ON SIMPLE ONE-CLASS CLASSIFICATION METHODS

Zineb Noumir, Paul Honeine

Institut Charles Delaunay
Université de technologie de Troyes
10010 Troyes, France

Cédric Richard

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

ABSTRACT

The one-class classification has been successfully applied in many communication, signal processing, and machine learning tasks. This problem, as defined by the one-class SVM approach, consists in identifying a sphere enclosing all (or the most) of the data. The classical strategy to solve the problem considers a simultaneous estimation of both the center and the radius of the sphere. In this paper, we study the impact of separating the estimation problem. It turns out that simple one-class classification methods can be easily derived, by considering a least-squares formulation. The proposed framework allows us to derive some theoretical results, such as an upper bound on the probability of false detection. The relevance of this work is illustrated on well-known datasets.

1. INTRODUCTION

The one-class classification machines has become a very active research domain in machine learning [1, 2], providing a detection rule based on recent advances in learning theory. In one-class classification, the problem consists in covering a single target class of samples, represented by a training set, and separate it from *any* novel sample not belonging to the same class, *i.e.*, an outlier sample. It has been successfully applied in many novelty detection and classification tasks, including communication network performance [3], wireless sensor networks [4], forensic science [5], detection of handwritten digits [6] and objet recognition [7], only to name a few. Moreover, it has been extended naturally to binary and multiclass classification tasks, by applying a single one-class classifier to each class and subsequently combining the decision rules [8].

Since only a single class is identified, it is essentially a data domain description or a class density estimation problem, while it provides a novelty detection rule. Different methods to solve the one-class problem have been developed, initiated from the so-called one-class support vector machines (SVM) [9, 2]. The one-class classification task consists in identifying a sphere of minimum volume that englobes all (or most of) the training data, by estimating jointly its center and its radius. These methods exploit many features from conventional SVM [10], including a nonlinear extension thanks to the concept of reproducing kernels. They also inherit the robustness to outliers in the training set, by providing a sparse solution of the center. This sparse solution explores a small fraction of the training samples, called support vectors (SVs), and lying outside or on the sphere.

In one-class SVM as defined in [9, 2], the resulting convex optimization problem is often solved using a quadratic programming technique. Several efforts have been made in order to derive one-class classification machines with low computational complexity [11]. In the same sense as least-squares SVM is derived from the classical SVM method [12, 13], some attempts have been made to derive from the one-class SVM a least-squares variant, such as in [14]. However, unlike the former, the latter do not have a decision function, thus inappropriate for novelty detection.

In this paper, we propose to solve the one-class problem by decoupling the estimation of the center and the radius of the sphere englobing all (or most of) the training samples. In the same spirit as the classical one-class SVM machines, we consider a sparse solution with SVs lying outside or on the sphere. It turns our that the optimal sparse solution can be defined using a least-squares optimization problem, thus leading to a low computational complexity problem. This framework allows us to derive some theoretical results. We give an upper bound on the probability of false detection, *i.e.,* probability that a new sample is outside the sphere defined by the sparse solution.

As opposed to the jointly optimization of the center and radius by the classical one-class SVM approach, our strategy decouples the estimation problem, thus provides a sub-optimal solution. Consequently, the proposed approach should degrade the performance. In practice, we found that the performance is essentially equivalent to the classical technique, while operating a dramatic speed-up. This is illustrated on experiments from a well-known benchmark for one-class machines [11].

The rest of the paper is organized as follows. Section 2 outlines the classical one-class SVM. We describe our approach in Section 3, and derive theoretical results in Section

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

4. Section 5 illustrates the relevance of our approach on real datasets. Conclusion and further directions are given in Section 6.

2. CLASSICAL ONE-CLASS SVM

Thanks to the concept of reproducing kernels [15], a (positive semi-definite) kernel function $\kappa(\cdot,\cdot)$ defines a nonlinear transformation $\Phi(\cdot)$ of the input space into some feature space. A sphere defined in the latter corresponds (*is pre-imaged* [16]) to a nonlinear characteristics in the input space. It turns our that only the inner product is often required, which can be evaluated using a kernel function, $\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for any $\boldsymbol{x}_i, \boldsymbol{x}_j$ from the input space $\mathcal{X}$.

The one-class SVM was initially derived in [2] for the estimation of the support of a distribution with the ν-SVM, and in [9] for novelty detection with the so-called "support vector data description". The principle idea is to find a sphere, of minimum volume, containing all the training samples. This sphere, described by its center $\boldsymbol{c}$ and its radius r, is obtained by solving the constrained optimization problem

$$\begin{array}{ll} \min\limits_{r,\boldsymbol{c}} & r^2 \\ \text{subject to} & \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\|^2 \leq r^2 \text{ for } i = 1, 2, \ldots, n \end{array}$$

While the above constraint may be too restrictive, one may tolerate a small fraction of the samples to be outside the sphere. This yields robustness, in the sense that it is less sensitive to the presence of outliers in the training dataset. For this purpose, let ν be a positive parameter that specifies the tradeoff between the sphere volume and the number of outliers. Then the problem becomes the estimation of $\boldsymbol{c}$, r, and a set of non-negative slack variables $\zeta_1, \zeta_2, \ldots, \zeta_n$:

$$\begin{array}{l} \min\limits_{r,\boldsymbol{c},\boldsymbol{\zeta}} \; r^2 + \dfrac{1}{\nu n} \displaystyle\sum_{i=1}^{n} \zeta_i \\ \text{subject to } \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\|^2 \leq r^2 + \zeta_i \text{ for all } i = 1, 2, \ldots, n \end{array}$$

By introducing the Karush-Kuhn-Tucker (KKT) optimality conditions, we get

$$\boldsymbol{c} = \sum_{i=1}^{n} \alpha_i \, \Phi(\boldsymbol{x}_i), \tag{1}$$

where the α_i's are the solution to the optimization problem:

$$\begin{array}{l} \max\limits_{\boldsymbol{\alpha}} \displaystyle\sum_{i=1}^{n} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{subject to } \displaystyle\sum_{i=1}^{n} \alpha_i = 1 \text{ and } 0 \leq \alpha_i \leq \frac{1}{\nu n} \text{ for all } i = 1, 2, \ldots, n. \end{array} \tag{2}$$

In accordance with the KKT conditions, each sample $\boldsymbol{x}_i$ can be classified into three categories: $\alpha_i = 0$ corresponds to a sample lying inside the sphere, samples with $0 < \alpha_i < \frac{1}{\nu n}$ lie on the sphere boundary, and samples with $\alpha_i = \frac{1}{\nu n}$ lie outside the sphere, *i.e.*, are outliers. The samples with non-zero α_i are called support vectors (SVs) since they are sufficient to describe the center as defined in expression (4). In practice, only a very small fraction of the data are SV. Let $\mathcal{I}_{\text{sv}}$ be the set of indices associated to SV, namely

$$\begin{cases} \alpha_i \neq 0 & \text{if } i \in \mathcal{I}_{\text{sv}}; \\ \alpha_i = 0 & \text{otherwise.} \end{cases}$$

Finally, the optimal radius is obtained from any SV lying on the boundary, namely any $\boldsymbol{x}_i$ with $0 < \alpha_i < \frac{1}{\nu n}$, since in this case $\|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\| = r$. This is equivalent to

$$r = \min_{i \in \mathcal{I}_{\text{sv}}} \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\|.$$

Therefore, the decision rule that any new sample $\boldsymbol{x}$ is not an outlier is given as $\|\Phi(\boldsymbol{x}) - \boldsymbol{c}\| < r$, where the distance is computed by using

$$\|\Phi(\boldsymbol{x}) - \boldsymbol{c}\|^2 = \sum_{i,j \in \mathcal{I}_{\text{sv}}} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) - 2 \sum_{i \in \mathcal{I}_{\text{sv}}} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + \kappa(\boldsymbol{x}, \boldsymbol{x}). \tag{3}$$

3. SIMPLE ONE-CLASS METHODS

Back to basics, the center (or empirical first moment) of a set of samples is defined by

$$\boldsymbol{c}_n = \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{x}_i), \tag{4}$$

and the radius of the sphere englobing all (or most of) the samples can be easily considered, where the distance is evaluated using (3) where $\alpha_i = 1/n$ for all $i = 1, 2, \ldots, n$ and $\mathcal{I}_{\text{sv}} = \{1, 2, \ldots, n\}$. While the sphere defined by the above full-model is extremely sensitive to outliers, one may consider a sparse solution by incorporating a small number of relevant samples in the model. This is essentially the spirit of the classical one-class SVM, which estimates jointly the center and the radius, by identifying the SVs. Our approach towards a sparse solution is based on three steps:

- Determine the full-model center from (4);
- Identify the SVs as the farthest samples from the center;
- Estimate accordingly the sparse model parameters.

3.1. Sparsification rule

The classical one-class SVM method provides a sparse model for the center, where only samples outside or lying on the sphere are SVs. Inspired by this result, we consider in our approach the distance criterion to identify this subset.

The set of SVs can be identified by considering the distance of each sample to the center, namely

$$\begin{aligned}\mathcal{I} &= \arg\max_{k\in\mathcal{I}} \|\Phi(\boldsymbol{x}_k) - \boldsymbol{c}_n\|^2 \\ &= \arg\max_{k\in\mathcal{I}} -2\sum_{i=1}^{n}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_k) + n\kappa(\boldsymbol{x}_k,\boldsymbol{x}_k),\end{aligned}$$

where the number of SVs is fixed in advance. Once the set $\{\boldsymbol{x}_i \mid i \in \mathcal{I}\}$ is determined, the radius is given as

$$r = \min_{i\in\mathcal{I}} \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_{\mathcal{I}}\|,$$

where $\boldsymbol{c}_{\mathcal{I}}$ is the sparse model of the center defined by

$$\boldsymbol{c}_{\mathcal{I}} = \sum_{i\in\mathcal{I}} \alpha_i \Phi(\boldsymbol{x}_i), \tag{5}$$

the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ being estimated as follows.

3.2. Sparse formulation of the center

Consider the error of approximating $\boldsymbol{c}_n$ with the sparse model $\boldsymbol{c}_{\mathcal{I}}$, $\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|$, which indicates the wellness of such approximation using a small subset of the training data. The coefficients in (5) are estimated by minimizing this error, with

$$\boldsymbol{\alpha} = \arg\min_{\alpha_1,\ldots,\alpha_n} \left\| \frac{1}{n}\sum_{i=1}^{n}\Phi(\boldsymbol{x}_i) - \sum_{i\in\mathcal{I}}\alpha_i\Phi(\boldsymbol{x}_i)\right\|^2, \tag{6}$$

where $\boldsymbol{\alpha}$ is a column vector of the optimal coefficients α_k's for $k \in \mathcal{I}$. Taking the derivative of this cost function with respect to each α_k, namely $-2\left\langle \Phi(\boldsymbol{x}_k), \boldsymbol{c}_n - \sum_{i\in\mathcal{I}}\alpha_i\Phi(\boldsymbol{x}_i)\right\rangle$, and setting it to zero, we get

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

In matrix form, we obtain

$$\boldsymbol{\alpha} = \boldsymbol{K}^{-1}\boldsymbol{\kappa}, \tag{7}$$

where $\boldsymbol{K}$ is the kernel matrix, with entries $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for $i,j \in \mathcal{I}$ and $\boldsymbol{\kappa}$ is a column vector with entries $\frac{1}{n}\sum_{i=1}^{n}\kappa(\boldsymbol{x}_k,\boldsymbol{x}_i)$ for $k \in \mathcal{I}$. To make this problem well-posed in practice, we include a regularization parameter ν, namely $\boldsymbol{\alpha} = (\boldsymbol{K} + \nu\boldsymbol{I})^{-1}\boldsymbol{\kappa}$, where $\boldsymbol{I}$ is the identity matrix of appropriate size. The error of approximating the center with the above solution is

$$\begin{aligned}\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|^2 &= \|\boldsymbol{c}_n\|^2 - 2\boldsymbol{\alpha}^\top\boldsymbol{\kappa} + \boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) - \boldsymbol{\kappa}^\top\boldsymbol{K}^{-1}\boldsymbol{\kappa}. \end{aligned}\tag{8}$$

3.3. Constrained sparse formulation of the center

While the box constraint on the coefficients requires advanced optimization techniques, it is easy to satisfy the equality constraint (see (2)). The constrained optimization problem becomes

$$\begin{aligned}&\boldsymbol{\alpha}_{\text{eq}} = \arg\min_{\alpha_1\ldots,\alpha_n} \left\| \frac{1}{n}\sum_{i=1}^{n}\Phi(\boldsymbol{x}_i) - \sum_{i\in\mathcal{I}}\alpha_i\Phi(\boldsymbol{x}_i)\right\|^2 \\ &\text{subject to } \mathbf{1}^\top\boldsymbol{\alpha}_{\text{eq}} = 1,\end{aligned}$$

where $\mathbf{1}$ is a column vector of 1's. By using the Lagrangian multipliers, we obtain

$$\boldsymbol{\alpha}_{\text{eq}} = \boldsymbol{\alpha} - \frac{\boldsymbol{K}^{-1}\mathbf{1}(\mathbf{1}^\top\boldsymbol{\alpha} - 1)}{\mathbf{1}^\top\boldsymbol{K}^{-1}\mathbf{1}}, \tag{9}$$

where $\boldsymbol{\alpha}$ is the unconstrained solution, as given in (7). One may also include a regularization term, as above.

4. SOME THEORETICAL RESULTS

Independently of the algorithm, one is considering a set of samples in order to estimate the true expectation. Let

$$\boldsymbol{c}_\infty = \int_{\mathcal{X}} \Phi(\boldsymbol{x})\, dP(\boldsymbol{x})$$

be the true expectation, where $P(\boldsymbol{x})$ is the probability distribution generating the samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. From these samples, one can give an estimate of $\boldsymbol{c}_\infty$ by using the empirical first moment $\boldsymbol{c}_n$, as defined in expression (4). The accuracy of such approximation is

$$\epsilon_0 = \|\boldsymbol{c}_n - \boldsymbol{c}_\infty\|.$$

Based on the Hoeffding's inequality, it is shown in [17] (see also [18]) that, with probability at least $1-\delta$ over the choice of a random set of n samples, we have

$$n\epsilon_0^2 \le \sup_{\boldsymbol{x}\in\mathcal{X}}\kappa(\boldsymbol{x},\boldsymbol{x})\left(2 + \sqrt{-2\ln\delta}\right)^2.$$

By the symmetry of the i.i.d assumption, we can bound the probability that a new sample $\boldsymbol{x}$, generated from the same probability distribution, is beyond the boundary defined by a one-class classification method, as given by the following proposition:

Proposition 1. *Consider the sphere centered on $\boldsymbol{c}_{\mathcal{I}}$ with radius $\max_{i=1,\ldots,n}\|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_{\mathcal{I}}\| + 2\epsilon_0 + 2\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|$. Then, with probability at least $1-\delta$ over the choice of a random set of n samples, we can bound the probability that a new sample $\boldsymbol{x}$ is outside this sphere, with*

$$\begin{aligned}&P\big(\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_{\mathcal{I}}\| > \max_{i=1,\ldots,n}\|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_{\mathcal{I}}\| + 2\epsilon_0 + 2\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|\big) \\ &\qquad \le \frac{1}{n+1}.\end{aligned}$$

Proof. To show this, we consider $\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_{\mathcal{I}}\|$ and apply the triangle inequality twice, we get

$$\begin{aligned}\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_{\mathcal{I}}\| &\leq \|\Phi(\boldsymbol{x}) - \boldsymbol{c}_n\| + \|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\| \\ &\leq \|\Phi(\boldsymbol{x}) - \boldsymbol{c}_\infty\| + \epsilon_0 + \|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|,\end{aligned}$$

where the first inequality follows from approximating the full-model center by a subset of samples, while the second inequality from estimating the expected center by a finite set of n samples. Equivalently, we have for any $\boldsymbol{x}_i$:

$$\begin{aligned}\|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_{\mathcal{I}}\| &\geq \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_n\| - \|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\| \\ &\geq \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_\infty\| - \epsilon_0 - \|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|.\end{aligned}$$

Therefore, we get

$$\begin{aligned}&P\big(\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_{\mathcal{I}}\| > \max_{i=1,\ldots,n} \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_{\mathcal{I}}\| + 2\epsilon_0 + 2\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|\big) \\ &\quad \leq P\big(\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_\infty\| > \max_{i=1,\ldots,n} \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_\infty\|\big) \\ &\quad \leq \frac{1}{n+1}\end{aligned}$$

where the first inequality follows from the above inequalities, and the last inequality is due to the symmetry of the i.i.d assumption, considering $n+1$ samples drawn from the same distribution. □

As a special case of this proposition, consider the full-model for the empirical center, namely $\mathcal{I} = \{1, 2, \ldots, n\}$. In this case we get the relation given in [17, Chapter 5]:

$$\mathbb{P}\big(\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_n\| > \max_{i=1,\ldots,n} \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_n\| + 2\epsilon_0\big) \leq \frac{1}{n+1}.$$

We can extend this result to the solution defined by considering that the samples defined by indices $\mathcal{I}$ are outliers, thus not inside the sphere. The following proposition can be easily proven using the same steps as in the proof of Proposition 1.

Proposition 2. *Consider the same setting as in Proposition 1, where the indices in $\mathcal{I}$ define the outliers with $|\mathcal{I}|$ the number of outliers. Then, with probability at least $1-\delta$ over the choice of a random set of n samples, we can bound the probability that a new sample $\boldsymbol{x}$ is outside the sphere excluding outliers, with*

$$\begin{aligned}&P\big(\|\Phi(\boldsymbol{x}) - \boldsymbol{c}_{\mathcal{I}}\| > \min_{i\in\mathcal{I}} \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}_{\mathcal{I}}\| + 2\epsilon_0 + 2\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|\big) \\ &\quad \leq \frac{|\mathcal{I}|}{n+1}.\end{aligned}$$

It is worth noting that in both propositions, the error $\|\boldsymbol{c}_n - \boldsymbol{c}_{\mathcal{I}}\|$ is minimized by our approach, as given by (8).

5. EXPERIMENTS

Since there are few benchmark datasets for one-class classification methods, multiclass tasks are often considered. A multiclass classification task can be tackled by using one-class machines: each class is defined by a one-class classifier, and subsequently we get the decision rule by combining these classifiers. In practice, the model parameters are estimated by considering a subset of the target class (n_{train} samples), and tested over the remaining samples (n_{test}), some from the target class and all the samples from the other classes.

To illustrate the relevance of the proposed approach, we have tested the proposed methods on real datasets well-known in the literature of one-class machines [19]: the IRIS dataset with 150 samples in 3 classes and 4 features, and WINE with 178 samples in 3 classes and 13 features. These datasets are available from the UCI machine learning repository. For experiments on IRIS data, we used only third and fourth features, as often investigated.

The Gaussian kernel was applied, with $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2)$. To estimate the classification error, a ten-fold cross-validation was used, with parameters optimized by a grid search over $\nu \in \{2^{-5}; 2^{-4}; \cdots; 2^4; 2^5\}$ and $\sigma \in \{2^{-5}; 2^{-4}; \cdots; 2^4; 2^5\}$. In order to provide comparable results, the number of SVs was fixed for all methods, by considering the optimal configuration for the classical one-class SVM. Table (4) gives the results in terms of classification error for each of the proposed methods, and compared to the classical one-class SVM. We also included the ratio of common SVs between the latter and each of the proposed methods, as well as the mean values. The computational cost of these machines, using the best configuration, are illustrated in terms of CPU time, as estimated on a 64-bit Matlab running on a Mackbook Pro with a 2.53 GHz Intel Core 2 Duo processor and 4 GB RAM.

6. CONCLUSION

In this paper, we studied the problem of one-class classification. By offering three one-class classification methods, we have shown that we can achieve a classification one-class while minimizing the classification error and especially with less computing time. The relevance of our approach is illustrated by experiments on well-known datasets. In future works, we study an online strategy for one-class classification, as well as other sparsification rules such as the coherence criterion.

7. REFERENCES

[1] D. de Ridder, D. Tax, and R. P. W. Duin, "An experimental comparison of one-class classification methods," in *Proc. 4th Annual Conference of the Advanced School for Computing and Imaging*. Delft, The Netherlands: ASCI, 1998.

			Classical	Simple one-class methods (this paper)					
			one-class SVM	Full model c_n in (4)		Sparse model $\boldsymbol{c}_{\mathcal{I}}$ in (5)-(7)		Constrained model $\boldsymbol{c}_{\mathcal{I}}$ in (5)-(9)	
Datasets	class #	$n_{\text{train}}\|n_{\text{test}}$	error	error	shared SVs	error	shared SVs	error	shared SVs
IRIS	0	45|105	0.28	0	88%	0.38	90%	0.38	90%
	1	45|105	0.86	2.09	76%	1.04	70%	1.14	70%
	2	45|105	1.62	3.90	56%	1.23	58%	1.23	58%
		total error	**0.92**	**1.99**	73%	**0.88**	72%	**0.91**	72%
		(total time)	*(3.42)*	*(0.52)*		*(0.66)*		*(0.66)*	
WINE	0	53|125	2.96	2.72	81%	3.04	81%	3.04	81%
	1	64|114	3.50	4.20	75%	4.29	78%	4.64	78%
	2	43|135	2.81	3.26	85%	3.33	84%	3.41	84%
		total error	**3.09**	**3.39**	80%	**3.55**	81%	**3.69**	81%
		(total time)	*(3.65)*	*(0.75)*		*(1.00)*		*(1.01)*	

Table 1. Experimental results with the classification error for each one-class classifier, and the total error, as well as the total computational time (in seconds). The ratio of common SVs, with respect to the results obtained from the classical one-class SVM, for each of the proposed methods is given, as well as the mean ratio of common SVs.

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

[3] R. Zhang, S. Zhang, S. Muthuraman, and J. Jiang, "One class support vector machine for anomaly detection in the communication network performance data," in *Proc. of the 5th conference on applied electromagnetics, wireless and optical communications*, Stevens Point, Wisconsin, USA, 2007, pp. 31–37.

[4] 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. of the IEEE 23rd International Conference on Advanced Information Networking and Applications Workshops/Symposia*, Bradford, United Kingdom, May 2009, pp. 990–995.

[5] F. Ratle, M. Kanevski, A.-L. Terrettaz-Zufferey, P. Esseiva, and O. Ribaux, "A comparison of one-class classifiers for novelty detection in forensic case data," in *Proc. 8th international conference on Intelligent data engineering and automated learning*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 67–76.

[6] D. M. J. Tax and P. Juszczak, "Kernel whitening for one-class classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, pp. 333–347, 2003.

[7] M. Kemmler, E. Rodner, and J. Denzler, "One-class classification with gaussian processes," in *Proc. Asian Conference on Computer Vision*, 2010, pp. 489–500.

[8] K. Hempstalk and E. Frank, "Discriminating against new classes: One-class versus multi-class classification," in *Proc. 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 325–336.

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

[10] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[11] Y.-H. Liu, Y.-C. Liu, and Y.-J. Chen, "Fast support vector data descriptions for novelty detection," *IEEE Trans. on Neural Networks*, vol. 21, no. 8, pp. 1296–1313, aug. 2010.

[12] J. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[13] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least squares classification," in *Advances in Learning Theory: Methods, Model and Applications*, ser. NATO Science Series III: Computer and Systems Sciences, J. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, Eds., vol. 190. Amsterdam, Netherlands: VIOS Press, May 2003, pp. 131–154.

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

[15] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, 1950.

[16] P. Honeine and C. Richard, "Preimage problem in kernel-based machine learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 77–88, 2011.

[17] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, UK: Cambridge University Press, 2004.

[18] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to Statistical Learning Theory," in *Advanced Lectures on Machine Learning*, 2004, pp. 169–207.

[19] D. Wang, D. S. Yeung, and E. C. C. Tsang, "Structured one class classification," *IEEE Trans. on systems, Man, and Cybernetics, Part B*, vol. 36, pp. 1283–1295, 2006.