Semi-supervised optimal transport methods for detecting anomalies

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

SEMI-SUPERVISED OPTIMAL TRANSPORT METHODS FOR DETECTING ANOMALIES

Amina Alaoui-Belghiti[⋆†] *Sylvain Chevallier*[⋆] *Eric Monacelli*[⋆] *Guillaume Bao*[‡] *Eric Azabou*[‡]

[⋆] Nexeya, Hensoldt, France
[†] LISV, UVSQ, France
[‡] Garches Neuro-Physio-Lab, AP-HP, Inserm 1173, UVSQ, France

ABSTRACT

Building upon advances on optimal transport and anomaly detection, we propose a generalization of an unsupervised and automatic method for detection of significant deviation from reference signals. Unlike most existing approaches for anomaly detection, our method is built on a non-parametric framework exploiting the optimal transportation to estimate deviation from an observed distribution. We described the theoretical background of our method and demonstrate its effectiveness on two datasets: an industrial predictive maintenance task based on audio recording and a detection of anomalous breathing relying on brain signals. In this type of problem, no negative or faulty samples are seen during training and the objective is to detect any abnormal sample without raising false alarm. The proposed approach outperforms all state-of-the-art methods for anomaly detection on the two considered datasets.

Index Terms— Anomaly detection, optimal transport, unsupervised learning

1. INTRODUCTION

In industrial or medical environment, a common situation were only positive and unlabeled samples are available, is often called anomaly or novelty detection. However this terminology is less precise as it often encompasses several distinct problems. The positive-unlabeled learning considers situations where the only labeled samples are positive [1], no negative or outlier samples are known during the training. This is a semi-supervised approach. The outlier detection problem deals with unsupervised learning: only unlabeled samples are available for training and those training data contain outliers [2, 3]. Novelty detection may refer to a more specific class of problem than the positive-unlabeled learning, for example to detect emergent patterns like new topics is text analysis. The newly detected patterns are integrated in the training dataset to train new models [4].

We consider here positive-unlabeled learning for anomaly detection based on time series data. This type of data is common in industrial context and especially for predictive maintenance task. In predictive maintenance, a continuous monitoring of the equipment components is required to detect abnormal behavior before they turn in faulty situation [5]. Using a network of sensors that report temperature and humidity levels, pressure, vibration or acoustic noise [6, 7].

Existing works are based on classification, nearest neighbors or partitioning methods. One-class SVM [8, 9] is most popular classification-based method for positive-unlabeled method. Neighbors-based methods are best known as Local Outlier Factor [10] and its variants [11, 12, 13], introducing the idea of local anomalies. Isolation Forest is a very efficient partition-based method [14]. It introduces the use of isolation as a more effective means of detecting anomalies.

Advances in optimal transport [15] allow to define metrics and topological spaces for quantifying the variation between known samples and potential anomalies. Metrics for handling probability measures are characterized by a transport plan from one probability space to another according to a cost matrix. Approaches have been applied to a wide variety of tasks [16], but often requires large amount of computational resources. A new algorithm for fast evaluation of transportation distance, known as the Sinkhorn distance [17], mitigates computational cost issues. Adding entropic regularization to transportation distance, the Sinkhorn distance has several interesting properties as it is a scale free, non-Euclidean formulation which is less subject to the curse of dimensionality. The implementation has been parallelized on GPU, speeding the computation [18].

This paper focus on positive-unlabeled learning tasks with algorithms based on optimal transport. Our main contribution is the introduction of novel semi-supervised algorithms for anomaly detection. This algorithms are compared with the state-of-the-art algorithms on two real datasets. The remainder of the paper is structured as follows. Section 2 describes the state of the art methods and the proposed approaches by providing a formal description of the system. In Sect. 3, the proposed methods are evaluated on two real datasets and compared with Isolation Forest, Local Outlier Factor and One Class-SVM. The Sect. 4 concludes this paper.

2. ALGORITHMS FOR POSITIVE-UNLABELED LEARNING

In the following, we will consider a set of k initial signals $\mathbf{X} = \{X_i\}_{i=1\dots k}, X \in \mathbb{R}^t$ and signal to evaluate $\hat{X}$. In this study, the signals are analyzed in the frequency domain, estimated with the power spectral density. It is evaluated with the Welch's method $F(\cdot)$, where partially overlapping segments are combined with a Hamming window function to estimate an FFT average. The resulting signals are $F(X) \in \mathbb{R}^n$.

2.1. State-of-the-art models

Three main approaches are developed in the literature for positive-unlabeled learning. First, classification methods rely on machine learning techniques for capturing the class of positive examples. The most robust method is the one-class SVM, with the ν margin parameter that defines if a new sample is considered as abnormal or not.

Approached with a partition-based techniques, the positive-unlabeled learning yields a multivariate outlier detection called Isolation Forest. Using random forest approach, the number of levels of the decision trees is a direct indicator of the abnormality of a sample: if a sample deviate from what has been observed it could be ruled out using only low number of trees.

Local Outlier Factor is a widely know method for positive-unlabeled learning that makes use of nearest neighbor approach. The local outlier factor compares the density of the k-neighbor distance for a given sample to the density of all its k-neighbors.

2.2. Proposed approaches

We proposed here two algorithms, a first which is simple and parametric and a second one that is more complex and non-parametric.

In the proposed approach, two metric spaces $\mathcal{X}_\infty$ and $\mathcal{X}_\in$, with the set $\mathcal{M}(X_1)$ of discrete probability measures. This approach could be extended to the continuous case with no further assumptions. The set of coupling matrices, for discrete measure α_1 (respectively α_2) on n locations on $\mathcal{X}_\infty$ (resp. $\mathcal{X}_\in$) with weight a_1 (resp. a_2), is defined as:

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

where $\mathbf{1}_n$ is the n-dimensional vector of ones.

A cost matrix $C \in \mathbb{R}^{n\times n}$ holds the mapping cost from a_1 to a_2 based on the transport matrix P that is the dot product $\langle P, C\rangle$. The optimal transport problem is defined as:

$$d_C(a_1, a_2) = \min_{P\in U(a_1,a_2)} \langle P, C\rangle. \tag{2}$$

The entropy of the coupling matrix is

$$H(P) = -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1). \tag{3}$$

The optimal transport problem with an entropic regularization thus writes:

$$d_C^\epsilon(a_1, a_2) = \min_{P\in U(a_1,a_2)} \langle P, C\rangle - \epsilon H(P). \tag{4}$$

The unique solution of this problem is of the form $P_{i,j} = u_i K_{i,j} v_j$ with $u_i, v_j \in \mathbb{R}_+^n$. The Sinkhorn algorithm is solved by applying iteratively the following update function for iteration $l + 1$:

$$u^{(l+1)} = \frac{a_1}{Kv^{(l)}} \text{ and } v^{(l+1)} = \frac{a_2}{K^T v^{(l)}}. \tag{5}$$

The Sinkhorn algorithm is easily parallelizable and could be executed on GPU.

2.3. Parametric algorithm

The signals $\mathbf{X}$ are averaged to obtain a barycenter that defines a reference $F(\bar{\mathbf{X}}) = \frac{1}{k}\sum_k F(X_k)$. The distances between the reference PSD $F(\bar{\mathbf{X}})$ and all the PSD samples $F(X_k)$ are estimated with the Sinkhorn distance $d_C^\epsilon(F(\bar{\mathbf{X}}), F(X_k))$, with C a Chebyshev cost function. For the test signal, the distances are estimated with $d_C^\epsilon(F(\bar{\mathbf{X}}), F(\tilde{X}))$.

```
Data: set of reference signals $\mathbf{X}$, signal to evaluate $\hat{X}$
Result: binary classification, 1 if normal signal, -1 if
        abnormal
$F(\bar{\mathbf{X}}) \leftarrow \frac{1}{k}\sum_k F(X_k)$
for $i \leftarrow 1$ to $k$ do
    $d_i \leftarrow d_C^\epsilon(F(\bar{\mathbf{X}}), F(X_i))$
end
Set threshold $\vartheta$ from LogNormal fit on $\{d_i\}_{i=1\dots k}$
$\hat{d} \leftarrow d_C^\epsilon(F(\bar{\mathbf{X}}), F(\hat{X}))$
if $\hat{d} > \vartheta$ then
    return $-1$
end
else
    return $1$
end
```

Algorithm 1: Parametric algorithm for anomaly detection

Under the assumption that the distance follows a Log-Normal distribution, it is possible to determine a threshold for predicting if the test signal. The different steps of the algorithm are shown in Algorithm 1. In Sect. 3, this algorithm is called *OT*.

2.4. Nonparametric algorithm

A more robust version of the previous algorithm is proposed hereafter. The assumption that the distribution of the distances between the barycenter and the training signal follows a Log-Normal distribution may be violated. In case of a bi-modal distribution or more complex ones, a decision based on wrong assumption may induce poor decision. Another problem arises when the anomaly to detect is restricted to a specific bandwidth. In that case, the anomaly may be undetected as the induced variation is ”diluted” in the whole frequency spectrum.

To mitigate these issues, we propose to rely on non-parametric statistics computed on a filter bank signal decomposition. The PSD of the signal is independently analyzed for f different frequency bands $B = b_1, \ldots, b_f$, this allows to detect abnormal variations occurring within narrow bandwidth. The lower and upper bounds of the distribution of the distances for each frequency bands b are estimated as the first and last percentile, that is $p^b_{0.01} = d^\epsilon_C(F(\bar{\mathbf{X}}^b), F(\mathbf{X}^b_{0.01}))$ and $p^b_{0.99} = d^\epsilon_C(F(\bar{\mathbf{X}}^b), F(\mathbf{X}^b_{0.99}))$. Anomaly scores for a frequency band b are computed as:

$$A^b_{\text{lower}} = \frac{d^\epsilon_C(F(\bar{\mathbf{X}}^b), F(\tilde{X}^b))}{p_{0.01}}, A^b_{\text{upper}} = \frac{d^\epsilon_C(F(\bar{\mathbf{X}}^b), F(\tilde{X}^b))}{p_{0.99}} \tag{6}$$

a value above 1 indicates an abnormal sample for the considered band. The decision function $g(\tilde{X})$ rely on a combination of the score for all f frequency band:

$$g(\tilde{X}) = \begin{cases} -1 \text{ if } \frac{1}{f}\sum_i A^{b_i}_{\text{lower}} > 1 \text{ or if } \frac{1}{f}\sum_i A^{b_i}_{\text{upper}} > 1 \\ 1 \text{ otherwise} \end{cases} \tag{7}$$

The different steps are described in Algorithm 2.

```
Data: set of reference signals $\mathbf{X}$, signal to evaluate $\hat{X}$
Result: binary classification, 1 if normal signal, -1 if
        abnormal
for $j \leftarrow 1$ to $f$ do
    $F(\bar{\mathbf{X}}^{b_j}) \leftarrow \frac{1}{k}\sum_k F(X^{b_j}_k)$
    Get $p^{b_j}_{0.01}$ and $p^{b_j}_{0.99}$
    Compute $A^b_{\text{lower}}$ and $A^b_{\text{upper}}$ from Eq. (6)
end
return $g(\tilde{X})$, as in Eq. (7)
```

Algorithm 2: Robust and non-parametric algorithm

In the experimental part, this algorithm is referred to as *multiband-OT*.

3. EXPERIMENTS

This section describes the experiments conducted on two different datasets. The first one is an audio recording of an industrial machine with and without an abnormal noise. The second dataset is based on recordings of brain signals in different breathing conditions.

3.1. Sound anomaly detection

This first dataset is dedicated to evaluate the performance of anomaly detection algorithms in the context of predictive maintenance [19].

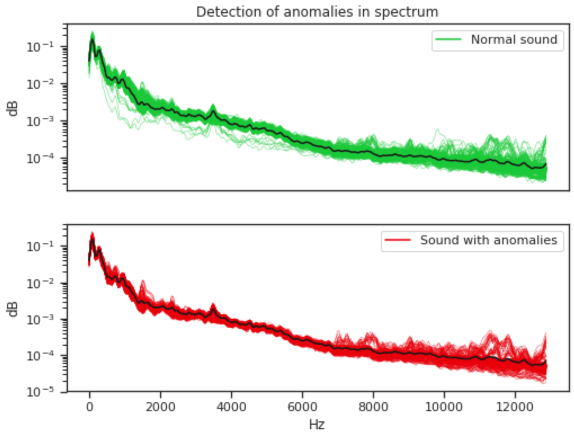

Fig. 1. Examples of normal (top) and abnormal (bottom) sounds extracted from the first dataset.

The signal is 15 minutes long recording in monaural at 44100 Hz[1]. Two qualitatively different kinds of faulty mechanical parts are considered: the sound of a light, high pitched whistling and a cyclical low-pitched sound, similar to a faulty ball bearing. The Fig. 1 shows the considered signals in the frequency domain. Each line is a measure in the frequency domain, estimated with the Welch’s method.

A repeated k-fold cross-validation is used to separate the dataset in training (500 samples) and test data (500 samples). Anomaly detection models are calibrated on training data and also to compute the reference signal $F(\bar{\mathbf{X}})$. The dataset includes 3 levels of anomalous sounds to detect, the normal machine behavior being mixed with faulty mechanical noise. The higher the noise level, the easier it is to detect.

The models are evaluated based on the achieved F-measure (or F1-score). This measure combines both precision and recall, as:

$$F_1 = 2\frac{P \cdot R}{P + R}, \tag{8}$$

where P is the precision and R is the recall. It is thus well suited to estimate the performance for anomaly detection, as one want to avoid false alarm as much as missed detection.

3.2. Detecting abnormal breathing in EEG

This dataset is based on the brain signals observed when subjects have trouble to breathe [20]. This is an important topic for the automatic detection of patient’s incorrect ventilation in intensive care units. These experiments are approved by the local Ethics committee, under number 11073 on the 24th November 2011, and registered in the public trials registry, under number NCT01548586 Subjects are exposed to different levels of inspiratory resistive load, as they are either able

[1] All of these recordings are available as well upon request

to breathe normally or need to inspire through a resistive system.

When the subjects are breathing with a resistive load, their brain generates specific activity called preinspiratory potentials. Several electrodes, here 16, are placed on the head of the subject. Preinspiratory potentials are characterized by variations in the μ frequency band (8-12 Hz).

Time-frequency plots, such as the ones shown on Fig. 2, highlights those variations that take place few seconds before breathing. Here, it could be seen as a decrease of activity around 10 Hz between 2.5 and 1 s before the inspiration.

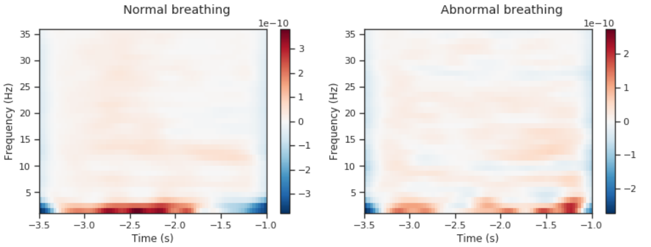

Fig. 2. Time-frequency analysis of EEG signals, with control condition (left) and abnormal breathing (right).

Three representative subjects are selected to conduct this experiment, with 100 samples for each subject and each condition. The number of samples in each condition is adjusted to ensure that the evaluation is made on balanced classes. As for the previous dataset, a cross-validation with a repeated k-fold is used to separate the training and testing samples. After a calibration on training data, the models are evaluated on test data with the F1-score.

3.3. Results

For the first dataset, it could be seen on Fig. 3 that all evaluated models are achieving correct results. The One-Class SVM has the lowest results, obtaining a score around 0.6 that increase with the noise level. The Local Outlier Factor obtains a stable score of 0.67 for all noise level, while Isolation Forest is around 0.75. The robustness of isolation forest is outperformed by the proposed methods, OT algorithm achieving between 0.77 and 0.88. The multiband-OT method introduced in this paper reaches the highest score, around 0.93.

For the EEG dataset, the models have more difficulties to correctly detect anomalies (Fig. 4). This is expected, as EEG is notoriously difficult to process. The One-Class SVM fails completely to detect abnormal breathing, labeling all samples as abnormal, hence obtaining a score of 0. Successful SVM approaches for EEG are highly specific and require a delicate parameter selection, as shown in the works of [21] and [22]. The OT model obtains between 0.2 and 0.4 depending of the subject. Isolation forest achieves scores of 0.4 to 0.64 with a wide intra-subject variability. Local Outlier Factor yields more stable results, between 0.5 and 0.64. The multiband-OT outperforms all others methods, with a score between 0.6 and 0.78.

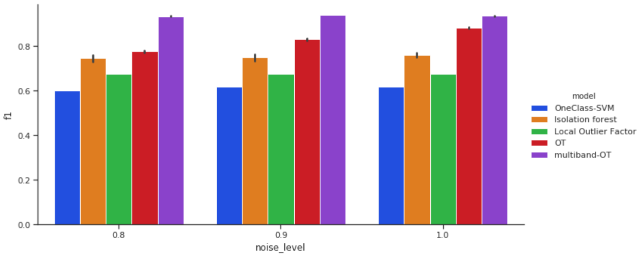

Fig. 3. Comparison of anomaly detection for dataset 1: microphone recording of normal/abnormal machine behavior.

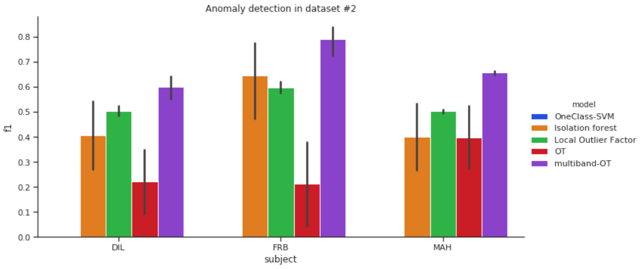

Fig. 4. Detecting abnormal breathing from time-frequency features of EEG signal.

It should be noted that the methods evaluated here are unsupervised and not specifically tuned for processing EEG. The obtained results have thus a lower accuracy that those that could be obtained with EEG-specific supervised methods.

F1-score is a balanced measure, that takes into account both type I and type II errors. It is an adequate choice for assessing the anomaly detection models as it summarizes the precision and recall performance in one indicator.

4. CONCLUSION

This paper focuses on a new method of semi-supervised and unsupervised anomaly detection with a positive-unlabeled learning on acoustic and EEG signals, by comparing them with reference signals through calculating Sinkhorn distances in optimal transport.

The contribution of this paper has been devoted to acoustic and EEG data, but we want to spread the study of anomaly detection in future work on other types of data with a possible extension for online implementation.

5. REFERENCES

[1] Victoria Hodge and Jim Austin, “A survey of outlier detection methodologies,” *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004.

[2] Annick M Leroy and Peter J Rousseeuw, “Robust regression and outlier detection,” *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1987*, 1987.

[3] Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad, and Mustafa Mat Deris, “A comparative study for outlier detection techniques in data mining,” in *IEEE CCIS*, 2006, pp. 1–6.

[4] Markos Markou and Sameer Singh, “Novelty detection: a review—part 1: statistical approaches,” *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[5] Mina Abdel-Sayed, Daniel Duclos, Gilles Faÿ, Jérôme Lacaille, and Mathilde Mougeot, “Dictionary comparison for anomaly detection on aircraft engine spectrograms,” in *Machine Learning and Data Mining in Pattern Recognition*, pp. 362–376. Springer, 2016.

[6] Eamonn Keogh, Stefano Lonardi, and Bill’Yuan-chi’ Chiu, “Finding surprising patterns in a time series database in linear time and space,” in *ACM SIGKDD*, 2002, pp. 550–556.

[7] Ryohei Fujimaki, Takehisa Yairi, and Kazuo Machida, “An approach to spacecraft anomaly detection problem using kernel feature space,” in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 401–410.

[8] Larry M Manevitz and Malik Yousef, “One-class svms for document classification,” *Journal of machine Learning research*, vol. 2, no. Dec, pp. 139–154, 2001.

[9] Junshui Ma and Simon Perkins, “Time-series novelty detection using one-class support vector machines,” in *IJCNN*, 2003, vol. 3, pp. 1741–1745.

[10] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander, “Lof: identifying density-based local outliers,” in *ACM Sigmod Record*, 2000, vol. 29, pp. 93–104.

[11] Hehe Ma, Yi Hu, and Hongbo Shi, “Fault detection and identification based on the neighborhood standardized local outlier factor method,” *Industrial & Engineering Chemistry Research*, vol. 52, no. 6, pp. 2389–2402, 2013.

[12] Mahsa Salehi, Christopher Leckie, James C Bezdek, Tharshan Vaithianathan, and Xuyun Zhang, “Fast memory efficient local outlier detection in data streams,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3246–3260, 2016.

[13] Zonghai Chen, Ke Xu, Jingwen Wei, and Guangzhong Dong, “Voltage fault detection for lithium-ion battery pack using local outlier factor,” *Measurement*, 2019.

[14] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, “Isolation forest,” in *IEEE Conf. on Data Mining*, 2008, pp. 413–422.

[15] Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.

[16] Gabriel Peyré, Marco Cuturi, et al., “Computational optimal transport,” Tech. Rep., 2017.

[17] Marco Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Advances in neural information processing systems*, 2013, pp. 2292–2300.

[18] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach, “Stochastic optimization for large-scale optimal transport,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3440–3448.

[19] Amina Alaoui-Belghiti, Sylvain Chevallier, and Eric Monacelli, “Unsupervised anomaly detection using optimal transport for predictive maintenance,” in *ICANN*. Springer, 2019, p. tbp.

[20] Sylvain Chevallier, Guillaume Bao, Mayssa Hammami, Fabienne Marlats, Louis Mayaud, Djillali Annane, Frédéric Lofaso, and Eric Azabou, “Brain-machine interface for mechanical ventilation using respiratory-related evoked potential,” in *ICANN*. Springer, 2018, pp. 662–671.

[21] A. Rakotomamonjy, V. Guigue, G. Mallet, and V. Alvarado, “Ensemble of SVMs for improving brain-computer interface P300 speller performances,” *15th International Conference on Artificial Neural Networks*, pp. 45–50, 2005.

[22] N. Jrad, M. Congedo, R. Phlypo, S. Rousseau, R. Flamary, F. Yger, and A. Rakotomamonjy, “sw-SVM: sensor weighting support vector machines for EEG-based brain-computer interfaces,” *Journal of Neural Engineering*, vol. 8, no. 5, pp. 056004, 2011.