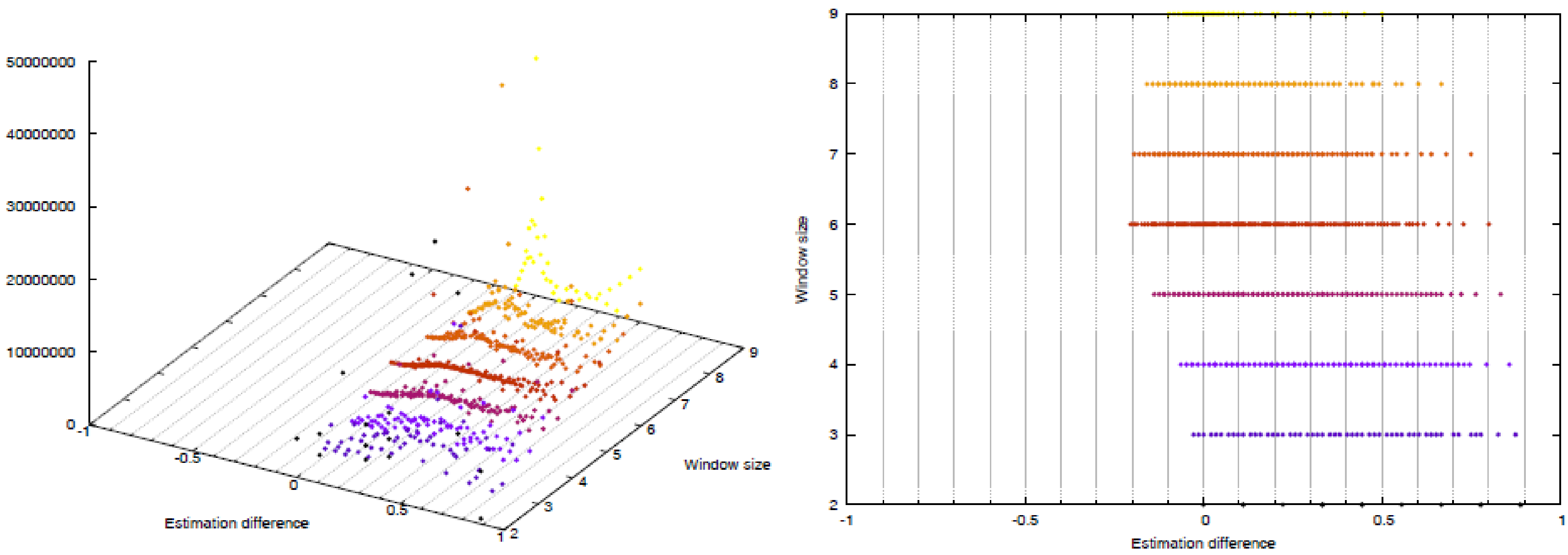
Block error 34% minority



(a) 3-dimensional plot (b) 2-dimensional view

Figure 6.8: Differences between block and batch AUC for different window sizes on the largest dataset with medium class imbalance (34% minority class examples)

Prequential error 14% minority

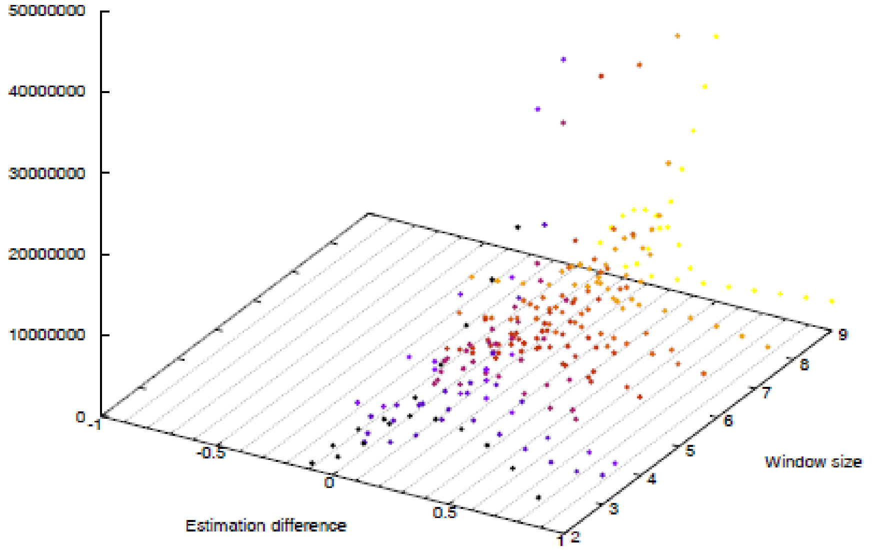


(a) 3-dimensional plot

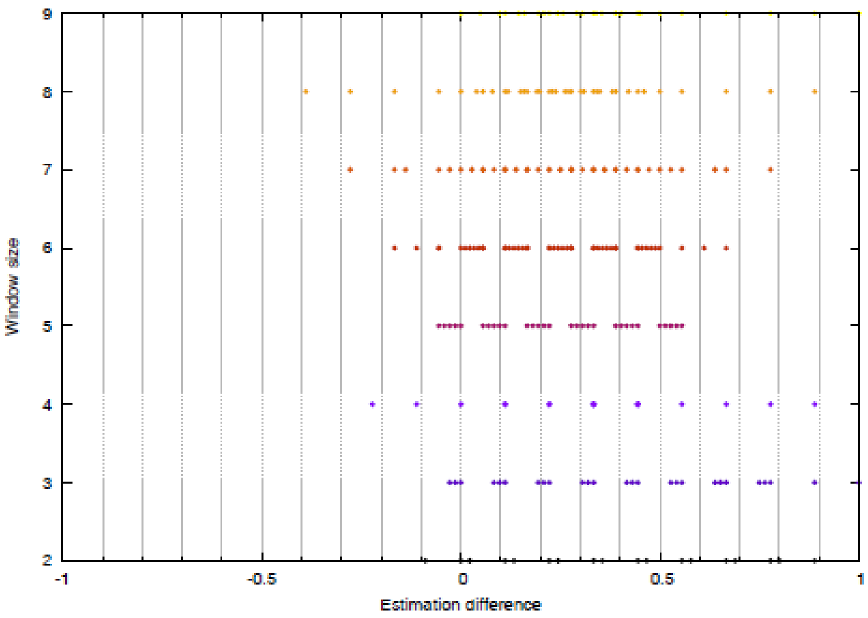
(b) 2-dimensional view

Figure 6.9: Differences between prequential and batch AUC for different window sizes on the largest dataset with high class imbalance (14% minority class examples)

Block error 14% minority



(a) 3-dimensional plot



(b) 2-dimensional view

Figure 6.10: Differences between block and batch AUC for different window sizes on the largest dataset with high class imbalance (14% minority class examples)