Semi-supervised optimal transport methods for detecting anomalies

Amina Alaoui-Belghiti, **Sylvain Chevallier**, Eric Monacelli, Guillaume Bao, Eric Azabou

Nexeya, Hensoldt, France
LISV - Université Paris-Saclay, France
Garches Neuro-Physio-Lab, AP-HP, Inserm 1173, UVSQ, France

ICASSP'2020
6 May 2020

Positive-unlabeled learning

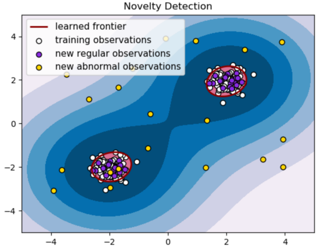
Semi-supervised approach

⇒ Only positive samples are available

⇒ No negative or outliers known during training

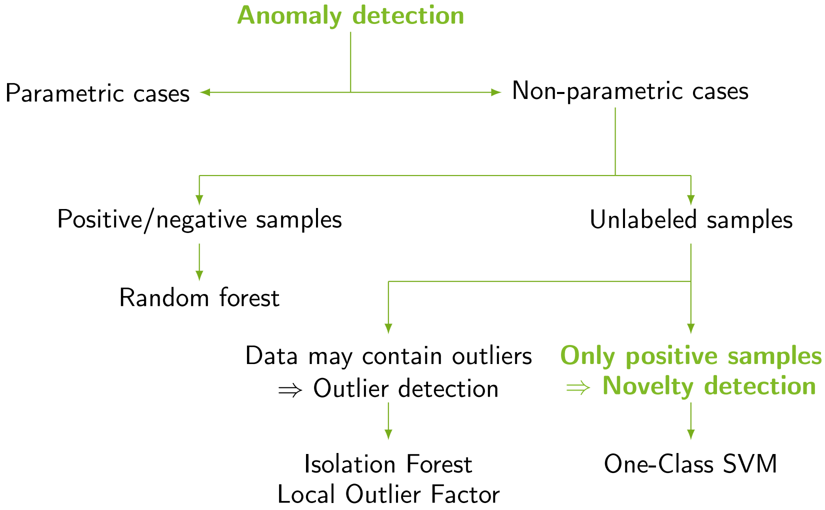
Applicative context:

- Surveillance of stabilized patients in hospital
- Monitoring of a newly calibrated industrial machine



Anomaly detection for time series data

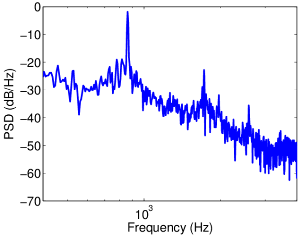
Novelty detection?

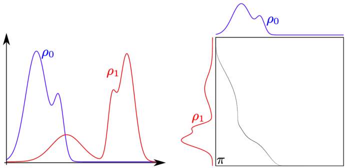


Proposed approach

- No dedicated algorithm for time series
- Specific shifts in frequency domain

⇒ Need a metric to quantify signal variations



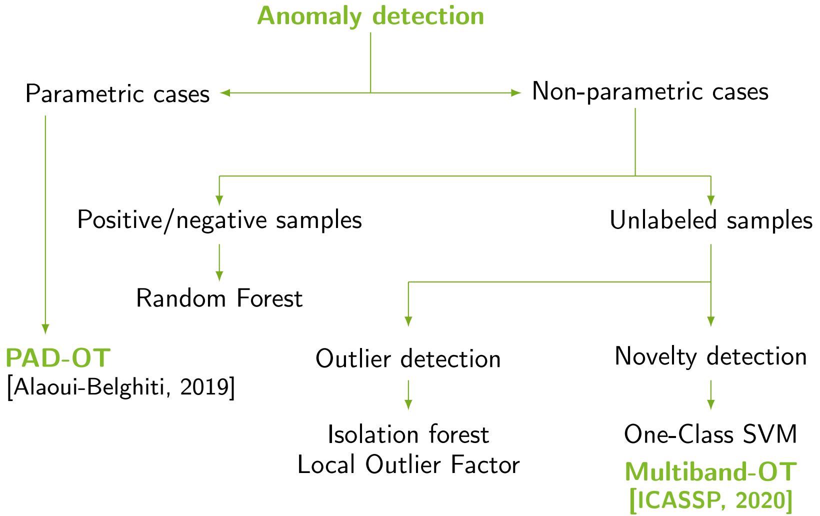


[Solomon, 2018]

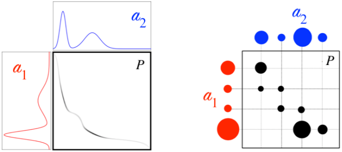
Optimal transport

- Cost of moving from one probability distribution to another
- Application to power spectral density of signal

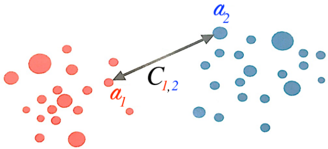
Proposed algorithms



A primer on Optimal Transport



Continuous and discrete transport



Coupling

Coupling: $U(a_1, a_2) = \left\{P \in \mathbb{R}_+^{n \times n} : P\mathbf{1}_n = a_1 \text{ and } P^T\mathbf{1}_n = a_2\right\}$

$$d_C^{\epsilon}(a_1, a_2) = \min_{P \in U(a_1, a_2)} \langle P, C \rangle - \epsilon H(P)$$

with transport matrix P and cost matrix C

$\Rightarrow$ Efficient implementation with entropic regularization [Cuturi, 2013]

Proposed algorithm

PAD-OT [Alaoui-Belghiti, 2019]

1. Estimate average PSD $F(\bar{X})$
2. Compute distance between samples and average $d_C^{\epsilon}(F(\bar{\mathbf{X}}), F(X_k))$
3. Threshold-based detection, assumptions on the distance distribution

Limitations:

- Assumptions are too restrictive
- Problem to detect anomalies occuring in a narrow frequency band

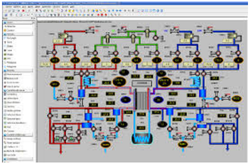
Multiband-OT

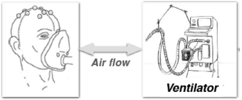
1. Filterbank decomposition $B = b_1, \ldots, b_f$ of PSD samples
2. Compute upper and lower bounds with first and last percentiles
3. Decision based on these bounds

Signal processing for emerging industry applications

Predictive maintenance

- Detecting abnormal behavior for decision on maintenance actions
- Large demand for adaptive and robust algorithms
- Improves product life span





[Navarro-Sune, 2017]

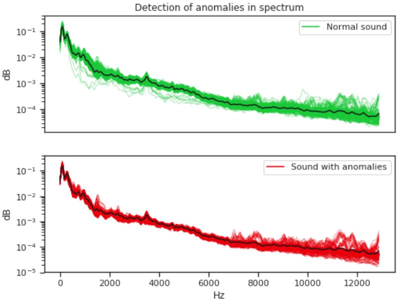
Patient ventilation in hospital

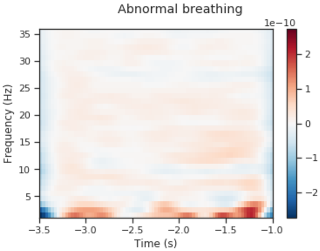
- Mechanical ventilator in ICU, with intubation
- Desynchronization: patient could fight the ventilator
- Source of stress, psychological consequences

Experiments

Sound anomaly detection

- 15 minutes 44 kHz recording of mechanical sound
- 2 kinds of faulty sounds to detect
- repeated k-fold, 500 training, 500 testing

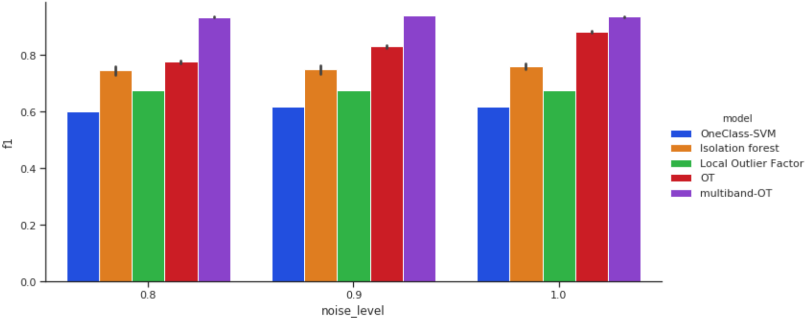




Detecting abnormal breathing in EEG

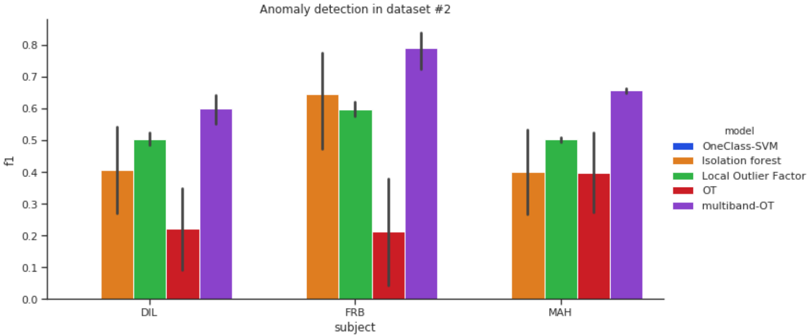
- EEG recorded from subjects in hospital
- Breathing normally and through resistive system
- Equilibrated class for the 2 conditions

Results on acoustic dataset for machine behavior



- OC-SVM has lowest results, Isolation Forest stable around 0.75,
- Best performances by PAD-OT and Multiband-OT
- Stable performance, even when faulty noise is difficult to detect

Application to respiratory-based EEG dataset



- OC-SVM fails to detect abnormal breathing
- Isolation Forest and Local Outlier Factor have some difficulties
- Multiband-OT outperforms other methods
- Only unsupervised methods, not tuned for EEG

Conclusion

Contributions

- New method for semi-supervised anomaly detection for time series
- Non-parametric and more sensitive to local changes
- Decision based on optimal transport cost between PSD
- Application to synthetic and real datasets
- Outperforms existing methods (OC-SVM, LOF, IF)

Future works

- Application to different datasets
- Evaluation in industrial context
- Automated machine learning (AutoML) for Multiband-OT

Thank you !

Annexes