Universität Potsdam

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Course: Large-Scale Time Series Analytics

Seminar Introduction

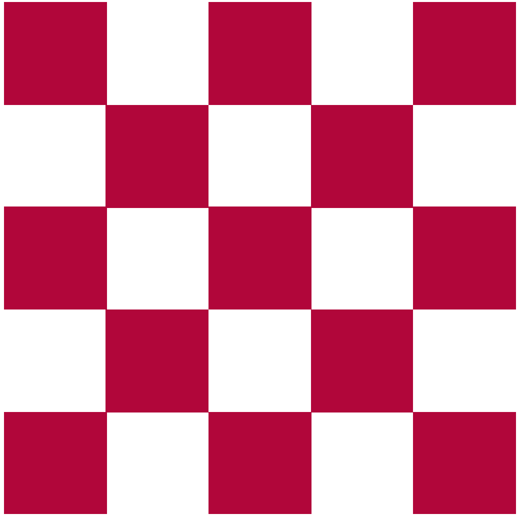
Sebastian Schmidl
Phillip Wenig

Covid Regulations

Please, register to this event with the Corona Warn-App!

Large-Scale Time Series Analytics
F-E.06, Campus II, HPI

Please, always wear your mask and sit in a checkboard fashion.

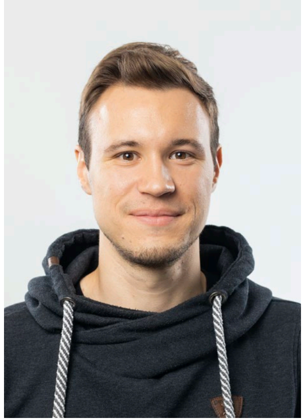


When you sit, you can take the mask off.

About Us

Sebastian Schmidl

PhD Student
Information
Systems Group, HPI



Phillip Wenig

PhD Student
Information
Systems Group, HPI

Prof. Dr. Thorsten Papenbrock

Department of Mathematics &
Computer Science,
Philipps-University of Marburg

Prof. Dr. Felix Naumann

Information Systems Group, HPI

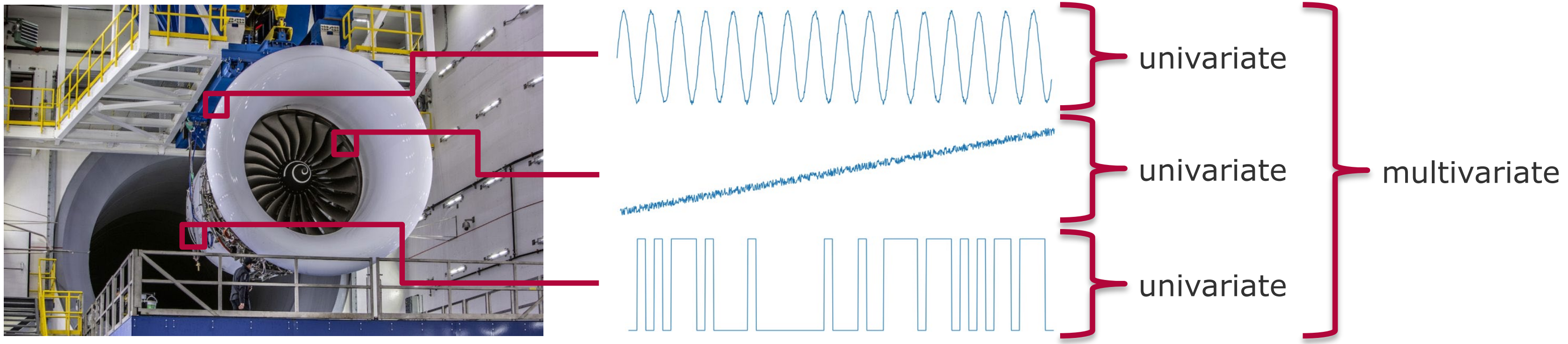
Description

In this project seminar, we investigate and improve **anomaly detection** algorithms for **multivariate time series**. You will receive a broad selection of state-of-the-art anomaly detection algorithms (with code and papers), various real-world and synthetic datasets, and information about the evaluation of time series anomaly detection (TSAD) approaches. Your are then challenged to **beat these approaches in runtime and/or quality**. Techniques that we consider for this task involve, i.a.,

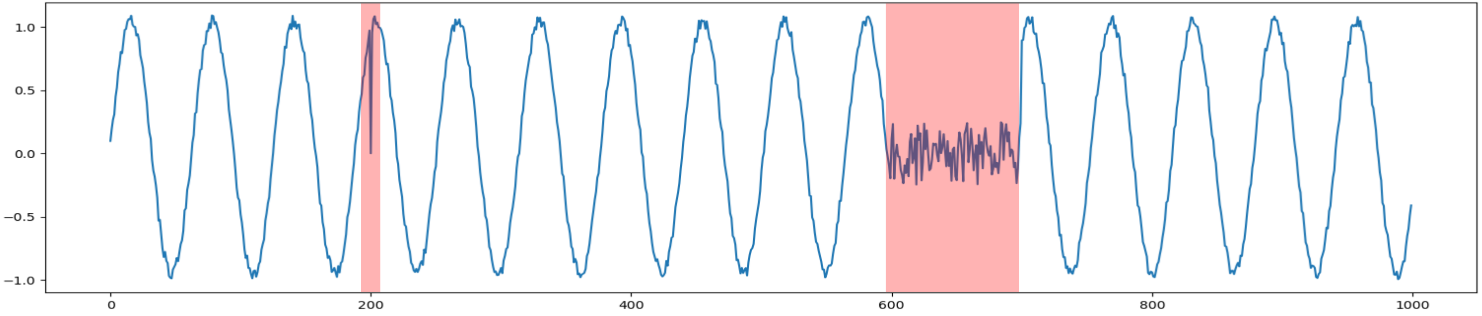
- workload parallelization and distribution,
- streaming,
- ensambling,
- machine learning, and
- hybridization.

Background

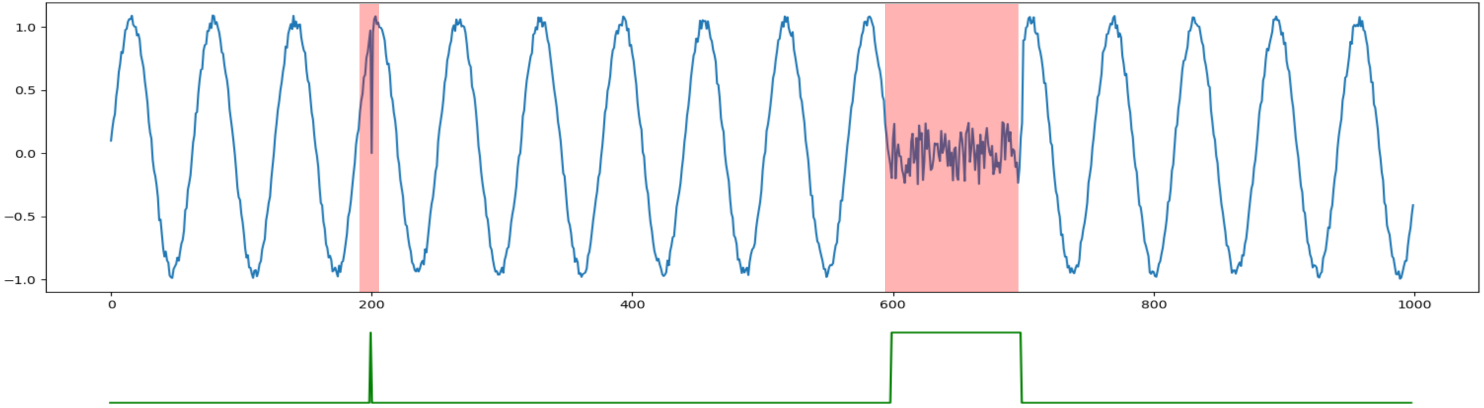
Time Series

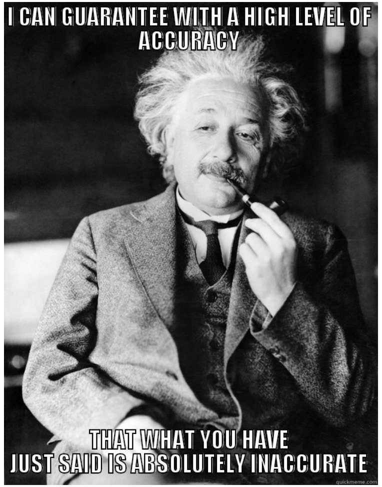


Anomalies



Background





Measuring Anomaly Detection Quality

Algorithm	Accuracy	AUROC	Hit-Precision	Hit-Recall
No anomaly found	89.9 %	50.0 %	0.0 %	0.0 %
Only point found	90.0 %	50.5 %	100.0 %	50.0 %
Only sequence found	99.9 %	99.5 %	100.0 %	50.0 %
All anomalies found	100.0 %	100.0 %	100.0 %	100.0 %

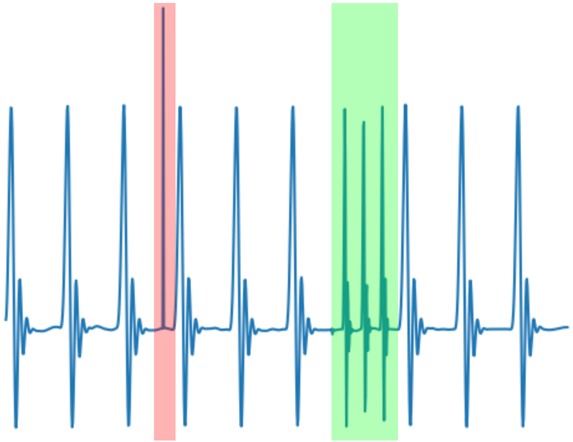
Why is this hard?

Searching for unknowns

?

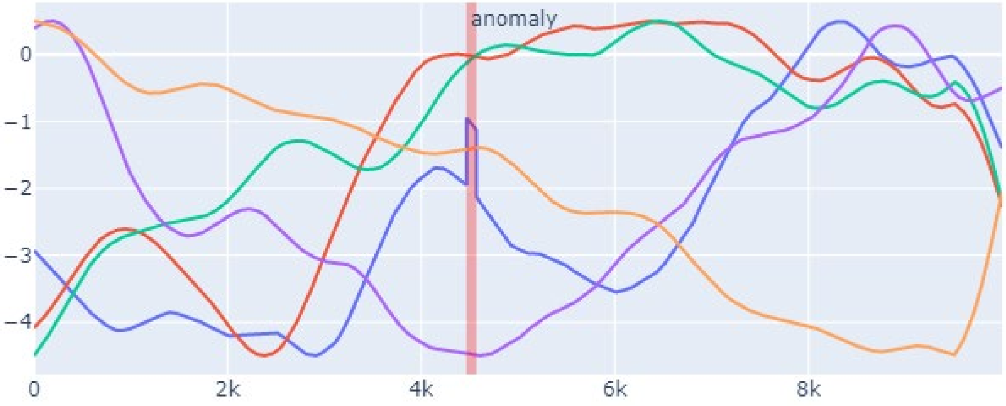
How does an anomaly look like?

Different Semantics of Anomalies



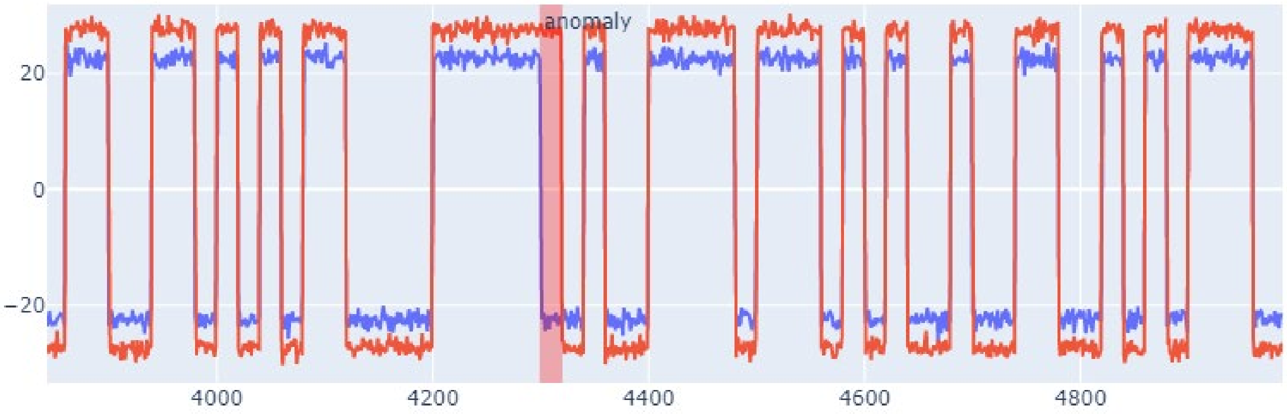
Motivation

- Multivariate time series anomaly detection (TSAD) challenges
  - **Localization**
    Anomalies can appear in only a single channel, in multiple channels, and in all channels at the same time.
  - **Correlation**
  - **Dimensionality**
  - **Complexity**



Motivation

- Multivariate time series anomaly detection (TSAD) challenges
  - **Localization**
  - **Correlation**
    Anomalies can appear as correlation anomalies, in which all individual channels behave normally but some subset of channels is out-of-sync.



Motivation

- Multivariate time series anomaly detection (TSAD) challenges
  - **Localization**
  - **Correlation**
  - **Dimensionality**
    Due to the curse of dimensionality[1], anomalies become very hard to detect on multivarite datasets, even for short datasets.
  - **Complexity**

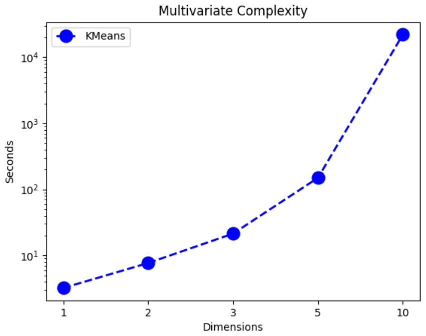
[2]

- Irrelevant attributes
- Interpretability of scores
- Exponential search space
- ML: increased number of training samples required
- Distances: difference between sample pairs gets very small
- kNN: emergence of hubs (=samples that appear more frequently in neighbor lists than others)

[1]: doi:10.1007/978-0-387-39940-9_133
[2]: doi:10.1002/sam.11161

Motivation

- Multivariate time series anomaly detection (TSAD) challenges
  - **Localization**
  - **Correlation**
  - **Dimensionality**
  - **Complexity**
    Mutlivariate time series are not only long (high number of data points), but also wide (high number of channels / dimensions), which in many cases leads to huge amounts of data that need to be processed within certain time and memory limits.



What We Provide

Overview

Algorithms

Datasets

Dataset generators

- GutenTAG
- CoMuT

Algorithm & dataset evaluation

- Evaluation Framework: TimeEval
- First results of our large evaluation

Icons made by Freepik from www.flaticon.com.

What We Provide

Algorithm overview

Stochastic Learning

I-HMM, SmartSifter, MultiHMM, LaserDBN, EM-HMM, GLA, EDBN, HSMM, CxDBN, HMAD, FuzzyDNBC, U-GMM-HMM

Data Mining

STOMP, Series2Graph, NorM, STAMP, TwoFinger, HOT SAX, DissimilarityAlgo, MERLIN, Left STAMPi, SCRIMP++, BoehmerGraph, SSA, DAD, VALMOD, ILOF, GrammarViz, PST, TSBitmap, TARZAN, DADS, SurpriseEncoding, KnorrSeq2, Ensemble GI, NormA

Classic Machine Learning

Random Black Forest, NetworkSVM, MS-SVDD, k-Means, PCA, sequenceMiner, NoveltySVR, SLADE-MTS, PCC, HBOS, Hybrid K-means, PhaseSpace-SVM, RobustPCA, KNN, SLADE-TS, S-SVM, EIF, XGBoosting, RUSBoost, Random Forest Regressor, Eros-SVMs, OC-KFD, AOSVM

Deep Learning

DeepAnT, Normalizing Flows, DeepNAP, SR-CNN, Hybrid KNN, STORN, CoalESN, Deep OCSVM, Telemanom, Robust Deep AutoEncoder, LAMP, MultiHTM, TCN-AE, EncDec-AD, Torsk, AD-LTI, VELC, AE, Donut, TAnoGAN, LSTM-AD, LSTM, BAGEL, HealthESN, NumentaHTM, DAE, Deep K-Means, Sparse AutoEncoder, RADM, MSCRED, StrDAE, MoteESN, MAD-GAN, MTAD-GAT, LSTM-based VAE-GAN, DeepPCA, OmniAnomaly, Image-embedding-CAE, DeepLSTM, OceanWNN

Statistics

MCD, FAST-MCD, Poly, Simple ES (EWMA), S-H-ESD (Twitter), ARIMA, MGDD, pEWMA, ARMA, EWMA-STR, Double ES (Holt's), Triple ES (Holt-Winter's), AR, MedianMethod, SARIMA, MA, Box Plot, CHEB, SH-ESD+, ANODE, Subsequence Fast-MCD, Yesterday, DSE, Kalman Filter, RePAD, AMD Segmentation, DSPOT, ConInd, PCI

Outlier Detection

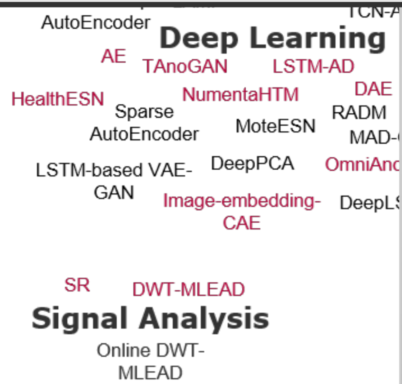
SALOF, Subsequence IF, CBLOF, COPOD, LOCI/aLOCI, MCOD, BLOF, LOF, DBStream, GeckoFSM, DILOF, Subsequence LOF, GridLOF, COF, TOLF, Isolation Forest, IF-LOF, Hybrid Isolation Forest, K-LOF, HSDE

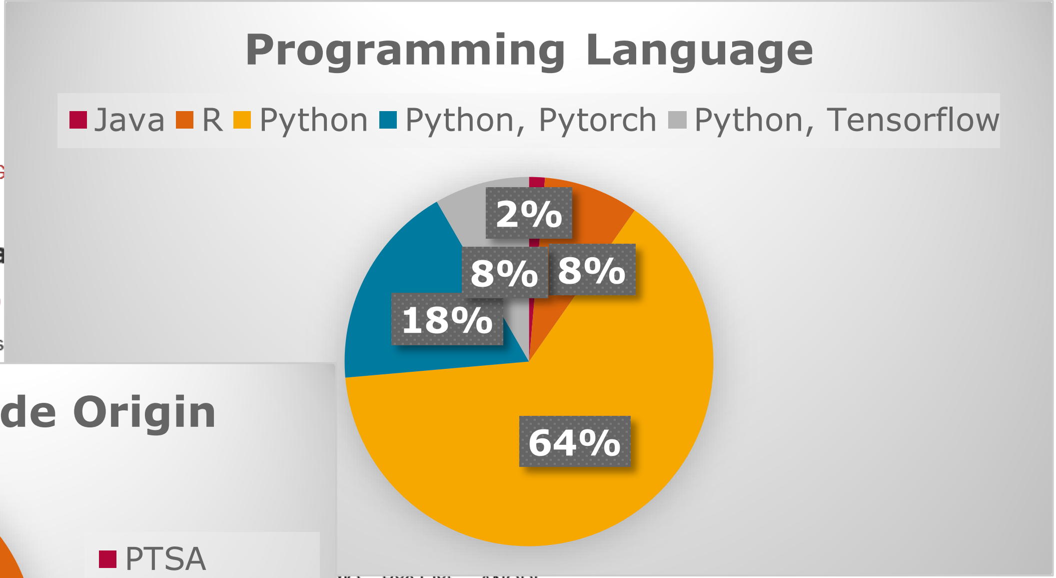
Signal Analysis

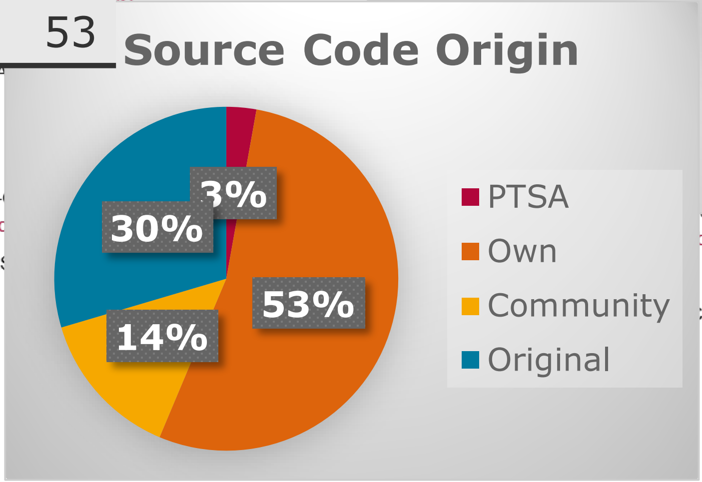
SR, DWT-MLEAD, Online DWT-MLEAD

What We Provide

Metric	Value
Unsupervised	65
Semi-supervised	53
Supervised	7
Univariate	71
Multivariate	53



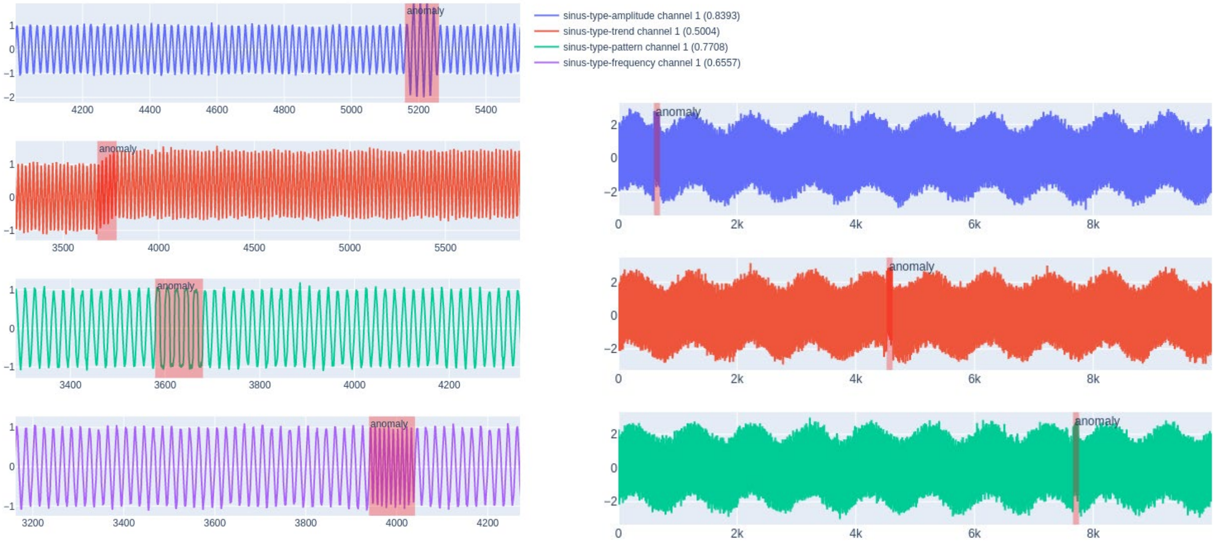


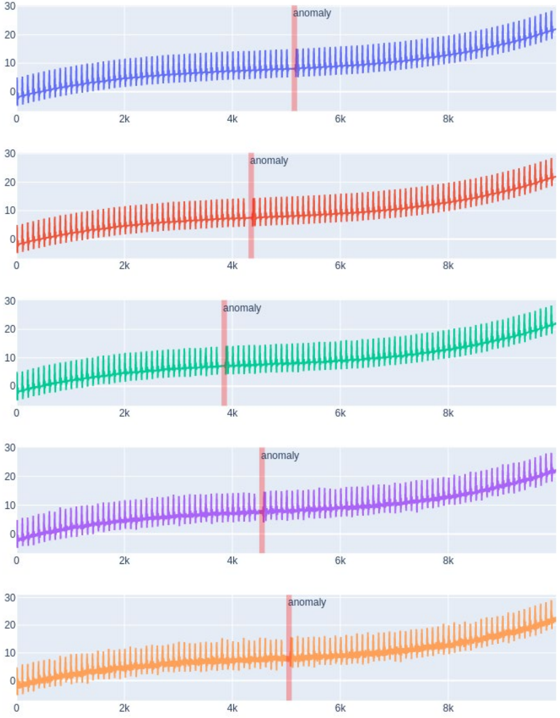


What We Provide

Datasets

- 580 test case datasets (generated with GutenTAG)
- 1143 benchmark datasets (~30% multivariate)





What We Provide

GutenTAG

Synthetic test case datasets

- Variations in base curve, noise, trend, dimensions
- Variations in anomaly position, number of similar anomalies, different anomalies
- Different **anomaly types**: local & global extremums, frequency shifts, amplitude change, jumps/platforms, mode/pattern/state change regions, delayed or premature patterns, variance change, noise change

GutenTAG

What We Provide

CoMuT

Synthetic multivariate test case datasets

- Different number of dimensions
- Always the same step function with random noise
- Anomalies are steps where one or more channels don't follow their switching behavior

C CoMuT
Project ID: 4197

Star 0 Fork 0

7 Commits 1 Branch 0 Tags 297 KB Files 297 KB Storage

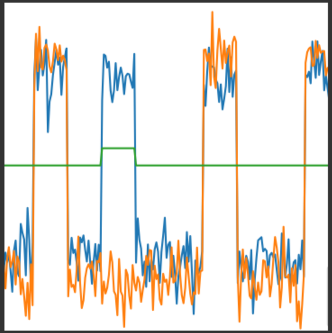
Correlated Multiple Time-Series Generator

main comut / + History Find file Web IDE Clone

README.md

Correlated Multivariate Time Series Generator (CoMuT)

This script can generate multivariate time series that contain anomalies only detectable when analyzing multiple dimensions at once. The time series' values fluctuate in a normal distribution around a given value (`value-offset`). After each `step` (a user-defined number of time points), the time series changes its algebraic sign based on a random boolean. The related dimensions can have different values but follow the same switching of algebraic signs. Hence, the dimensions have always the same or always the opposite algebraic sign. When an anomaly is inserted, this algebraic sign is changed to its opposite for a whole step.

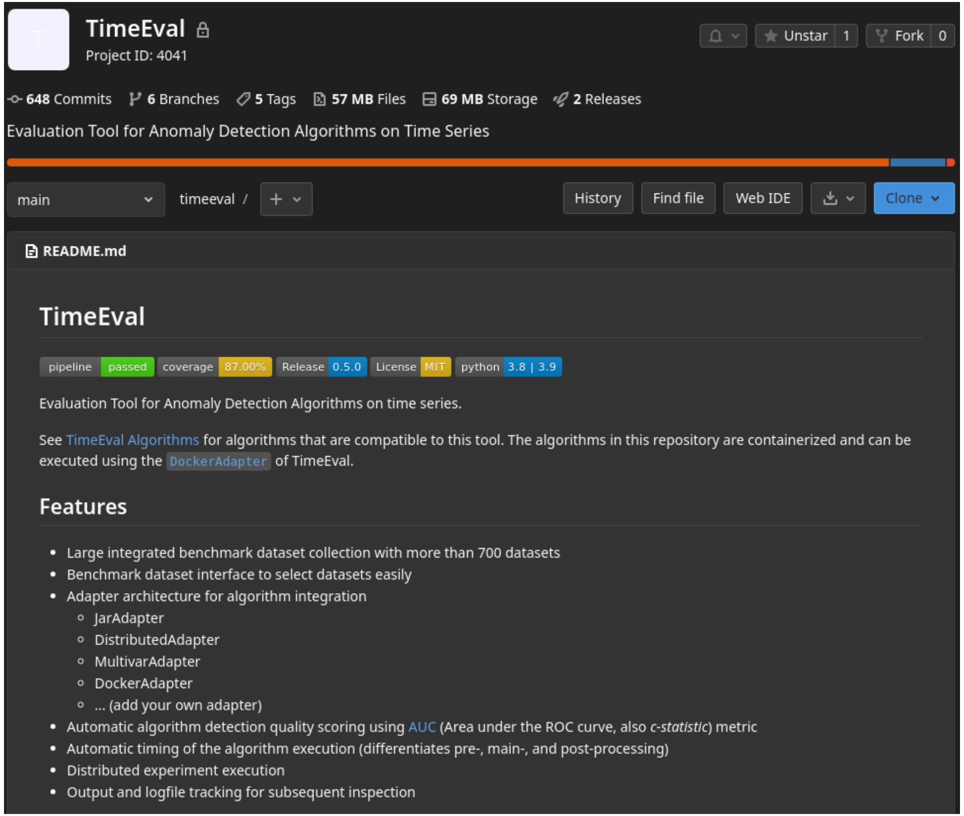


In the plot above, you see the dimensions plotted in **blue** and **orange**. The green line indicates if there is an anomaly. The anomaly in the blue dimension is `step-length` long and does not correlate with the orange algebraic sign anymore (which is negative in this case).

What We Provide

TimeEval

- Evaluation tool
- Canonical algorithm interface and dataset format
- Parallelized & distributed execution of experiments
- Automatic result collection and quality and runtime assessment


TimeEval
Project ID: 4041

Unstar 1 Fork 0

648 Commits 6 Branches 5 Tags 57 MB Files 69 MB Storage 2 Releases

Evaluation Tool for Anomaly Detection Algorithms on Time Series

main timeeval / + History Find file Web IDE Clone

README.md

TimeEval

pipeline passed coverage 87.00% Release 0.5.0 License MIT python 3.8 | 3.9

Evaluation Tool for Anomaly Detection Algorithms on time series.

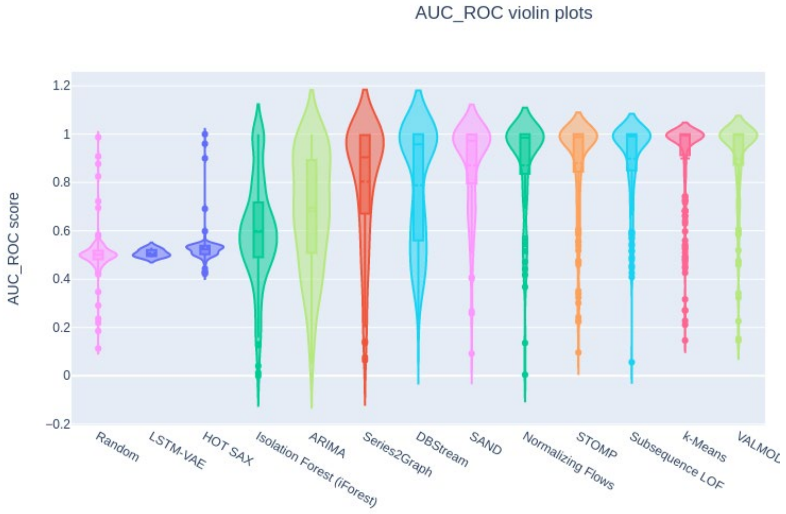
See TimeEval Algorithms for algorithms that are compatible to this tool. The algorithms in this repository are containerized and can be executed using the `DockerAdapter` of TimeEval.

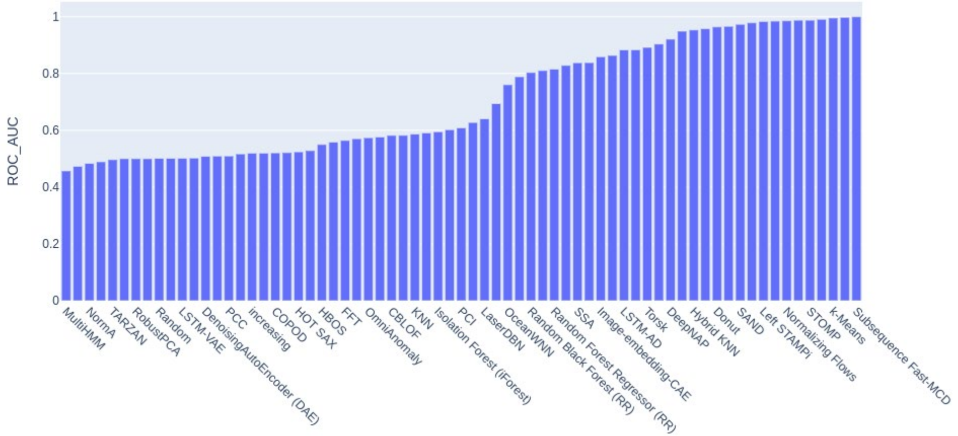
Features

- Large integrated benchmark dataset collection with more than 700 datasets
- Benchmark dataset interface to select datasets easily
- Adapter architecture for algorithm integration
  - JarAdapter
  - DistributedAdapter
  - MultivarAdapter
  - DockerAdapter
  - ... (add your own adapter)
- Automatic algorithm detection quality scoring using AUC (Area under the ROC curve, also *c-statistic*) metric
- Automatic timing of the algorithm execution (differentiates pre-, main-, and post-processing)
- Distributed experiment execution
- Output and logfile tracking for subsequent inspection

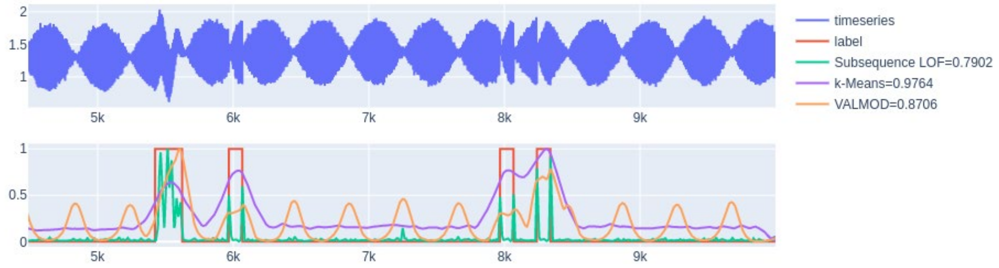
What We Provide

Evaluation results (preview)





Results of Subsequence LOF,k-Means,VALMOD on sinus-combined-diff-2



Goals

- Each team develops an improved multivariate TSAD algorithm
- Beats state-of-the-art for a specific use case / scenario
  - **More reliable**
    The developed algorithm is more robust against uncommon data formats and values, missing data points, etc. It can produce results, where other algorithms give up.
  - **More accurate**
  - **More efficient**
  - **More capable**

Goals

- Each team develops an improved multivariate TSAD algorithm
- Beats state-of-the-art for a specific use case / scenario
  - **More reliable**
  - **More accurate**
    The developed algorithm can produce qualitatively better results according to quality metrics, such as area under the ROC-curve (ROC-AUC), PR-AUC, RANGE-PR-AUC, or average precision (AP).
  - **More efficient**
  - **More capable**

Goals

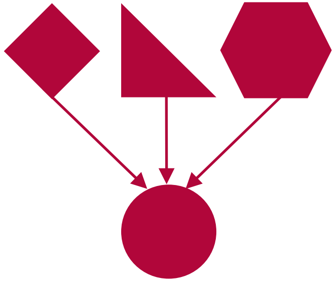
- Each team develops an improved multivariate TSAD algorithm
- Beats state-of-the-art for a specific use case / scenario
  - **More reliable**
  - **More accurate**
  - **More efficient**
    The developed algorithm can process larger datasets in shorter time and/or with lower memory requirements than the existing approaches while not (significantly) falling behind on result quality.
  - **More capable**

Goals

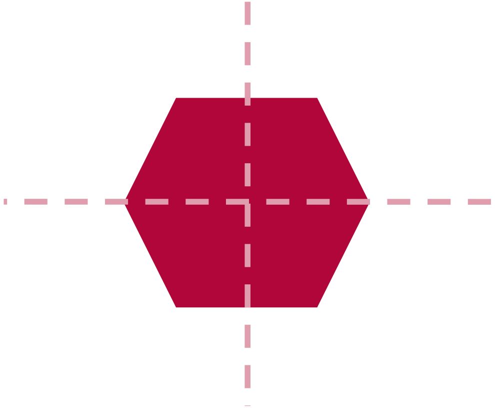
- Each team develops an improved multivariate TSAD algorithm
- Beats state-of-the-art for a specific use case / scenario
  - **More reliable**
  - **More accurate**
  - **More efficient**
  - **More capable**
    The developed algorithm can detect anomalies in certain datasets or of certain types that no existing algorithms can detect.

Ideas and Starting Points

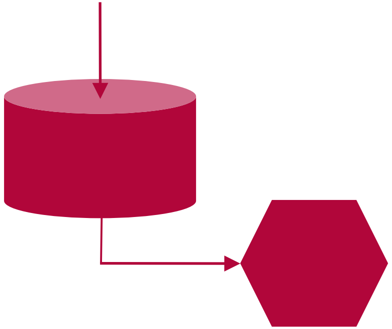
Ensembling



Distribution



Novel preprocessing



Organizational

Metadata

- Project seminar for master students
- Extent: 6 credit points, 4 SWS
- Location: F-E.06, Campus II, HPI
- Dates: Wednesdays, 17:00 – 18:30
- Class: At most 8 participants (4 teams á 2 students)
- Supervisors: Sebastian Schmidl, Phillip Wenig, Thorsten Papenbrock (remote), and Felix Naumann
- Website: https://hpi.de/naumann/teaching/current-courses/ws-21-22/large-scale-time-series-analytics.html

We probably meet in our seminar room during the semester

We can vote for a different time

Team Meetings

- Regular meetings with supervisors in the teams of 2 (bi-weekly?)
- On-demand meetings

Organizational

Registration

- until **29.10.2021 12:00**
- Send e-mail to sebastian.schmidl(at)hpi.de
  - Subject: "Registration to Large-Scale Time Series Analytics seminar"
  - Content:
    - Prior knowledge, courses taken
    - (optional) which person to team up with (both have to write an e-mail)

- Notification about participation on **Friday, 29.10.2021, afternoon**!

Important Dates

Date	Topic
27.10.2021 (F-E.06)	Seminar introduction
29.10.2021 12:00	Deadline registration
29.10.2021 (afternoon)	Acceptance notification
03.11.2021	Kick-off & introduction to state-of-the-art
10.11.2021	Topic selection & team building
Week of 10.01.2022	Midterm presentation
March 2022 (based on students' voting)	Final presentation
March 2022 (based on students' voting)	Artifacts & report submission

Grading

- Oral assessment
  - (10%) Active participation during all seminar events.
  - (30%) Presentations including:
    - (15%) Midterm presentation
    - (15%) Final presentation
- Demonstration of a developed software program
  - (20%) Implementation & Documentation
  - (20%) Evaluation
  - (20%) Technical report writing
    - ~6 pages per team / ~3 pages per person
    - 2-column ACM template

Starting Literature

- Reviews / Surveys
  - Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. **Anomaly detection: A Survey**. *ACM Computing Surveys 41, 3, Article 15* (July 2009), 58 pages. DOI:https://doi.org/10.1145/1541880.1541882
- Series2Graph
  - Paul Boniol and Themis Palpanas. 2020. **Series2Graph: Graph-based subsequence anomaly detection for time series**. *Proc. VLDB Endow. 13, 12* (August 2020), 1821–1834. DOI:https://doi.org/10.14778/3407790.3407792
- K-Means
  - Takehisa Yairi and Yoshikiyo Kato and Koichi Hori. 2001. **Fault detection by mining association rules from house-keeping data**. *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.76.2665
- Matrix Profile (STOMP)
  - Y. Zhu *et al*. 2016. **Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins**. *IEEE International Conference on Data Mining (ICDM)*, pp. 739-748, DOI:https://doi.org/10.1109/ICDM.2016.0085
- Current Benchmarks are flawed
  - Wu, Renjie, and Eamonn Keogh. 2021. **Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress**. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*