Calibration of One-Class SVM for MV set estimation

Albert Thomas, Vincent Feuillard, Alexandre Gramfort

Calibration of One-Class SVM for MV set estimation

Albert Thomas *†, Vincent Feuillard * and Alexandre Gramfort †
*Airbus Group Innovations, first.last@airbus.com, 12 rue Pasteur, 92150, Suresnes, France
†LTCI, CNRS, Télécom Paris-Tech, Université Paris-Saclay, first.last@telecom-paristech.fr, 75013, Paris, France

***Abstract*—A general approach for anomaly detection or novelty detection consists in estimating high density regions or Minimum Volume (MV) sets. The One-Class Support Vector Machine (OCSVM) is a state-of-the-art algorithm for estimating such regions from high dimensional data. Yet it suffers from practical limitations. When applied to a limited number of samples it can lead to poor performance even when picking the best hyperparameters. Moreover the solution of OCSVM is very sensitive to the selection of hyperparameters which makes it hard to optimize in an unsupervised setting. We present a new approach to estimate MV sets using the OCSVM with a different choice of the parameter controlling the proportion of outliers. The solution function of the OCSVM is learnt on a training set and the desired probability mass is obtained by adjusting the offset on a test set to prevent overfitting. Models learnt on different train/test splits are then aggregated to reduce the variance induced by such random splits. Our approach makes it possible to tune the hyperparameters automatically and obtain nested set estimates. Experimental results show that our approach outperforms the standard OCSVM formulation while suffering less from the curse of dimensionality than kernel density estimates. Results on actual data sets are also presented.**

I. Introduction

An anomaly is defined as any observation that does not conform to the expected normal behavior [1]. The goal of anomaly detection also referred as novelty detection is to identify abnormal observations without previously knowing them. Applications include machine fault detection, network intrusion detection in cybersecurity or fraud detection in finance. Given observations $X_1, \dots, X_n \in \mathbb{R}^d$, $d \geq 1$, independent and identically distributed realizations of an unknown probability distribution P, we would like to learn a subset of $\mathbb{R}^d$ such that points lying inside this set will be considered as normal and points lying outside will be considered as anomalies. The implicit hypothesis made in this context is that anomalies correspond to rare events and are located in the tail of the distribution. A possible approach is to estimate the subset corresponding to the region where the data are most concentrated. Such a region is called a Minimum Volume (MV) set, i.e., the set of minimum volume with probability mass at least α, with α close to 1.

The notion of MV sets has been introduced by Polonik [2]. Let μ be the Lebesgue measure and $\alpha \in (0, 1)$. A MV set with mass at least α is a solution of the following optimization problem

$$\min_{G \in \mathcal{B}(\mathbb{R}^d)} \mu(G) \quad \text{such that } P(G) \geq \alpha \ , \tag{1}$$

where $\mathcal{B}(\mathbb{R}^d)$ is the set of all measurable subsets of $\mathbb{R}^d$.

We assume that the probability measure P has a density h with respect to the Lebesgue measure μ and that h has no flat parts, i.e., $\mu(\{x, h(x) = \tau\}) = 0$ for all $\tau > 0$. One can show that under regularity assumptions on h, the optimization problem (1) has an unique solution G^*_α (up to subsets of null μ-measure). This solution satisfies $P(G^*_\alpha) = \alpha$ and is a density level set, i.e., a set of the form $\{h \geqslant \tau\}, \tau > 0$ [3]. A MV set is thus a density level set. The converse holds with no assumption on the density: density level sets are MV sets.

There are essentially two different approaches to estimate a MV set. The first one is to resort to a plug-in approach where one first estimates the underlying density and then thresholds it at the level $\hat{\tau}_\alpha$ such that $P(\{\hat{h}_n \geq \hat{\tau}_\alpha\}) = \alpha$ where $\hat{h}_n$ is a density estimator. The main drawback of this approach is that plug-in estimators do not scale well with the dimension (for e.g. see [4]–[6]). Moreover the entire density is estimated while just a density level set is needed.

The second one is to resort to a direct approach by choosing the set of minimum volume containing a proportion α of the sample points among a class of sets such as Glivenko-Cantelli or Vapnik-Cervonenkis classes. Direct approach algorithms include algorithms from [7], [8] and the OCSVM [9], [10]. Scott and Nowak [7] introduce a framework analogous to the empirical risk minimization in binary classification to estimate a MV set. Davenport et al. [8] use a Neyman-Pearson classification approach to estimate MV sets with SVMs or any other classification algorithms. Tax and Duin [10] introduce the Support Vector Data Description (SVDD) algorithm to search for the hypersphere with the minimum volume containing at least a proportion α of data sample in a Reproducing Kernel Hilbert Space (RKHS). If the kernel used is the Gaussian kernel then the OCSVM and SVDD are equivalent [11].

While the problem of anomaly detection is unsupervised, it is known that an unsupervised problem can be transformed into a supervised one [12]. Steinwart et al. [13] introduce a classification framework for density level set estimation. The classification is performed between the original data and

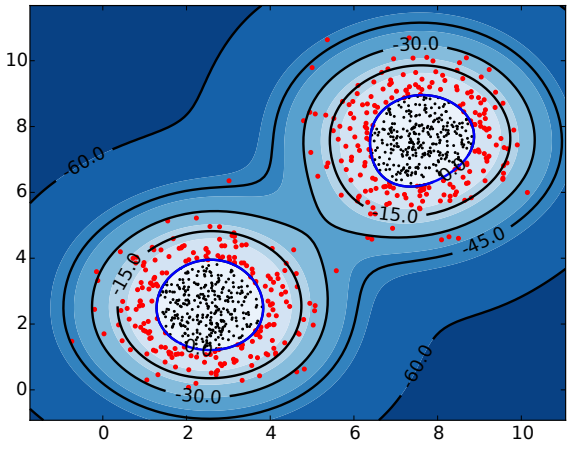

Fig. 1. Application of the OCSVM with $\nu = 0.4$ on a Gaussian mixture sample of size $n = 1000$. In blue the estimated set, in black the level sets of the solution function of the OCSVM, in red the support vectors. The solution function captures the structure of the tail of Gaussian mixture distribution.

an artificial second class. The density level set $\{h \geqslant \tau_\alpha\}$ can then be learnt with any classification algorithm without estimating the entire density h. However one still needs to choose the threshold corresponding to a mass α which can be computationally expensive.

The OCSVM algorithm introduced by Schölkopf et al. [9] is one of the most popular algorithm for anomaly and novelty detection. In [14], Vert and Vert show that the OCSVM is a consistent estimator of density level sets. In fact they give a more powerful result: the solution function returned by the OCSVM gives an estimate of the tail of the underlying density h. The OCSVM is mainly applied with the Gaussian kernel and the performance highly depends on the kernel bandwidth selection.

With the formulation introduced by Schölkopf et al. [9], the mass of the estimated set is controlled by a parameter ν specified by the user. The estimated set is guaranteed to contain at least a fraction $1 - \nu$ of the data. However simple simulations show that the OCSVM can perform very poorly to estimate a MV set for a finite data sample. For instance for a Gaussian mixture such as the one in Figure 1, no value of the kernel bandwidth gives a good approximation of the true MV set with mass at least 0.95 when the parameter ν is chosen such that the empirical probability of the estimated set is larger than α (see section III-A). However using a different value of ν, the set estimated by the OCSVM clearly differs from the true MV set but the solution function captures the structure of the tail of the underlying distribution as shown in Figure 1.

The approach we propose and describe in the second part of this paper consists in fixing ν at a value such that the proportion of points outside the estimated set will be strictly greater than $1-\alpha$. The solution function is learnt on a training set and then thresholded to obtain the desired probability mass on a test set to prevent overfitting. To reduce the variance induced by the random split of the data set into a training set and a test set we aggregate several models. Thresholding the solution function of the OCSVM to obtain the desired probability mass is an approach that has already been very briefly mentionned in [9] and in [15]. However, to the best of our knowledge, such an approach has never been considered thoroughly. In the second part of this paper we present the OCSVM and its properties before presenting our approach. In the last part we compare the performance of our approach with the OCSVM on simulated data sets and apply our approach to real data sets. Connections can be made between this paper and [16] in which Filipone et al. apply the possibilistic c-means algorithm in kernel-induced spaces.

II. Method

A. Background on One-Class SVM

The OCSVM was introduced by Schölkopf et al. [9] to estimate high density regions from a data sample. After mapping the data in a feature space through a function Φ determined by a specific kernel k the OCSVM finds a separating hyperplane between the origin and the mapped data. The separating hyperplane defined by a vector w and an offset ρ is given by the solution of the following optimization problem

$$\begin{aligned} \min_{w,\xi,\rho} \quad & \frac{1}{2}\|w\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \langle w, \Phi(x_i)\rangle \geq \rho - \xi_i \quad , \quad 1 \leq i \leq n \\ & \xi_i \geq 0 \quad , \quad 1 \leq i \leq n \end{aligned} \tag{2}$$

where $\nu \in (0, 1)$ is a parameter specified by the user. This problem is convex and as strong duality holds it is solved through its dual

$$\begin{aligned} \min_{\gamma} \quad & \frac{1}{2}\sum_{1\leq i,j\leq n}\gamma_i\gamma_j k(x_i, x_j) \\ \text{s.t.} \quad & 0 \leq \gamma_i \leq \frac{1}{\nu n} \quad , \quad 1 \leq i \leq n \\ & \sum_{i=1}^{n}\gamma_i = 1 \end{aligned} \tag{3}$$

The resulting solution function is given by

$$x \mapsto \sum_{i=1}^{n}\gamma_i k(x, x_i)$$

and the resulting estimated MV set by

$$\hat{G} = \{x, \sum_{i=1}^{n}\gamma_i k(x, x_i) - \rho_\nu \geq 0\} \tag{4}$$

where ρ_ν denotes the ρ solution of (2).

As with SVM in supervised settings, not all the γ_i are non-zero. The points x_i such that $\gamma_i > 0$ are called support vectors (SVs). Support vectors are exactly the samples located outside or on the border of the set $\hat{G}$:

$$\{x_j, 1 \leq j \leq n, \sum_{i=1}^{n}\gamma_i k(x_j, x_i) - \rho_\nu \leq 0\} \ .$$

Outliers are exactly the samples that are located strictly outside the set $\hat{G}$:

$$\{x_j, 1 \leq j \leq n, \sum_{i=1}^{n} \gamma_i k(x_j, x_i) - \rho_\nu < 0\} \ .$$

The parameter ν needs to be chosen by the user. We have the following property [9]:

Proposition 1: Assuming the solution of (2) satisfies $\rho_\nu > 0$ the following statement holds

i) ν is an upper bound on the fraction of outliers and a lower bound on the fraction of SVs

$$\frac{\text{Outliers}}{n} \leq \nu \leq \frac{\text{SV}}{n} \ .$$

ii) If the data were generated independently from a distribution P absolutely continuous with respect to the Lebesgue measure and if the kernel k is analytic and non constant then

$$\frac{\text{SV}}{n} \longrightarrow \nu \quad \text{almost surely}$$
$$\frac{\text{Outliers}}{n} \longrightarrow \nu \quad \text{almost surely}$$

This property is of great interest in practice. It gives the user some insights on how to choose the parameter ν. Indeed the empirical probability of the estimated set is greater than $1-\nu$ and the probability of the estimated set converges almost surely to $1 - \nu$ as n tends to infinity. Hence one possible approach is to choose $\nu = 1 - \alpha$ to estimate a MV set with mass at least α [17], [18].

In the following the kernel k is the Gaussian kernel $k_\sigma, \sigma > 0$, and is defined as

$$k_\sigma(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right) \ .$$

We denote by f_σ the solution function

$$f_\sigma(x) = \sum_{i=1}^{n} \gamma_i k_\sigma(x, x_i) \ .$$

The paper of Vert and Vert [14] proves the consistency of the OCSVM for density level sets estimation and hence for MV sets estimation. The optimization problem associated with the OCSVM studied in their paper is the following

$$\min_{f \in H_\sigma} \quad \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - f(x_i)) + \lambda \|f\|^2_{H_\sigma} \qquad (5)$$

where H_σ is the RKHS associated to the normalized Gaussian kernel and $\lambda > 0$ a regularization parameter.

Vert and Vert [14] prove that for a well calibrated kernel bandwidth σ, the OCSVM is a consistent estimator of every density level sets of level $\tau \in (0, 2\lambda)$. To show such a result they prove that the solution of the OCSVM when a normalized Gaussian kernel is used converges in norm L^2 and in probability to the underlying density truncated at 2λ:

$$\lim_{n \to +\infty} \|f_\sigma - h_\lambda\|_{L_2} = 0 \quad \text{in probability}$$

where

$$h_\lambda = \begin{cases} \dfrac{h(x)}{2\lambda} & \text{if } h(x) \leq 2\lambda \\ 1 & \text{otherwise.} \end{cases}$$

Remark 1 (Connection with kernel smoothing): If $\nu = 1$ the constraints of the dual problem (3) give $\gamma_i = \frac{1}{n}$ for all $i \in \{1, \ldots, n\}$. This means that all the samples are taken into account in the solution and the solution function is

$$f_\sigma(x) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(x, x_i) \ .$$

This function is the one we recover when performing a kernel smoothing with the same kernel bandwidth σ in all the directions.

The advantage of the OCSVM over a kernel smoothing is that the estimated set is only characterized by the support vectors which, for small values of ν, represent a small fraction of the sample size: the solution is sparse. This property is useful when performing the prediction task which is therefore less expensive than when using a kernel smoothing approach. Besides the solution function gives an approximation of the tail of the underlying density and, unlike a kernel smoothing, the approximation given by the solution function can be very bad elsewhere. This is why classification is sometimes said to be easier than regression [19]: we only want to be good in a neighborhood of the border of the set of interest and not elsewhere.

Eventually, parametrization of the mass of the MV set estimated by the OCSVM via the parameter ν does not allow to obtain nested set estimates as the mass α increases. For each ν a new optimization problem is solved and nothing ensures that the different set estimates are nested. Variants of the OCSVM that ensure this property have been introduced [20], [21]. With our approach, the mass of the MV set is parametrized through the offset and this allows us to produce nested sets in a neighborhood of the estimated MV set with mass at least α. For the same solution function, we select different offsets ρ, one for each mass.

B. Automatic Calibration of OCSVM

We want to estimate a MV set with mass at least α with α close to 1 from the sample $X_1, \ldots, X_n$. Thanks to the result of Vert and Vert [14], we know that the solution function of the OCSVM gives an approximation of the tail of the underlying distribution. More precisely in our approach we use the fact that f_σ is an approximation of the underlying density in a neighborhood of the border of the MV set. The algorithm we propose is described in Figure 2 and detailed hereafter.

First the data set $X = (X_1, \ldots, X_n)$ is randomly split in a training set X_{train} and a test set X_{test} respectively of size n_{train} and n_{test}. Let $\hat{G}$ be the set estimated by the OCSVM on the training set. The parameter ν is chosen such that we are able to estimate the underlying distribution for the interval of masses $[\alpha - c, \alpha + c]$ where $c > 0$. Therefore ν must be chosen such that $P_{n_{train}}(\hat{G}) \leqslant \alpha - c$, where $P_{n_{train}}$ denotes

Input: parameter ν, mass α, data set X, kernel bandwidths set Σ, $c > 0$
Randomly split X in a training set X_{train} and a test set X_{test}
for kernel bandwidth σ in Σ **do**
  $f_\sigma = \text{OCSVM}(\nu, \sigma, X_{train})$
  for β in $[\alpha - c, \alpha + c]$ **do**
    Bisection search to find $\hat{\rho}_\beta$ such that

$$P_{n_{test}}(\hat{G}^\sigma_{\hat{\rho}_\beta}) = \beta$$

where $\hat{G}^\sigma_{\hat{\rho}_\beta} = \{x, f_\sigma(x) - \hat{\rho}_\beta \geq 0\}$
    Computation of $\mu^\sigma_{\hat{\rho}_\beta} = \mu(\hat{G}^\sigma_{\hat{\rho}_\beta})$ by Monte Carlo integration
  end for
end for
Compute Area under the Mass Volume curve $(\beta, \mu^\sigma_{\hat{\rho}_\beta})$ for each σ: AMV(σ)
$\sigma_{opt} = \text{argmin}_{\sigma \in \Sigma} \text{AMV}(\sigma)$
return $\hat{G}^{\sigma_{opt}}_{\hat{\rho}_\alpha} = \{x, f_{\sigma_{opt}}(x) - \hat{\rho}_\alpha \geq 0\}$

Fig. 2. Algorithm of the OCSVM with a calibrated offset and the selection of the optimal kernel bandwidth

the empirical probability measure based on the training set. $P_{n_{train}}(\hat{G}) \leqslant \alpha - c$ is equivalent to a fraction of outliers, points lying outside $\hat{G}$, greater than $1 - (\alpha - c)$. What we have from proposition 1 is that the fraction of outliers is less than ν for all n and converges almost surely to ν as n tends to infinity. The closer ν is to 1, the more outliers we allow the OCSVM to find. If ν has been set such that the fraction of outliers is less than $1-(\alpha-c)$, then a higher value should be chosen. As we only consider values of α close to 1, we do not need ν to be too close to 1 and can therefore preserve the sparsity of the OCSVM. In our algorithm we assume that a good value for ν is known and is set independently of the data set.

The function f_σ gives an approximation of the tail of the distribution. Consequently thresholding it at $\hat{\rho}_\alpha$ such that $P_{n_{test}}(f_\sigma \geq \hat{\rho}_\alpha) = \alpha$ should offer an approximation of the MV set with mass at least α, where $P_{n_{test}}$ denotes the empirical probability measured based on the test set.

Remark 2: Let $\alpha_1 < \cdots < \alpha_N$ be N values in $[\alpha - c, \alpha + c]$ and let $\hat{\rho}_1 \geq \cdots \geq \hat{\rho}_N$ be such that for all $i \in \{1, \ldots, N\}$ we have $P_{n_{test}}(f_\sigma \geq \hat{\rho}_i) = \alpha_i$. Let $\hat{G}_i$ be the set $\hat{G}_i = \{x, f_\sigma(x) \geq \hat{\rho}_i\}$, then by construction the following holds

$$\hat{G}_1 \subset \cdots \subset \hat{G}_N \ .$$

C. Performance metric and kernel bandwidth selection

To assess the performance of our approach and select the kernel bandwidth we need a performance metric. The kernel bandwidth parameter selection is an important task in practice as the solution of OCSVM highly depends on its choice. Low values of σ lead to overfitting. On the contrary, high values of σ lead to underfitting.

A performance metric used for the theoretical study of MV sets or density level set estimators is the Lebesgue measure of the symmetric difference between the true MV set G^*_α and the estimate $\hat{G}$, $\mu(G^*_\alpha \Delta \hat{G})$ where $A \Delta B = (A \backslash B) \cup (B \backslash A)$ [7], [13], [22].

This performance metric depends on the true MV set G^*_α. We use it to assess the performance of our approach and select the optimal kernel bandwidth when we have access to the true MV set.

Several performance metrics have been used to assess the quality of one-class classification algorithms and select the optimal hyperparameters (see among others [8], [20], [21], [23]). It is noteworthy to say that all these metrics require to sample points uniformly, either to compute the volume of the estimated set or to generate an artificial second class. Therefore both method suffer from the curse of dimensionality. First, the proportion of points uniformly sampled in the hypercube enclosing the data lying in the estimated set can decrease exponentially to 0 with the dimension. Second, for high dimensions, data are expected to be very sparse and to be very easily separated, leading classification solutions to overfit. We must therefore limit the use of these metrics to data sets of low dimension, for e.g. $d \leq 10$. This has been mentionned by Tax in [11], [23].

The performance metric we decide to use in our algorithm to select the kernel bandwidth is the Mass Volume curve introduced by Clémençon and Jakubowicz [24] and defined as $\{(\alpha, \mu(G^*_\alpha)), \alpha \in (0, 1)\}$. To use this performance metric, we still need to sample points uniformly to compute the volume. The Mass Volume curve is a functional criterion that can be used to assess the quality of a scoring rule in the unsupervised setting. The Mass Volume curve of the true underlying distribution is the lowest Mass Volume curve that can be obtained. Clémençon and Robbiano [25] give the explicit relation between the well known area under the ROC curve (AUC) and the area under the Mass Volume curve. Minimizing the area under the Mass Volume curve is equivalent to maximizing the AUC when the second class has been generated from a uniform distribution.

The Mass Volume curve is suited to assess the quality of scoring rules whereas the first purpose of the OCSVM is not to estimate a scoring rule. Indeed, the OCSVM with $\nu = 1 - \alpha$ gives an estimated set of the form $\{x, f_\sigma(x) \geq \rho_\nu\}$. However there is no guarantee that for all $\rho \neq \rho_\nu$, sets of the form $\{x, f_\sigma(x) \geq \rho\}$ are good approximations of MV sets. Our approach estimates a scoring rule for the points located in the tail of the distribution and we use the area under the Mass Volume curve for masses in a neighborhood of α as a performance metric to select the best kernel bandwidth.

To compute the Mass Volume curve, $\{(P(\hat{G}_\beta), \mu(\hat{G}_\beta)), \beta \in [\alpha - c, \alpha + c]\}$, we need to compute the probability and the volume of the estimated set. The probability is estimated on the test set and is thus equal to β as we choose the offset such that the empirical probability of the estimated set on the test set equals β. We estimate the volume by Monte Carlo estimation.

Volume computation: The volume of a set $G = \{x, f_\sigma(x) \geq \rho\}$ is defined as

$$\mu(G) = \int 1_G(x)\mu(dx) \ . \tag{6}$$

This integral cannot be computed exactly so we resort to Monte Carlo estimation. As we do not know how to sample uniformly in the set G either we resort to importance sampling rewriting (6) as

$$\mu(G) = \int \frac{1_G(x)}{q(x)} q(x)\mu(dx) \tag{7}$$

where q must be a well chosen distribution.

The most popular distribution used in the literature is the uniform distribution over the hypercube G_c enclosing the data. Let V_c be the volume of G_c then the density of such a distribution is $q_c(x) = \frac{1}{V_c} 1_{G_c}(x)$ and

$$\begin{aligned}\mu(G) &= V_c \int \frac{1_G(x)}{1_{G_c}(x)} q_c(x)\mu(dx) = V_c \int 1_G(x) q_c(x)\mu(dx) \\ &= V_c \mathbb{E}_{q_c}[1_G(Z)] \ .\end{aligned}$$

Thanks to the Law of Large Numbers the volume $\mu(G)$ is estimated by

$$\hat{\mu}_c(G) = \frac{V_c}{m} \sum_{i=1}^{m} 1_G(Z_i) \qquad Z_i \sim q_c.$$

Sampling uniform data is an issue worth mentioning as it is the factor limiting the estimation of Minimum Volume sets in a high dimension setting.

D. Aggregation

In section II-B we presented our approach consisting in the following:

1) Randomly split the data set in a training set and a test set
2) Train the OCSVM on the training set to obtain f_σ
3) Find the offset $\hat{\rho}_\alpha$ such that $P_{n_{test}}(\{f_\sigma \geq \hat{\rho}_\alpha\}) = \alpha$ on the test set

Randomly splitting the data set in training and test sets introduces variance in the result. To reduce the variance we aggregate several models based on B train/test splits. Let $(f_\sigma^b, \hat{\rho}_\alpha^b)$, $1 \leq b \leq B$ be the models obtained, where $\hat{\rho}_\alpha^b$ is such that

$$P^b_{n_{test}}(\{x, f_\sigma(x) - \hat{\rho}_\alpha^b \geq 0\}) = \alpha \ .$$

Averaging all the models we obtain

$$F_\sigma^B(x) = \frac{1}{B} \sum_{b=1}^{B} (f_\sigma^b(x) - \hat{\rho}_\alpha^b) \ .$$

The final estimated set is given by

$$\hat{G}_\alpha^B = \{x, F_\sigma^B(x) \geq 0\} \ .$$

Input: parameter ν, mass α, data set X, kernel bandwidths set Σ, $c > 0$, number of models B
for b in $\{1, \ldots, B\}$ **do**
  Randomly split X in a training set X_{train} and a test set X_{test}
  for kernel bandwidth σ in Σ **do**
    $f_\sigma^b = \text{OCSVM}(\nu, \sigma, X_{train})$
    for β in $[\alpha - c, \alpha + c]$ **do**
      Bisection search to find $\hat{\rho}_\beta^b$ such that

$$P_{n_{test}}(\hat{G}^\sigma_{\hat{\rho}_\beta^b}) = \beta$$

where $\hat{G}^\sigma_{\hat{\rho}_\beta^b} = \{x, f_\sigma^b(x) - \hat{\rho}_\beta^b \geq 0\}$
    end for
  end for
end for
For all β and all σ, compute the volume μ_β^σ of the set $\{x, F^B_{\sigma,\beta}(x) \geq 0\}$ where $F^B_{\sigma,\beta}(x) = \frac{1}{B}\sum_{b=1}^B (f_\sigma^b(x) - \hat{\rho}_\beta^b)$
Compute Area under the Mass Volume curve $(\beta, \mu_\beta^\sigma)$ for each σ: AMV(σ)
$\sigma_{opt} = \text{argmin}_{\sigma \in \Sigma} \text{AMV}(\sigma)$
return $\hat{G}_\alpha^B = \{x, F^B_{\sigma_{opt},\alpha}(x) \geq 0\}$ where $F^B_{\sigma_{opt},\alpha} = \frac{1}{B}\sum_{b=1}^B (f^b_{\sigma_{opt}}(x) - \hat{\rho}_\alpha^b)$

Fig. 3. Aggregation of the models learnt on different train/test splits

The algorithm is described in Figure 3.

Proposition 2 (Nested sets): Considering several values $0 < \alpha_1 < \cdots < \alpha_N < 1$, we can construct nested sets $\hat{G}^B_{\alpha_1} \subset \cdots \subset G^B_{\alpha_N}$.

Proof: For $i \in \{1, \ldots, N\}$, let $(f_\sigma^{b,i}, \hat{\rho}_i^b)$, $1 \leq b \leq B$ be the models obtained on the sequence of training and test sets for the mass α_i. We have $f_\sigma^{b,i} = f_\sigma^{b,j}$ for all $i, j \in \{1, \ldots, N\}$ as $f_\sigma^{b,i}$ only depends on the train and test split. By construction we also have $\hat{\rho}_1^b \geq \cdots \geq \hat{\rho}_N^b$ for all b. Then for all b,

$$f_\sigma^{b,1} - \hat{\rho}_1^b \leq \cdots \leq f_\sigma^{b,N} - \hat{\rho}_N^b \ .$$

By summing

$$F_\sigma^{B,1} \leq \cdots \leq F_\sigma^{B,N}$$

and if $\hat{G}^B_{\alpha_i} = \{x, F_\sigma^{B,i}(x) \geq 0\}$ then

$$\hat{G}^B_{\alpha_1} \subset \cdots \subset \hat{G}^B_{\alpha_N} \ .$$

■

III. Experiments

For all the experiments we choose $\nu = 1 - \alpha$ for the OCSVM and $\nu = 0.4$ for our approach. With our approach 80% of the data set is used as the training set and the other 20% as the test set. Unless stated otherwise, $\alpha = 0.95$, the Mass Volume curves are made from 10 masses equally spaced between 0.91 and 0.99 and we uniformly sample 10000 points in the smallest hypercube enclosing the data to compute the volumes. All the experimental work was done with Scikit-learn [26] using the underlying LIBSVM library [27].

A. Simulation with bimodal distribution

We sample $n = 1000$ points from a two-dimensional Gaussian mixture of density $h(x) = \frac{1}{2}\mathcal{N}((2.5, 2.5), I)(x) + \frac{1}{2}\mathcal{N}((7.5, 7.5), I)(x)$ where I denotes the identity matrix and $\mathcal{N}(m, \Sigma)(x)$ the density of the Gaussian distribution with mean m and covariance Σ. We want to estimate the MV set with mass at least 0.95 from this sample. Knowing the density, we only need the level τ_α such that $P(h(X) \geqslant \tau_\alpha) = \alpha$ to know the true MV set G^*_α. τ_α is the $1 - \alpha$ quantile of the distribution of $h(X)$. We estimate such a quantile with 1 million points generated from h. To compute the volume of the symmetric difference between the estimated set and the true MV set we sample points uniformly in the hypercube enclosing the data. Our approach is implemented with an aggregation of 10 models. The comparison of the performance as a function of σ between the OCSVM and our approach is shown in Figure 4. We observe that the performance of the OCSVM obtained for the best value of σ, i.e., the value of σ minimizing this performance, is worse than the performance reached for a wide range of values of σ with our approach. We represent the sets obtained for the values of σ giving the best performance for each approach in Figures 5 and 6. The solution obtained with our approach is clearly better. Besides, even the solution obtained for the best σ of OCSVM tends to overfit (Figure 5).

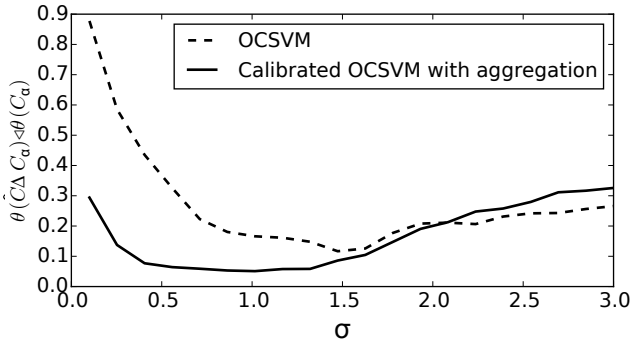

Fig. 4. Performance as a function of σ: OCSVM (dashed line) and our approach with an aggregation of 10 models (solid line)

In Figure 7 we show the evolution of the measure of the symmetric difference between the true and the estimated MV set with mass at least 0.95 as a function of the number of samples. The results are averaged over 100 repetitions. For each sample size, the best σ is computed by minimization of the area under the Mass Volume curve for our approach and through minimization of the measure of the symmetric difference for the OCSVM. Again in the case of OCSVM the ground truth is assumed to be known for parameter tuning while our approach automaticaly tune σ without the knowledge of the ground truth. Despite this, our approach outperforms the OCSVM when we consider the measure of the symmetric difference between the true and the estimated MV set metric. The approach with aggregation further improves the performance without.

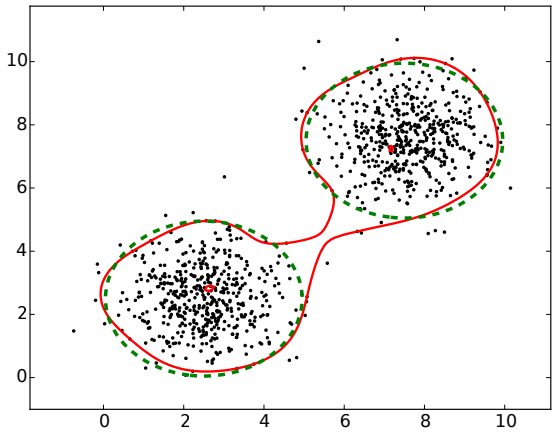

Fig. 5. In dashed line the true MV set. In solid line the estimated MV set for the best σ of the OCSVM with respect to the measure of the symmetric difference between the true and the estimated MV set shown in Figure 4.

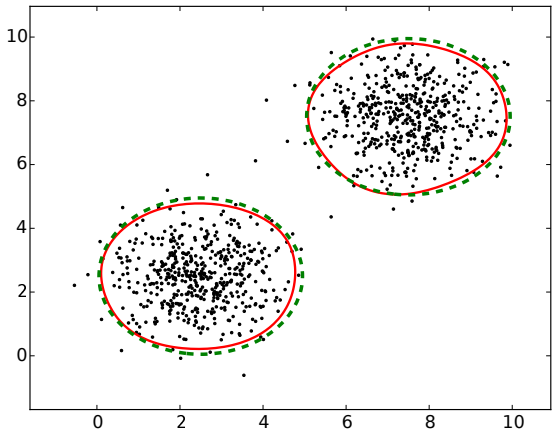

Fig. 6. In dashed line the true MV set. In solid line the estimated MV set for the best σ of our approach with respect to the measure of the symmetric difference between the true and the estimated MV set shown in Figure 4.

B. Bimodal distribution with outliers

We now considered a two-dimensional Gaussian mixture sample to which we add 5% outliers uniformly distributed over an hypercube enclosing the data. We thus sample $n = 1000$ points from the distribution with density $h(x) = 0.475 \cdot \mathcal{N}((2.5, 2.5), I)(x) + 0.475 \cdot \mathcal{N}((7.5, 7.5), I)(x) + \frac{0.05}{V_C} 1_C(x)$ where $C = [-2, 12] \times [-2, 12]$ and V_C is the volume of C. Knowing the density, we proceed as in section III-A to compute the true MV set G^*_α of such a distribution. For the OCSVM we choose the value of σ minimizing the measure

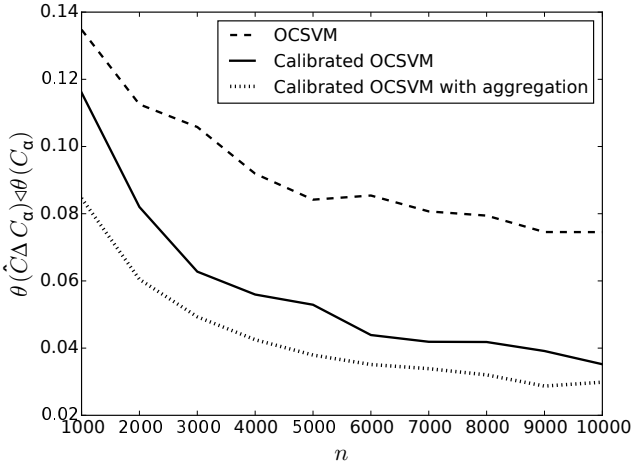

Fig. 7. Performance as a function of the number of samples n. OCSVM (dashed line), our approach without aggregation (solid line) and our approach with an aggregation of 3 models (dotted line).

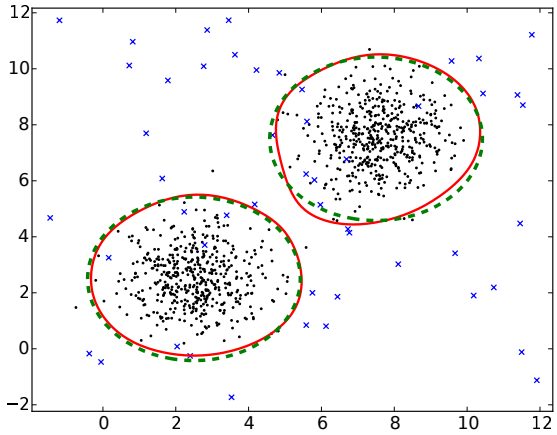

Fig. 9. In dashed line the true MV set. In solid line the estimated MV set. Outliers are represented by crosses.

of the symmetric difference between the estimated set and the true MV set. The estimated set is shown in Figure 8. Our approach is implemented with an aggregation of 10 models. We consider 20 values of σ equally spaced between 0.01 and 3. The best σ is obtained by minimization of the area under the Mass Volume curve. The estimated set is shown in Figure 9. This experiment suggests that our approach is more robust to outliers than the OCSVM.

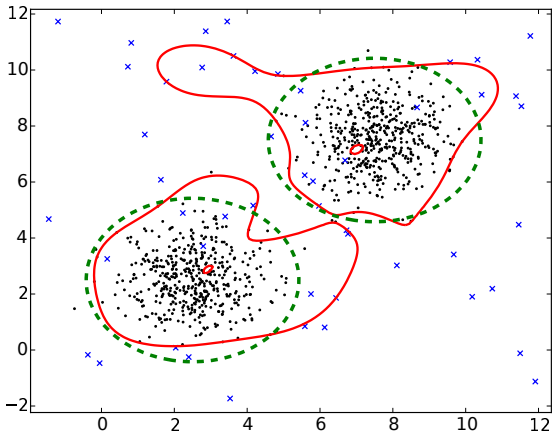

Fig. 8. In dashed line the true MV set. In solid line the estimated MV set. Outliers are represented by crosses.

C. Comparison with plug-in approach

In this section we compare the performance of the plug-in approach with our approach with respect to the number of features d for a Gaussian mixture. We recall here that the plug-in approach consists in estimating the underlying density and then thresholding it at the level $\hat{\tau}_\alpha$ such that $P(\{\hat{h}_n \geq \hat{\tau}_\alpha\}) = \alpha$. The performance metric used to compare both approach is the measure of the symmetric difference between the true and the estimated MV set with mass at least 0.95. We generate a Gaussian mixture sample of size n=500 with density $h(x) = \frac{1}{2}\mathcal{N}(2.5\cdot\mathbf{1}_d, I_d)(x) + \frac{1}{2}\mathcal{N}(7.5\cdot\mathbf{1}_d, I_d)(x)$, $\mathbf{1}_d$ denoting the vector of $\mathbb{R}^d$ with all its components equal to 1 and I_d denoting the identity matrix of dimension d. For the plug-in approach we use a kernel density estimator $\hat{h}_n$ to estimate h and a bisection search to estimate $\hat{\tau}_\alpha$. The kernel used is the Gaussian kernel with same bandwidth s in all the directions. The bandwidth s is selected through a 4-fold cross validation among 15 values equally spaced between 0.1 and 10. Then we threshold $\hat{h}_n$ at $\hat{\tau}_\alpha$ such that $P_n(\hat{h}_n \geq \hat{\tau}_\alpha) = \alpha$ where P_n is the empirical probability measure based on the sample of size n. Our approach is performed with an aggregation of 5 models and the kernel bandwidth is automaticaly selected through minimization of the area under the Mass Volume Curve. In Figure 10 we show the evolution of the performance for both approach. The results are averaged over 100 repetitions. Even though the performance of both approach is quite similar for $d = 2$ and $d = 3$, for $d > 3$ we observe that the performance of the plug-in approach deteriorates much more faster than the performance of our approach. We limit this experiment to $d = 8$ because of the difficulty to compute volumes in high dimension.

D. Two moons data set

We generate a two-dimensional two moons data set of size $n = 2000$ and try to estimate a MV set with mass at least 0.95. We choose 30 values of σ equally spaced between 0.01 and 0.5. We average 25 models based on 25 train/test random splits of the data set. The best σ obtained by minimization of the area under the Mass Volume curve is $\sigma = 0.15$ (see Figure 11). The estimated set is represented in Figure 12. Its empirical mass on the whole data set is 0.96.

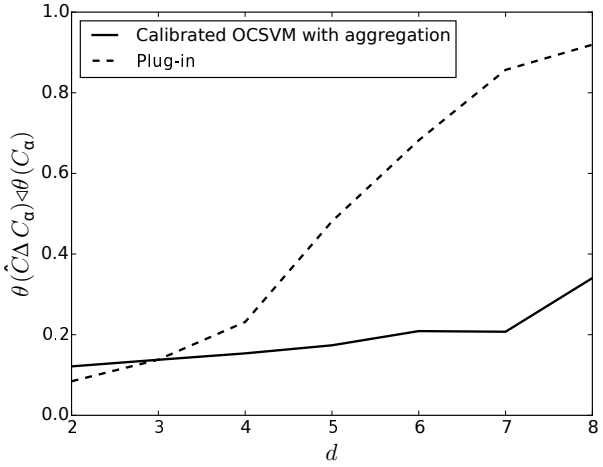

Fig. 10. Performance as a function of the number of features d. Plug-in approach (dashed line) and our approach with an aggregation of 5 models (solid line). Our approach clearly outperforms the plug-in approach as soon as dimension d increases.

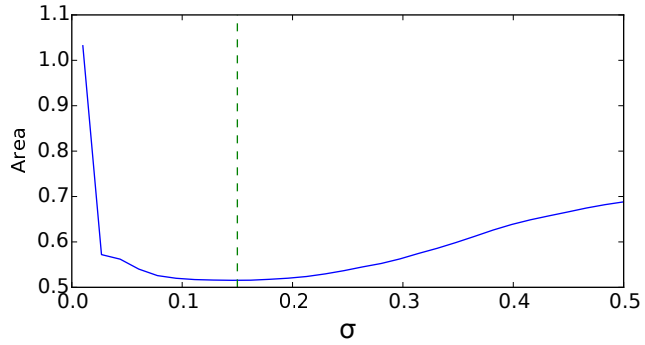

Fig. 11. Area under the mass volume curve as a function of σ for the two moons data set. The minimum is reached at $\sigma = 0.15$.

E. Real data set

We consider here the Boston housing data set [28] from the UCI machine learning repository. This data set concerns housing values in suburbs of Boston and consists in $n = 506$ samples and $d = 14$ features which can be either categorical, integer or real. We only consider two of the features for a better representation of our approach: the average number of rooms per dwelling and the percentage lower status of the population. We first standardize the features, i.e., component wise center and scale to unit variance, and then apply our approach to estimate MV sets. We choose 30 values of σ equally spaced between 0.01 and 4. We average 25 models based on 25 train/test random splits of the data set. The best σ obtained by minimization of area under the Mass Volume curve is $\sigma = 0.42$ (see Figure 13). The estimated sets are represented in Figure 14. The estimated MV set with mass at least 0.90 has an empirical mass of 0.91 on the whole data set and the estimated MV set with mass at least 0.95 has en

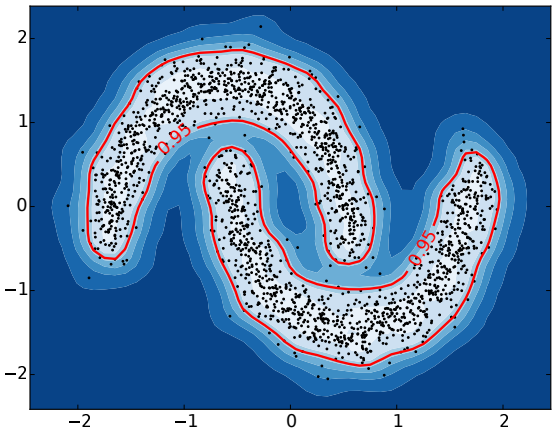

Fig. 12. Estimated MV set with mass at least 0.95 for a generated two moons data set

empirical mass of 0.95. We observe that the estimated sets are nested.

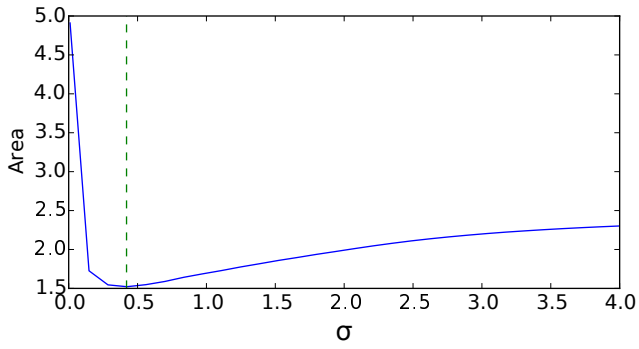

Fig. 13. Area under the mass volume curve as a function of σ for the Boston housing data set. The minimum is reached at $\sigma = 0.42$.

IV. Conclusion

This paper presents a new approach to estimate MV sets using the OCSVM algorithm. Results show that it outperforms the standard way to use the OCSVM. Our approach is based on the calibration of the offset of the solution function to obtain the desired probability mass on a test set. It allows to compute nested set estimates without the need to add any condition ensuring this property and consider several regularization parameters. Moreover it provides a scoring rule for samples located in the tail of the underlying distribution. The computed Mass Volume curve allows to assess the performance of the approach and to select the kernel bandwidth automatically. Our solution inherits the sparsity of the OCSVM which is a computational advantage over kernel smoothing.

The kernel bandwidth selection requires to compute the volume of the estimated set which suffers from the curse of dimensionality. This issue is still an open research area.

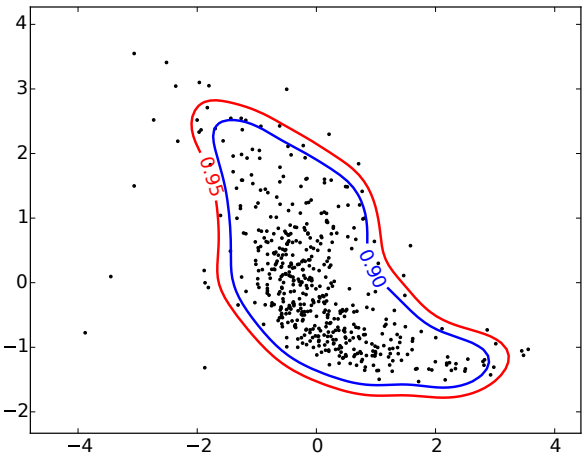

Fig. 14. MV sets with mass at least 0.90 and 0.95 respectively in blue and in red estimated from the two features, average number of rooms per dwelling (x axis) and percentage lower status of the population (y axis), of the Boston housing data set. The features have been standardized.

Sampling more precisely in the region where the data lives instead of sampling in the hypercube enclosing the data is a possible approach to scale to higher dimensions.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection : A survey," *ACM Computing Surveys*, vol. 41, pp. 1–58, 2009.

[2] W. Polonik, "Minimum volume sets and generalized quantile processes," *Stochastic Processes and their Applications*, vol. 69, pp. 1–24, 1997.

[3] J. H J Einmahl and D. M Mason, "Generalized quantile processes," *The Annals of Statistics*, vol. 20, pp. 1062–1078, 1992.

[4] A. Baíllo, "Total error in a plug-in estimator of level sets," *Statistics & Probability Letters*, vol. 65, no. 4, pp. 411–417, 2003.

[5] B. Cadre, "Kernel estimation of density level sets," *Journal of Multivariate Analysis*, vol. 97, no. 4, pp. 999 – 1023, 2006.

[6] B. Cadre, B. Pelletier, and P. Pudlo, "Estimation of density level sets with a given probability content," *Journal of Nonparametric Statistics*, vol. 25, no. 1, pp. 261–272, 2013.

[7] C. D. Scott and R. D. Nowak, "Learning minimum volume sets," *Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[8] M. A Davenport, R. G Baraniuk, and C. D Scott, "Learning minimum volume sets with support vector machines," *Machine Learning for Signal Processing*, pp. 301–306, 2006.

[9] B. Schölkopf, J. Platt, A. J. Shawe-Taylor, J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13(7), 2001.

[10] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, pp. 45–66, 2004.

[11] D. M. J. Tax, "One-class classification," *Ph.D Dissertation, Delft University of Technology*, 2001.

[12] T. Hastie, R. Tibshirani, and F. J, *The Elements Of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009.

[13] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *Journal of Machine Learning Research*, vol. 6, pp. 211–232, 2005.

[14] R. Vert and J.-P. Vert, "Consistency and convergence rates of one-class SVMs and related algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 1469–1480, 2006.

[15] M. P. Martínez, E. Vazquez, E. Walter, and G. Fleury, "RKHS classification for multivariate extreme-value analysis," 2008.

[16] F. Filippone, M. ans Masulli and S. Rovetta, "Applying the possibilistic c-means algorithm in kernel-induced spaces," *Fuzzy Systems, IEEE Transactions on*, vol. 18, no. 3, pp. 572–584, 2010.

[17] A. Glazer, M. Lindenbaum, and M. S, "Learning high-density regions for a generalized kolmogorov-smirnov test in high-dimensional data," *In Proceedings of The 26th Conference on Neural Information Processing Systems*, 2012.

[18] A. Glazer, M. Lindenbaum, and S. Markovitch, "q-OCSVM: A q-quantile estimator for high-dimensional distributions," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 503–511.

[19] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[20] G. Lee and C. Scott, "The one class support vector machine solution path," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2, April 2007, pp. II–521–II–524.

[21] ——, "Nested support vector machines," *IEEE Transactions On Signal Processing*, vol. 58, pp. 1648–1660, 2010.

[22] P. Rigollet and R. Vert, "Optimal rates for plug-in estimators of density level sets," *Bernouilli*, vol. 15, pp. 1154–1178, 2009.

[23] D. M. J. Tax and R. P. W. Duin, "Uniform object generation for optimizing one-class classifiers," *Journal for Machine Learning Research*, pp. 155–173, 2001.

[24] S. Clémençon and J. Jakubowicz, "Scoring anomalies: a m-estimation formulation," in *AISTATS '13: Sixteenth international conference on Artificial Intelligence and Statistics*, vol. 31. Microtome Publishing, 2013, pp. 659–667.

[25] S. Clémençon and S. Robbiano, "Anomaly ranking as supervised bipartite ranking," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 343–351.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dufour, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[28] D. Harrison and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978.