Institute for Program Structures and Data Organization (IPD)
Department of Informatics

Efficient Subspace Search in Data Streams

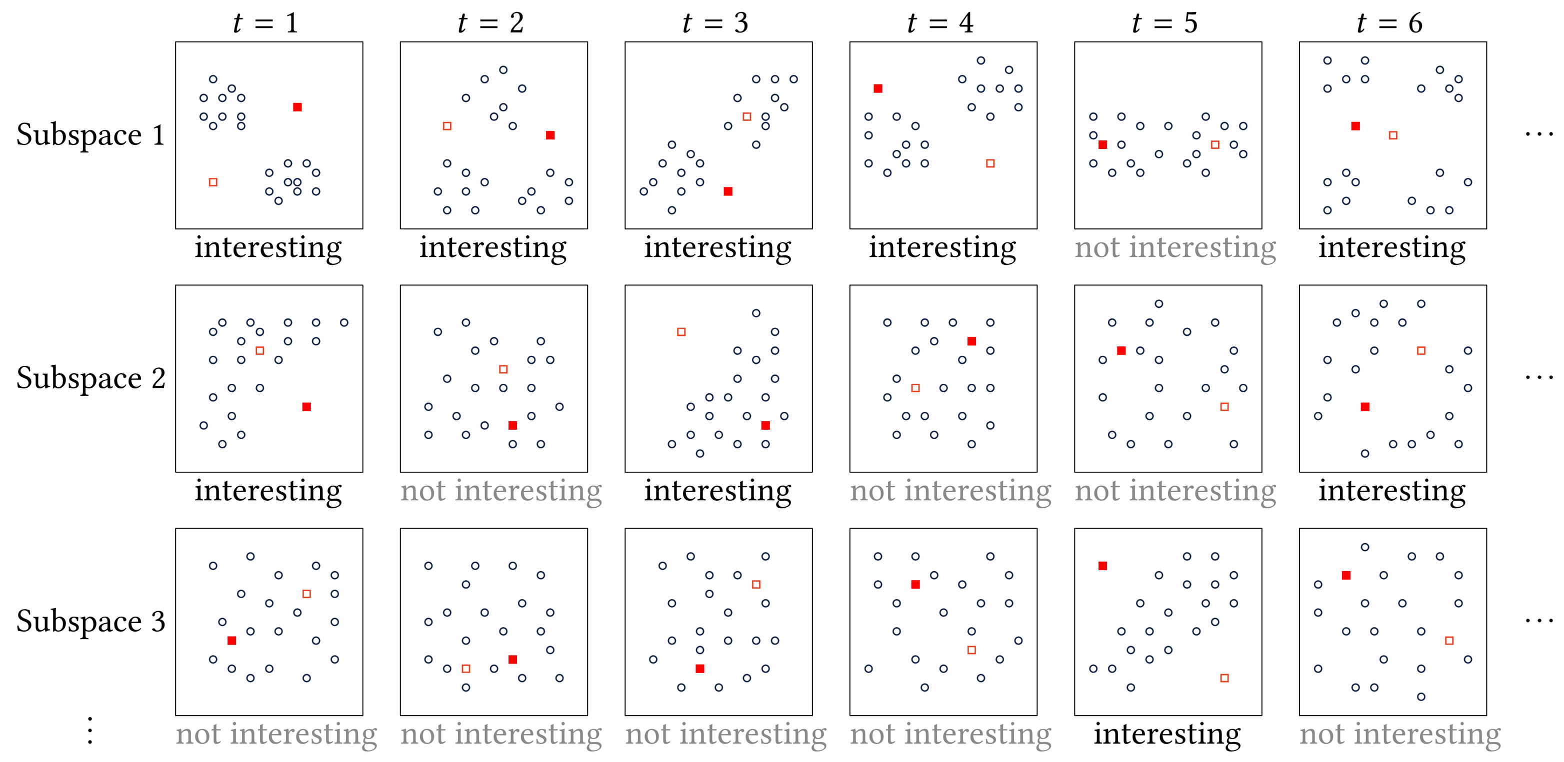
Edouard Fouché, Florian Kalinke, Klemens Böhm

Information Systems, Elsevier (In press). Preprint: arXiv:2011.06959

Motivation

- **Data streams are everywhere**
  - Network traffic, sensor data, financial transactions, …
- **Mining patterns (outliers, clusters) must take place in real time**
  - Difficulties: curse of dimensionality & concept drift
- **"Ensemble" Feature Selection: Subspace Search**
  - Goal: find relevant projections, as in [2]
  - So far only works for specific algorithms, or static data
- **Subspace Search requires:**
  - A quality measure (how "good" is a given projection)
  - A search scheme (explore the exponential set of subspaces)
  - → We extend this idea to the streaming setting
- **Stream constraints:** Efficiency, Single Scan, Adaptation, Anytime

- *"Searching for outliers in high-dimensional data is like searching for a needle in a haystack, while the haystack "hides" among an exponential number of haystacks"* [3].
- **Data streams:** The haystacks and needle location can also change.



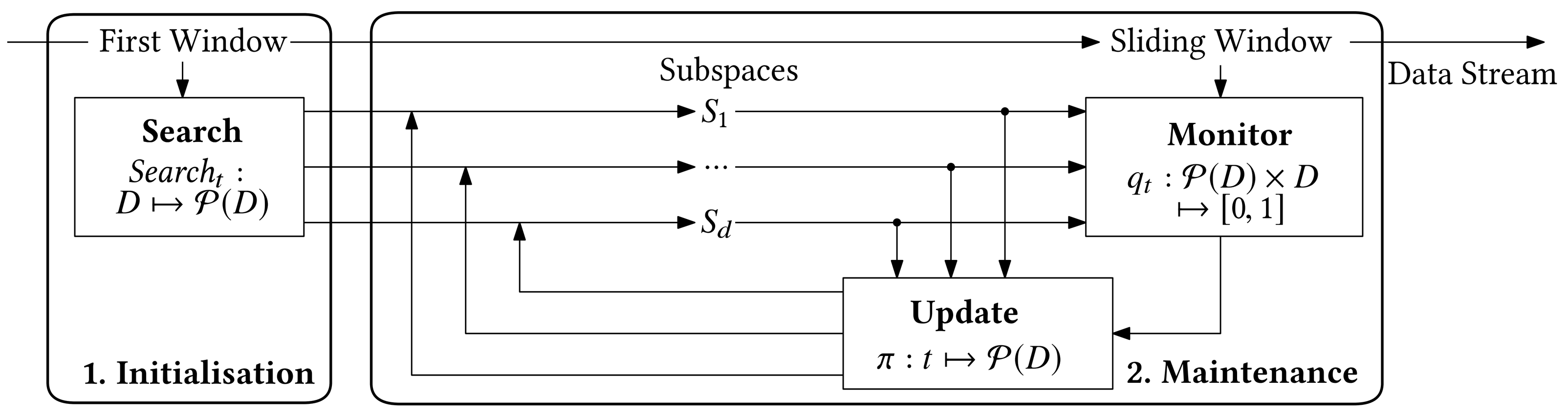
Related Work

- Subspace search for static data: [2, 5, 6]
- Subspace search for data streams: [7, 8]
  - Only for specific static Data Mining algorithms
- Closest competitor: StreamHiCS [9]
  - Boils down to a periodic repetition of the procedure in [2]
- The most similar approach to ours, but for static data: GMD [6]
- Outlier detectors: xStream [10], RS-Stream [11]

Our Contributions

- **SGMRD: Streaming Greedy Maximum Random Deviation**
  - A new method for "general" subspace search in data streams
  - SGMRD leverages novel multivariate dependency measures and Multi-Armed Bandit (MAB) algorithms
  - Monitoring subspaces in data streams improves the performance of subsequent mining tasks, e.g., outlier detection
- **Code, data:** `https://github.com/edouardfouche/SGMRD`

Our Approach & Experiments



Search: Greedy dimension-based scheme
- Returns one single subspace per dimension d, as in [6]
- This subspace maximizes quality w.r.t. d

Monitor: Quality as a "contrast" measure
- **Contrast:** $q_t(S, s_i) = 1 - \frac{1}{M}\sum_{m=1}^{M} T(\hat{p}(S|s_i), \hat{p}(S|\bar{s_i})) \in [0,1]$
- **Smoothing:** $Q_{t+1}(s_i) = \gamma * q_t(S, s_i) + (1 - \gamma) * q_{t+1}(S, s_i)$

Update: Policy based on (Multiple-Play) Multi-Armed Bandit
- **Success:** The search w.r.t. s_i yields a better subspace (1)
- **Failure:** The search w.r.t. s_i did not yield a better subspace (0)
- We use a strategy based on Thompson Sampling (TS) [12]

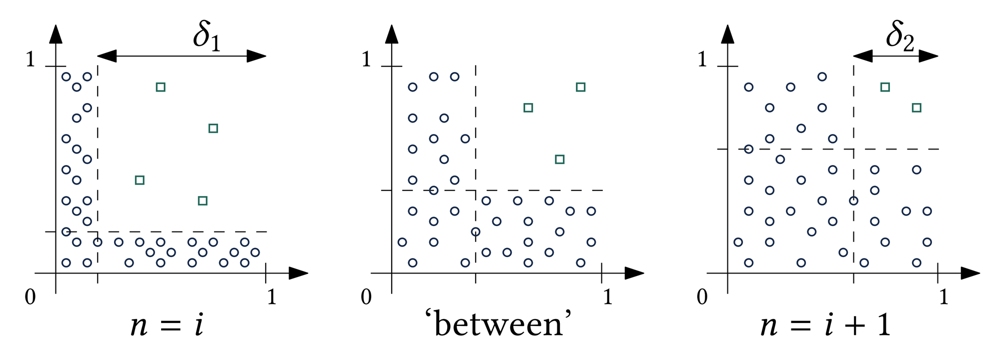
Downstream Data Mining
- Virtually any task: outlier detection, clustering, predictions…
- We focus on outlier detection (with Local Outlier Factor (LOF))

→ For more details and experiments, see [1]

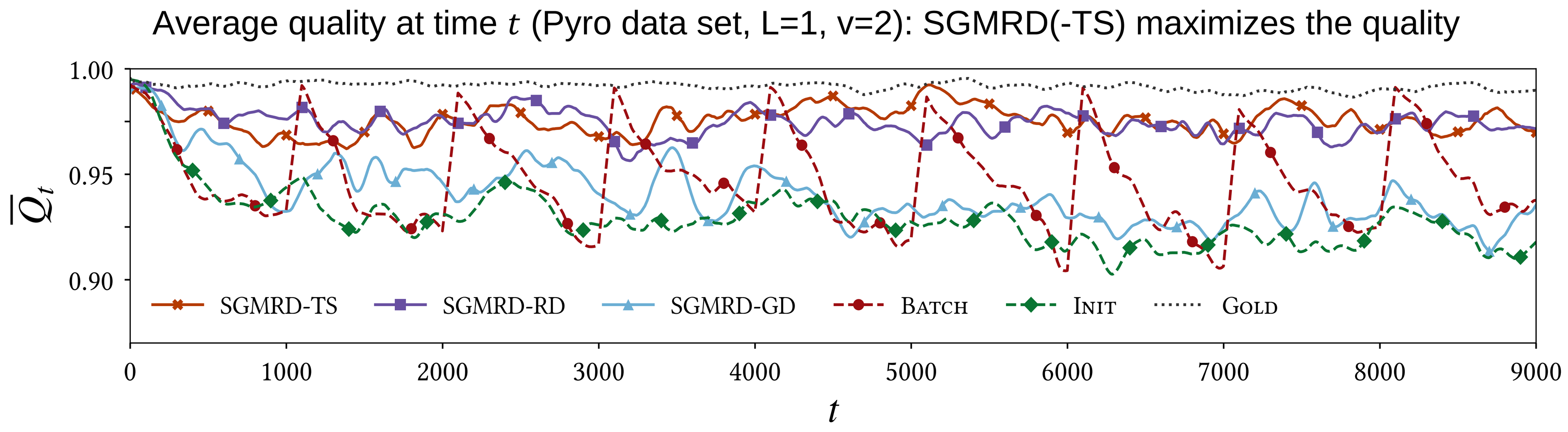
Characteristics of the benchmark data sets

Benchmark	# Instances	# Dimensions	% Outliers
Pyro	10,000	100	NA
KDDCup99	25,000	38	7.12
Activity	22,253	51	10
Backblaze	12,600	44	1
Credit	284,807	29	0.17
Synth10	10,000	10	0.86
Synth20	10,000	20	0.88
Synth50	10,000	50	0.81

Our synthetic benchmark generation



Average quality at time t (Pyro data set, L=1, v=2): SGMRD(-TS) maximizes the quality



Outlier detection: SGMRD outperforms its competitors in terms of ROC AUC, precision, recall

	Approach	AUC	AP	P1%	P2%	P5%	R1%	R2%	R5%
Activity	**SGMRD**	**97.32**	**85.39**	**94.59**	**94.83**	**94.24**	**9.44**	**18.97**	**47.10**
	LOF	93.93	61.80	74.32	64.72	64.03	7.42	12.94	32.00
	StreamHiCS	88.52	47.38	70.72	54.61	51.89	7.06	10.92	25.93
	RS-Stream	95.95	68.23	71.62	72.58	75.00	7.15	14.52	37.48
	xStream	77.71	20.41	3.60	10.14	16.31	0.36	2.02	8.13
KddCup99	**SGMRD**	**69.98**	**10.29**	**0.00**	**0.20**	**0.56**	**0.00**	**0.06**	**0.39**
	LOF	65.07	9.57	0.00	0.00	0.08	0.00	0.00	0.06
	StreamHiCS	57.11	7.89	0.00	0.00	0.08	0.00	0.00	0.06
	RS-Stream	43.21	5.73	0.00	0.00	0.08	0.00	0.00	0.06
	xStream	52.70	8.23	0.00	0.20	0.08	0.00	0.06	0.06
Backblaze	**SGMRD**	**90.91**	13.31	7.14	18.65	**15.40**	7.14	37.30	**76.98**
	LOF	56.92	1.65	2.38	3.57	2.38	2.38	7.14	11.90
	StreamHiCS	79.22	**40.07**	**50.79**	**26.59**	10.95	**50.79**	**53.17**	54.76
	RS-Stream	80.55	7.17	7.14	13.49	9.37	7.14	26.98	46.83
	xStream	76.86	3.69	1.59	3.17	6.19	1.59	6.35	30.95
Credit	**SGMRD**	**95.06**	**15.87**	10.69	**7.47**	**3.22**	55.51	**77.57**	**83.65**
	LOF	91.50	4.67	6.22	5.20	2.69	32.32	53.99	69.96
	StreamHiCS	89.21	3.48	3.59	2.93	2.28	18.63	30.42	59.32
	RS-Stream	85.13	1.63	2.27	2.49	1.86	11.79	25.86	48.29
	xStream	94.62	9.10	**10.83**	6.48	3.18	**56.27**	67.30	82.51

	Approach	AUC	AP	P1%	P2%	P5%	R1%	R2%	R5%
Synth10	**SGMRD**	**92.70**	**59.93**	**50.00**	**26.00**	**12.00**	**58.14**	**60.47**	**69.77**
	LOF	88.77	31.44	33.00	18.50	10.40	38.37	43.02	60.47
	StreamHiCS	88.81	31.16	33.00	19.00	10.40	38.37	44.19	60.47
	RS-Stream	71.23	1.87	0.00	0.00	2.80	0.00	0.00	16.28
	xStream	68.51	2.58	5.00	3.00	4.00	5.81	6.98	23.26
Synth20	**SGMRD**	**85.05**	**41.19**	**36.00**	**19.50**	**9.20**	**40.91**	**44.32**	**52.27**
	LOF	72.55	5.57	8.00	6.00	4.40	9.09	13.64	25.00
	StreamHiCS	71.71	5.37	8.00	6.00	4.00	9.09	13.64	22.73
	RS-Stream	48.39	0.80	0.00	0.00	0.00	0.00	0.00	0.00
	xStream	63.64	1.58	1.00	1.50	2.20	1.14	3.41	12.50
Synth50	**SGMRD**	**75.87**	**31.27**	**27.00**	**16.00**	**7.60**	**33.33**	**39.51**	**46.91**
	LOF	61.38	1.08	0.00	0.50	0.60	0.00	1.23	3.70
	StreamHiCS	63.90	12.00	11.00	6.00	3.40	13.58	14.81	20.99
	RS-Stream	46.52	0.73	0.00	0.00	0.00	0.00	0.00	0.00
	xStream	48.43	0.90	1.00	0.50	1.40	1.23	1.23	8.64

References

[1] Fouché, E., Kalinke, F., and Böhm, K. (2020). Efficient Subspace Search in Data Streams. Information Systems, Elsevier. In press.

[2] Keller, F. et al. (2012). HiCS: High Contrast Subspaces for Density-Based Outlier Ranking. In ICDE '12.

[3] Aggarwal, C. (2013). Outlier Analysis, Springer.

[5] Wang, Y. et al. (2017). Unbiased multivariate correlation analysis. In AAAI'17.

[6] Trittenbach, H. and Böhm, K. (2019). Dimension-based subspace search for outlier detection. International Journal of Data Science and Analytics.

[7] Aggarwal, C. (2009). On high dimensional projected clustering of uncertain data streams. In ICDE'09.

[8] Zhang, J. et al. (2008). SPOT: A system for detecting projected outliers from high-dimensional data streams. In ICDE'08.

[9] Becker, V. (2016). Concept change detection in correlated subspaces in data streams, Master's thesis, KIT.

[10] Manzoor, E. et al. (2018). xStream: Outlier Detection in Feature-Evolving Data Streams. In KDD'18.

[11] Sathe, S. and Aggarwal, C. (2018). Subspace histograms for outlier detection in linear time. Knowledge and Information Systems.

[12] Komiyama, J. et al. (2015) Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In ICML'15.