ONLINE ONE-CLASS MACHINES BASED ON THE COHERENCE CRITERION

Zineb Noumir, Paul Honeine

Institut Charles Delaunay (CNRS)
Université de technologie de Troyes
10010 Troyes, France

Cédric Richard

Laboratoire H. Fizeau (CNRS)
Université de Nice Sophia-Antipolis
06108 Nice, France

ABSTRACT

In this paper, we investigate a novel online one-class classification method. We consider a least-squares optimization problem, where the model complexity is controlled by the coherence criterion as a sparsification rule. This criterion is coupled with a simple updating rule for online learning, which yields a low computational demanding algorithm. Experiments conducted on time series illustrate the relevance of our approach to existing methods.

Index Terms— support vector machines, kernel methods, one-class classification, online learning, coherence parameter

1. INTRODUCTION

One-class classification for novelty detection has recently generated a great interest in the machine learning community. When time series are considered, an adaptive scheme is required for online detection. In an online learning scenario, training data are available one sample at a time, as opposed to the batch mode where all the samples are presented to the system at the same time. An online learning is also advantageous when dealing with very large datasets. Several applications of interest in signal processing include audio and speech segmentation [1] and wireless sensor networks [2].

Several methods have been developed to solve the one-class problem, the most widely studied being the one-class support vector machines (SVM) [3]. It determines a sphere of minimum volume that encloses all (or most of) the available data, by estimating its center and radius. One-class SVM takes advantage of many properties from SVM literature, such as the nonlinear extension by using kernel functions and the sparseness of the center. The sparsity property states that the center of the sphere explores only a small fraction of the training samples, known as support vectors (SVs). Quadratic programming techniques are often applied to solve this problem. Such approach is inappropriate for online learning.

Many online learning methods have been proposed for SVM in binary classification problems; see for instance [4, 5] and references therein. So far, there are few attempts to establish online versions of the one-class SVM. In [6], it is argued that the binary classification algorithm in [5] cannot be directly implemented for the one-class problem. In [7], a modified formulation of the one-class SVM is presented, following [8] where an exponential window is applied to the data. Still, this technique is based on the slow-varying assumption, and several approximations were applied. In another approach, a novelty detection test can also be defined by comparing two one-class SVMs, one trained on the sliding window before the present instance, and one on the sliding window after it. A similar approach is considered in [2] with an application on wireless sensor networks. Nevertheless, this approach is not an online one-class technique.

In order to derive an online version of the one-class SVM machines, the main difficulty remains in the nature of the constrained quadratic optimization problem. Inspired from the least-squares SVM, a one-class technique is proposed in [9]. However, as pointed out by the authors, it does not have a decision function, and moreover, it loses the sparseness, thus inappropriate for online detection. In this paper, we revisit the one-class problem by considering a least-squares estimation of the center of the sphere. The sparsity of the solution is controlled online by the coherence parameter, borrowed from the literature on sparse approximation problems with dictionaries [10]. The coherence parameter designates the greatest correlation between elements of a dictionary (or two dictionaries) [11, 12]. In signal processing, the quality of representing a signal using a dictionary is studied in [13], while in [14] the authors use the coherence parameter for signal processing in multichannel transmissions and for source separation problems. More recently, we have adapted the coherence parameter to a dictionary of kernel functions, and applied it with success for online prediction of time series data with kernel functions [15, 16]. The coherence criterion provides an elegant model reduction criterion with a less computationally demanding procedure. In this paper, we associate this criterion with a one-class classification algorithm by solving a least-squares optimization problem.

The rest of the paper is organized as follows. Section 2 outlines the one-class SVM. We present our approach in Sec-

This work was supported by ANR-08-SECU-013-02 VigiRes'Eau.

© EURASIP, 2012 - ISSN 2076-1465

tion 3. Algorithms for offline and online learning are described in Section 4. Section 5 provides an experimental study on time series.

2. ONE-CLASS SVM

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be the set of training samples, et $\Phi(\cdot)$ be a nonlinear transformation defined by the use of a kernel, $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$, such as the Gaussian kernel $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2)$ where σ is the bandwidth parameter. The one-class SVM[1] finds a sphere, of minimum volume, containing most of the samples. This sphere, described by its center $\boldsymbol{c}$ and its radius r, is obtained by solving the constrained optimization problem:

$$\min_{r,\boldsymbol{c},\boldsymbol{\zeta}} \; r^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \zeta_i$$
$$\text{subject to } \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\|^2 \leq r^2 + \zeta_i \text{ for all } i$$

In this expression, ν is a positive parameter that specifies the tradeoff between the sphere volume and the number of outliers, i.e., samples lying outside the sphere. By introducing the KKT optimality conditions, we get $\boldsymbol{c} = \sum_{i=1}^{n} \alpha_i \, \Phi(\boldsymbol{x}_i)$, where the α_i's are the solution to the optimization problem:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
$$\text{subject to} \sum_{i=1}^{n} \alpha_i = 1 \text{ and } 0 \leq \alpha_i \leq \frac{1}{\nu n} \text{ for all } i \qquad (1)$$

It is well known that only a small fraction of the training samples contributes to the above model. These samples, called support vectors (SVs), have non-zero α_i's.

In order to provide an online algorithm for one-class SVM, the main difficulty remains in the constrained optimization problem. For instance, upon arrival of a new sample, the coefficients α_i's should be updated subject to constraints (1). However, the upper bound on the α_i's depends on n, thus should be updated and accordingly the values of these coefficients. This problem is studied in [6], where 6 conditions are considered. This increases the computational complexity, which depends on the stability of the solution when new samples are added.

[1]There exists another formulation of the one-class classification problem: Let $\boldsymbol{w}$ defines the hyperplane separating the origin from the projections of the samples, in the feature space, then the optimization problem is

$$\min_{r,\boldsymbol{w},\boldsymbol{\zeta}} \|\boldsymbol{w}\|^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \zeta_i + r \quad \text{subject to } \langle \boldsymbol{w}, \Phi(\boldsymbol{x}_i) \rangle \geq r - \zeta_i \text{ for all } i$$

with the decision function given by $\langle \boldsymbol{w}, \Phi(\boldsymbol{x}_n) \rangle$. The equivalence between both formulations, i.e., between $\boldsymbol{w}$ and $\boldsymbol{c}$, is studied in [3].

3. PROPOSED ONE-CLASS APPROACH

Consider the estimation of the center from n available samples, namely

$$\boldsymbol{c}_n = \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i). \qquad (2)$$

Then, any new sample $\boldsymbol{x}$ that satisfies $\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_n\| > r$ can be considered as an outlier, where the distance is

$$\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_n\|^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) - \frac{2}{n} \sum_{i=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + \kappa(\boldsymbol{x}, \boldsymbol{x}).$$

This formulation is inappropriate for large scale data, and unsuitable for online learning as n grows infinitely. The use of a sparse solution for the center provides a robust formulation, appropriate for simple online and offline learning algorithms.

Let $\boldsymbol{c}_{\mathcal{I}}$ denotes a sparse model of the center $\boldsymbol{c}_n$ by using a small subset of the available samples with

$$\boldsymbol{c}_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \alpha_i \, \Phi(\boldsymbol{x}_i), \qquad (3)$$

where $\mathcal{I} \subset \{1, 2, \ldots, n\}$, and let $|\mathcal{I}|$ denotes the cardinality of this subset. The distance of any $\Phi(\boldsymbol{x})$ to $\boldsymbol{c}_{\mathcal{I}}$ is

$$\|\Phi(\boldsymbol{x}) - \boldsymbol{c}\|^2 = \sum_{i,j \in \mathcal{I}} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) - 2 \sum_{i \in \mathcal{I}} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + \kappa(\boldsymbol{x}, \boldsymbol{x}). \qquad (4)$$

The optimization problem consists of properly identifying this subset and estimating the optimal coefficients α_i's. A joint optimization procedure requires advances techniques, as illustrated above with the one-class SVM. In this paper, we propose a separate optimization scheme:

Step 1 Select the most relevant samples in the expansion (3), by using the "coherence" as a sparsification criterion.

Step 2 Estimate the optimal coefficients α_i's, in the sense of the least-squares sense, namely $\min_{\boldsymbol{\alpha}} \|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|^2$.

Next, we propose two different settings, an offline and an online scheme, and derive appropriate learning algorithms.

4. OFFLINE AND ONLINE ONE-CLASS METHODS

4.1. Coherence parameter

The coherence is a fundamental quantity for characterizing dictionaries in sparse approximation problems [10]. It designates the greatest correlation between the elements of a dictionary. For a dictionary of unit-norm[2] elements, $\{\Phi(\boldsymbol{x}_i) \mid i \in \mathcal{I}\}$, the coherence parameter is defined by [16]

$$\mu = \max_{\substack{i,j \in \mathcal{I} \\ i \neq j}} |\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle| = \max_{\substack{i,j \in \mathcal{I} \\ i \neq j}} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)|. \qquad (5)$$

[2]A unit-norm element satisfies $\|\Phi(\boldsymbol{x})\| = 1$ for every $\boldsymbol{x}$, namely $\kappa(\boldsymbol{x}, \boldsymbol{x}) = 1$; otherwise, substitute $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) / \sqrt{\kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) \, \kappa(\boldsymbol{x}_j, \boldsymbol{x}_j)}$ for $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in the expression of the coherence parameter.

This parameter corresponds to the largest absolute value of the off-diagonal entries in the Gram (kernel) matrix, i.e., matrix with entries $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i, j \in \mathcal{I}$. We say that the dictionary is μ-coherent. Accordingly, an orthonormal basis is 0-coherent, while dictionaries with at least two identical elements are 1-coherent.

In an offline learning scheme, we consider the subset with the least coherence, the number of elements being fixed in advance. One can for instance consider the Gram matrix of all entries, followed by a pruning procedure by removing the entries with the largest off-diagonal values.

4.2. Optimal parameters

Once the elements of the expression (3) identified, we estimate the optimal coefficients α_i's for $i \in \mathcal{I}$. Let $\boldsymbol{\alpha}$ be the column vector of these coefficients. We consider the minimization of the approximation error $\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|$, namely

$$\boldsymbol{\alpha} = \arg\min_{\substack{\alpha_i \\ i \in \mathcal{I}}} \left\| \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i) - \sum_{i \in \mathcal{I}} \alpha_i \Phi(\boldsymbol{x}_i) \right\|^2.$$

By taking the derivative with respect to each α_k, and setting it to zero, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \kappa(\boldsymbol{x}_k, \boldsymbol{x}_i) = \sum_{i \in \mathcal{I}} \alpha_i \, \kappa(\boldsymbol{x}_k, \boldsymbol{x}_i), \text{ for each } k \in \mathcal{I}.$$

Written in matrix form, we get $\boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{\kappa}$, where $\boldsymbol{K}$ is the Gram (kernel) matrix with entries $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i, j \in \mathcal{I}$ and $\boldsymbol{\kappa}$ is the column vector with entries $\frac{1}{n} \sum_{i=1}^{n} \kappa(\boldsymbol{x}_k, \boldsymbol{x}_i)$ for each $k \in \mathcal{I}$. The final solution is given by

$$\boldsymbol{\alpha} = \boldsymbol{K}^{-1} \boldsymbol{\kappa}. \tag{6}$$

This problem is well-posed[3], since the Gram matrix $\boldsymbol{K}$ is non-singular. In fact, for a μ-coherent dictionary of $|\mathcal{I}|$ elements, the eigenvalues of its Gram matrix are greater than or equal to $1 - (|\mathcal{I}| - 1)\mu$. This property was previously derived by two of the authors of this paper [15, Proposition 1].

4.3. Online Coherence Criterion

In an online learning scheme, we have a new sample at each time step. The sparsification rule determines whether, at time step $n+1$, the candidate $\Phi(\boldsymbol{x}_{n+1})$ can be well approximated by a combination of the elements of the dictionary. If not, it is added to the dictionary. The coherence-based sparsification criterion consists of inserting $\Phi(\boldsymbol{x}_{n+1})$ in the dictionary provided that its coherence remains below a given threshold μ_0, namely

$$\max_{i \in \mathcal{I}} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{n+1})| \leq \mu_0. \tag{7}$$

[3]The well-posedness of our approach show its relevance to other methods that are numerically unstable such as [6, See Section 2.4].

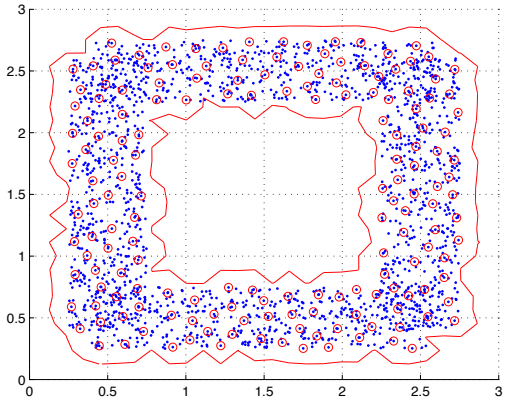

Fig. 1. The online method applied on the "frame" distribution. Samples are given in dots ·, and elements of the dictionary are shown with circles ○. The contour is given by the distance to the estimated center.

The value of the parameter $\mu_0 \in [0, 1[$ determines the level of sparsity. In [16, Proposition 2], we have demonstrated that, for a compact subspace, the dimension of the dictionary determined with the above sparsification rule remains finite as n goes to infinity.

Let $\boldsymbol{\alpha}_n$ be the coefficients estimated at time step n, and $\boldsymbol{K}_n$ and $\boldsymbol{\kappa}_n$ the corresponding (Gram kernel) matrix and vector, respectively. Then, the optimal solution from (6) is

$$\boldsymbol{\alpha}_n = \boldsymbol{K}_n^{-1} \boldsymbol{\kappa}_n. \tag{8}$$

4.4. Online update scheme

Upon the arrival of a new sample at time $n+1$, we consider the coherence criterion (7), to determine whether the model order remains unchanged or is incremented by including the new element in the dictionary.

First case: $\max_{i \in \mathcal{I}} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{n+1})| > \mu_0$

In this case, the new candidate entry $\Phi(\boldsymbol{x}_{n+1})$ is not included in the dictionary. Accordingly, the coefficients α_i's are updated, in order to approximate the discarded new entry by the elements of the the expansion (3). Since the dictionary remains unchanged, we have $\boldsymbol{K}_{n+1} = \boldsymbol{K}_n$. The only change is within the vector $\boldsymbol{\kappa}_n$ in (8), which becomes

$$\boldsymbol{\kappa}_{n+1} = \frac{1}{n+1} \big(n \, \boldsymbol{\kappa}_n + \boldsymbol{b} \big)$$

where $\boldsymbol{b}$ is the column vector with entries $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{n+1})$ for all $i \in \mathcal{I}$. We get the updating rule of $\boldsymbol{\alpha}_{n+1}$ from $\boldsymbol{\alpha}_n$ as follows:

$$\begin{aligned} \boldsymbol{\alpha}_{n+1} &= \boldsymbol{K}_{n+1}^{-1} \boldsymbol{\kappa}_{n+1} \\ &= \frac{n}{n+1} \boldsymbol{\alpha}_n + \frac{1}{n+1} \boldsymbol{K}_n^{-1} \boldsymbol{b}. \end{aligned}$$

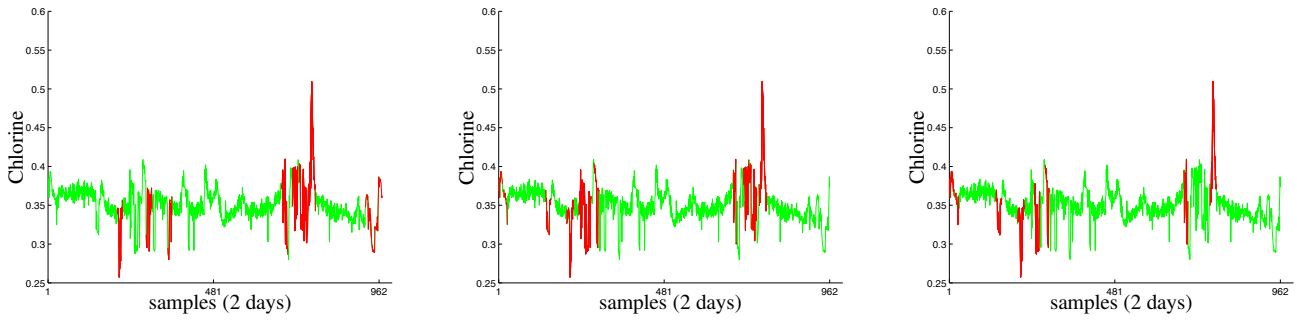

Fig. 2. The time series, with model elements illustrated (in red), for the proposed offline (left figure) and online (middle figure) algorithms, and the adaptive one-class SVM (right figure).

Second case: $\max_{i\in\mathcal{I}} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{n+1})| \leq \mu_0$

In this case, the new candidate $\Phi(\boldsymbol{x}_{n+1})$ is included into the dictionary. The number of terms in the model (3) is incremented, with the dictionary being determined by $\mathcal{I}\cup\{n+1\}$. This leads to the following expressions of the new Gram matrix

$$\boldsymbol{K}_{n+1} = \begin{bmatrix} \boldsymbol{K}_n & \boldsymbol{b} \\ \boldsymbol{b}^\top & \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1}) \end{bmatrix}.$$

In order to determine $\boldsymbol{K}_{n+1}^{-1}$, we use the the Woodbury matrix identity to get the inverse of $\boldsymbol{K}_{n+1}$ from $\boldsymbol{K}_n^{-1}$:

$$\boldsymbol{K}_{n+1}^{-1} = \begin{bmatrix} \boldsymbol{K}_n^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{bmatrix} + \frac{1}{c} \begin{bmatrix} -\boldsymbol{K}_n^{-1}\boldsymbol{b} \\ 1 \end{bmatrix} \begin{bmatrix} -\boldsymbol{b}^\top \boldsymbol{K}_n^{-1} & 1 \end{bmatrix},$$

where $c = \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1}) - \boldsymbol{b}^\top \boldsymbol{K}_n^{-1}\boldsymbol{b}$ and $\boldsymbol{0}$ is a column vector of zeros of appropriate size.

The vector $\boldsymbol{\kappa}_{n+1}$ is updated from $\boldsymbol{\kappa}_n$, with

$$\boldsymbol{\kappa}_{n+1} = \frac{1}{n+1} \begin{bmatrix} n\,\boldsymbol{\kappa}_n + \boldsymbol{b} \\ \kappa_{n+1} \end{bmatrix}$$

where $\kappa_{n+1} = \sum_{i=1}^{n+1} \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_i)$. The latter expression requires having all the samples in memory. Still one can overcome this difficulty by considering an instant estimation, with $\kappa_{n+1} = (n+1)\,\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1})$. It is worth noting that the second case, i.e., the incrementation of the model order, seldom occurs.

Finally, by combining these expressions, we get the following update of $\boldsymbol{\alpha}_{n+1}$ from $\boldsymbol{\alpha}_n$:

$$\begin{aligned}\boldsymbol{\alpha}_{n+1} =& \frac{1}{n+1} \begin{bmatrix} n\boldsymbol{\alpha}_n + \boldsymbol{K}_n^{-1}\boldsymbol{b} \\ 0 \end{bmatrix} \\ & - \frac{1}{(n+1)\,c} \begin{bmatrix} \boldsymbol{K}_n^{-1}\boldsymbol{b} \\ 1 \end{bmatrix} [n\boldsymbol{b}^\top \boldsymbol{\alpha}_n + \boldsymbol{b}^\top \boldsymbol{K}_n^{-1}\boldsymbol{b} - \kappa_{n+1}].\end{aligned}$$

5. EXPERIMENTATIONS

Toy dataset

In order to illustrate our approach, we considered a 2D toy dataset. The training set consisted of 2 000 samples drawn from a "frame" distribution, as illustrated in Figure 1. For such large scale dataset, classical one-class SVM algorithms cannot be applied. In our experiments, we used the Gaussian kernel, with its bandwidth set to $\sigma = 0.5$. We applied the proposed online algorithm, as given in Section 4.4. The coherence threshold[4] in (7) was set to $\mu_0 = 0.01$, which led to a model with 167 elements (8% of the training data). The distance, computed with (4) and shown with the contours in Figure 1, illustrates the boundary of the (hyper)sphere.

Time series domain description

We conducted some experiments on a time series. It consists of the variation in chlorine concentration at a given node in a water network. Chlorine is a highly efficient disinfectant, injected in water supplies to kill residual bacteria. Chlorination is the process of adding chlorine to water as a method of water purification to make it fit for human consumption as drinking water. The measurements of chlorine were taken from the public water supplies of the Cannes city in France. We considered 2 days of chlorine concentration measures, with measurements sampled at the rate of a sample every 3 minutes. These time series exhibit large fluctuations due to the variations in water consumption and an inefficient control system. See Figure 2.

To capture the structure of the time series, a 10-length sliding window was used, with $\boldsymbol{x}_i = [x_{i-9} \;\cdots\; x_{i-1} \;\; x_i]$, where the Gaussian kernel was applied and $\sigma = 0.2$ for all algorithms. We compared the proposed algorithms with two algorithms from the literature: the one-class SVM algorithm [3]

[4]The coherence threshold plays the same role as the ν parameter in SVM. Preliminary experiments are often conducted to select the appropriate values. A comparative study with different values is not provided due to space restrictions.

and the adaptive one-class SVM [7]. Figure 2 shows the elements retained in the model, obtained either by pruning as given by the coherence parameter in (5) (offline algorithm) or by using the coherence criterion in (7) (online algorithm). For the online algorithm, the coherence threshold was set to $\mu_0 = 0.5$, which led to a model with 49 elements. To get comparable results, the number of elements was set to 49 for the offline algorithm, and the parameters of one-class SVM and adaptive one-class SVM were set accordingly. It is obvious from Figure 2 that the retained elements (in red) are relevant in describing the fluctuations in the time series. The pertinence of the proposed model update rules are shown with the the decision function (distance computed with (4)), as given in Figure 3. This illustrates that our approach, online and offline, is comparable to the one-class SVM and adaptive one-class SVM results, with lower computational complexity. See Table 1 for the computational cost of these algorithms.

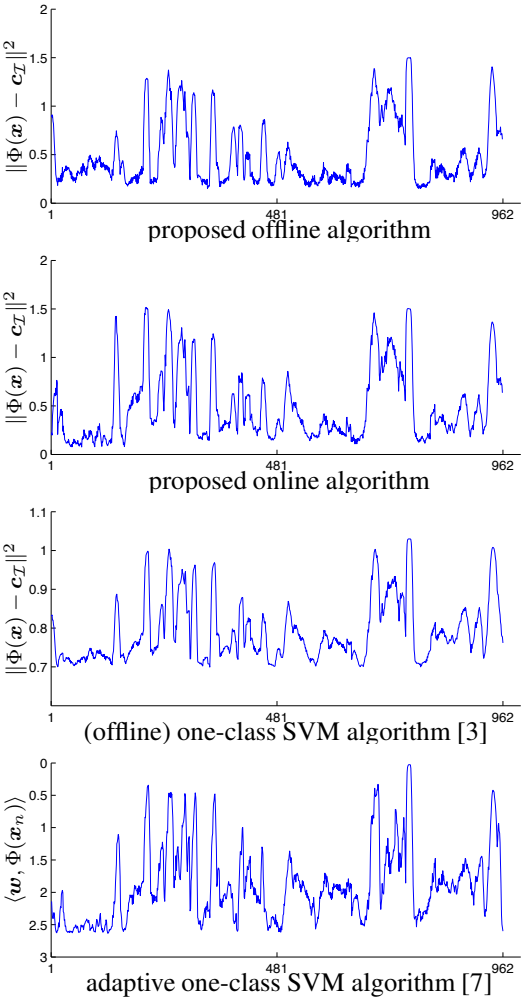

Fig. 3. The decision functions estimated by several algorithms. A threshold can be introduced for novelty detection (not illustrated here).

offline algorithm	online algorithm	one-class SVM [3]	adaptive [7]
3.7	0.6	141.4	1.8

Table 1. Estimated computational time (in seconds) of different algorithms for the real time series.

6. CONCLUSION

In this paper, we investigated a novel online one-class classification method, by associating the coherence parameter with a least-squares optimization problem. Future works include a convergence study and the use of a step-size parameter.

7. REFERENCES

[1] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation," in *IEEE ICASSP-02*, 2002, pp. 1313–1316.

[2] Y. Zhang, N. Meratnia, and P. Havinga, "Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks," in *Proc. International Conference on Advanced Information Networking and Applications Workshops*, Washington, DC, USA, 2009, pp. 990–995.

[3] D. Tax, "One-class classification," PhD thesis, Delft University of Technology, Delft, June 2001.

[4] J. K. Anlauf and M. Biehl, "The AdaTron: an adaptive perceptron algorithm," *Europhys. Letters*, vol. 10, 1989.

[5] G. Cauwenberghs and T. Poggio, "Incremental and decremental support vector machine learning," in *Advances in Neural Information Processing Systems*, vol. 13, 2001.

[6] M. Davy, F. Desobry, A. Gretton, and C. Doncarli, "An Online Support Vector Machine for Abnormal Events Detection," *Signal Processing*, vol. 86, no. 8, pp. 2009–2025, 2006.

[7] V. Gómez-Verdejo, J. Arenas-García, M. Lázaro-Gredilla, and A. Navia-Vázquez, "Adaptive one-class support vector machine," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2975–2981, 2011.

[8] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," vol. 52, no. 8, Aug 2004.

[9] Y.-S. Choi, "Least squares one-class support vector machine," *Pattern Recogn. Lett.*, vol. 30, pp. 1236–1240, October 2009.

[10] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, pp. 2231–2242, 2004.

[11] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, pp. 2845–2862, 2001.

[12] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2558–2567, 2002.

[13] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, "Approximation of functions over redundant dictionaries using coherence," in *Proc. of SODA*, 2003, pp. 243–252.

[14] R. Gribonval and P. V, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Trans. Inform. Theory*, vol. 52, pp. 255–261, 2006.

[15] P. Honeine, C. Richard, and J. C. M. Bermudez, "On-line nonlinear sparse approximation of functions," in *Proc. IEEE International Symposium on Information Theory*, Nice, France, June 2007, pp. 956–960.

[16] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, March 2009.