DICTIONARY-DECOMPOSITION-BASED ONE-CLASS SVM FOR UNSUPERVISED DETECTION OF ANOMALOUS TIME SERIES

*Marco Bevilacqua**, *Sotirios A. Tsaftaris**†

* IMT Institute for Advanced Studies
PRIAn Research Unit
Lucca, Italy

† Northwestern University
Electrical Engineering and Computer Science
Evanston, IL, United States of America

ABSTRACT

In this paper, we propose a new unsupervised method to detect anomalous time series, given as little as one test data set. Particularly, the method proposed consists in an iterative procedure with two alternating steps: a dictionary learning (DL) step, to learn a pattern for the "normal" time series, and a modified One-class Support Vector Machines (OCSVM) classifier, which finds the anomalies by taking into account how much each time series deviates from the pattern found. The algorithm that arises from the original combination of these two steps represents a general framework, where any DL algorithm can be plugged in according to the model desired. When tested on two different applications (detection of shape anomalies and analysis of cardiac MRI time series), our proposed method is shown to outperform other unsupervised approaches, presenting higher detection accuracy.

***Index Terms*—** Anomaly detection, Dictionary Learning, Support Vector Machines

1. INTRODUCTION

Anomaly detection, in machine learning or in the *Knowledge discovery in databases* (KDD) field, refers to the problem of finding patterns in data that do not conform to an expected model [1, 2]. This problem finds a direct application in a variety of domains, e.g. health-care, fraud detection, intrusion detection, etc., where an anomaly is considered a highly critical event that has not been observed in the data. In many of these applications, data is collected in the form of time series. Two typical cases are then observable: we can have process anomalies, if the whole time series deviates from a typical pattern, or subsequence anomalies, when the abnormal events occur in specific time frames.

Traditional approaches to time series anomaly detection follow a supervised scheme: a database of annotated time series is available, and classification-based or nearest-neighbor-based techniques are generally employed. We are instead interested in an unsupervised approach, where anomalies are detected having available as little as a single test data set. To this end, we propose a new method to detect process anomalies, which is fully unsupervised and requires no prior knowledge on the number and type of anomalies. Our only assumption is that most of the time series (the "normal" ones) conform with a dictionary-based sparse decomposition model ($\mathbf{Y} \approx \mathbf{DX}$), whereas a little fraction of them significantly deviates from this model. The same assumption is made in [3], where the authors develop a method for anomaly detection with an ad-hoc sparse coding. Unlike [3], where the dictionary $\mathbf{D}$ is assumed known, however, we perform dictionary learning (DL) in the presence of outliers, i.e. the model is trained directly from the given data set. As a second step, then, a modified One-class Support Vector Machines (OCSVM) classification is performed, to partition the data set into normal and anomalous time series. OCSVM is a state-of-the-art SVM-like classifier [4–6], which aims at finding the boundaries to separate data points related to a single dominant class from the rest of the data points, considered as outliers. We propose a modification to the original OCSVM classifier [4], which penalizes each data point according to the related representation error originated from the DL step.

The contributions of this paper are then twofold: (i) the modification proposed on the OCSVM algorithm, and (ii) the whole scheme presented, which represents a new framework for unsupervised anomaly detection in combining DL and OCSVM as alternating steps of an iterative procedure. The method can be referred to as a *dictionary-decomposition-based One-class SVM*. The first step of the proposed method (DL) can be customized according to a specific sparse decomposition model chosen; independently of the choice made, the modified OCSVM will take benefit of the dictionary-based decomposition made to better identify the anomalies.

The rest of the paper is organized as follows. In Section 2, our modification to the OCSVM algorithm (first contribution) is outlined. Then, in Section 3, the whole method is presented (second contribution), by detailing the procedure followed. Before drawing conclusions, Section 4 presents some

This work was supported in part by the US National Institutes of Health (grant no. 2R01HL091989-05), a Marie Curie Action "Reintegration Grant" (grant no. 256534) of the EU's Seventh Framework Programme (FP7), and the project WB@Lucca funded by Fondazione Cassa Di Risparmio Di Lucca (CaRiLu).

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

experimental results on two distinct applications.

2. ONE-CLASS SUPPORT VECTOR MACHINES (OCSVM)

A One-class Support Vector Machines (OCSVM) classifier is a machine learning technique that aims at describing a unique "normal" class of data points by learning a compact model such that as many as possible data points are included. The two models usually considered are hyperplanes [4] and hyperspheres [5], which eventually lead to an equivalent formulation. In the former case, the goal is to learn a decision boundary (a hyperplane) that achieves the maximum separation between the points and the origin. When training and testing an OCSVM model on the same data set, we actually perform outlier/anomaly detection.

Like traditional SVMs algorithms, OCSVM uses an implicit transformation function $\phi(\cdot)$ to project the data onto a higher-dimensional kernel space. In this space, the function that defines the decision boundary is the following:

$$g(\boldsymbol{y}) = \boldsymbol{\omega}^{\mathrm{T}}\phi(\boldsymbol{y}) - \rho \ , \tag{1}$$

where $\boldsymbol{\omega}$ is the vector perpendicular to the decision boundary and ρ is the bias term. The classification of a particular data point $\hat{\boldsymbol{y}}$ is given by the sign of the function $g(\hat{\boldsymbol{y}})$. The vector $\boldsymbol{\omega}$ and the bias ρ are found by solving the well-known optimization operation:

$$\begin{array}{l} \min\limits_{\boldsymbol{\omega},\rho,\boldsymbol{\xi}} \dfrac{\|\boldsymbol{\omega}\|^2}{2} - \rho + \dfrac{1}{\nu N}\sum\limits_{i=1}^{N}\boldsymbol{\xi}_i \\ \text{s.t.} \quad \omega^{\mathrm{T}}\phi(\boldsymbol{y}_i) \geq \rho - \boldsymbol{\xi}_i \ , \quad \boldsymbol{\xi}_i \geq 0 \end{array} \ , \tag{2}$$

where $\boldsymbol{\xi}_i$ is the *slack variable* related to the data point $\boldsymbol{y}_i$ that plays as a margin for it to lie on the other side of the decision boundary, and ν is a regularization parameter.

2.1. Modified OCSVM

As other methods that introduce modifications to the OCSVM classifier [7, 8], we propose to change the role of the slack variables $\boldsymbol{\xi}$. In particular, we want to assign a larger or smaller margin to each data point according to how bad or well, respectively, it is represented by a dictionary-based sparse decomposition model, which is supposed to characterize the normal class. In other words, we want data points with a large reconstruction error to be assigned a large slack variable that allows them to be outliers.

Given the generic dictionary-based sparse model [9]

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{x}_i\|_0 \leq L \ , \tag{3}$$

where we can possibly have other constraints on the sparse representation coefficients or the dictionary atoms to design, a measure of the reconstruction quality of a particular data point $\boldsymbol{y}_i$ can be defined as:

$$r_i = \|\boldsymbol{y}_i - \mathbf{D}\boldsymbol{x}_i\|^2 \ . \tag{4}$$

As easily provable, the metric in Equation (4), which is the Euclidean distance between a data point and its projection onto the dictionary subspace (which is, in turn, a point that approximately lies within the normal class), reported in the kernel space, can be expressed as follow:

$$d_i = 2(C - \phi(\boldsymbol{y}_i) \cdot \phi(D\boldsymbol{x}_i)) \ , \tag{5}$$

where C is a constant measuring the product in the kernel space between a data point and itself.

The distances $\{d_i\}$ can now be used in the optimization function as a direct expression of the slack variables. The original OCSVM objective becomes than the following:

$$\min_{\boldsymbol{\omega},\rho} \frac{\|\boldsymbol{\omega}\|^2}{2} - \rho \quad \text{s.t.} \quad \boldsymbol{\omega}^{\mathrm{T}}\phi(\boldsymbol{y}_i) \geq \rho - \lambda d_i \ , \tag{6}$$

where λ is a regularization parameter that weights each distance to provide the actual margin.

3. PROPOSED METHOD: DICTIONARY-DECOMPOSITION-BASED ONE-CLASS SVM

In Section 2 we proposed a modified OCSVM classifier, which, under the assumption that a dictionary-based sparse decomposition model fits better the normal data points, uses the obtained reconstruction errors to guide the construction of the decision boundary, in a similar way as in [10] a "sparse construction cost" is used for the detection of abnormal events.

The whole procedure consists then of two steps:

- A dictionary learning (DL) algorithm, to learn a sparse decomposition model aiming at characterizing the normal class; and
- Our modified OCSVM classifier, proposed in Section 2, to detect the anomalies by tracing the boundary of the normal class.

The two steps are alternated in an iterative fashion: at each iteration, DL is performed only on the supposedly normal data points (given the classification at the previous iteration); with a possibly better learned pattern, the following classification is refined. The workflow of the whole method is depicted in Fig. 1.

The label vector is initialized with all ones (1 means "normal", -1 means "anomaly"), i.e., at the first run of the DL algorithm, all time series are considered normal. At the next iterations, the label vector changes, possibly converging towards the optimal solution, thus improving the pattern discovery step via DL. Details of each step of the proposed

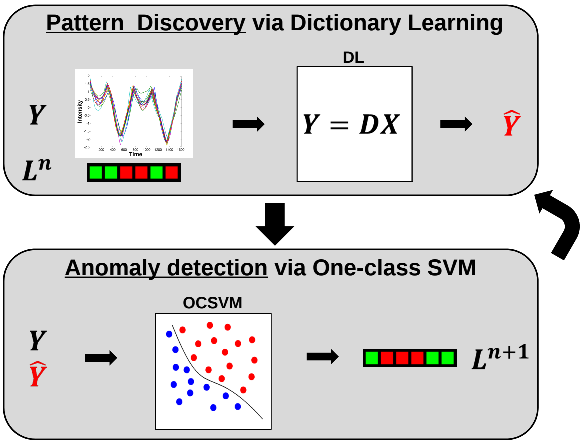

Fig. 1. Workflow of the the proposed method.

method, consisting in a dictionary-decomposition-based one-class SVM (for brevity, the acronym "DOCS" is used), are reported in Algorithm 1.

It is important to note that in the DL step non overcomplete dictionary models are preferable. In fact, we want to have a compact model as large as necessary to represent well the normal class, without taking the risk to also accommodate for the anomalies.

Algorithm 1 Details of the Dictionary-decomposition-based One-class SVM (DOCS) method.

1: **procedure** DOCS($\boldsymbol{Y}$)

2: Initialization of the labels:

$$\boldsymbol{l}_i = 1 \quad \forall i = 1, \dots N$$

3: Dictionary Learning algorithm:

$$\min_{\mathbf{D}, \boldsymbol{X}} \|\boldsymbol{Y}_{\boldsymbol{l}_i=1} - \boldsymbol{D}\boldsymbol{X}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{x}_i\|_0 \leq L$$

4: Compute distances in the Kernel space:

$$d_i = 2(-\phi(\boldsymbol{y}_i) \cdot \phi(\mathbf{D}\boldsymbol{x}_i))$$

5: Solve the modified One-class SVM:

$$\min_{\boldsymbol{\omega}, \rho} \frac{\|\boldsymbol{\omega}\|^2}{2} - \rho \quad \text{s.t.} \quad \boldsymbol{\omega}^{\mathrm{T}} \phi(\boldsymbol{y}_i) \geq \rho - \lambda d_i$$

6: Determine the labels:

$$\boldsymbol{l}_i = \operatorname{sgn}(\boldsymbol{\omega}^{\mathrm{T}} \phi(\boldsymbol{y}_i) - \rho)$$

7: If num. max iterations not reached **go to** Step 3.

8: **return** $\boldsymbol{l}$ ▷ Output labels

9: **end procedure**

4. EXPERIMENTAL RESULTS

In this section, we test the proposed method for time series anomaly detection with two applications: detection of shape anomalies (Section 4.1), and ischemia detection using CP–BOLD time series (Section 4.2). In both cases, we define the dictionary-based decomposition model used, and compute the performance of our method in terms of detection accuracy w.r.t. other state-of-the-art approaches.

4.1. Application 1: shape anomaly detection

For the detection of shape anomalies, given an object, a shape 1-D sequence can be extracted by moving along the contour of the object and registering the distance of the current point to the image center (Fig. 2) [11].

For this problem, we considered two different data sets: *Diatoms* and *MixedBag*[1]. The two data sets are composed by several classes of shapes. An anomaly detection test can then be set up by taking all instances from one particular class and add random anomalies coming from other classes. Fig. 3 shows two examples of 1-D sequences for two different classes of the *MixedBag* data set.

As we can see in Fig. 3, 1-D sequences related to shapes are typically characterized by a single underlying pattern with possible shifts. To learn a representation of them that is invariant to shifts, we decide to adopt the shift-invariant dictionary learning algorithm of Rusu *et al.* [12]. The algorithm aims at learning the *circulant dictionary* than can best represent a given input data matrix (each input data vector is seen as a weighted version of the shifting pattern with a strict unitary ℓ^0-sparsity constraint). In addition, we propose to add a non-negative constraint for the sparse coefficient vectors. This drives to the learning of a pattern with no sign ambiguity (otherwise, a time series, although maybe behaving very differently than the pattern, could still be assigned a large coefficient).

$$\min_{\mathbf{C}, \boldsymbol{X}} \|\boldsymbol{Y}_{\boldsymbol{l}_i=1} - \boldsymbol{C}\boldsymbol{X}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{x}_i\|_0 = 1 \ , \ \mathbf{1}^T \boldsymbol{x}_i \geq 0 \quad (7)$$

Our proposed method is compared with regular One-class SVM (OCSVM) [4], where we choose the optimal value of ν by cross-correlation and perform the classifier directly on the raw 1-D sequences, and with OCSVM performed in the frequency domain, i.e. on the magnitude vector of the FFT coefficients (this in order to correct for shifts in the data). Results for several runs, across all classes (at each time, one class plays as the dominant one), and for two different percentage of anomalies (10% and 20%), are reported in Table 1.

As we can observe from the table, our method turns out to be the best performing in 5 cases out of 6, presenting the highest detection accuracy.

[1] The data sets can be downloaded at the *URL* `http://www.cs.ucr.edu/~eamonn/shape/shape.htm`.

	10% anomalies			20% anomalies		
Dataset	**OCSVM**	**FD-OCSVM**	**Proposed**	**OCSVM**	**FD-OCSVM**	**Proposed**
Diatoms	89.1%	89.1%	**90.8%**	83.5%	**83.9%**	82.5%
MixedBag	74.9%	86.4%	**91.4%**	78.1%	80.9%	**82.2%**

Table 1. Performance results in terms of detection accuracy of the proposed method w.r.t. simple One-class SVM (OCSVM) and OCSVM in the frequency domain.

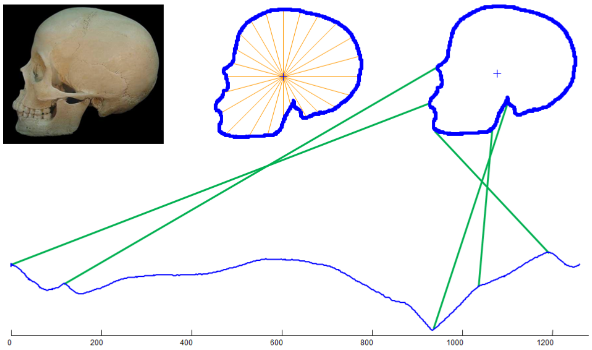

Fig. 2. Generation process of a 1-D sequence from a shape: moving along the contour of the image, the distance to the image center is encoded.

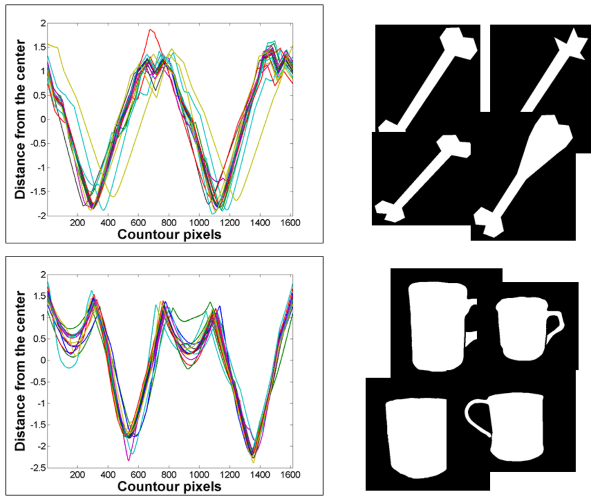

Fig. 3. Examples of 1-D sequences and original shape images for two classes of the *MixedBag* data set.

4.2. Application 2: ischemia detection with CP–BOLD time series –a study on synthetic data

As another application, we consider Cardiac Phase-resolved Blood Oxygen Level-Dependent (CP–BOLD) Magnetic Resonance Imaging (MRI), which is a current state-of-the-art imaging technique that examines changes in myocardial oxygenation without the use of contrast media and pharmacological stress agents. Signal intensity, i.e. pixel intensity values in the images acquired, can be locally averaged to analyze specific patterns occurring in a given region of the myocardium: by taking local averages across time we can then extract time series. If we look at the time series related to a patient affected by ischemia (Fig. 4), we can notice that some of them roughly follow a specific pattern (by presenting mutual shifts), whereas others appear more irregular: the former correspond to myocardial territories remote to ischemia, while the latter relate to the ischemic area [13]. Since under the hypothesis of single-vessel disease the remote time series are larger in number, we can then use an anomaly detection algorithm, as the proposed method, to detect ischemia.

To test unsupervised ischemia detection algorithms, we used the CP–BOLD time series simulator described in [14], by setting different numbers of myocardial segments (N) and different values of “ischemic extent” (IE), which is the equivalent to the percentage of anomalies. A comparison of the proposed algorithm with One-class SVM and *Independent Component Analysis* (ICA) is reported in Table 2.

In this case, our method outperforms in every test the two other state-of-the-art approaches considered, showing a considerably improved ischemia detection accuracy.

	N=150			N=50		
IE	**ICA**	**OCSVM**	**Proposed**	**ICA**	**OCSVM**	**Proposed**
40%	92±6	89±1	**98±1**	79±6	87±1	**89±5**
33%	87±6	82±1	**97±2**	84±6	79±2	**88±8**
25%	87±5	74±1	**97±2**	85±7	72±1	**93±5**

Table 2. Accuracy (mean ± std) in % of ICA, OCSVM, and proposed method, with variable N and IE.

In the tests made, an optimal empirical formula for the regularization parameter is found via grid search as $\lambda = 1/(2M)$, where M is the dimension of the data vectors. The maximum number of iterations is instead set to 10; however, the algorithm possibly stops before when the same vector of labels is produced for two consecutive runs. E.g., in the case of Application 1, it is observed that the algorithm converges after 3.3 iterations (on average).

5. CONCLUSION

In this paper, we presented a new method to detect anomalous time series in an unsupervised fashion, given a single

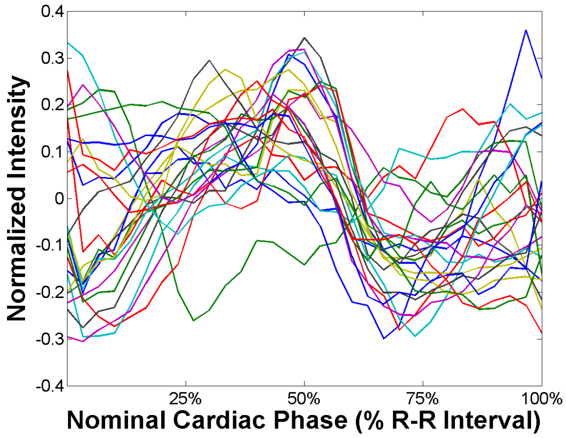

Fig. 4. Examples of CP–BOLD time series extracted from a subject .

data set and no prior knowledge. The method consists in a dictionary-decomposition-based One-class SVM: it relies on a dictionary (DL) step, which aims at learning a model for the normal time series, and a proposed ad-hoc One-class SVM classifier, which detects the anomalous ones. The proposed method iterates the two steps in a way that they can take advantage of each other, and it is flexible in the sense that any DL algorithm can be used as for the first step. When comparing our method with other state-of-the-art approaches, it is shown to reach higher averages in detection accuracy. The method presented is expected to be useful in several other applications beyond those ones discussed in the paper, where anomalies in time series data are to be detected, e.g. leaf detection and analysis for plant phenotyping and trajectory analysis in transportation.

REFERENCES

[1] Varun Chandola, *Anomaly detection for symbolic sequences and time series data*, Ph.D. thesis, University of Minnesota, 2009.

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

[3] Amir Adler, Michael Elad, Yacov Hel-Or, and Ehud Rivlin, "Sparse Coding with Anomaly Detection," *Journal of Signal Processing Systems*, pp. 1–10, July 2014.

[4] Bernhard Schölkopf, John Shawe-Taylor, John C. Platt, Alex J. Smola, and Robert C. Williamson, "Estimating the Support of a High-Dimensional Distribution," Tech. Rep., 2000.

[5] David M.J. Tax and Robert P.W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[6] J. Ma and S. Perkins, "Time-series novelty detection using one-class support vector machines," in *Proceedings of the International Joint Conference on Neural Networks*, July 2003, vol. 3, pp. 1741–1745.

[7] Mennatallah Amer, Markus Goldstein, and Slim Abdennadher, "Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection," in *ACM SIGKDD Workshop on Outlier Detection and Description (ODD)*, New York, New York, USA, 2013, pp. 8–15, ACM Press.

[8] Wei Liu, Gang Hua, and John R. Smith, "Unsupervised One-Class Learning for Automatic Outlier Removal," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014, pp. 3826–3833, Ieee.

[9] Alfred M. Bruckstein, David L. Donoho, and Michael Elad, "From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images," *Society for Industrial and Applied Mathematics (SIAM) review*, vol. 51, no. 1, pp. 34–81, 2009.

[10] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *IEEE Computer Science Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[11] Li Wei, Eamonn Keogh, and Xiaopeng Xi, "SAXually Explicit Images: Finding Unusual Shapes," in *Proceedings of the 6th International Conference on Data Mining (ICDM)*, Washington, DC, USA, 2006, pp. 711–720, IEEE Computer Society.

[12] C. Rusu, B. Dumitrescu, and S.A. Tsaftaris, "Explicit Shift-Invariant Dictionary Learning," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 6–9, Jan 2014.

[13] Sotirios A. Tsaftaris, Xiangzhi Zhou, Richard Tang, Debiao Li, and Rohan Dharmakumar, "Detecting Myocardial Ischemia at Rest With Cardiac Phase–Resolved Blood Oxygen Level–Dependent Cardiovascular Magnetic Resonance," *Circulation: Cardiovascular Imaging*, vol. 6, no. 2, pp. 311–319, 2013.

[14] Cristian Rusu, Rita Morisi, Davide Boschetto, Rohan Dharmakumar, and Sotirios A. Tsaftaris, "Synthetic Generation of Myocardial Blood-oxygen-level-dependent MRI Time Series via Structural Sparse Decomposition Modeling," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1422–1433, 2014.