Detection of Periodical Patterns and Contextual Anomalies in Data Streams

Alexander Hartl[1], Félix Iglesias Vázquez[1,2,*] and Tanja Zseby[1]

[1]Inst. of Telecom., TU Wien, Gusshausstraße 25 / E389, 1040 Vienna, Austria
[2]Le Studium, 1 Rue Dupanloup, 45000 Orléans, France

Abstract

We present SDOoop, a streaming data analysis algorithm that spots contextual anomalies undetectable by traditional methods, while enabling the inspection of data geometries, clusters and temporal patterns. We used SDOoop to model real network communications in critical infrastructures. We also evaluated SDOoop with data from intrusion detection and natural science domains and obtained performances equivalent or superior to state-of-the-art approaches. SDOoop is ideal for big data, being able to instantly process large volumes of information.

Keywords

Contextual Anomalies, Streaming Data Analysis

1. Introduction

A contextual (aka. conditional or out-of-phase) anomaly "occurs if a point deviates in its local context" [1], i.e., if it happens outside its usual time. Consider a method whose observation horizon spans a one-week period. If a cluster occurs exclusively during weekends, but a data point of this cluster accidentally appears on Wednesday, this method *will not* identify it as an anomaly, but as a normal inlier instead. Most traditional approaches are blind to identify contextual anomalies, which have been tackled mainly in time series analysis [2], but here experts also emphasize the low attention given to them despite its relevance for cybersecurity, healthcare and fraud detection [3].

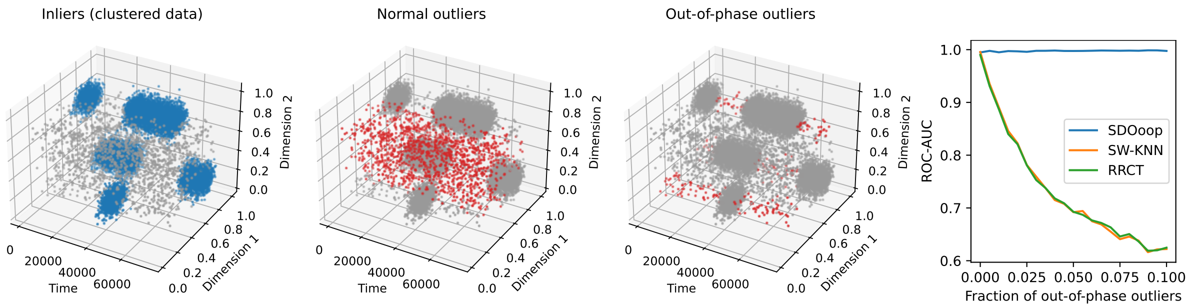
SDOoop (SDO out-of-phase) is an algorithm for streaming anomaly/outlier detection (SAD) whose models store temporal information. Based on SDO [4] and SDOstream [5], SDOoop builds models by sampling a fixed number of data points at representative locations in feature space, called *observers*. It uses an exponentially weighted moving average (EWMA) to estimate model information from the arriving data mass. In parallel, observers hold temporal information as coefficients of Fourier transforms (FT). Thus, for a specific time of interest t, observers "twinkle" to show only the most representative model for time t.

2. Methodology

We conducted exhaustive testing of SDOoop (described in [6] and https://github.com/CN-TU/tpsdos-experiments), including: (a) a proof of concept (PoC) of the contextual outlier detection, (b) anomaly detection comparisons with established algorithms on public datasets, and (c) evaluations of SDOoop ability to discover and model temporal patterns in real communications from critical infrastructures (smart metering) and the darkspace [7].

In Fig. 1 we can see the distinctive ability of SDOoop to detect contextual outliers. Table 1 compares accuracy (AAP and ROC-AUC [8]) of consolidated SAD algorithms for the SWAN-SF [9] and KDD Cup'99 [10] datasets, related to solar flares and network security respectively. SDOoop performances

Discovery Science - Late Breaking Contributions 2024
*Corresponding author.
alexander.hartl@tuwien.ac.at (A. Hartl); felix.iglesias@tuwien.ac.at (F. Iglesias Vázquez); tanja.zseby@tuwien.ac.at (T. Zseby)
0000-0003-4376-9605 (A. Hartl); 0000-0001-6081-969X (F. Iglesias Vázquez); 0000-0002-5391-467X (T. Zseby)



(a) Dataset example with 0.5% of contextual outliers (b) Performance

Figure 1: PoC. SDOoop keeps high accuracy regardless of the contextual oulier rate.

are excellent in both cases. While the anomalies defined in the SWAN-SF dataset are not contextual, some of the U2R (User to Root) attacks in the KDD Cup'99 dataset are, hence the notable advantage of SDOoop. Table 2 shows a qualitative comparison of main SDA methods, SW-*k*NN and SW-LOF being the streaming (i.e., sliding window) versions of the popular *k*NN [11] and LOF [12] algorithms[1].

Table 1: SAD accuracy.

	SWAN-SF		KDDCup99	
	AAP	AUC	AAP	AUC
SW-*k*NN	0.69	**0.91**	0.07	0.72
SW-LOF	0.15	0.58	-0.00	0.67
LODA [15]	0.72	**0.91**	0.10	0.92
RS-Hash [16]	**0.73**	**0.91**	0.13	0.95
RRCT [17]	0.23	0.69	0.07	0.85
SDOoop	**0.73**	**0.91**	**0.33**	**0.97**

Table 2: Qualitative comparison.

	SW-*k*NN	SW-LOF	LODA [15]	xStream [18]	RS-Hash [16]	RRCT [17]	SDOst [5]	SDOoop
Fixed time complex.	~	×	✓	✓	✓	~	✓	✓
Fixed space complex.	×	×	×	✓	✓	×	✓	✓
Interpretability	✓	~	×	×	×	×	✓	✓
Detect temp. patterns	×	×	×	×	×	×	×	✓
Detect context. anom.	×	×	×	×	×	×	×	✓

In tests with real communications, SDOoop discovered and modeled main temporal patterns of traffic from critical infrastructures, corresponding to: ICMP pings (device checking), DNS lookups (name resolution for meter reading transmissions), DNS caching, and heartbeat messages. As for the darkspace, SDOoop captured anomalies through their diurnal and semi-diurnal periodicities, identified in previous research [19] with Conficker.C worms, BitTorrent misconfigurations, horizontal scan, vertical scan and UDP probing activities.

3. Conclusions

SDOoop conforms to next-generation machine learning, which, besides accuracy and speed, must provide interpretable and informative models.

References

[1] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, K.-R. Müller, A unifying review of deep and shallow anomaly detection, Proceedings of the IEEE 109 (2021) 756–795.

[2] K. Shaukat, T. M. Alam, S. Luo, S. Shabbir, I. A. Hameed, J. Li, S. K. Abbas, U. Javed, A review of time-series anomaly detection techniques: A step to future perspectives, in: K. Arai (Ed.), Adv. in Inf. & Com., Springer, 2021, pp. 865–877.

[3] K. Golmohammadi, O. R. Zaiane, Time series contextual anomaly detection for detecting market

[1]Algorithm implementations used in the evaluation are from the dSalmon Python package [13], while synthetic data have been generated with MDCgen [14].

manipulation in stock market, in: IEEE Int. Conf. on Data Sci. and Adv. Analytics (DSAA), 2015. doi:10.1109/DSAA.2015.7344856.

[4] F. Iglesias, T. Zseby, A. Zimek, Outlier detection based on low density models, in: ICDMW, 2018, pp. 970–979. doi:10.1109/ICDMW.2018.00140.

[5] A. Hartl, F. Iglesias, T. Zseby, SDOstream: Low-density models for streaming outlier detection, in: ESANN 2020 proceedings, 2020, pp. 661–666.

[6] A. Hartl, F. Iglesias, T. Zseby, SDOoop: Capturing periodical patterns and out-of-phase anomalies in streaming data analysis (2024). arXiv:2409.02973, arXiv, eprint: 2409.02973, https://arxiv.org/abs/2409.02973.

[7] CAIDA, The UCSD network telescope "patch tuesday" dataset, http://www.caida.org/data/passive/telescope-patch-tuesday_dataset.xml, ???? Acc.: 2021-03-09.

[8] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, M. E. Houle, On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study, DAMI 30 (2016) 891–927. doi:10.1007/s10618-015-0444-8.

[9] R. A. Angryk, P. C. Martens, B. Aydin, D. Kempton, S. S. Mahajan, S. Basodi, A. Ahmadzadeh, X. Cai, S. Filali Boubrahimi, S. M. Hamdi, M. A. Schuh, M. K. Georgoulis, Multivariate time series dataset for space weather data analytics, Scientific Data 7 (2020).

[10] KDD Cup 1999 data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, ????

[11] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, SIGMOD Rec. 29 (2000) 427–438.

[12] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, Lof: Identifying density-based local outliers, SIGMOD Rec. 29 (2000) 93–104.

[13] A. Hartl, F. Iglesias, T. Zseby, dSalmon: High-speed anomaly detection for evolving multivariate data streams, in: Performance Evaluation Methodologies & Tools, Springer, 2024, pp. 153–169. doi:10.1007/978-3-031-48885-6_10.

[14] F. Iglesias, T. Zseby, D. Ferreira, A. Zimek, Mdcgen: Multidimensional dataset generator for clustering, Journal of Classification 36 (2019) 599–618. doi:10.1007/s00357-019-9312-3.

[15] T. Pevnỳ, Loda: Lightweight on-line detector of anomalies, Machine Learning 102 (2016) 275–304. doi:10.1007/s10994-015-5521-0.

[16] S. Sathe, C. C. Aggarwal, Subspace outlier detection in linear time with randomized hashing, in: IEEE 16th ICDM, IEEE, 2016, pp. 459–468.

[17] S. Guha, N. Mishra, G. Roy, O. Schrijvers, Robust random cut forest based anomaly detection on streams, in: Int. Conf. on Mach. Learn., PMLR, 2016, pp. 2712–2721.

[18] E. Manzoor, H. Lamba, L. Akoglu, xStream: Outlier detection in feature-evolving data streams, in: 24th ACM SIGKDD, 2018, p. 1963–1972.

[19] F. Iglesias, T. Zseby, Pattern discovery in internet background radiation, IEEE Trans. on Big Data 5 (2017) 467–480. doi:10.1109/TBDATA.2017.2723893.