A survey on unsupervised learning algorithms for detecting abnormal points in streaming data

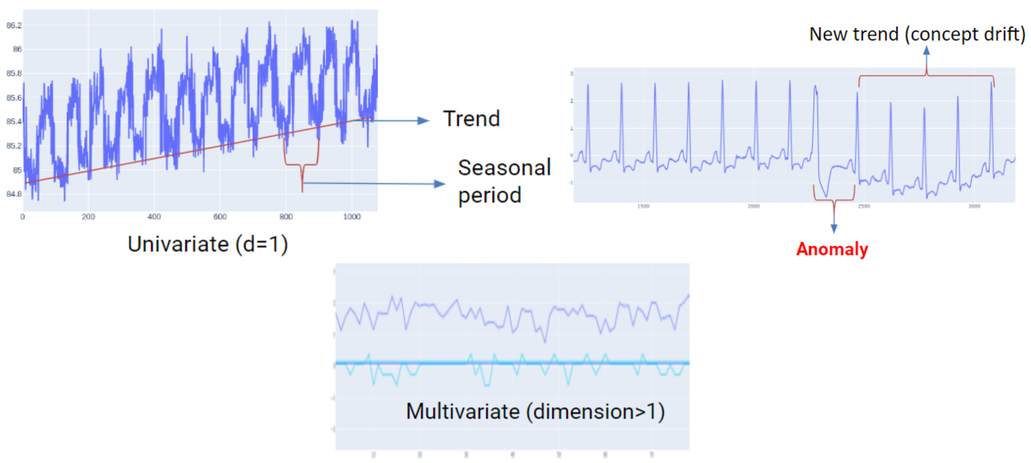
Sophie NGO BIBINBE, Michael MBOUOPDA, Raïssa MBIADOU SALEU, Engelbert MEPHU NGUIFO

Université Clermont Auvergne, Clermont Auvergne INP, CNRS, Mines Saint-Etienne, LIMOS, 63000 Clermont-Ferrand, France
{anne.ngo_bibinbe, michael.mbouopda, gertrude_raissa.mbiadou_saleu, engelbert.mephu_nguifo}@uca.fr

Abstract

In this work, we compared unsupervised data stream abnormal point detection methods on various datasets with emphasis on their performance and runtime, as well as the presence of concept drift, seasonality, trend and cycle as a characteristic of the dataset.

Background

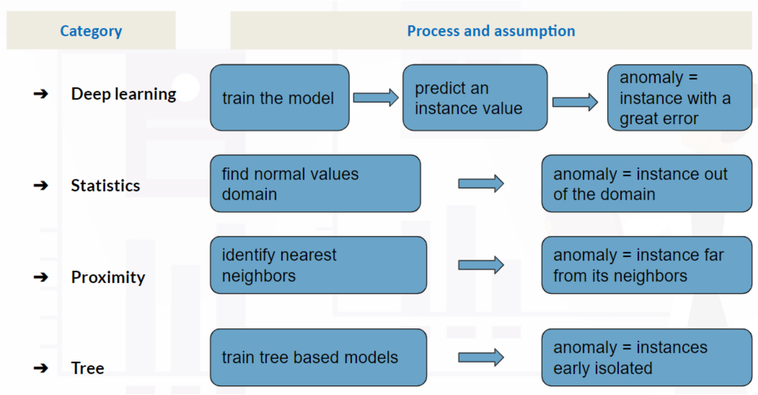


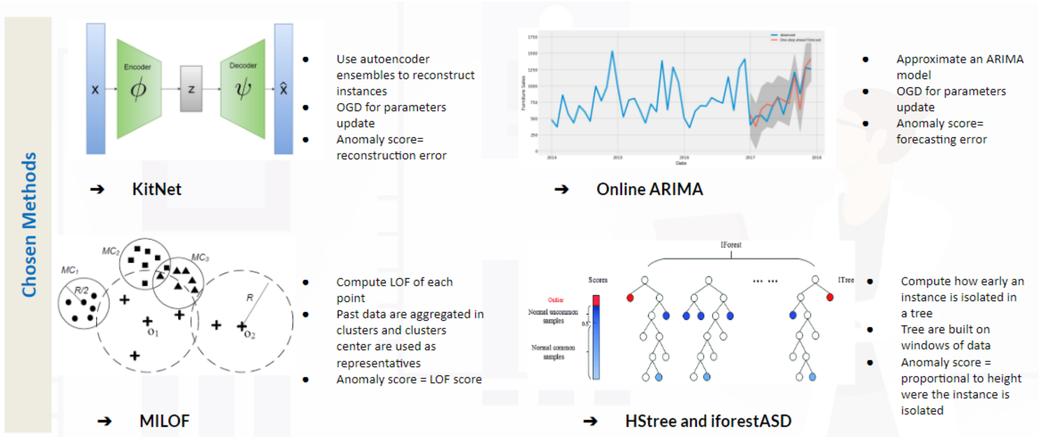
- To tackle abnormal point in data stream, several methods based on different assumptions have been proposed in the literature (ILOF, DILOF, MILOF, KitNet, HTM, IforestASD, KNNCAD, Expose, Twitter ADVec, etc.). However, there is still a lack of experimental comparisons of those methods, which makes it difficult to choose a specific one.
- How to choose a method depending on the context and the type of data? The only existing benchmark compares only statistical methods and a deep learning method on univariate data without linking the study to the data characteristics.

Objective

To guide in the choice of a method or a type of method according to the arrival speed of the stream and the characteristics of the stream.

Methods





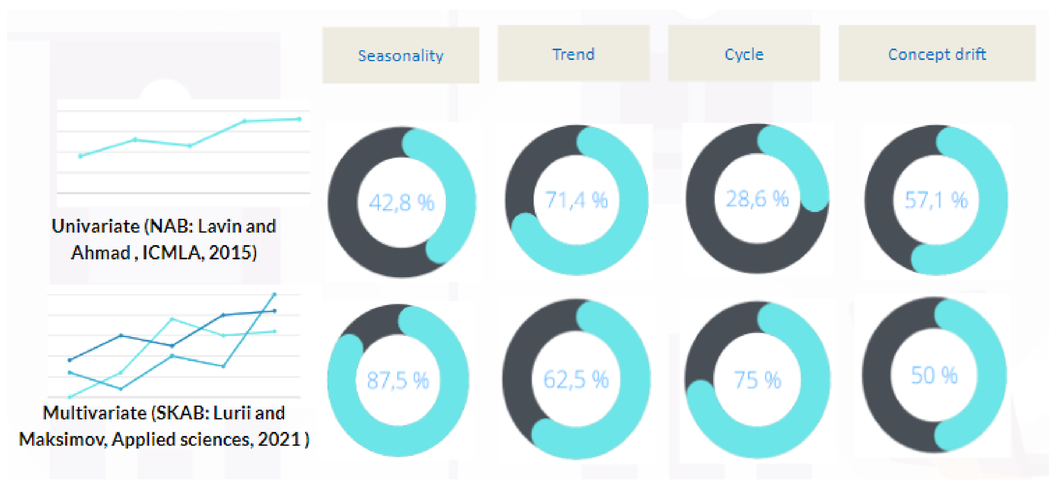
These methods have been chosen for their efficiency reported in the state of the art.

Experimental protocol

- **Evaluation metric**: a method finds an anomaly (TP) if it detects one within 1% (of the series length) of the defined anomaly position [1].
  $F1\_score = \frac{2 \times recall \times precision}{recall + precision}; recall = \frac{TP}{TP+FN}; precision = \frac{TP}{TP+FP}$. Multiple anomalies identified in close proximity to the same true anomaly are considered a single TP.
- Best hyperparameters chosen by Bayesian optimization.

Datasets

Data stream can be characterized by the presence of seasonality, trends, cycles, and conceptual drifts. We identified these 4 features on each of the datasets taken from the SKAB [2] (multivariate) and NAB [3] (univariate) for our study.



Details of the identified characteristics are available on [4]

Results

The analysis was made with an emphasis on the latency time and the performance of the methods according to the characteristics of the datasets. For more details see [4].

On 7 datasets on which HStree got the best score, 4 of them had seasonality

Methods	best score	Concept drift	Seasonality	Trend	Cycle	Univariate average score	Multivariate average score
MILOF	1/14	0/1	0/1	1/1	1/1	0.33	0.22
HS-tree	7/14	5/7	4/7	5/7	2/7	0.47	0.504
iForestASD	3/14	2/3	1/3	2/3	0/3	0.54	0.39
Online ARIMA	3/7	2/3	3/3	3/3	2/7	0.56	-
KitNet	2/7	0/2	2/2	0/2	0/2	-	0.503

Summary of observations (scores, characteristics)

Methods:	MILOF	HS-tree	iForestASD	Online ARIMA	KitNet
Latency(ms) Univarite	22.5	222.5	27.8	11.06	-
Latency(ms) Multivariate	9.5	80.7	31.9	-	0.3

Summary of methods Latency

- Online ARIMA: good performance in the presence of seasonality and trends but not in the presence of frequent conceptual drifts.
  - best score on 42.8% of the tested datasets.
  - average F1-score on datasets with frequent conceptual drifts: **11%**.
  - average F1-score on the other datasets: **74.6%**.
- HStree and IforestASD: good overall performance for all characteristics with stable scores.
- Online gradient descent methods: Online ARIMA and KitNet
  - shorter latency time.
  - slow re-learning in the presence of conceptual drifts.

References

[1] T. Nakamura, M. Imamura, R. Mercer, and E. Keogh, "Merlin: Parameter-free discovery of arbitrary length anomalies in time series archives," in *IEEE ICDM*, 2020, pp. 1190–1195.

[2] V. L. Iurii Katser Viacheslav Kozitsin and I. Maksimov, "Unsupervised offline changepoint detection ensembles," *Applied sciences*, vol. 11, p. 4280, May 2021.

[3] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms – the numenta anomaly benchmark," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 38–44. DOI: 10.1109/ICMLA.2015.141.

[4] A. M. S. Ngo Bibinbe, M. F. Mbouopda, G. R. Mbiadou Saleu, and E. Mephu Nguifo, url https://github.com/nams2000/anomaly-detection-in-data-stream, 2021.

Acknowledgements

This work was partially supported by the IMobS3 "Laboratoire d'Excellence" (LabEx for "Innovative Mobility : Smart and Sustainable Solutions"), and the project IT2.fr (https://cordis.europa.eu/project/id/875999/fr) under the DASMA subproject (dasma.limos.fr) funded by Bpifrance.