SHRINKAGE METHODS FOR ONE-CLASS CLASSIFICATION

Patric Nader, Paul Honeine, Pierre Beauseroy

Institut Charles Delaunay (CNRS), Université de Technologie de Troyes, France
{patric.nader, paul.honeine, pierre.beauseroy}@utt.fr

ABSTRACT

Over the last decades, machine learning techniques have been an important asset for detecting nonlinear relations in data. In particular, one-class classification has been very popular in many fields, specifically in applications where the available data refer to a unique class only. In this paper, we propose a sparse approach for one-class classification problems. We define the one-class by the hypersphere enclosing the samples in the Reproducing Kernel Hilbert Space, where the center of this hypersphere depends only on a small fraction of the training dataset. The selection of the most relevant samples is achieved through shrinkage methods, namely Least Angle Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Net. We modify these selection methods and adapt them for estimating the one-class center in the RKHS. We compare our algorithms to well-known one-class methods, and the experimental analysis are conducted on real datasets.

Index Terms— One-class classification, kernel methods, shrinkage methods

1. INTRODUCTION

Statistical machine learning, such as kernel methods, have been widely used in the past decades to discover nonlinear relations and hidden patterns in data, and they have been applied in many fields for classification and regression problems [1]. In particular, one-class classification algorithms gained a lot of interest specifically in industrial applications, where the data related to the malfunctioning modes are difficult to obtain, and the only available data designate the normal functioning modes of the studied system. One-class classifiers learn the normal behavior of the system, and provide decision functions in a way to accept as many normal samples as possible and to reject the outliers [2]. One-class algorithms have been applied in many fields, namely for face recognition applications [3], seizure analysis from EEG signals [4], and recently for intrusion detection in industrial systems [5, 6].

This work is supported by the French "Agence Nationale de la Recherche"(ANR), grant SCALA.

The authors would like to thank Thomas Morris and the Mississippi state university SCADA Laboratory for providing the real datasets.

Several formulations have been proposed to solve one-class classification problems. Schölkopf *et al.* proposed in [7] the one-class Support Vector Machines (SVM), in which they separate the data from the origin with maximum margin using a hyperplane. This approach is greedy in terms of computational cost since it requires to solve a constrained quadratic programming problem. Tax *et al.* introduced in [8] the Support Vector Data Description (SVDD) which estimates the hypersphere with minimum radius enclosing most of the training data. This approach requires also to solve a constrained quadratic programming problem, and it is equivalent to SVM when the Gaussian kernel is used. A fast approach was introduced in [9] to overcome the computational cost of quadratic programming problems, but the use of the Euclidean distance as a novelty measure leads to a high sensitivity towards outliers. A sparse approach was proposed in [10], in which the selection of the most relevant samples is based on the coherence criterion. This approach does not require the same computational costs as one-class SVM, but the sensitivity towards outliers remains unchanged. Hoffman used in [11] the Kernel Principal Component Analysis (KPCA) for one-class classification, where the data were projected into the subspace spanned by the most relevant eigenvectors of the covariance matrix. Despite the relatively low computational complexity of the reconstruction error used as a novelty measure, this approach loses the sparsity of SVM and SVDD.

In this paper, we propose a sparse approach for one-class classification, where our one-class is defined by a hypersphere enclosing the samples in a RKHS. The center of the hypersphere is the approximation of the empirical center of the data in that space, and this sparse center depends only on a small fraction of the training dataset. Since a good selection of these samples is crucial to obtain good results in sparse approaches, this selection is achieved by adapting well-known shrinkage methods, namely Least Angle Regression [12], Least Absolute Shrinkage and Selection Operator [13], and Elastic Net [14]. We modify these algorithms to the computation of the center in the RKHS. The remainder of this paper is organized as follows. Section 2 describes the proposed one-class framework and the adapted shrinkage methods. An extension to the Mahalanobis distance in proposed in Section 3. Section 4 discusses the results on real datasets, and Section 5 provides conclusion and future works.

2. PROPOSED ONE-CLASS FRAMEWORK

Consider a training set of samples $\boldsymbol{x}_i$, for $i = 1, 2, \ldots, n$, in a d-dimensional input space $\mathcal{X}$. Let $\boldsymbol{K}$ be the $n \times n$ kernel matrix with entries $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ for $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathcal{X}$, where $\phi(\boldsymbol{x})$ is the mapping function to a RKHS of some given reproducing kernel $k(\cdot, \cdot)$. The expectation of the mapped samples, namely $\mathrm{E}[\phi(\boldsymbol{x})]$, can be estimated with the empirical center in the RKHS, namely $\boldsymbol{c}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i)$. We define the one-class by the hypersphere enclosing the samples in the RKHS, and we approximate $\boldsymbol{c}_n$ with a sparse representation $\boldsymbol{c}_{\mathcal{A}}$. The sparse center $\boldsymbol{c}_{\mathcal{A}}$ is a linear combination of some of the mapped samples, namely $\boldsymbol{c}_{\mathcal{A}} = \sum_{j=1}^{n} \beta_j \phi(\boldsymbol{x}_j)$, where the coefficients β_j are obtained as detailed next. We also define the decision function of any sample $\boldsymbol{x}$ by its distance in the RKHS to the center $\boldsymbol{c}_{\mathcal{A}}$, and we fix a threshold in order to classify it as outlier or normal. The expression of the Euclidian distance in the RKHS is given by:

$$\|\phi(\boldsymbol{x}) - \boldsymbol{c}_{\mathcal{A}}\|_2^2 = k(\boldsymbol{x}, \boldsymbol{x}) - 2 \sum_{i=1}^{n} \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}) + \sum_{i,j=1}^{n} \beta_i \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

In order to obtain a sparse approach, only a small fraction of the coefficients β_j in the center's expression has to be nonzero. We need to minimize the error of approximating $\boldsymbol{c}_n$ with $\boldsymbol{c}_{\mathcal{A}}$ in a way to get a sparse representation of the training dataset. Therefore, we solve the following optimization:

$$\arg\min_{\beta_j} \Big\| \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i) - \sum_{j=1}^{n} \beta_j \phi(\boldsymbol{x}_j) \Big\|_2^2, \qquad (1)$$

subject to some sparsity-inducing constraints. We propose to revisit three well-known shrinkage approaches, namely Least Angle Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Net. These shrinkage approaches have been used for feature selection in regression to solve optimization problems of the form: $\arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ subject to some constraints, such as $\sum |\beta|$ cannot exceed some predefined threshold. These methods used for the sample selection induce sparsity, and only a small number of the coefficients remains nonzero. In the following, we adapt these methods to solve the optimization problem of equation (1), thus selecting the most relevant samples among the training dataset, where only the corresponding coefficients remain nonzero.

2.1. Least Angle Regression

The Least Angle Regression (LARS) is a model selection algorithm that builds a model sequentially by augmenting the set of the most relevant samples one sample at a time. Let $\widehat{\boldsymbol{c}}_{\mathcal{A}_k}$ be the estimation of the sparse center in the subspace $\mathcal{A}$ of the most relevant samples at step k, and $(\boldsymbol{c}_n - \widehat{\boldsymbol{c}}_{\mathcal{A}_k})$ the current residual. LARS finds the sample having the largest absolute correlation with the residual $(\boldsymbol{c}_n - \widehat{\boldsymbol{c}}_{\mathcal{A}_k})$, and projects

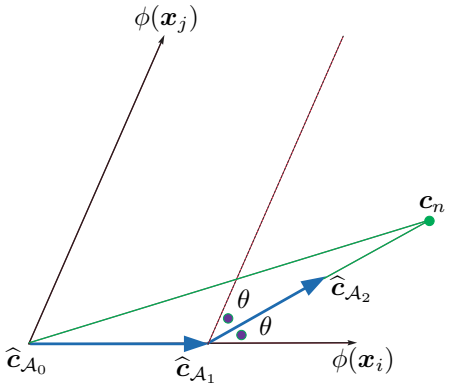

Fig. 1. An illustration of the successive LARS estimates in a simple 2-dimensional space, where the algorithm starts at $\widehat{\boldsymbol{c}}_{\mathcal{A}_0}$. In this example, the first residual $(\boldsymbol{c}_n - \widehat{\boldsymbol{c}}_{\mathcal{A}_0})$ makes a smaller angle with $\phi(\boldsymbol{x}_i)$ than with $\phi(\boldsymbol{x}_j)$, so we start moving in the direction of $\phi(\boldsymbol{x}_i)$. The next step at $\widehat{\boldsymbol{c}}_{\mathcal{A}_1}$, the current residual $(\boldsymbol{c}_n - \widehat{\boldsymbol{c}}_{\mathcal{A}_1})$ makes equal angles θ with $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x}_j)$, so we have to move in a direction that preserves this equiangularity such as $\widehat{\boldsymbol{c}}_{\mathcal{A}_2}$.

the other samples on this first one. LARS repeats the selection process until a new sample has the same correlation level with the residual, and continues in a direction equiangular between these samples until a third one enters the set of the most correlated samples, and so on. An example of the successive LARS estimates is illustrated in figure 1, where the algorithm starts at $\widehat{\boldsymbol{c}}_{\mathcal{A}_0}$, and the equiangular vectors are updated in a way to preserve equal angles with the original axes.

Let $\boldsymbol{X} = (\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_n))$, and let $\boldsymbol{X}_{\mathcal{A}}$ denotes the samples of the set $\mathcal{A}$ having the greatest absolute current correlations, and $\boldsymbol{K}_{\mathcal{A}}$ the $|\mathcal{A}| \times |\mathcal{A}|$ corresponding kernel matrix, with $|\mathcal{A}|$ the cardinality of $\mathcal{A}$. The expression of the current estimate of the sparse center has this form: $\widehat{\boldsymbol{c}}_{\mathcal{A}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. The algorithm begins with $\widehat{\boldsymbol{c}}_{\mathcal{A}_0} = \boldsymbol{0}$, and updates the center once at a time. The vector of current correlations is defined as follows:

$$\widehat{\boldsymbol{corr}} = \boldsymbol{X}^T(\boldsymbol{c}_n - \widehat{\boldsymbol{c}}_{\mathcal{A}}) = \frac{1}{n} \sum_{i,j=1}^{n} k(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i,j=1}^{n} \widehat{\beta}_j k(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

The equiangular vector needed for the projection operation has the following form:

$$\boldsymbol{u}_{\mathcal{A}} = \boldsymbol{X}_{\mathcal{A}} \boldsymbol{w}_{\mathcal{A}},$$

where $\boldsymbol{w}_{\mathcal{A}} = A_{\mathcal{A}} \boldsymbol{G}_{\mathcal{A}}^{-1} \boldsymbol{1}_{\mathcal{A}}$ is the weight vector making equal angles with the columns of $\boldsymbol{X}_{\mathcal{A}}$, $\boldsymbol{G}_{\mathcal{A}} = \boldsymbol{s}^T \boldsymbol{K}_{\mathcal{A}} \boldsymbol{s}$ is a matrix related to the set $\mathcal{A}$, $\boldsymbol{s}$ denotes the vector of the signs of the current correlations, and $A_{\mathcal{A}} = (\boldsymbol{1}_{\mathcal{A}}^T \boldsymbol{G}_{\mathcal{A}}^{-1} \boldsymbol{1}_{\mathcal{A}})^{-\frac{1}{2}}$. After finding $\boldsymbol{X}_{\mathcal{A}}$, $A_{\mathcal{A}}$, and $\boldsymbol{u}_{\mathcal{A}}$, the current estimate $\widehat{\boldsymbol{c}}_{\mathcal{A}}$ is updated to $\widehat{\boldsymbol{c}}_{\mathcal{A}+} = \widehat{\boldsymbol{c}}_{\mathcal{A}} + \widehat{\gamma} \boldsymbol{u}_{\mathcal{A}}$ using the equiangular vector, where

$$\widehat{\gamma} = \min_{j=1,\cdots,|\mathcal{A}|} \Big\{ \frac{\widehat{C} - \widehat{corr}_j}{A_{\mathcal{A}} - a_j}, \frac{\widehat{C} + \widehat{corr}_j}{A_{\mathcal{A}} + a_j} \Big\},$$

having $\min$ the minimum over the positive components, a_j an element of the inner product vector defined by

$$\boldsymbol{a} = \boldsymbol{X}^T \boldsymbol{u}_{\mathcal{A}} = \boldsymbol{X}^T \boldsymbol{X}_{\mathcal{A}} \boldsymbol{w}_{\mathcal{A}} = \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{A}|} k(\boldsymbol{x}_i, \boldsymbol{x}_j) \boldsymbol{w}_{\mathcal{A}},$$

and $\widehat{C} = \max_j\{|\widehat{corr}_j|\}$. Finally, the coefficients $\boldsymbol{\beta}$ are updated as follows:

$$\boldsymbol{\beta}_{new} = \widehat{\boldsymbol{\beta}} + \widehat{\gamma} \boldsymbol{s}^T \boldsymbol{w}_{\mathcal{A}}. \quad (2)$$

The main drawback of LARS is when dealing with highly correlated samples, which may limit its application to high dimensional data.

2.2. Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO) minimizes the residual sum of squares under a constraint on the ℓ_1-norm of the coefficient vector. LASSO solves the following optimization problem:

$$\arg\min_{\beta_j} \big\| \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i) - \sum_{j=1}^{n} \beta_j \phi(\boldsymbol{x}_j) \big\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1, \quad (3)$$

for a given $\lambda > 0$, where the ℓ_1-based regularization term induces sparsity in the solution. LASSO shrinks the estimated coefficients towards the origin and sets some of them to zero, in a way to retain the most relevant samples and to discard the other ones. In fact, the LASSO solutions can be generated by some modifications of the LARS algorithm. Unlike in LARS, the coefficients in LASSO cannot change signs during the update since they are piecewise linear, and the sign of any nonzero coefficient β_j must agree with the sign s_j of the current correlation $\widehat{corr}_j$ for any j [12]. To update the coefficients as in equation (2), we have $\beta_j(\gamma) = \widehat{\beta}_j + \gamma s_j w_{\mathcal{A}j}$ for any j. Therefore, $\beta_j(\gamma)$ changes sign at:

$$\gamma_j = -\frac{\beta_j}{s_j w_{\mathcal{A}j}},$$

having the first such change occurring at $\widetilde{\gamma} = \min_{\gamma_j > 0}\{\gamma_j\}$. The sign restriction is violated when $\widetilde{\gamma} < \widehat{\gamma}$, and $\beta_j(\gamma)$ cannot be a LASSO solution; $\beta_j(\gamma)$ has changed sign while $c_j(\gamma)$ has not. The sample having the corresponding index j is removed from the set of the most relevant samples, namely $\mathcal{A} = \mathcal{A} \setminus \{\boldsymbol{x}_j\}$, and the algorithm moves to the next equiangular direction. Therefore, this modification allows the active set to increase or decrease one at a time until the LARS algorithm leads to all LASSO solutions.

2.3. Elastic Net

The elastic net (LARSEN) is a LARS-derived regularization method that linearly combines the ℓ_1 and ℓ_2 penalties of LASSO and ridge methods. Similarly to LASSO, LARSEN does both continuous shrinkage and variable selection, and it produces a sparse model. In addition, unlike LASSO, LARSEN has a grouping effect where strongly correlated samples are in or out of the model together.

The entire LARSEN solution path can be directly computed from the LARS algorithm. Indeed, the naïve LARSEN optimization problem is defined as follows:

$$\arg\min_{\beta_j} \big\| \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i) - \sum_{j=1}^{n} \beta_j \phi(\boldsymbol{x}_j) \big\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2,$$

for $\lambda_1, \lambda_2 > 0$. This problem becomes a pure LASSO optimization when $\lambda_2 = 0$, and a simple ridge regression when $\lambda_1 = 0$. The naïve LARSEN problem can be transformed into an equivalent LASSO problem as in equation (3), where the parameter λ is replaced by $\lambda_1/\sqrt{1+\lambda_2}$ [14]. Therefore, as detailed in the previous section, a simple modification in the LARS algorithm leads the LASSO solution path. This kind of approximation incurs a double amount of shrinkage introducing unnecessary extra bias, compared with pure LASSO or ridge shrinkage. In order to improve the prediction performance and undo shrinkage, the coefficients of the naïve version of LARSEN are rescaled to obtain the LARSEN coefficients as follows:

$$\boldsymbol{\beta}_{(\text{LARSEN})} = (1+\lambda_2)\boldsymbol{\beta}_{(\text{naïve LARSEN})}.$$

Therefore, the LARS algorithm leads to the LARSEN solution paths. An example that highlights the differences in the solution paths of LARS, LASSO and LARSEN algorithms is illustrated in figure 2.

3. EXTENSION TO THE MAHALANOBIS DISTANCE

Since the Euclidian distance is sensitive to the scale in each direction, we also use the Mahalanobis distance in the decision function of the classifier. In fact, the Mahalanobis distance takes into account the covariance in each feature direction and the different scaling of the coordinate axes [15]. The Mahalanobis distance is computed in the RKHS as detailed in [16]:

$$\sum_{k=1}^{n} \frac{1}{\lambda_k} \Big(\sum_{i=1}^{n} \alpha_i^k k(\boldsymbol{x}_i, \boldsymbol{x}) - \sum_{i=1}^{n} \alpha_i^k \frac{1}{n} \sum_{j=1}^{n} k(\boldsymbol{x}_j, \boldsymbol{x}) - \sum_{i,j=1}^{n} \alpha_i^k \beta_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n} \alpha_i^k \frac{1}{n} \sum_{j,l=1}^{n} \beta_l k(\boldsymbol{x}_j, \boldsymbol{x}_l) \Big)^2,$$

where $n\lambda_k$ and $\boldsymbol{\alpha}^k$ are the eigenvalues and eigenvectors of the centered version of $\boldsymbol{K}$. We make use of the advantages in KPCA, and only the most relevant eigenvectors are taken into consideration while the remaining ones are considered as noise. We also adopt the kernel whitening normalization of the eigenvectors as proposed in [17], where the variance of the mapped data is constant for all the feature directions.

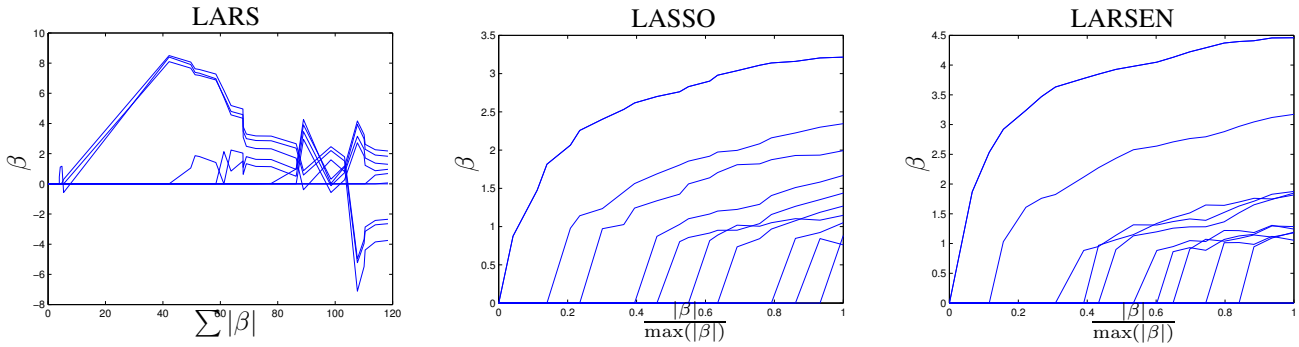

Fig. 2. The solution paths of LARS, LASSO and LARSEN algorithms. The LARS solution paths are the most unstable, while LARSEN has smoother solution paths that clearly show the "grouping effect" advantage of correlated variables over the LASSO.

4. EXPERIMENTAL RESULTS

The proposed one-class approach is tested on two real datasets from the Mississipi State University SCADA Laboratory, the gas pipeline and the water storage tank testbeds [18]. The gas pipeline is used to move petroleum products to the market, and the water storage tank is similar to the oil storage tanks found in the petrochemical industry. These real datasets raise many challenges, where each input sample consists of 27 attributes for the gas pipeline and 24 attributes for the water storage tank, i.e., gas pressure, water level, pump state, target gas pressure/water level, PID's parameters, time interval, length of the packets, and command functions. Furthermore, 28 types of attacks are injected into the network traffic of the system in order to hide its real functioning state and to disrupt the communication. These attacks are arranged into 7 groups: Naive Malicious Response Injection (NMRI), Complex Malicious Response Injection (CMRI), Malicious State Command Injection (MSCI), Malicious Parameter Command Injection (MPCI), Malicious Function Command Injection (MFCI), Denial of Service (DOS) and Reconnaissance Attacks (RA). See [18] for more details.

The Gaussian kernel is used in this paper, for it is the most common kernel for one-class problems. The expression of this kernel is given by $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2})$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are two input samples, and $\|\cdot\|_2$ represents the l_2-norm in the input space. The bandwidth parameter σ is computed as proposed in [6], namely $\sigma = \frac{d_{\max}}{\sqrt{2M}}$, where $d_{\max}$ refers to the maximal distance between any two samples in the input space, and M represents the upper bound on the number of outliers among the training dataset. We compared our approach with two other approaches, Support Vector Data Description and Kernel Principal Component Analysis. The selection of the most relevant samples in the proposed approach is performed via the aforementioned shrinkage algorithms, namely LARS, LASSO and LARSEN. In each case of these three approaches, the decision function of the classifier is defined using the Euclidean distance and the Mahalanobis distance. The sparse center in the proposed approach depends only on 10% of the training samples. We tested these algorithms on nearly 100 000 samples related to the aforementioned attacks, and the detection rates are given in Tables 1 and 2. The Mahalanobis distance outperforms in general the Euclidean distance, due to the scale sensitivity of the latter one. LARS and LASSO have nearly the same results, where LARSEN outperforms both shrinkage algorithms. The best results are achieved when LARSEN is used to select the most relevant samples, and the used norm in the decision function is the Mahalanobis distance. The latter combination gives better detection rates than the other approaches, especially SVDD and KPCA, for several types of attacks. Table 3 shows the estimated time for each algorithm, and it indicates that the proposed approach is faster than SVDD and KPCA regardless of the used shrinkage algorithm. The fastest algorithm is LARS, and the slowest one is SVDD in which a quadratic programming problem has to be resolved. Therefore, combining the Mahalanobis distance with LARSEN leads to the best detection rates, and it is faster than both SVDD and KPCA.

5. CONCLUSION

In this paper, we proposed a sparse one-class classification approach, where the selection of the most relevant samples among the training dataset is achieved through well-known shrinkage methods, namely LARS, LASSO and LARSEN. We modified these methods and we adapted their algorithms to estimate a sparse center in a RKHS. We tested our algorithms on real datasets from the Mississipi State University SCADA Laboratory, and we compared the results with well-known one-class classification approaches, namely SVDD and KPCA. The tests showed that combining LARSEN with the Mahalanobis distance results in an approach having the best detection rates and the fastest algorithm.

For future works, a further and more detailed study on the existing subset selection algorithms is required, since it is a very important step in sparse approaches in order to obtain significant results. In addition, multimode classification algorithms could be studied for intrusion detection in industrial

Table 1. Error detection probabilities for the gas pipeline testbed.

			Euclidean distance			Mahalanobis distance		
	SVDD	KPCA	LARS	LASSO	LARSEN	LARS	LASSO	LARSEN
NMRI	98.1	98.7	98.3	98.7	99.1	99.1	98.9	**99.2**
CMRI	99.5	**99.8**	98.1	98.3	99.2	98.7	98.8	99.5
MSCI	**89.1**	86.2	55.8	57.3	68.1	71.1	74.5	79.3
MPCI	98.2	98.6	97.1	96.7	97.8	98.2	97.6	**98.9**
MFCI	**89.9**	89.3	77.8	80.1	83.6	81.3	82.7	85.9
DOS	96.1	96.8	96.1	96.9	97.1	97.3	97.2	**97.5**
RA	**99.8**	**99.8**	99.1	99.5	**99.8**	99.6	99.7	**99.8**

Table 2. Error detection probabilities for the water storage testbed.

			Euclidean distance			Mahalanobis distance		
	SVDD	KPCA	LARS	LASSO	LARSEN	LARS	LASSO	LARSEN
NMRI	95.1	97.1	93.4	91.7	94.7	97.4	94.1	**98.1**
CMRI	61.2	**75.3**	59.1	62.4	69.2	71.8	67.7	74.1
MSCI	97.3	98.1	97.1	97.4	97.9	98.1	98.1	**98.3**
MPCI	98.6	99.5	98.9	97.9	99.1	99.1	98.4	**99.7**
MFCI	97.9	**99.9**	97.1	98.4	99.1	99.1	99.3	99.8
DOS	71.7	79.9	72.3	71.2	74.7	81.1	79.1	**82.6**
RA	97.8	**99.5**	98.1	98.4	98.7	99.1	99.3	**99.5**

systems. Furthermore, one could investigate online versions of the proposed algorithms in order to improve live detection in real time applications.

Table 3. Estimated time (in seconds) of each approach.

			In this paper		
	SVDD	KPCA	LARS	LASSO	LARSEN
gas	70.23	18.31	**9.85**	13.93	14.22
water	123.72	20.1	**11.79**	14.10	15.72

REFERENCES

[1] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Annals of Statistics*, vol. 36, pp. 1171–1220, 2008.

[2] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Proceedings of the 20th Irish conference on Artificial intelligence and cognitive science*, ser. AICS'09, 2010, pp. 188–197.

[3] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T. Huang, "One-class classification for spontaneous facial expression analysis," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, April 2006, pp. 281–286.

[4] A. B. Gardner, A. M. Krieger, G. Vachtsevanos, and B. Litt, "One-class novelty detection for seizure analysis from intracranial eeg," *Journal of Machine Learning Research*, vol. 7, pp. 1025–1044, 2006.

[5] P. Nader, P. Honeine, and P. Beauseroy, "Intrusion detection in scada systems using one-class classification," in *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco, 9–13 September 2013.

[6] ——, "l_p-norms in one-class classification for intrusion detection in scada systems," *Industrial Informatics, IEEE Transactions on*, vol. 10, no. 4, pp. 2308–2317, Nov 2014.

[7] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[8] D. M. J. Tax and R. P. W. Duin, "Data domain description using support vectors," in *Proceedings of the European Symposium on Artificial Neural Networks*, 1999, pp. 251–256.

[9] Z. Noumir, P. Honeine, and C. Richard, "On simple one-class classification methods," in *Proc. IEEE International Symposium on Information Theory*, MIT, Cambridge (MA), USA, 1–6 July 2012.

[10] ——, "One-class machines based on the coherence criterion," in *Proc. IEEE workshop on Statistical Signal Processing*, Ann Arbor, Michigan, USA, 5–8 August 2012, pp. 600–603.

[11] H. Hoffmann, "Kernel pca for novelty detection," *Pattern Recognition*, vol. 40, no. 3, pp. 863 – 874, 2007.

[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[15] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, vol. 2, no. 1, Apr. 1936, pp. 49–55.

[16] P. Nader, P. Honeine, and P. Beauseroy, "Mahalanobis-based one-class classification," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Reims, France, 21–24 September 2014.

[17] D. M. J. Tax and P. Juszczak, "Kernel whitening for one-class classification," in *SVM*, 2002, pp. 40–52.

[18] T. Morris, A. Srivastava, B. Reaves, W. Gao, K. Pavurapu, and R. Reddi, "A control system testbed to validate critical infrastructure protection concepts," *International Journal of Critical Infrastructure Protection*, vol. 4, no. 2, pp. 88 – 103, 2011.