Pattern Recognition 132 (2022) 108930

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

ℓ_p-Norm Support Vector Data Description

Shervin Rahimzadeh Arashloo*

Department of Computer Engineering, Faculty of Engineering, Bilkent University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 15 March 2022
Revised 22 June 2022
Accepted 21 July 2022
Available online 23 July 2022

Keywords:
One-class classification
Kernel methods
Support vector data description
ℓ_p-norm penalty

ABSTRACT

The support vector data description (SVDD) approach serves as a de facto standard for one-class classification where the learning task entails inferring the smallest hyper-sphere to enclose target objects while linearly penalising the errors/slacks via an ℓ_1-norm penalty term. In this study, we generalise this modelling formalism to a general ℓ_p-norm ($p \geq 1$) penalty function on slacks. By virtue of an ℓ_p-norm function, in the primal space, the proposed approach enables formulating a non-linear cost for slacks. From a dual problem perspective, the proposed method introduces a dual norm into the objective function, thus, proving a controlling mechanism to tune into the intrinsic sparsity/uniformity of the problem for enhanced descriptive capability.

A theoretical analysis based on Rademacher complexities characterises the generalisation performance of the proposed approach while the experimental results on several datasets confirm the merits of the proposed method compared to other alternatives.

1. Introduction

One-class classification (OCC) addresses the problem of recognising patterns that adhere to a specific condition presumed as normal, and identifying them from any other object violating the normality criterion. OCC stands apart from the conventional two-/multi-class classification paradigm [1] in that it primarily uses observations from a single, very often the target class for training. One-class classification acts as an essential building block in a diverse range of practical systems including presentation attack detection in biometrics [2], audio or video surveillance [3,4], intrusion detection [5], social network [6], *etc.*

As with many other machine learning problems, state-of-the-art OCC algorithms are built on the premise of deep learning methodology [7] using massive labelled datasets, typically containing millions of samples. Although deep structures have led to breakthroughs in one-class learning and classification, their reliance on huge sets of data may pose certain limitations in practice. In this context, collecting sufficiently large sets of training observations for certain applications can be a challenge, hindering a full exploitation of the expressive capacity of deep networks. Even if sufficient data is gathered, labelling such huge amounts of data may be a bigger challenge. Whilst crowd-sourcing may be considered as an applicable strategy to label huge sets of data in some fields, for a variety of different reasons including level of knowledge, data privacy, time required to produce accurate labels, *etc.* it may not serve as a viable option in the domains such as defence, security, healthcare. Although certain techniques such as active learning may be instrumental in reducing the quantity of necessary labelling/labelled data, they still demand the time and domain expertise of a human operator. In the absence of large training sets required by deep nets, and specifically for small to moderate-sized datasets containing hundreds or thousands of training samples, kernel-based methods offer a very promising methodology of classification. Moreover, unlike deep networks that incorporate many heuristics with regards to their structure and the corresponding (hyper)parameters, kernel methods are based on solid foundations and are characterised by strong bases in optimisation and statistical learning theory.

The support vector data description (SVDD) approach [8] which is proposed as an adaptation of support vector machines to the one-class setting, presents a very popular kernel-based method for one-class classification. Although designed for one-class setting, the SVDD approach does not require the training data to be exclusively and purely normal/positive which can be regarded as a quite appealing property in practical applications where the data is very often contaminated with noise and outliers. Furthermore, it provides an intuitive geometric characterisation of a predominantly positive dataset without making any specific assumption regarding the underlying distribution. Moreover, the SVDD decision making

* Corresponding author. Shervin Rahimzadeh Arashloo, Turkey
E-mail address: s.rahimzadeh@cs.bilkent.edu.tr

https://doi.org/10.1016/j.patcog.2022.108930
0031-3203/

process entails computing a simple distance between the centre of the target class and a test observation to label it as either normal (*i.e.* positive/target) or as anomaly (*i.e.* negative/outlier). And finally, when large sets of training data are available, the SVDD method may be extended to a deep structure to directly learn features from the data for improved performance [7]. These properties make SVDD a highly favoured method of practice in a variety of one-class classification applications where it serves as one of the most widely used techniques, if not the most.

The underlying idea in the SVDD approach is to determine the smallest hyper-sphere enclosing the data. While its hard-margin formulation requires all target data to be strictly encapsulated within the inferred hyper-spherical boundary, in the soft margin SVDD approach, in order to take into account the possibility of a contaminated dataset, the distance from each training object to the centre of the hyper-sphere need not be strictly smaller than the radius but larger distances are penalised. In order to encode and penalise violations from the hyper-spherical decision boundary, non-negative slack variables measuring the extent of violation of each object from the decision boundary are introduced. The optimisation problem is then modified to reflect such violations and penalise an ℓ_1-norm term on the slacks. In other words, the conventional SVDD method, and also the standard two-class SVM classifier are founded on the idea of minimising an ℓ_1-norm risk over the set of non-negative slack variables. In the context of two-class classification, very recently [9], the classical ℓ_1-norm penalty term in the SVM formulation has been revisited to consider two alternative slack penalties defined by the ℓ_2- and the ℓ_∞-norms to formulate new SVM algorithms. A reformulation of the standard two-class SVM to the ℓ_2 and ℓ_∞ penalty terms has been verified to improve the classification performance, sometimes significantly [9].

The standard SVDD approach operates on an ℓ_1-norm penalty over slacks. This represents a simplistic and restricting assumption since the object function in this case is linear w.r.t. the errors. In real-world problems, non-linear objective functions are widely deployed for learning purposes as they may better suit the non-linear characteristics of real data. Furthermore, in the standard SVDD formulation the fixed ℓ_1-norm may not effectively capture the inherent sparsity of the data and there exists no explicit controlling mechanism over sparsity of the solution. These represent some important limitations of the existing SVDD formulations which may inhibit their widespread deployment in practical problems. In spite of these limitations, a large body of the pursuant and ongoing work on SVDD has mainly focused on modifying the standard SVDD formulation via linear reweighting schemes. In this work, we address the aforementioned limitations and study the merits of different norm risks for "one-class" classification in the context of the SVDD approach. For this purpose, we consider a general ℓ_p-norm ($p \geq 1$) slack penalty term where p serves as a free parameter of the algorithm. As such, while in the standard SVDD method the slacks are penalised linearly, by introducing an ℓ_p-norm function, non-linear cost functions of the slacks may be optimised where the degree of the non-linearity (*i.e.* p) may be tuned on the data. By varying parameter $p \geq 1$, the relative magnitudes of errors corresponding to different samples may be modified, and thus, the relative importance of objects with larger errors w.r.t. others with smaller slacks in the objective function may be controlled. The reflection of the ℓ_p-norm penalty term onto the dual space formulation of the problem turns out to be a dual ℓ_q-norm ($q = \frac{p}{p-1}$), thus, providing the method the capability to tune into the inherent sparsity of the problem. By virtue of using $p \geq 1$, the objective function of the proposed ℓ_p-norm SVDD approach is convex, facilitating an effective optimisation. Through experiments on different datasets, it is shown that the introduction of a variable ℓ_p-norm penalty into the objective function is effective and possesses the potential to improve the generalisation capability, sometimes significantly.

1.1. Contributions

The major contributions of the current study may be summarised as listed below.

- We generalise the SVDD formulation from an ℓ_1 to an ℓ_p-norm penalty function and illustrate that the proposed generalisation may lead to significant improvements in the performance of the algorithm;
- We extend the proposed ℓ_p-norm formulation from a pure one-class setting to the training scenario where labelled negative objects are also available and illustrate the merits offered by the proposed extension;
- Based on Rademacher complexities, we theoretically study the generalisation performance of the proposed ℓ_p-norm approach and derive bounds on its error;
- And we carry out an experimental evaluation of the proposed method on multiple OCC datasets and provide a comparison to the original SVDD method and its different variants, as well as other OCC techniques from the literature.

1.2. Organisation

The remainder of the paper is structured as follows. In Section 2, the relevant literature with a particular emphasis on different variants of the SVDD formalism is reviewed. In Section 3, once a short overview of the support vector data description (SVDD) approach [8] is given, we present our proposed ℓ_p-norm SVDD approach for the pure one-class setting and then derive its extension for labelled negative training observations. Section 4 studies the generalisation error bound of the proposed approach based on Rademacher complexities. We present and analyse the results of an experimental evaluation of the proposed method in Section 5 where possible extensions of the proposed approach are also discussed. Finally, Section 6 concludes the paper.

2. Prior work

Although different categorisations of OCC methods exist in different studies [8], the one-class classification techniques may be roughly identified as either generative or non-generative [10], the latter best represented by discriminative approaches. While in the generative techniques, the objective is to model the underlying generative process of the data, the discriminative methods try to directly partition the observation space into different regions for classification. Discriminative approaches tend to yield better performance in practice since they try to explicitly solve the OCC problem without attempting to solve an intermediate and more general task of inferring the underlying distribution or generative process.

Generative OCC approaches encompass the methods that try to estimate the underlying distribution using, for example, Gaussian distribution modelling, or those which use a mixture of distributions [11]. A different sub-category of generative approaches includes methods that for decision making use the residual of reconstructing a test sample with respect to a hypothesised model, some instances of which are the kernel principal component analysis (KPCA) and its variants [12], or the autoencoder-based techniques. Discriminative methods constitute a strong alternative to the generative one-class learners. As an instance, based on a variant of the Fisher classification principle, the kernel null space method tries to map positive objects onto a single point in a discriminative subspace, obtaining very competitive results compared to some

other alternatives [13]. Another successful discriminative one-class method focuses on the use of Gaussian Process (GP) priors [14] trying to directly infer the a posteriori class probability of the target class. The work in [15] presents an incremental convex-concave hull one-class method, shown to be able to reduce the computational time while expanding the boundary of the normal class. Among others, a widely applied discriminative one-class classification method is that of support vector data description (SVDD) approach [8] that tries to estimate the smallest volume surrounding the positive objects. In the case of the existence of labelled negative training objects, the decision boundary is refined by requiring the negative objects to lie outside the hyper-spherical boundary. The soft version of this approach allows the positive and negative (if any) training objects to violate the boundary criterion but subject to a linear penalty on the extent of the violation (called *slack*) where a parameter controls the trade-off between the volume and such errors in the objective function. Due to its success in data description and its intuitive geometrical interpretation and the ability to benefit from a kernel-based representation, the SVDD approach serves as a widely used technique in the OCC literature, motivating many subsequent research. As an instance, in [16], based on the observation that the SVDD centre and the volume are sensitive to the parameter controlling the trade-off between the errors (slacks) and the volume, a method called GL-SVDD is proposed where local and global probability densities are used to derive sample-adaptive errors via associating weights to the slacks corresponding to different objects. In [17], a different sample-specific weighting approach (P-SVDD) based on the position of the feature space image is proposed to adaptively regularise the complexity of the SVDD sphere. The authors in [18] define a density-based distance between a sample point and the centre of the hyper-sphere to adjust the constraint set of the SVDD optimisation problem by re-weighting training objects. Apart from the research focused on improving the performance of the SVDD method in a one-class setting, there also exist other studies where the SVDD approach is generalised to two [19], or to multiple classes [20,21]. The work in [22], tries to improve the performance of the SVDD method by discovering the characteristics of the data and tuning model parameters via the chaotic bat algorithm. Other study [23], drawing on ergodicity of chaotic functions via switching automatically between global and local searches of the Bat algorithm, presented the so-called Chaotic Bat SVDD approach.

Considering the body of work discussed above, one observes that the majority of the existing studies tries to modify the slack error term by introducing an adaptive weighting for each data sample based on different cues. Clearly, a simple linear weighting scheme does change the linearity of the objective function with respect to the slacks. The exception to the studies above is the work in [24] where instead of an ℓ_1-norm penalty, an ℓ_2-norm penalty is considered over the slacks. As will be demonstrated in the subsequent sections, an ℓ_2 slack penalty may not always yield an optimal performance for data description. The current study is a generalisation of the existing SVDD formulations as it considers an ℓ_p ($p \geq 1$) slack norm penalty where p serves as a free parameter of the algorithm allowing for different non-linear penalties to be optimised w.r.t. slacks while at the same time providing the opportunity to tune into the inherent sparsity characteristics of the data.

3. Methodology

In this section, first, we briefly review the SVDD method [8] and then present the proposed approach.

3.1. Preliminaries

The Support Vector Data Description (SVDD) approach [8] tries to estimate the smallest hyperspherical volume that encloses normal/target data in some pre-determined feature space. As a hypersphere is characterised by its centre $\mathcal{O}$ and its radius $R > 0$, the learning problem in the SVDD method is defined as minimising the radius while requiring the hypersphere to encapsulate all normal objects $\mathbf{x}_i$'s, that is

$$\min_{R,\mathcal{O}} E(R,\mathcal{O}) = R^2$$
$$\text{s.t. } \|\mathbf{x}_i - \mathcal{O}\|_2^2 \leq R^2, \ \forall i \tag{1}$$

In practice, however, the training data might be contaminated with noise and outliers. In order to handle possible outliers in the training set and derive a solution with a better generalisation capability, the objective function in the SVDD method is modified so that the distance from the centre $\mathcal{O}$ to each training observation $\mathbf{x}_i$ need not be strictly smaller than R, rather, larger distances are penalised. For this purpose, using non-negative slack variables ζ_i's, the SVDD optimisation task is modified as

$$\min_{R,\mathcal{O},\boldsymbol{\zeta}} E(R,\mathcal{O},\boldsymbol{\zeta}) = R^2 + c\sum_i \zeta_i$$
$$\text{s.t. } \|\mathbf{x}_i - \mathcal{O}\|_2^2 \leq R^2 + \zeta_i, \ \zeta_i \geq 0, \ \forall i \tag{2}$$

where $\boldsymbol{\zeta}$ denotes a vector collection of ζ_i's and the trade-off between the sum of errors (*i.e.* ζ_i's) and the squared radius is controlled using parameter c. The optimisation problem above corresponds to the case where only normal samples (and possibly a minority of noisy objects) are presumed to exist in the training set. When labelled negative training objects are also available, the learning problem in the SVDD method is modified to enforce positive samples to lie within the hyper-sphere while negative samples are encouraged to fall outside its boundary.

The SVDD objective function in Eq. 2 depends on an ℓ_1-norm of the slack variables as $\sum_i \zeta_i = \|\boldsymbol{\zeta}\|_1$, and consequently, all errors/slacks are penalised linearly. Although penalising all errors linearly in their magnitudes is a plausible option, it is by no means the only possibly. An an instance, a different alternative may be to penalise only the maximum error/slack which can be achieved by incorporating an ℓ_∞-norm on the slacks as $\max_i \zeta_i = \|\boldsymbol{\zeta}\|_\infty$. Any other penalty which would lie between penalising all the slacks linearly and penalising only the maximum error may then be characterised using a general ℓ_p-norm on the errors, *i.e.* via $\sum_i \zeta_i^p = \|\boldsymbol{\zeta}\|_p^p$. In particular, introducing a variable norm parameter p opens the door to consider non-linear penalties on the errors compared with the original SVDD method which is limited to a linear penalty on the slacks. From a dual problem viewpoint, introducing an ℓ_p norm penalty on the slacks provides a mechanism to control the relative sparsity/uniformity of the solution in order to better consider the intrinsic sparsity of the problem. As such, in the proposed approach, we generalise the SVDD error function using an ℓ_p-norm function of slacks, discussed next.

3.2. ℓ_p-Norm SVDD

By replacing the ℓ_1-norm term on the slack variables in Eq. 2 with a function of ℓ_p-norm, the optimisation problem in the proposed approach is defined as

$$\min_{R,\mathcal{O},\boldsymbol{\zeta}} E(R,\mathcal{O}) = R^2 + c\sum_i \zeta_i^p$$
$$\text{s.t. } \|\mathbf{x}_i - \mathcal{O}\|_2^2 \leq R^2 + \zeta_i, \ \zeta_i \geq 0, \ \forall i \tag{3}$$

In the objective function above, by modifying parameter p the relative magnitudes of slacks with respect to each other may be controlled. In particular, by increasing p, the relative importance of objects with a larger error compared to those with a smaller error increases. In the limit when $p \to +\infty$, the largest error will have a

dominant impact on the objective function. In order to solve the optimisation problem above, the Lagrangian is formed as

$$\mathcal{L} = R^2 + c\sum_i \zeta_i^p - \sum_i \alpha_i[R^2 + \zeta_i - (\|\mathbf{x}_i\|_2^2 - 2\mathcal{O}^\top\mathbf{x}_i + \|\mathcal{O}\|_2^2)] - \sum_i \gamma_i\zeta_i \tag{4}$$

where α_i's and γ_i's are non-negative Lagrange multipliers. In order to derive the dual function, the Lagrangian should be minimised with respect to the primal variables R, $\mathcal{O}$, ζ_i. Setting the partial derivatives of $\mathcal{L}$ w.r.t. R, $\mathcal{O}$, and ζ_i to zero yields

$$\frac{\partial\mathcal{L}}{\partial R} = 0 \quad \Rightarrow \quad \sum_i \alpha_i = 1 \tag{5a}$$

$$\frac{\partial\mathcal{L}}{\partial \mathcal{O}} = 0 \quad \Rightarrow \quad \mathcal{O} = \sum_i \alpha_i\mathbf{x}_i \tag{5b}$$

$$\frac{\partial\mathcal{L}}{\partial \zeta_i} = 0 \quad \Rightarrow \zeta_i = \left(\frac{\alpha_i+\gamma_i}{cp}\right)^{\frac{1}{p-1}} \tag{5c}$$

Substituting the relations above into $\mathcal{L}$, the Lagrangian is obtained as

$$\mathcal{L} = (cp)^{\frac{-1}{p-1}}(1/p-1)\|\boldsymbol{\alpha}+\boldsymbol{\gamma}\|_{p/(p-1)}^{p/(p-1)} + \sum_i \alpha_i\mathbf{x}_i^\top\mathbf{x}_i - \sum_i\sum_j \alpha_i\alpha_j\mathbf{x}_i^\top\mathbf{x}_j \tag{6}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ denote vector collections of α_i's and γ_i's. Furthermore, one can easily check that the Slater's condition is satisfied, and thus, the following complementary conditions also hold at the optimum:

$$\gamma_i\zeta_i = 0, \forall i \tag{7a}$$

$$\alpha_i(R^2 + \zeta_i - \|\mathbf{x}_i - \mathcal{O}\|_2^2) = 0, \forall i \tag{7b}$$

Using Eq. 5c and Eq. 7a, it must hold that $\gamma_i\left(\frac{\alpha_i+\gamma_i}{cp}\right)^{\frac{1}{p-1}} = 0$. Since $\alpha_i \geq 0$ and $\gamma_i \geq 0$, one concludes that $\gamma_i = 0$, $\forall i$. As a result, the Lagrangian in Eq. 6 would be simplified as

$$\mathcal{L} = (cp)^{\frac{-1}{p-1}}(1/p-1)\|\boldsymbol{\alpha}\|_{p/(p-1)}^{p/(p-1)} + \sum_i \alpha_i\mathbf{x}_i^\top\mathbf{x}_i - \sum_i\sum_j \alpha_i\alpha_j\mathbf{x}_i^\top\mathbf{x}_j \tag{8}$$

The dual problem entails maximising $\mathcal{L}$ in $\boldsymbol{\alpha}$:

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{L}$$
$$\text{s.t. } \boldsymbol{\alpha} \geq 0, \|\boldsymbol{\alpha}\|_1 = 1 \tag{9}$$

Note that, for $p \geq 1$, we have $p/(p-1) \geq 1$, and consequently, the term $\|\boldsymbol{\alpha}\|_{p/(p-1)}^{p/(p-1)}$ in the Lagrangian is convex w.r.t. $\boldsymbol{\alpha}$. Note also that the other terms in $\mathcal{L}$ are either linear or quadratic functions of $\boldsymbol{\alpha}$, and hence, are convex while the constraints are affine. As a result, the optimisation problem above is a convex optimisation task.

3.3. ℓ_p-norm SVDD with negative samples

In the proposed ℓ_p-norm approach, similar to the original SVDD method [8], when labelled non-target/negative training observations are available, they may be utilised to refine the description. In this case, as opposed to the positive samples that should be enclosed within the hypersphere, the non-target objects should lie outside its boundary. In what follows, the normal/positive samples are indexed by i, j and the negative objects by l, m. In order to allow for possible errors in both the positive and the negative training samples, slack variables ζ_i's and ζ_l's are introduced. The optimisation problem when labelled negative samples are available is then defined as

$$\min_{R,\mathcal{O},\boldsymbol{\zeta}} E(R,\mathcal{O},\boldsymbol{\zeta}) = R^2 + c_1\sum_i \zeta_i^p + c_2\sum_l \zeta_l^p$$
$$\text{s.t. } \|\mathbf{x}_i - \mathcal{O}\|_2^2 \leq R^2+\zeta_i, \ \|\mathbf{x}_l - \mathcal{O}\|_2^2 \geq R^2 - \zeta_l, \ \zeta_i \geq 0, \ \zeta_l \geq 0, \ \forall i,l \tag{10}$$

In the objective function above, while c_1 may be used to control the fraction of positive training objects that fall outside the hypersphere boundary, c_2 may be adjusted to regulate the fraction of negative training samples that will lie within the hypersphere. By introducing Lagrange multipliers $\alpha_i \geq 0$, $\alpha_l \geq 0$, $\gamma_i \geq 0$, $\gamma_l \geq 0$, the Lagrangian of Eq. 10 is formed as

$$\mathcal{L} = R^2 + c_1\sum_i \zeta_i^p + c_2\sum_l \zeta_l^p - \sum_i \gamma_i\zeta_i - \sum_l \gamma_l\zeta_l - \sum_i \alpha_i[R^2 + \zeta_i - (\|\mathbf{x}_i\|_2^2 - 2\mathcal{O}^\top\mathbf{x}_i + \|\mathcal{O}\|_2^2)] - \sum_l \alpha_l[(\|\mathbf{x}_l\|_2^2 - 2\mathcal{O}^\top\mathbf{x}_l + \|\mathcal{O}\|_2^2) - R^2 + \zeta_l] \tag{11}$$

In order to form the dual function, the Lagrangian should be minimised w.r.t. R, $\mathcal{O}$, ζ_i's, and ζ_l's. Setting the partial derivatives of $\mathcal{L}$ w.r.t. to R, $\mathcal{O}$, ζ_i, and ζ_l to zero yields

$$\frac{\partial\mathcal{L}}{\partial R} = 0 \quad \Rightarrow \quad \sum_i \alpha_i - \sum_l \alpha_l = 1 \tag{12a}$$

$$\frac{\partial\mathcal{L}}{\partial \mathcal{O}} = 0 \quad \Rightarrow \quad \mathcal{O} = \sum_i \alpha_i\mathbf{x}_i - \sum_l \alpha_l\mathbf{x}_l \tag{12b}$$

$$\frac{\partial\mathcal{L}}{\partial \zeta_i} = 0 \quad \Rightarrow \zeta_i = \left(\frac{\alpha_i+\gamma_i}{c_1 p}\right)^{\frac{1}{p-1}} \tag{12c}$$

$$\frac{\partial\mathcal{L}}{\partial \zeta_l} = 0 \quad \Rightarrow \zeta_l = \left(\frac{\alpha_l+\gamma_l}{c_2 p}\right)^{\frac{1}{p-1}} \tag{12d}$$

Resubstituting the relations above into Eq. 11 gives

$$\mathcal{L} = (c_1 p)^{\frac{-1}{p-1}}(1/p-1)\|\boldsymbol{\alpha}_T + \boldsymbol{\gamma}_T\|_{p/(p-1)}^{p/(p-1)} + (c_2 p)^{\frac{-1}{p-1}}(1/p-1)\|\boldsymbol{\alpha}_N + \boldsymbol{\gamma}_N\|_{p/(p-1)}^{p/(p-1)} + \sum_i \alpha_i\mathbf{x}_i^\top\mathbf{x}_i - \sum_l \alpha_l\mathbf{x}_l^\top\mathbf{x}_l - \sum_i\sum_j \alpha_i\alpha_j\mathbf{x}_i^\top\mathbf{x}_j - \sum_l\sum_m \alpha_l\alpha_m\mathbf{x}_l^\top\mathbf{x}_m + 2\sum_i\sum_l \alpha_i\alpha_l\mathbf{x}_i^\top\mathbf{x}_l \tag{13}$$

where $\boldsymbol{\alpha}_T$ and $\boldsymbol{\alpha}_N$ respectively stand for vector collections of α_i's and α_l's. Similarly, $\boldsymbol{\gamma}_T$ and $\boldsymbol{\gamma}_N$ denote vector collections of γ_i's and γ_l's, respectively. Since the Slater's condition holds, the following complementary conditions are also satisfied at the optimum:

$$\gamma_i\zeta_i = 0, \forall i \tag{14a}$$

$$\gamma_l\zeta_l = 0, \forall l \tag{14b}$$

$$\alpha_i(R^2 + \zeta_i - \|\mathbf{x}_i - \mathcal{O}\|_2^2) = 0, \forall i \tag{14c}$$

$$\alpha_l(R^2 - \zeta_l - \|\mathbf{x}_l - \mathcal{O}\|_2^2) = 0, \forall l \tag{14d}$$

Using Eqs. 12c and 14a, and also Eqs. 12d and 14b, one concludes that $\gamma_i = 0$, $\forall i$ and $\gamma_l = 0$, $\forall l$. As a result, the Lagrangian in Eq. 13 would be

$$\mathcal{L} = (c_1 p)^{\frac{-1}{p-1}} (1/p-1) \|\boldsymbol{\alpha}_T\|_{p/(p-1)}^{p/(p-1)} + (c_2 p)^{\frac{-1}{p-1}} (1/p-1) \|\boldsymbol{\alpha}_N\|_{p/(p-1)}^{p/(p-1)} + \sum_i \alpha_i \mathbf{x}_i^\top \mathbf{x}_i - \sum_l \alpha_l \mathbf{x}_l^\top \mathbf{x}_l - \sum_i \sum_j \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_l \sum_m \alpha_l \alpha_m \mathbf{x}_l^\top \mathbf{x}_m + 2 \sum_i \sum_l \alpha_i \alpha_l \mathbf{x}_i^\top \mathbf{x}_l \tag{15}$$

The dual problem then reads

$$\begin{aligned} &\max_{\boldsymbol{\alpha}_T, \boldsymbol{\alpha}_N} \quad \mathcal{L} \\ &\text{s.t. } \boldsymbol{\alpha}_T \geq 0, \boldsymbol{\alpha}_N \geq 0, \\ &\quad \|\boldsymbol{\alpha}_T\|_1 - \|\boldsymbol{\alpha}_N\|_1 = 1 \end{aligned} \tag{16}$$

Since $p \geq 1$ leads to $p/(p-1) \geq 1$, the terms $\|.\|_{p/(p-1)}^{p/(p-1)}$ in the Lagrangian are convex while the remaining terms are either linear or quadratic functions and the constraint sets are affine. Subsequently, the maximisation problem in Eq. 16 is convex.

3.4. Joint formulation

As discussed earlier, when only positive labelled training observations are available, in the proposed approach one solves the optimisation problem in Eq. 9 with the Lagrangian given in Eq. 8. When in addition to the positive training samples, labelled negative training objects are also available, the problem to be solved is expressed as the optimisation task in Eq. 16 with the corresponding Lagrangian given in Eq. 15. Although the optimisation tasks corresponding to the pure positive case and that of the second scenario where negative training samples are also available may appear different, nevertheless, both optimisation problems can be expressed compactly using a joint formulation as follows. Let us assume that vector $\mathbf{y}$ corresponds to the labels of training samples where for positive objects the label is $+1$ while for any possible non-target training samples the corresponding label is -1. Furthermore, suppose the Lagrange multipliers associated with the negative and positive samples are all collected into a single vector $\boldsymbol{\alpha}$. In order to reduce the clutter in the formulation, let us further assume $q = p/(p-1)$, $\bar{c_1} = \frac{1}{2}(c_1 p)^{\frac{-1}{p-1}}(1 - 1/p)$ and $\bar{c_2} = \frac{1}{2}(c_2 p)^{\frac{-1}{p-1}}(1 - 1/p)$. With these definitions, the Lagrangian in Eq. 15 may be expressed as

$$\mathcal{L} = -\bar{c_1} \|\boldsymbol{\alpha} \odot (1 + \mathbf{y})\|_q^q - \bar{c_2} \|\boldsymbol{\alpha} \odot (1 - \mathbf{y})\|_q^q + \sum_i \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_i - \sum_{i,j} \alpha_i y_i \alpha_j y_j \mathbf{x}_i^\top \mathbf{x}_j \tag{17}$$

where $\odot$ denotes hadamard/elementwise product. It may be easily verified that when only positive training samples are available, the Lagrangian above correctly recovers that of Eq. 8 while in the existence of labelled negative training objects, it matches that of Eq. 15. As a result, in the proposed approach, the generic optimisation problem to solve can be expressed as

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \quad \mathcal{L} \\ &\text{s.t. } \boldsymbol{\alpha} \geq 0, \mathbf{y}^\top \boldsymbol{\alpha} = 1 \end{aligned} \tag{18}$$

where $\mathbf{y}$ is the vectors of labels and the Lagrangian $\mathcal{L}$ is given as Eq. 17.

3.5. Kernel space representation

In many practical applications, instead of a rigid boundary, a more elastic description is favoured. In such cases, a reproducing kernel Hilbert space representation may be adopted. Inspecting the Lagrangian in Eq. 17, it can be observed that the training samples only appear in terms of inner products which facilitates deriving a kernel-space representation for the proposed approach. Since in the kernel space it holds that $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ where $\kappa(.,.)$ is the kernel function, the Lagrangian in the reproducing kernel Hilbert space may be written as

$$\mathcal{L} = -\bar{c_1} \|\boldsymbol{\alpha} \odot (1 + \mathbf{y})\|_q^q - \bar{c_2} \|\boldsymbol{\alpha} \odot (1 - \mathbf{y})\|_q^q + \sum_i \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}_i) - (\boldsymbol{\alpha} \odot \mathbf{y})^\top \mathbf{K} (\boldsymbol{\alpha} \odot \mathbf{y}) \tag{19}$$

where $\mathbf{K}$ denotes the kernel matrix. If additionally, all objects have unit length in the feature space, *i.e.* if $\phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_i) = 1$, one may further simplify the Lagrangian. For this propose, note that as for normalised feature vectors we have $\sum_i \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}_i) = \mathbf{y}^\top \boldsymbol{\alpha}$ and since due to the constraints imposed it must hold that $\mathbf{y}^\top \boldsymbol{\alpha} = 1$, the term $\sum_i \alpha_i y_i \kappa(\mathbf{x}_i, \mathbf{x}_i)$ can be safely dropped from the objective function without affecting the result. As a result, the optimisation problem for unit-length features shall be

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \bar{c_1} \|\boldsymbol{\alpha} \odot (1 + \mathbf{y})\|_q^q + \bar{c_2} \|\boldsymbol{\alpha} \odot (1 - \mathbf{y})\|_q^q + (\boldsymbol{\alpha} \odot \mathbf{y})^\top \mathbf{K} (\boldsymbol{\alpha} \odot \mathbf{y}) \\ & \text{s.t. } \boldsymbol{\alpha} \geq 0, \mathbf{y}^\top \boldsymbol{\alpha} = 1 \end{aligned} \tag{20}$$

As a widely used kernel function, the Gaussian kernel by definition, yields unit-length feature vectors in the kernel space, and the formulation above is applicable. Note that when $p \to +1$, then $q \to +\infty$, and the terms based on ℓ_q-norm tend to functions of $\ell_{+\infty}$-norm which is equal to the maximum element of a vector. Using the equivalency between an Ivanov and a Tikhonov regularisation, and the fact that bounding the maximum element of a vector bounds all the elements of the vector with the same constant, the norm terms appear as upper bound constraints on the dual variable which yields exactly the same constrained optimisation problem as that of the standard SVDD approach. Alternatively, when $p = 1$, setting the gradient of the Lagrangian in Eq. 11 w.r.t. ζ_i's and ζ_l's to zero will yield upper bound constraints on $\boldsymbol{\alpha}$, similar to the standard SVDD method in [8].

3.6. Decision strategy

Similar to the conventional SVDD approach, for decision making in the proposed ℓ_p-norm method, the distance of an object to the centre of the description is gauged and employed as a dissimilarity criterion. The distance of an object $\mathbf{z}$ to the centre of the hypersphere $\mathcal{O}$ in the kernel space is

$$f(\mathbf{z}) = \|\phi(\mathbf{z}) - \phi(\mathcal{O})\|_2^2 = \kappa(\mathbf{z}, \mathbf{z}) - 2 \sum_i \alpha_i y_i \kappa(\mathbf{z}, \mathbf{x}_i) + (\boldsymbol{\alpha} \odot \mathbf{y})^\top \mathbf{K} (\boldsymbol{\alpha} \odot \mathbf{y}) \tag{21}$$

In order to compute the radius of the description, note that the complementary conditions in Eqs. 14c and 14d may be compactly represented as $\alpha_j (R^2 + y_j \zeta_j - \left\|\phi(\mathbf{x}_j) - \phi(\mathcal{O})\right\|_2^2) = 0$. As a result, if for an object $\mathbf{x}_j$ the corresponding Lagrange multiplier α_j is non-zero, it must hold that $R^2 + y_j \zeta_j - \left\|\phi(\mathbf{x}_j) - \phi(\mathcal{O})\right\|_2^2 = 0$, and hence, the radius of the description may be computed as

$$\begin{aligned} R^2 &= \left\|\phi(\mathbf{x}_j) - \phi(\mathcal{O})\right\|_2^2 - y_j \zeta_j \\ &= \kappa(\mathbf{x}_j, \mathbf{x}_j) - 2 \sum_i \alpha_i y_i \kappa(\mathbf{x}_j, \mathbf{x}_i) + (\boldsymbol{\alpha} \odot \mathbf{y})^\top \mathbf{K} (\boldsymbol{\alpha} \odot \mathbf{y}) - y_j \zeta_j \end{aligned} \tag{22}$$

where j indexes an object whose corresponding Lagrange multiplier α_j is non-zero. The objects whose distance to the centre of the hyper-sphere is larger than the radius (subject to some margin) would be classified as novel.

4. Generalisation error bound

In this section, using the Rademacher complexities, we characterise the generalisation error bound for the proposed ℓ_p-norm SVDD approach.

Theorem 1. *Let us assume $\mathcal{F}$ corresponds to a class of kernel-based linear functions:*

$$\mathcal{F} = \{x \to \mathbf{w}^\top \phi(x), \|\mathbf{w}\|_2 \leq B\} \tag{23}$$

then the empirical Rademacher complexity of function class $\mathcal{F}$ over samples $(\mathbf{x}_i)_{i=1}^n$, denoted as $\hat{\mathcal{R}}_n(\mathcal{F})$, is bounded as [25]

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{2B}{n}\sqrt{tr(\mathbf{K})} \leq \frac{2BB_\kappa}{\sqrt{n}} \tag{24}$$

where $tr(.)$ denotes matrix trace and $\mathbf{K}$ stands for the kernel matrix associated with the feature mapping $\phi(.)$ and B_κ^2 is an upper bound on the kernel function $\kappa(.,.)$.

Next, we present the main theorem concerning the generalisation performance of the proposed approach.

Theorem 2. *In the proposed approach, assuming that υ is a margin parameter, with confidence greater than $1-\Delta$, a test point $\mathbf{x}$ is incorrectly classified with the probability bounded as*

$$P[y\big(f(\mathbf{x}) - R^2\big) > \upsilon] \leq \frac{1}{n\upsilon^p}\|\boldsymbol{\zeta}\|_p^p + \frac{4pBB_\kappa}{\upsilon^p\sqrt{n}}(B^2 + 3B_\kappa^2 + R^2)^{p-1} + 3\sqrt{\frac{\ln(2/\Delta)}{2n}} \tag{25}$$

where y is the ground truth label for observation $\mathbf{x}$.

For the proof of Theorem 2, first, we review a few relevant theories and then present the proof.

Theorem 3. *Assume $\delta \in (0,1)$ and suppose $\mathcal{G}$ is a function class from X to [0,1]. Let $(\mathbf{x}_i)_{i=1}^n$ be independent samples that are drawn according to a probability distribution $\mathcal{D}$. Then with a probability higher than $1-\Delta$ over $(\mathbf{x}_i)_{i=1}^n$, for each $g \in \mathcal{G}$ it holds that [25]*

$$\mathbb{E}_{\mathcal{D}}[g(\mathbf{x})] \leq \hat{\mathbb{E}}[g(\mathbf{x})] + \hat{\mathcal{R}}_n(\mathcal{G}) + 3\sqrt{\frac{\ln(2/\Delta)}{2n}} \tag{26}$$

where $\hat{\mathbb{E}}[g(\mathbf{x})]$ is the empirical expectation of $g(\mathbf{x})$ on the random sample set $(\mathbf{x}_i)_{i=1}^n$ and $\hat{\mathcal{R}}_n(\mathcal{G})$ denotes the empirical Rademacher complexity of the function class $\mathcal{G}$.

Theorem 4. *If $\mathcal{A}: \mathbb{R} \to \mathbb{R}$ is L-Lipschitz and satisfies $\mathcal{A}(0) = 0$, then the empirical Rademacher complexity of the composition function class $\mathcal{A} \circ \mathcal{F}$ satisfies $\hat{\mathcal{R}}_n(\mathcal{A} \circ \mathcal{F}) \leq 2L\hat{\mathcal{R}}_n(\mathcal{F})$ [25].*

Towards the proof of Theorem 2, we present the following theorem.

Theorem 5. *Let us consider $h(\mathbf{x})$ as the hypothesis function defined as $h(\mathbf{x}) = f(\mathbf{x}) - R^2$ where $f(\mathbf{x})$ measures the distance of sample $\mathbf{x}$ with label y to the centre of the hypersphere in the feature space (see Eq. 21). For some fixed margin $\upsilon > 0$, we define $g(.)$ as*

$$g(\mathbf{x}) = \mathcal{A}(yh(\mathbf{x})) = \begin{cases} 0 & \text{if } yh(\mathbf{x}) \leq 0; \\ \big(yh(\mathbf{x})/\upsilon\big)^p & \text{if } 0 \leq yh(\mathbf{x}) \leq \upsilon; \\ 1 & \text{else.} \end{cases} \tag{27}$$

$\mathcal{A}: \mathbb{R} \to [0,1]$ is L-Lipschitz and satisfies $\mathcal{A}(0) = 0$. Then with a probability higher than $1-\Delta$ over $(\mathbf{x}_i)_{i=1}^n$ it holds

$$\mathbb{E}_{\mathcal{D}}[g(\mathbf{x})] \leq \frac{1}{\upsilon^p n}\|\boldsymbol{\zeta}\|_p^p + \frac{4Bp}{n\upsilon^p}(B^2 + 3B_\kappa + R^2)^{p-1}\sqrt{tr(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\Delta)}{2n}} \tag{28}$$

Proof. We have

$$\hat{\mathbb{E}}[g(\mathbf{x})] = \frac{1}{n}\sum_i g(\mathbf{x}_i) \leq \frac{1}{n\upsilon^p}\sum_i \big(y_i(f(\mathbf{x}_i) - R^2)\big)_+^p = \frac{1}{n\upsilon^p}\sum_i \zeta_i^p = \frac{1}{n\upsilon^p}\|\boldsymbol{\zeta}\|_p^p \tag{29}$$

where $(z)_+ = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{otherwise.} \end{cases}$ and $\boldsymbol{\zeta}$ stands for a vector collection of all ζ_i's. Note that $\mathcal{A}(.)$ is Lipschitz with constant L. As with the zero-one loss, the margin loss above penalises any misclassified objects but also penalises h when it correctly classifies an object with low confidence. In order to derive an upper bound on L, observe that $\frac{\partial \mathcal{A}}{\partial\big(yh(\mathbf{x})\big)} = \frac{p}{\upsilon^p}\big(y(f(\mathbf{x}) - R^2)\big)^{p-1}$, and consequently, we have

$$\left\| \frac{\partial \mathcal{A}}{\partial\big(yh(\mathbf{x})\big)} \right\|_2 = \frac{p}{\upsilon^p}\big\|(f(\mathbf{x}) - R^2)\big\|_2^{p-1} \leq \frac{p}{\upsilon^p}(\|f(\mathbf{x})\|_2 + R^2)^{p-1} \tag{30}$$

Since the kernel function is bounded by B_κ^2, using Eq. 21, and the fact that $\|\mathbf{w}\|_2^2 = \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}$ [25], we have

$$\|f(\mathbf{x})\|_2 \leq B^2 + 3B_\kappa^2 \tag{31}$$

and hence

$$\left\| \frac{\partial \mathcal{A}}{\partial\big(yh(\mathbf{x})\big)} \right\|_2 \leq \frac{p}{\upsilon^p}(B^2 + 3B_\kappa^2 + R^2)^{p-1} \tag{32}$$

As a result, $\mathcal{A}(.)$ is Lipschitz with constant $L = \frac{p}{\upsilon^p}(B^2 + 3B_\kappa^2 + R^2)^{p-1}$.

Next, using Theorem 4 and Theorem 1, we have

$$\hat{\mathcal{R}}_n(\mathcal{G}) \leq 2L\hat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{4BB_\kappa L}{\sqrt{n}} \leq \frac{4pBB_\kappa}{\upsilon^p\sqrt{n}}(B^2 + 3B_\kappa^2 + R^2)^{p-1} \tag{33}$$

Using Eq. 29 and Eq. 33 in Theorem 3, the proof to Theorem 5 is complete. Since we have $P[y\big(f(\mathbf{x}) - R^2\big) > \upsilon] \leq \mathbb{E}_{\mathcal{D}}[g(\mathbf{x})]$, using Theorem 5, the proof to Theorem 2 is completed. □

As may be observed from Eq. 25, parameter p directly affects the expected loss on the training set (the first term on the RHS of the equation) and also controls the Rademacher complexity (the second term on the RHS of Eq. 25) of the proposed method. As the error probability varies as a function of p, the utility of a free norm parameter in the proposed approach is justified. Note that depending on $\boldsymbol{\zeta}$ and the margin parameter υ, setting $p = 1$ may not necessarily minimise the RHS in Eq. 25, and hence, may lead to an increased probability of misclassification in the proposed approach. In practice, the norm parameter p may be adjusted according to the characteristics of the data to optimise the performance or to control the false acceptance/rejection rate. Note also, since parameter p appears in the dual problem as $\|.\|_{p/(p-1)}^{p/(p-1)}$ terms (see Eq. 20), it also affects the sparsity of $\boldsymbol{\alpha}$.

5. Experiments

In this section, an experimental evaluation of the proposed approach is conducted where we provide a comparison to some other variants of the SVDD approach as well as to baseline approaches on multiple datasets. The rest of this section is organised as detailed next.

- In Section 5.1, we visualise the decision boundaries inferred by the proposed approach for different p's for synthetic data.
- In Section 5.2, the implementation details, the experimental set-up, and the standard datasets used in the experiments are discussed.

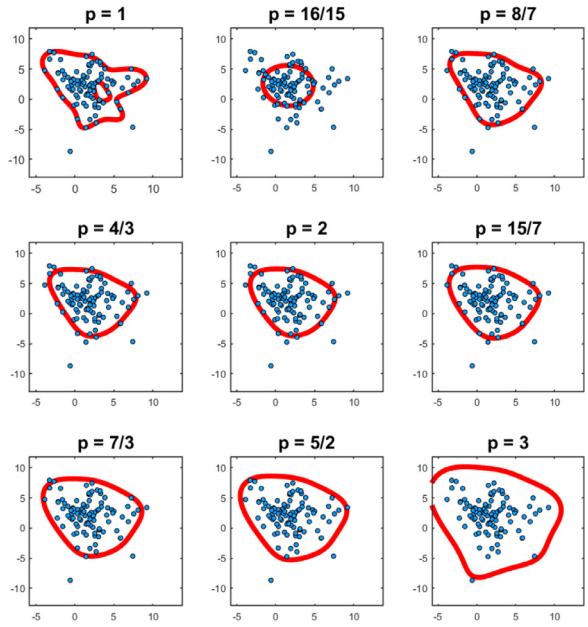

Fig. 1. Decision boundaries for the proposed ℓ_p-SVDD approach with a Gaussian kernel function for different values of p for 100 normally distributed random samples with mean of 2 and standard deviation of 3 in each dimension. ($p = 1$ corresponds to the original SVDD method).

- In Section 5.3, the results of an experimental evaluation of the proposed approach in a pure one-class setting (labelled negative objects unavailable) are presented and compared with other methods on multiple datasets.
- Section 5.4 provides the results of an experimental evaluation of the proposed approach in the presence of negative training samples along with a comparison against other methods on multiple datasets.

5.1. Decision boundaries

In order to visualise the effect of norm parameter p on the inferred decision boundaries, we randomly generate 100 normally distributed 2D samples with a mean of 2 and standard deviation of 3 in each direction. Using a Gaussian kernel function, the proposed approach is then run to derive a description of the data. The experiment is repeated for different values of $p \in \{1, 16/15, 8/7, 4/3, 2, 15/7, 7/3, 5/2, 3\}$ where $p = 1$ corresponds to the original SVDD method in [8]. The decision boundaries superimposed on the data are visualised in Fig. 1. From the figure, it may be observed that for the case of $p = 1$ the method has inferred a boundary which separates a region of relatively low density in the middle of the distribution from the rest of the 2D space. For the random data samples generated in this experiment with a mean of (2,2) this clearly indicates a case of over-fitting. By increasing the norm parameter above 1, the decision boundary better covers the mean of the distribution. More specifically, while for smaller values of p the boundary is tighter, for larger values the description tends to encapsulate a higher percentage of the normal samples. As will be discussed in the following sections, in the proposed method, we tune parameter p using the validation set corresponding to each dataset.

5.2. Implementation details

In the experiments that follow, the features are first standardised by subtracting the mean computed over all positive training samples and then dividing by the standard deviation followed by normalising each feature vector to have a unit ℓ_2-norm. For datasets D1 : 20, the positive samples are divided randomly into three non-overlapping equal size subsets to form the training, validation, and the test sets. Similarly, the negative samples are divided randomly into three disjoint equal size subsets for training, validation and testing purposes. In order to minimise possible effects of random data partitioning on the performance, we repeat the procedure above 10 times, and record the mean along with the standard deviation of the performance over these 10 trials. For dataset D21, in order to make the results consistent and comparable with the relevant literature, we follow the standard zero-shot evaluation protocol for this dataset [26]. We set the parameters of all methods on the corresponding validation subset of each dataset. In particular, for the proposed approach $p \in \{32/31, 16/15, 8/7, 6/5, 4/3, 3/2, 2, 5/2, 5, 20\}$ and $c_1, c_2 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. In all experiments, we use a Gaussian kernel the width of which is set to half of the average pairwise Euclidean distance among all training objects. As the dual problem in Eq. 20 is convex, one may use different algorithms [27] for optimisation. In this work, we use CVX [28], a package for solving convex programmes.

In order to evaluate the proposed approach, 20 benchmark databases from the UCI repository [29], TUDelf University [30], KEEL repository [31], CENPARMI [32], Statlib [33], and Zalando [34] are used. Furthermore, we use the Replay-Mobile face presentation attack detection dataset [26] which is a relatively larger dataset where GoogleNet representations extracted from the face region are used as features [35]. The datasets used in the experiments correspond to different application domains from varied sources. The statistics of the datasets used in the experiments are reported in Table 1. For the evaluation of the proposed approach, we conduct two sets of experiments. The first set follows a pure one-class classification paradigm, *i.e.* only positive samples are used to train the models. In the second set of experiments, negative objects are also deployed for model training. For comparison, we report the performance of the original SVDD approach of [8] denoted as "ℓ_1-SVDD" and also its alternative variant which considers squared errors in the objective function, denoted as "ℓ_2-SVDD" [24]. The proposed approach is denoted as "ℓ_p-SVDD" in the corresponding tables. We also provide a comparison of the proposed ℓ_p-SVDD method to some linear re-weighting variants of the SVDD approach including P-SVDD [17], DW-SVDD [36], and GL-SVDD [16] as well as state-of-the-art OCC techniques. In particular, we have included kernel-based one-class classifiers which are applicable to moderately-sized datasets. These are the kernel Gaussian Process method (GP) [14], the Kernel Null Foley-Sammon Transform (KNFST) [13,37], and the Kernel Principal Component Analysis (KPCA) [38].

Following the common approach in the literature and in order to facilitate the comparison of the performances of different methods independent from a specific operating threshold, we report the performances in terms of the AUC measure which is the area under the Receiver Operating Characteristic curve (ROC). The ROC curve characterises the true positive rate against the false positive rate at various operating thresholds. A higher AUC indicates a better performance for the system.

5.3. Pure one-class setting

In this setting, only positive objects are used for training. Table 2 reports the performances of different methods in this setting where we set parameter p on the validation subset of each dataset to maximise the performance. A number of observations from the table are in order. First, on all datasets the proposed ℓ_p-SVDD approach yields a superior performance compared to its ℓ_1-SVDD and ℓ_2-SVDD variants. In particular, on some datasets such

Table 1
Statistics of the datasets used.

Abbrev.	Dataset	#total	#positive	dim.	Source
D1	Iris (virginica)	150	50	4	UCI
D2	Hepatitis (normal)	155	123	19	UCI
D3	Ecoli (periplasm)	336	52	7	UCI
D4	Concordia16 (digit 1)	4000	400	256	CENPARMI
D5	Delft pump (5x1)	336	133	64	Delft
D6	Balance-scale (middle)	625	49	4	UCI
D7	Wine (2)	178	71	13	UCI
D8	Waveform (0)	900	300	21	UCI
D9	Survival (<5yr)	306	81	3	UCI
D10	Housing (MEDV>35)	506	48	13	Statlib
D11	glass6	214	185	9	KEEL
D12	haberman	306	225	3	KEEL
D13	led7digit	443	406	7	KEEL
D14	pima	768	500	8	KEEL
D15	wisconsin	683	444	9	KEEL
D16	yeast(0-5-6-7-9-vs-4)	528	477	8	KEEL
D17	cleveland(0-vs-4)	177	164	13	KEEL
D18	Breast(benign)	699	458	9	UCI
D19	Survival (>5yr)	306	225	3	UCI
D20	FMNIST (Class "1")	1926	1027	784	Zalando
D21	Replay-Mobile	100885	45,233	1024	Idiap

Table 2
Comparison of the performance of OCC approaches on different datasets in a pure one-class scenario in terms of %AUC (mean±std).

	GP	KPCA	KNFST	P-SVDD	DW-SVDD	GL-SVDD	ℓ_1-SVDD	ℓ_2-SVDD	ℓ_p-SVDD
D1	68.03±16.51	78.17±22.69	68.51±16.46	78.98±15.58	67.42±16.75	67.70±16.39	67.42±16.75	67.37±16.52	**81.23 ± 11.26**
D2	58.78±4.37	66.46±6.75	58.65±4.39	62.51±5.98	59.36±4.37	61.20±6.06	59.98±4.31	59.48±4.26	**66.60 ± 6.80**
D3	61.30±11.33	53.06±13.66	61.32±11.15	53.98±12.31	61.67±11.66	61.08±10.49	61.08±11.36	61.20±11.46	**62.19 ± 10.53**
D4	92.96±2.60	94.10±1.73	92.94±2.62	93.24±2.50	93.31±2.54	93.21±2.53	94.36±1.84	93.58±2.42	**94.43 ± 1.58**
D5	87.87±3.43	86.07±3.29	87.85±3.45	84.78±3.92	87.85±3.45	88.02±3.33	87.85±3.45	87.90±3.40	**88.41 ± 2.77**
D6	94.52±3.57	90.77±2.46	94.66±3.52	92.94±3.16	94.68±3.54	94.64±3.54	94.70±3.55	94.60±3.56	**94.79 ± 3.59**
D7	60.82±11.03	72.30±11.10	60.71±10.99	65.64±9.04	61.12±11.57	64.12±10.85	60.79±11.04	60.90±11.19	**72.53 ± 11.36**
D8	64.15±4.66	61.47±1.90	64.10±4.64	64.09±5.91	65.37±5.09	65.39±5.41	64.91±5.53	65.10±5.48	**65.86 ± 4.23**
D9	53.41±18.99	47.37±14.10	59.66±16.34	52.29±19.26	58.28±21.15	66.40±12.11	64.43±14.87	60.72±15.26	**67.40 ± 10.21**
D10	87.73±6.00	78.70±8.49	87.76±6.02	81.28±11.89	87.77±6.11	87.82±6.03	87.77±6.11	87.78±6.08	**87.91 ± 5.92**
D11	87.81±10.11	95.44±2.59	87.55±10.65	95.48±3.53	86.90±10.06	96.58±2.18	94.24±2.53	90.68±5.67	**96.73 ± 1.38**
D12	59.70±5.90	70.66±5.79	55.54±5.51	59.75±7.79	55.84±7.12	**70.77 ± 5.96**	58.34±7.41	66.02±4.94	**70.77 ± 5.88**
D13	62.63±18.55	67.49±7.79	62.34±18.01	66.60±15.73	41.54±6.43	66.77±14.25	62.84±13.62	69.04±10.99	**69.41 ± 9.62**
D14	51.66±5.36	**71.43 ± 3.41**	51.46±5.32	53.47±5.42	54.93±4.78	56.04±4.46	61.24±4.95	59.83±3.75	**71.43 ± 3.57**
D15	52.70±12.37	95.88±1.05	53.51±12.56	65.21±12.28	47.10±14.97	70.68±10.68	93.39±1.73	68.75±10.83	**95.91 ± 1.03**
D16	57.74±9.93	63.71±7.39	58.56±8.80	62.94±8.48	61.02±10.33	63.26±9.98	60.93±9.65	61.52±10.68	**64.23 ± 8.29**
D17	51.49±13.99	79.13±9.71	51.78±14.04	57.93±13.99	51.09±13.90	51.85±14.47	51.09±13.90	51.35±14.06	**79.38 ± 9.27**
D18	**45.02 ± 3.25**	38.11±2.53	45.74±2.93	41.15±5.13	42.64±5.04	42.07±5.14	41.22±5.62	41.99±5.16	41.76±5.67
D19	55.16±13.36	37.01±8.20	44.70±11.81	51.49±13.02	61.38±14.80	61.01±9.08	60.28±9.47	58.64±11.56	**62.60 ± 9.83**
D20	94.82±1.47	94.86±10.9	94.78±1.49	94.96±1.23	95.37±0.87	**96.92 ± 0.73**	96.37±0.79	96.48±1.14	**96.92 ± 0.63**
D21	91.09±6.35	89.19±6.49	91.06±6.30	90.91±6.38	91.26±5.96	91.28±5.97	91.15±6.09	91.26±5.89	**91.41 ± 5.72**

as D1 and D14, the improvement in the performance offered by the proposed approach is substantial while on some other datasets such as D17 the improvement is huge and reaches 28%. It should be noted that the performance improvements offered by the proposed approach are obtained despite the fact that the validation sets of some datasets may not be very large, and hence, may not serve as a very good representative of the entire the distribution of samples for tuning parameter p. It is expected that a more representative validation set would lead to even further improvements in the performance. A statistical ranking of different methods in the pure one-class setting using the Friedmans test is provided in Table 3. From the table, it can be observed that while the proposed ℓ_p-SVDD approach ranks the best among other approaches while the standard ℓ_1-SVDD approach ranks much worst which underlines the significance of the proposed ℓ_p-norm approach. Furthermore, although the ℓ_2-SVDD method provides some improvement with respect to the original ℓ_1-SVDD approach, its performance is still inferior compared to the proposed method. The second best performing method (in terms of average ranking) corresponds to a sample re-weighting SVDD approach presented in [16] which uses global and local statistics to linearly weight slacks in the objective function.

Table 3
Average ranking of different OCC methods in a pure one-class setting using the Friedman's test. (p-value=$1.12e-10$)

Algorithm	Rank
GP	6.47
KPCA	5.45
KNFST	6.71
P-SVDD	5.95
DW-SVDD	5.61
GL-SVDD	3.54
ℓ_1-SVDD	5.19
ℓ_2-SVDD	4.73
ℓ_p-SVDD (this work)	**1.30**

5.4. Training in the presence of negative data

In this second evaluation setting, in addition to positive objects, labelled negative samples are also used for training. Table 4 reports the performances of different methods in this setting. Note that as in the case of pure one-class learning, the optimal p value for the proposed approach is determined on the validation set. From

Table 4
Comparison of the performance of OCC approaches on different datasets in the presence of negative training objects in terms of %AUC (mean±std).

Dataset	KNFST$^-$	P-SVDD$^-$	DW-SVDD$^-$	GL-SVDD$^-$	ℓ_1-SVDD$^-$	ℓ_2-SVDD$^-$	ℓ_p-SVDD$^-$
D1	95.90±5.48	39.45±21.77	74.38±24.80	64.62±19.09	58.53±21.63	59.79±22.50	**100.00 ± 0.00**
D2	69.25±11.91	55.88±4.63	58.98±5.02	66.41±6.91	59.38±4.99	59.60±5.12	**71.57 ± 7.92**
D3	70.80±4.94	52.53±8.94	60.70±7.55	60.71±8.34	59.19±8.86	59.95±8.56	**75.82 ± 5.37**
D4	**96.56 ± 1.50**	93.27±1.71	94.92±1.45	93.33±1.79	94.46±1.23	93.88±1.23	**96.56 ± 1.43**
D5	93.02±2.11	88.63±5.07	91.17±3.49	91.17±3.49	91.17±3.49	91.19±3.45	**93.32 ± 1.94**
D6	90.29±12.84	88.73±3.46	92.61±4.15	92.56±4.14	92.62±4.14	92.41±4.23	**96.11 ± 4.79**
D7	93.70±3.77	50.84±8.34	73.84±6.50	58.59±8.39	58.13±9.14	58.66±8.90	**94.83 ± 3.00**
D8	90.01±1.31	64.97±3.74	66.93±3.24	67.20±3.15	66.95±3.30	65.95±3.35	**91.51 ± 1.40**
D9	64.14±9.19	48.56±10.65	83.08±10.33	63.30±11.58	59.92±9.96	60.72±15.15	**96.44 ± 2.25**
D10	89.37±7.15	82.92±7.86	86.66±8.53	85.88±8.47	86.66±8.53	86.58±8.58	**89.81 ± 5.00**
D11	96.52±3.62	84.95±11.15	94.90±1.78	96.50±1.86	94.02±3.96	89.88±5.76	**97.12 ± 1.72**
D12	52.52±7.84	58.84±6.19	63.06±5.22	69.44±7.54	62.10±9.37	65.95±6.16	**71.23 ± 7.42**
D13	91.11±7.62	68.97±12.57	47.65±13.31	69.24±13.25	69.54±10.59	66.53±8.41	**95.20 ± 2.93**
D14	63.94±3.85	55.37±4.12	64.88±4.01	61.15±4.01	59.14±4.89	64.90±3.59	**79.75 ± 0.99**
D15	94.20±2.85	68.65±4.92	65.80±8.32	91.91±2.71	93.32±2.35	88.39±3.73	**98.69 ± 0.54**
D16	61.74±12.47	64.12±5.56	63.19±6.27	64.67±5.55	64.33±5.08	62.68±7.29	**84.70 ± 2.69**
D17	88.11±6.10	51.27±14.69	50.22±13.93	51.75±14.62	50.22±13.93	50.36±13.67	**92.80 ± 3.73**
D18	56.83±3.88	49.04±1.92	51.52±2.54	51.74±3.55	52.02±2.40	47.06±4.54	**66.10 ± 9.22**
D19	78.46±4.54	62.56±9.53	0.8586±10.37	65.30±9.56	66.69±9.38	50.44±12.80	**92.65 ± 2.13**
D20	**99.25 ± 0.40**	95.83±1.40	97.72±0.99	97.55±0.79	97.03±1.15	97.24±0.96	98.07±0.86

Table 5
Average ranking of different OCC methods in the presence of negative training samples using the Friedmans test). (p-value=$8.97e-14$)

Algorithm	Rank
KNFST$^-$	2.80
P-SVDD$^-$	6.350
DW-SVDD$^-$	4.45
GL-SVDD$^-$	3.90
ℓ_1-SVDD$^-$	4.55
ℓ_2-SVDD$^-$	4.90
ℓ_p-SVDD$^-$ (this work)	**1.05**

among the GP, KPCA and KNFST approaches, only the KNFST approach is able to directly deploy negative samples for training. In order to emphasise that a method uses negative objects for training, a negative exponent ("$^-$") is used in the table. We also include the P-SVDD, DW-SVDD, and the GL-SVDD approaches trained using both negative and positive samples and denote them as P-SVDD$^-$, DW-SVDD$^-$, and GL-SVDD$^-$. From Table 4, it can be observed that on all datasets the proposed ℓ_p-SVDD approach obtains a better performance as compared with its ℓ_1-SVDD and ℓ_2-SVDD variants. In particular, while on some datasets the ℓ_1-SVDD and ℓ_2-SVDD approaches are unable to effectively utilise negative training samples, the proposed ℓ_p-SVDD method can better benefit from such samples to refine the description for improved performance. When compared with linear sample re-weighting methods of P-SVDD$^-$, DW-SVDD$^-$, and GL-SVDD$^-$, the proposed approach also performs better. An average ranking of different methods in this evaluation setting is provided in Table 5. From Table 5 it may be seen that the proposed ℓ_p-SVDD$^-$ approach utilising negative objects for training ranks the best among other competitors. Furthermore, neither the ℓ_1-SVDD$^-$ nor the ℓ_2-SVDD$^-$ methods which use negative training samples do not rank the second. The second best performing method in this setting corresponds to the KNFST method [13,37].

5.4.1. Noise Analysis

In order to analyse the behaviours of different OCC methods in the presence of noise in the data we evaluate different approaches subject to different levels of attribute noise [39] on six sample datasets. The results of this experiment are visualised in Fig. 2. From the figure, it may be observed that, as expected, the performance of all methods deteriorate when the percentage of noise increases. However, on some datasets (D1 and D3), the performance of the proposed approach stays relatively more stable compared to other approaches. Moreover, the proposed approach performs comparatively better compared to other methods over the range of noise level.

5.4.2. Other kernel Functions

In this section, in order to examine the utility of the proposed approach using other kernel functions, we use the recently proposed Hermite orthogonal polynomial kernel function [40] and compare the performances of different approaches on six example datasets. The results of this experiment are reported in Table 6. From the table, it can be seen that the proposed approach performs relatively better compared to other OCC methods. Moreover, while on some datasets the performance of the Hermit orthogonal polynomial kernel is inferior compared to that of the Gaussian kernel, on other datasets it provides an edge over the Gaussian kernel function. It may be concluded that depending on the characteristics of the data, the Hermit kernel may provide performance advantages over the Gaussian kernel.

5.5. Running times

In this section, we provide a comparison of different methods in terms of their running times for training. Since all methods included in the comparison are kernel-based approaches, in order to provide a more accurate comparison, we report the running times excluding the computation time of the kernel matrix (incurring $\mathcal{O}(n^2)$ complexity) which is common to all methods. The results of this experiment are reported in Table 7. From the table, the following observations may be made. The fastest OCC methods in the comparison are those of GP and KNFST closely followed by the KPCA approach. The SVDD-based methods are typically computationally less efficient, partly due to the complexity of the corresponding constraint optimisation problem. Among the SVDD-based methods, the proposed ℓ_p-SVDD approach appears comparatively less efficient which may be attributed to the relatively more complex optimisation problem to be solved for this method.

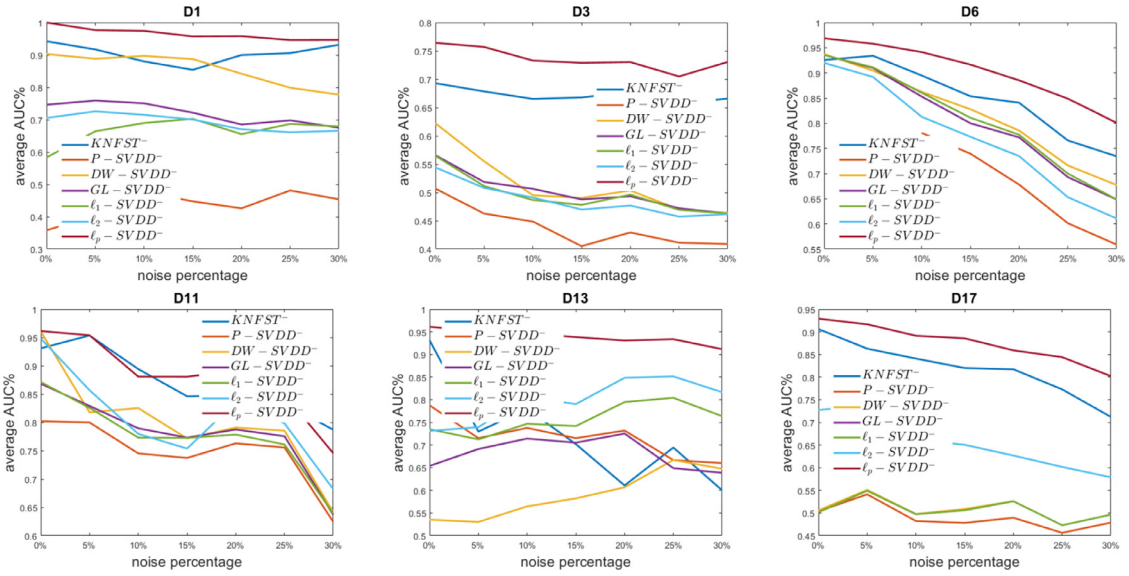

Fig. 2. Performances of different methods subject to noise in the dataset in terms of average AUC.

Table 6
Comparison of the performance of OCC approaches on different datasets using Hermit orthogonal polynomial kernel in terms of %AUC (mean±std).

Dataset	KNFST^{-}	P-SVDD^{-}	DW-SVDD^{-}	GL-SVDD^{-}	ℓ_1-SVDD^{-}	ℓ_2-SVDD^{-}	ℓ_p-SVDD^{-}
D1	76.62±16.34	54.05±18.28	92.63±9.12	84.92±13.22	66.14±17.83	76.87±17.14	**99.65 ± 1.15**
D3	61.28±10.39	59.89±10.23	71.25±8.24	68.19±8.72	66.21±8.91	64.55±8.97	**77.50 ± 6.44**
D6	57.34±9.16	61.31±11.49	58.76±8.24	86.02±15.70	87.97±14.16	41.51±12.34	**96.72 ± 3.36**
D11	92.90±7.45	85.45±7.14	92.76±3.79	93.73±4.37	92.37±4.85	95.15±2.75	**95.45 ± 2.21**
D13	89.74±6.23	67.73±19.47	53.32±5.44	70.06±11.95	78.45±9.45	81.73±6.03	**96.15 ± 0.76**
D17	88.39±7.66	49.88±17.13	51.81±16.97	50.08±17.08	50.56±17.29	82.29±7.91	**92.25 ± 3.75**

Table 7
Comparison of average running times for different methods in seconds.

	GP	KPCA	KNFST	P-SVDD	DW-SVDD	GL-SVDD	ℓ_1-SVDD	ℓ_2-SVDD	ℓ_p-SVDD
D1	0.0003	0.0004	0.0002	0.2016	0.1804	0.1792	0.1794	0.2469	0.3083
D2	0.0002	0.0022	0.0003	0.2105	0.1922	0.1896	0.1966	0.2884	0.3871
D3	0.0002	0.0012	0.0003	0.1996	0.1914	0.1794	0.1821	0.2473	0.3153
D4	0.0003	0.0017	0.0004	0.2316	0.2436	0.2231	0.2265	0.4587	0.7738
D5	0.0002	0.0014	0.0003	0.1978	0.1847	0.1902	0.1937	0.2874	0.4073
D6	0.0002	0.0013	0.0003	0.1995	0.1836	0.1835	0.1858	0.2455	0.3050
D7	0.0002	0.0012	0.0003	0.1984	0.1841	0.1846	0.1868	0.2622	0.3130
D8	0.0003	0.0023	0.0006	0.2571	0.2232	0.2192	0.2288	0.4661	0.6209
D9	0.0002	0.0013	0.0003	0.2052	0.1898	0.1893	0.1855	0.2663	0.3353
D10	0.0002	0.0011	0.0002	0.1975	0.1845	0.1839	0.1829	0.2487	0.2922
D11	0.0002	0.0034	0.0005	0.3005	0.2169	0.2152	0.2050	0.3947	0.5071
D12	0.0002	0.0016	0.0003	0.2268	0.2153	0.2137	0.2089	0.4067	0.5947
D13	0.0005	0.0024	0.0005	0.2543	0.2402	0.2455	0.2403	0.5062	0.8135
D14	0.0004	0.0032	0.0007	0.2711	0.2733	0.2596	0.2613	0.5942	0.9610
D15	0.0004	0.0030	0.0007	0.2756	0.2409	0.2465	0.2729	0.5709	0.8943
D16	0.0003	0.0029	0.0005	0.2733	0.2706	0.2554	0.2874	0.5871	0.9243
D17	0.0002	0.0016	0.0003	0.2096	0.1919	0.1918	0.1912	0.3433	0.4514
D18	0.0005	0.0027	0.0004	0.2852	0.2546	0.2778	0.2923	0.5853	0.9324
D19	0.0001	0.0008	0.0003	0.2316	0.2032	0.2042	0.2043	0.3613	0.534
D20	0.0011	0.0062	0.0012	0.4752	0.4683	0.5180	0.4554	1.1746	1.8325

6. Conclusion

We presented a generalisation of the SVDD approach from the conventional ℓ_1-norm to an ℓ_p-norm ($p \geq 1$) risk w.r.t. slacks. The proposed approach, in the primal space, enabled formulating non-linear loss functions over errors while in terms of the dual representation of the problem, it was illustrated that proposed approach leads to an optimisation task where ℓ_q-norm ($q \geq 1$) penalties over the dual variable (absent in the original SVDD formulation) are introduced into the objective function, allowing the algorithm to tune into the inherent sparsity of the data. We showed that the proposed approach leads to a convex optimisation task, and thus, may be optimised effectively. A theoretical analysis of the proposed method revealed the dependence of the generalisation error bound on parameter p. In particular, it was shown that the training error loss and the empirical Rademacher complexity of the algorithm need not be minimised for $p = 1$ which is the case considered in the standard SVDD formulation. The results of an experimental evaluation on several standard OCC datasets showed that the proposed approach leads to improvements upon the

existing ℓ_1- and ℓ_2-norm penalty functions in addition to some other linear sample re-weighting SVDD variants and also performs very favourably compared with other existing OCC methods.

The proposed method, similar to the standard SVDD approach, does not make any specific assumption regarding the distribution of the data. Nevertheless, if information regarding the distribution of the data is available, one may consider possible adaptations of the proposed approach where such information is reflected onto the objective function. One limitation of the proposed method is that all slacks are similarly penalised via a common ℓ_p-norm. In this context, one possible extension might be to consider sample-specific penalty functions. Other limitation of the proposed approach relates to its inability to directly learn features from the data. In this respect, a possible direction for future investigation is to couple the training stage of the proposed approach with that of a deep network for a joint learning of data representation and a one-class description for enhanced performance. While in this work we considered the case of $q \geq 1$, other possibilities may include the non-convex case of $q < 1$ to study and analyse its impact on the performance. Finally, a further possible direction for future investigation may include analysing and designing other potentially faster solutions to the convex optimisation problem of the proposed approach.

Data Availability

ℓ_p-Norm Support Vector Data Description (Mendeley Data)

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by The Scientific and Technological Research Council of Turkey (T8BTAK) under the grant no 121E465.

References

[1] B.B. Hazarika, D. Gupta, P. Borah, An intuitionistic fuzzy kernel ridge regression classifier for binary classification, Applied Soft Computing 112 (2021) 107816.
[2] S. Fatemifar, S.R. Arashloo, M. Awais, J. Kittler, Client-specific anomaly detection for face presentation attack detection, Pattern Recognition 112 (2021) 107696.
[3] A. Rabaoui, M. Davy, S. Rossignol, N. Ellouze, Using one-class svms and wavelets for audio surveillance, IEEE Transactions on Information Forensics and Security 3 (4) (2008) 763–775.
[4] X. Zhang, S. Yang, J. Zhang, W. Zhang, Video anomaly detection and localization using motion-field shape description and homogeneity testing, Pattern Recognition 105 (2020) 107394.
[5] P. Nader, P. Honeine, P. Beauseroy, lp-norms in one-class classification for intrusion detection in scada systems, IEEE Transactions on Industrial Informatics 10 (4) (2014) 2308–2317.
[6] R. Chaker, Z.A. Aghbari, I.N. Junejo, Social network model for crowd anomaly detection and localization, Pattern Recognition 61 (2017) 266–281.
[7] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, M. Kloft, Deep one-class classification, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, volume 80, PMLR, 2018, pp. 4393–4402.
[8] D.M. Tax, R.P. Duin, Support vector data description, Machine Learning 54 (1) (2004) 45–66.
[9] V. Vapnik, R. Izmailov, Reinforced svm method and memorization mechanisms, Pattern Recognition 119 (2021) 108018.
[10] J. Kittler, W. Christmas, T. de Campos, D. Windridge, F. Yan, J. Illingworth, M. Osman, Domain anomaly detection in machine perception: A system architecture and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (5) (2014) 845–859.
[11] E.-S. Platzer, J. Denzler, H. SuBe, J. Nagele, K.-H. Wehking, Challenging anomaly detection in wire ropes using linear prediction combined with one-class classification, in: O. Deussen, D.A. Keim, D. Saupe (Eds.), VMV, Aka GmbH, 2008, pp. 343–352.
[12] Y. Xiao, H. Wang, W. Xu, J. Zhou, L1 norm based kpca for novelty detection, Pattern Recognition 46 (1) (2013) 389–396.
[13] S.R. Arashloo, J. Kittler, Robust one-class kernel spectral regression, IEEE Transactions on Neural Networks and Learning Systems 32 (3) (2021) 999–1013.
[14] M. Kemmler, E. Rodner, E.-S. Wacker, J. Denzler, One-class classification with gaussian processes, Pattern Recognition 46 (12) (2013) 3507–3518.
[15] J. Hamidzadeh, M. Moradi, Incremental one-class classifier based on convex-concave hull, Pattern Analysis and Applications 23 (4) (2020) 1523–1549.
[16] W. Hu, T. Hu, Y. Wei, J. Lou, S. Wang, Global plus local jointly regularized support vector data description for novelty detection, IEEE Transactions on Neural Networks and Learning Systems (2021) 1–13.
[17] C.-D. Wang, J. Lai, Position regularized support vector domain description, Pattern Recognition 46 (3) (2013) 875–884.
[18] K. Lee, D.-W. Kim, D. Lee, K.H. Lee, Improving support vector data description using local density degree, Pattern Recognition 38 (10) (2005) 1768–1771.
[19] G. Huang, H. Chen, Z. Zhou, F. Yin, K. Guo, Two-class support vector data description, Pattern Recognition 44 (2) (2011) 320–329.
[20] M. Turkoz, S. Kim, Y. Son, M.K. Jeong, E.A. Elsayed, Generalized support vector data description for anomaly detection, Pattern Recognition 100 (2020) 107119.
[21] M. Turkoz, S. Kim, Multi-class bayesian support vector data description with anomalies, Annals of Operations Research Nov (2021).
[22] R. Sadeghi, J. Hamidzadeh, Automatic support vector data description, Soft Computing 22 (1) (2018) 147–158.
[23] J. Hamidzadeh, R. Sadeghi, N. Namaei, Weighted support vector data description based on chaotic bat algorithm, Applied Soft Computing 60 (2017) 540–551.
[24] I.W. Tsang, J.T. Kwok, P.-M. Cheung, Core vector machines: Fast svm training on very large data sets, Journal of Machine Learning Research 6 (13) (2005) 363–392.
[25] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.
[26] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, S. Marcel, The replay–mobile face presentation-attack database, in: Proceedings of the International Conference on Biometrics Special Interests Group (BioSIG), 2016.
[27] P. Borah, D. Gupta, Functional iterative approaches for solving support vector classification problems based on generalized huber loss, Neural Computing and Applications 32 (13) (2020) 9245–9265.
[28] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, 2014, (http://cvxr.com/cvx).
[29] D. Dua, C. Graff, UCI machine learning repository, 2017, http://archive.ics.uci.edu/ml.
[30] D.M.J. Tax, R.P.W. Duin, Characterizing one-class datasets, 2006.
[31] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework, J. Multiple Valued Log. Soft Comput. 17 (2-3) (2011) 255–287.
[32] S.-B. Cho, Neural-network classifiers for recognizing totally unconstrained handwritten numerals, Trans. Neur. Netw. 8 (1) (1997) 43–53.
[33] D. Harrison, D.L. Rubinfeld, Hedonic housing prices and the demand for clean air, Journal of Environmental Economics and Management 5 (1) (1978) 81–102.
[34] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017, cs.LG/1708.07747
[35] S.R. Arashloo, Matrix-regularized one-class multiple kernel learning for unseen face presentation attack detection, IEEE Transactions on Information Forensics and Security 16 (2021) 4635–4647.
[36] M. Cha, J.S. Kim, J.-G. Baek, Density weighted support vector data description, Expert Systems with Applications 41 (7) (2014) 3343–3350.
[37] Y. Lin, G. Gu, H. Liu, J. Shen, Kernel null foley-sammon transform, in: 2008 International Conference on Computer Science and Software Engineering, volume 1, 2008, pp. 981–984.
[38] H. Hoffmann, Kernel pca for novelty detection, Pattern Recognition 40 (3) (2007) 863–874.
[39] X. Zhu, X. Wu, Class noise vs. attribute noise: A quantitative study, Artif. Intell. Rev. 22 (3) (2004) 177–210.
[40] V. Hooshmand Moghaddam, J. Hamidzadeh, New hermite orthogonal polynomial kernel and combined kernels in support vector machine classifier, Pattern Recognition 60 (2016) 921–935.

Shervin Rahimzadeh Arashloo received the Ph.D. degree from the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K., in 2010. He is currently an Assistant Professor with the Department of Computer Engineering, Bilkent University, Ankara, Turkey, and a Visiting Research Fellow with CVSSP, University of Surrey. His research interests include pattern recognition, machine learning, and signal processing.