One Class Classification for Anomaly Detection: Support Vector Data Description Revisited

Eric J. Pauwels and Onkar Ambekar

Centrum Wiskunde & Informatica CWI,
Science Park 123, 1098 XG Amsterdam, The Netherlands
eric.pauwels@cwi.nl
http://www.cwi.nl

Abstract. The *Support Vector Data Description* (SVDD) has been introduced to address the problem of anomaly (or outlier) detection. It essentially fits the smallest possible sphere around the given data points, allowing some points to be excluded as outliers. Whether or not a point is excluded, is governed by a slack variable. Mathematically, the values for the slack variables are obtained by minimizing a cost function that balances the size of the sphere against the penalty associated with outliers. In this paper we argue that the SVDD slack variables lack a clear geometric meaning, and we therefore re-analyze the cost function to get a better insight into the characteristics of the solution. We also introduce and analyze two new definitions of slack variables and show that one of the proposed methods behaves more robustly with respect to outliers, thus providing tighter bounds compared to SVDD.

Key words: One class classification, outlier detection, anomaly detection, support vector data description, minimal sphere fitting

1 Introduction

In a conventional classification problem, the aim is to find a classifier that optimally separates two (or more) classes. The input to the problem is a labelled training set comprising a roughly comparable number of exemplars from each class. Howerever, there are types of problems in which this assumption of (approximate) equi-distribution of exemplars no longer holds. The prototypical example that springs to mind is *anomaly detection*. By its very definition, an anomaly is a rare event and training data will more often than not contain very few or even no anomalous exemplars. Furthermore, anomalies can often only be exposed when looked at in context, i.e. when compared to the majority of regular points. Anomaly detection therefore provides an example of so-called *one-class classification*, the gist of which amounts to the following: Given data points that all originated from a single class but are possibly contaminated with a small number of outliers, find the class boundary.

In this paper we will focus on an optimization approach championed by Tax [8] and Schölkopf *et.al.* [4]. The starting point is a classical problem in quadratic

programming: given a set of n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a p-dimensional space, find the most tightly fitting (hyper)sphere that encompasses all. Denoting the centre of this sphere by $\mathbf{a}$ and its radius by R, this problem boils down to a constrained minimization problem:

$$\min_{\mathbf{a},R} R^2 \quad \text{subject to} \quad \|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2, \quad \forall i = 1, \ldots, n. \tag{1}$$

However, if the possibility exist that the dataset has been contaminated with a small number of anomalies, it might prove beneficial to exclude suspicious points from the sphere and label them as *outliers*. This then allows one to shrink the sphere and obtain a better optimum for the criterion in eq.(1). Obviously, in order to keep the problem non-trivial, one needs to introduce some sort of penalty for the excluded points. In [8] and [4] the authors take their cue from standard support vector machines (SVM) and propose the use of non-negative slack variables meant to relax the inclusion criterion in eq.(1). More precisely, for each point they introduce a variable $\xi_i \geq 0$ such that

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i. \tag{2}$$

This relaxation of the constraints is then offset by adding a penalty term to the cost function:

$$\zeta(R, \mathbf{a}, \xi) := R^2 + C \sum_{i=1}^{n} \xi_i.$$

The constant C is a (pre-defined) *unit cost* that governs the trade-off between the size of the sphere and the number of outliers. After these modifications the authors in [8, 4] arrive at the following constrained optimization problem: given n data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and a pre-defined unit cost C, find

$$\min_{\mathbf{a},R,\xi} \{R^2 + C \sum_{i=1}^{n} \xi_i\} \quad \text{s.t.} \quad \forall i = 1, \ldots, n : \ \|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0. \tag{3}$$

The resulting data summarization segregates "regular" points on the inside from "outliers" on the outside of the sphere and is called *support vector data description* (SVDD).

Aim of this paper The starting point for this paper is the observation that the slack variables in eq.(3) lack a straightforward geometrical interpretation. Indeed, denoting $d_i = \|\mathbf{x}_i - \mathbf{a}\|$, it transpires that the slack variables can be represented explicitly as:

$$\xi_i = (d_i^2 - R^2)_+ = \begin{cases} d_i^2 - R^2 & \text{if } \ d_i > R, \\ 0 & \text{if } \ d_i \leq R. \end{cases} \tag{4}$$

However, except in the case where the dimension of the ambient space (p) equals two or three, these slack variables don't have an obvious geometric interpretation.

It would therefore be more natural to set the slack variable equal to $\varphi_i = (d_i - R)_+$ upon which the relaxed constraints can be expressed as:

$$\forall i : \quad \|\mathbf{x}_i - \mathbf{a}\| \leq R + \varphi_i, \quad \varphi_i \geq 0. \tag{5}$$

The corresponding penalized function would then take the form:

$$\zeta_2(\mathbf{a}, R) := R^2 + C \sum_i \varphi_i^2. \tag{6}$$

(Notice that we can drop φ from the list of arguments as it can be computed as soon as $\mathbf{a}$ and R are specified). For lack of space we will not be able to study this alternative in detail. Suffice it to say that the solution includes non-acceptable, trivial configurations. However, there is no obvious reason why the variables in the cost function should appear as squares. This suggests that we also should look at a second — completely linear — alternative:

$$\zeta_1(\mathbf{a}, R) := R + C \sum_i \varphi_i. \tag{7}$$

The *goal of this paper* is therefore twofold. Firstly, we want to re-analyze the original optimization problem (3) as introduced in [8] and [4]. However, in contradistinction to these authors, we will refrain from casting it in its dual form, but focus on the primal problem instead. This will furnish us with additional insights into the geometry and behaviour of the solutions. Secondly, we will then extend this analysis to the alternative ζ_1 (see eq. 7) mentioned above and conclude that, in some respects, it is preferable to the original. In fact the difference between these two solutions is not unlike the difference in behaviour between the *mean* and *median* (for a quick preview of this result, we suggest to have a peek at Fig. 2).

Related work Although lack of space precludes a comprehensive revision of all related work, it is fair to say that after the seminal papers [5, 8] most activity focussed on applications, in particular clustering, see e.g. [1]. In particular, a lot of research has gone into the appropriate choice of the Gaussian kernel size when using the kernelized version of this technique [3, 2], as well as efficient methods for cluster labeling. In [6] a different direction of generalization is pursued: rather than mapping the data into a high-dimensional feature space, the spherical constraints are relaxed into ellipsoidal ones in the original data space, thereby side-stepping the vexing question of kernel-choice.

2 Support Vector Data Description Revisited

In this section we will re-analyze the cost function (3) which lies at the heart of the SVDD classifier. However, rather than recasting the problem in its dual form (as is done in [8] and [4]), we will focus directly on the primal problem. This allows us to gain additional insight in the qualitative behaviour of the solutions (cf. section 2.2) as well as sharpen the bounds on the unit cost C (see item 3 of Prop. 1).

2.1 Outlier Detection as an Optimization Problem

Recall from (3) that the anomaly (a.k.a. outlier) detection problem has been recast into the following constrained optimization problem. As input we accept n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in p-dimensional space, and some fixed pre-defined unit cost C. In addition, we introduce a vector $\xi = (\xi_1, \ldots, \xi_n)$ of n slack variables in terms of which we can define the cost function

$$\zeta(\mathbf{a}, R, \xi) := R^2 + C \sum_i \xi_i. \tag{8}$$

The SVDD outlier detection (as introduced in [8] and [4]) now amounts to finding the solution to the following constrained minimization problem:

$$\min_{\mathbf{a},R,\xi} \ \zeta(\mathbf{a}, R, \xi) \quad \text{s.t.} \quad \forall i = 1, \ldots, n: \ \|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0. \tag{9}$$

If we denote the distance of each point $\mathbf{x}_i$ to the centre $\mathbf{a}$ as $d_i = \|\mathbf{x}_i - \mathbf{a}\|$ then it's straightforward to see that the slack variables can be explicified as $\xi_i := (d_i^2 - R^2)_+$, where the ramp function x_+ is defined by:

$$x_+ := \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \tag{10}$$

This allows us to rewrite the cost function in a more concise form:

$$\zeta(\mathbf{a}, R) = R^2 + C \sum_i (d_i^2 - R^2)_+. \tag{11}$$

Notice that the cost function is now a function of $\mathbf{a}$ and R only, with all other constraints absorbed in the ramp function x_+. From this representation it immediately transpires that ζ is continuous in its arguments, albeit not everywhere differentiable.

2.2 Properties of the solution

Proposition 1 *The solution of the (unconstrained) optimization problem*

$$(\mathbf{a}^*, R^*) := \arg\min_{\mathbf{a},R} \ \zeta(\mathbf{a}, R) \qquad \textit{where} \quad \zeta(\mathbf{a}, R) = R^2 + C \sum_{i=1}^{n} (d_i^2 - R^2)_+ \tag{12}$$

has the following qualitative properties:

1. **Behaviour of the marginal functions:**
   (a) *Keeping R fixed, ζ is a convex function of the centre $\mathbf{a}$.*
   (b) *Keeping $\mathbf{a}$ fixed, ζ is piecewise quadratic in R.*

2. **Location of the optimal centre $\mathbf{a}^*$:** *The centre of the optimal sphere can be specified as a weighted mean of the data points*

$$\mathbf{a}^* = \frac{\sum_i h_i \mathbf{x}_i}{\sum_i h_i} \tag{13}$$

where

$$h_i = \begin{cases} 1 & \text{if } d_i > R^* \\ 0 \leq \theta_i \leq 1 & \text{if } d_i = R^* \\ 0 & \text{if } d_i < R^*. \end{cases} \tag{14}$$

such that

$$\sum_i h_i = 1/C. \tag{15}$$

3. **Dependency on penalty cost C:**
The value of the unit cost C determines the qualitative behaviour of the solution. More precisely:
   - *If $C < 1/n$ then the optimal radius R^* will be zero, i.e. all points will reside outside of the sphere.*
   - *If $C \geq 1/2$ all points will be enclosed, and the sphere will be the minimum volume enclosing sphere.*
   - *For values $1/n \leq C \leq 1/2$, the qualitative shape of the solution changes whenever $C = 1/k$ for $k = 2, 3, \ldots n$.*

PROOF

1. Behaviour of the marginal functions

- **1.a: Keeping R fixed, ζ is a convex function of the centre a.** Assuming that in eq. (12) the radius R and cost C are fixed, the dependency of the cost functional is completely captured by second term:

$$\sum_i (d_i^2 - R^2)_+ \equiv \sum_i \max\{d_i^2 - R^2, 0\}.$$

  Convexity of ζ as a function of $\mathbf{a}$ is now immediate as each $d_i^2 \equiv d_i^2(\mathbf{a}) = \|\mathbf{x}_i - \mathbf{a}\|^2$ is convex and both the operations of maximization and summing are convexity-preserving.
- **1.b: Keeping a fixed, ζ is piecewise quadratic in R.** Introducing the auxiliary binary variables:

$$b_i(R) = \begin{cases} 1 \text{ if } d_i > R, \\ 0 \text{ if } d_i \leq R, \end{cases} \qquad \text{for } \mathbf{a} \text{ fixed}, \tag{16}$$

  allows us to rewrite $(d_i^2 - R^2)_+ \equiv b_i(R)(d_i^2 - R^2)$, from which

$$\zeta(R) = \left(1 - C \sum_{i=1}^{n} b_i(R)\right) R^2 + C \sum_{i=1}^{n} b_i(R) d_i^2, \tag{17}$$

or again,

$$\zeta(R) = \beta(R)\, R^2 + C\gamma(R). \tag{18}$$

where

$$\beta(R) := 1 - C\sum_i b_i(R) \quad \text{and} \quad \gamma(R) := \sum_i b_i(R) d_i^2. \tag{19}$$

As it is clear that the coefficients β and γ are piecewise constant, producing a jump whenever R grows beyond one of the distances d_i, it follows that $\zeta(R)$ is (continuous) piecewise quadratic. More precisely, if we assume that the points $\mathbf{x}_i$ have been (re-)labeled such that $d_1 \equiv \|\mathbf{x}_1 - \mathbf{a}\| \leq d_2 \equiv \|\mathbf{x}_2 - \mathbf{a}\| \leq \ldots \leq d_n \equiv \|\mathbf{x}_n - \mathbf{a}\|$, then for $0 \leq R < d_1$, all $b_i(R) = 1$ and hence $\beta(R) = 1 - nC$. On the interval $d_1 \leq R < d_2$ we find that $b_1 = 0$ while $b_2 = b_3 = \ldots b_n = 1$ implying that $\beta(R) = 1 - (n-1)C$, and so on. So we conclude that β is a piecewise constant function, making an upward jump of size C whenever R passes a d_i. This is illustrated in Fig. 1 where the bottom figure plots the piecewise constant coefficient β for two different values of C, while the corresponding ζ functions are plotted in the top graph. Clearly, every β-plateau gives rise to a different quadratic part of ζ. More importantly, as long as $\beta(R) < 0$ the resulting quadratic part in ζ is strictly decreasing. Hence we conclude that the minimum of ζ occurs at the point $R^* = \arg\min \zeta(R)$ where β jumps above zero. Indeed, at that point, the corresponding quadratic becomes strictly increasing, forcing the minimum to be located at the jump between the two segments.
From the above we can also conclude that the optimal radius $R^* = \arg\min \zeta(R)$ is unique except when $C = 1/k$ for some integer $1 \leq k \leq n$. In those instances there will be an R-segment on which $\sum b_i = k$, forcing the corresponding β coefficient to vanish. This then gives rise to a flat, horizontal plateau of minimal values for the ζ function. In such cases we will pick (arbitrarily) the maximal possible value for R, i.e.: $R^* := \sup\{R : \zeta(R) \text{ is minimal}\}$. Finally, we want to draw attention to the fact that the optimal sphere always passes through at least one data point, as the optimal radius R^* coincides with at least one d_i.

2. Location of the optimal centre Earlier we pointed out that the $\zeta(\mathbf{a}, R)$ is continuous but not everywhere differentiable. This means that we cannot simply insist on vanishing gradients to determine the optimum, as the gradient might not exist. However, we can take advantage of a more general concept that is similar in spirit: *subgradients*. Recall that for a differentiable convex function f, the graph of the function lies above every tangent. Mathematically this can be reformulated by saying that at any x:

$$f(\mathbf{y}) \geq \nabla f(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y}. \tag{20}$$

If f is not necessarily differentiable at x then we will say that any vector g_x is a *subgradient* at x if:

$$f(\mathbf{y}) \geq \mathbf{g_x} \cdot (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y}. \tag{21}$$

The collection of all subgradients at a point $\mathbf{x}$ is called the *subdifferential* of f at $\mathbf{x}$ and denoted by $\partial f(\mathbf{x})$. Notice that the subdifferential is a set-valued function! It is now easy to prove that the classical condition for x_* to be the minimum of a convex function f (i.e. $\nabla f(\mathbf{x}_*) = \mathbf{0}$) can be generalized to non-differentiable functions as:

$$\mathbf{0} \in \partial f(\mathbf{x}_*). \tag{22}$$

To apply the above the problem at hand, we first note that the subdifferential of the ramp function x_+ is given by:

$$\partial x_+ = \begin{cases} 0 & \text{if } x < 0 \\ [0,\,1] & \text{if } x = 0 \quad \text{(i.e. set-valued)} \\ 1 & \text{if } x > 0 \end{cases} \tag{23}$$

as at $x = 0$ any straight line with slope between 0 and 1 will be located under the graph of the ramp function. To streamline notation, we introduce (a version of) the Heaviside stepfunction

$$H(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 \leq h \leq 1 & \text{if } x = 0 \quad \text{(i.e. set-valued)} \\ 0 & \text{if } x < 0 \end{cases} \tag{24}$$

To forestall confusion we point out that, unlike when used as a distribution, this definition of the Heaviside function insists its value at the origin is between zero and one. Using this convention, we have the convenient shorthand notation:

$$\partial x_+ = H(x).$$

Computing the subgradients (for convenience we will drop the notational distinction between standard- and sub-gradients) we obtain:

$$\frac{\partial \zeta}{\partial R} = 2R - 2RC \sum_i H(d_i^2 - R^2) \tag{25}$$

$$\nabla_{\mathbf{a}} \zeta = -2C \sum_i H(d_i^2 - R^2)\,(\mathbf{x}_i - \mathbf{a}) \tag{26}$$

where we used the well-known fact:

$$\nabla_{\mathbf{a}}(d_i^2) = \nabla_{\mathbf{a}} ||\mathbf{x}_i - \mathbf{a}||^2 = \nabla_{\mathbf{a}}(\mathbf{x}_i \cdot \mathbf{x}_i - 2\mathbf{x}_i \cdot \mathbf{a} + \mathbf{a} \cdot \mathbf{a}) = -2(\mathbf{x}_i - \mathbf{a}). \tag{27}$$

Insisting that zero is indeed a subgradient means that we need to pick values $h_i := H(d_i^2 - R^2)$ such that:

$$0 \in \partial\zeta/\partial R \Rightarrow \sum_{i=1}^{n} h_i = 1/C \tag{28}$$

$$0 \in \nabla_{\mathbf{a}} \zeta \Rightarrow \sum_{i=1}^{n} h_i(\mathbf{x}_i - \mathbf{a}) = 0 \tag{29}$$

The above characterization allows us to draw a number of straightforward conclusions (for notational convenience we will drop the asterisk to indicate optimality, and simply write $\mathbf{a}^* = \mathbf{a}$ and $R^* = R$):

1. Combining eqs.(28) and (29) it immediately transpires that

$$\mathbf{a} = C \sum h_i \, \mathbf{x}_i, \tag{30}$$

or again, and more suggestively,

$$\mathbf{a} = \frac{\sum h_i \, \mathbf{x}_i}{\sum h_i}. \tag{31}$$

Furthermore, the sums in the RHS can be split into three parts depending on whether a point lies *inside* ($d_i < R$), *on* ($d_i = R$) or *outside* ($d_i > R$) the sphere, e.g.:

$$\begin{aligned} \sum_i h_i &= \sum_{i:d_i<R} H(d_i^2 - R^2) + \sum_{i:d_i=R} H(d_i^2 - R^2) + \sum_{i:d_i>R} H(d_i^2 - R^2) \\ &= 0 + \sum_{i:d_i=R} \theta_i + \sum_{i:d_i>R} 1 \end{aligned} \tag{32}$$

where $0 \leq \theta_i \equiv H(d_i - R = 0) \leq 1$. Hence:

$$\mathbf{a} = \frac{\sum\limits_{i:d_i=R} \theta_i \mathbf{x}_i + \sum\limits_{i:d_i>R} \mathbf{x}_i}{\sum\limits_{i:d_i=R} \theta_i + \sum\limits_{i:d_i>R} 1} \tag{33}$$

This representation highlights the fact that the centre $\mathbf{a}$ is a weighted mean of the points *on* or *outside* the sphere (the so-called *support vectors* (SV), [8]), while the points inside the sphere exert no influence on its position. Notice that the points outside of the sphere are assigned maximal weight.

2. If we denote the number of points *inside, on* and *outside* the sphere by n_{in}, n_{on} and n_{out} respectively, then by definition $\#SV = n_{on} + n_{out}$. Invoking eq. (28) and combining this with the fact that $0 \leq \theta_i \leq 1$ it follows that

$$1/C = \sum_i h_i = \sum_{i:d_i=R} \theta_i + \sum_{i:d_i>R} 1 \tag{34}$$

Hence, since $0 \leq \theta_i \leq 1$ it can be concluded that

$$n_{out} = \sum_{d_i>R} 1 \quad \leq \quad 1/C \quad \leq \quad n_{on} + n_{out} = \#SV \tag{35}$$

Put differently:
(a) $1/C$ is a lower bound on the number of support vectors ($\#SV$).
(b) $1/C$ is an upper bound on the number of outliers (n_{out}).
The same result was obtained by Schölkopf [4], who introduced the parameter $\nu = 1/nC$ as a bound on the *fraction* of support vectors ($\#SV/n$) and outliers (n_{out}/n).

3. Dependency on unit-cost C In this section we try to gain further insight into how the cost function determines the behaviour of the optimum. Let us assume that we have already minimized the cost function (11) and identified the optimal centre $\mathbf{a}^*$ and corresponding radius R^*. For convenience's sake, we again assume that we have relabeled the data points in such a way that the distances $d_i = \|\mathbf{x}_i - \mathbf{a}^*\|$ are ordered in ascending order: $0 \leq d_1 \leq d_2 \leq \ldots \leq d_n$. We now investigate how the total cost ζ depends on the unit cost C in the neighbourhood of this optimum.

Figure 1 nicely illustrate the influence of the unit cost C on the qualitative behaviour of the optimal radius R^*. Indeed, increasing C slightly has the following effects on the β-function:

- The values of the coefficients h_i will change (cf. eq. 28) which in turn will result in a shift of the optimal centre (through eq. 31). As a consequence the distances d_i to the data points $\mathbf{x}_i$ will slightly change, resulting in slight shifts of the step locations of the β-function. Since the position of the optimal radius R^* coincides with one of these step locations (viz. the jump from a negative to a positive β-segment), increasing C slightly will typically induces small changes in R^*. However, from time to time, one will witness a jump-like change in R^* as explained below.
- Since the size of a β-step equals the unit cost, slightly increasing C will push the each β-segment slightly downwards as the maximum of β remains fixed at one (i.e. $\lim_{R\to\infty} \beta(R) = 1$). As a consequence, β-segments that are originally positive, will at some point dip below the X-axis. As this happens, the corresponding quadratic segment will make the transition from *convex and increasing* to *concave and decreasing* forcing the minimum R^* to make a jump.

This now allows us to draw a number of straightforward conclusions about the constraints on the unit cost C.

- The first segment of the β function occurs for $0 \leq R < d_1$. On this segment $b_i = 1$ for all $i = 1, \ldots, n$ and hence $\beta(R) = 1 - C\sum_i b_i = 1 - nC$. If $C < 1/n$, then $\beta > 0$ on this first segment and hence on all the subsequent ones. In that case, $\zeta(R)$ is strictly increasing and has a single trivial minimum at $R^* = 0$. Put differently, in order to have a non-trivial optimization problem, we need to insist on $C \geq 1/n$ (cf. item 3 in proposition 1). icicic
- If, on the other hand, we want to make sure that there are no outliers, then the optimum R^* has to coincide with the last jump, i.e. $R^* = d_n$. This implies that the quadratic segment on the interval $[d_{n-1},\ d_n]$ has to be decreasing (or flat), and consequently $\beta(R) = 1 - C\sum_i b_i \leq 0$. Since on this last segment we have that all b_i vanish except for b_n, it follows that $\beta(R) = 1 - C \leq 0$ (and vice versa). We therefore conclude that for values $C \geq 1$ there will be no outliers.
  This result was also obtained in [8, 5] but we can now further tighten the above bound by observing that when the optimal sphere encloses all points,

it has to pass through *at least two* points (irrespective of the ambient dimension). This implies that $d_{n-1} = d_n$ and the first non-trivial interval preceding d_n is in fact $[d_{n-2}, d_{n-1}]$. Rerunning the above analysis, we can conclude that $C \geq 1/2$ implies that all data points are enclosed.

- Using the same logic, if we insist that at most k out of n are outside the circle, we need to make sure that the quadratic on $[d_{n-k}, d_{n-k+1}]$ is convex and increasing. On that interval we know that $\sum_i b_i = k$. Hence we conclude that on this interval $\beta(R) = 1 - kC > 0$ or again: $C < 1/k$. Hence, $\nu = 1/nC > k/n$ is an upper bound on the fraction of points outside the descriptor (cf. [4]).
- In fact, by incorporating some straightforward geometric constraints into the set-up we can further narrow down the different possible configuration. As a simple example, consider the case of a *generic* 2-dimensional data set. The sphere then reduces to a circle and we can conclude that – since we assume the data set to be generic – the number of points on the optimal circle (i.e. n_{on}) either equals 1 (as the optimal circle passes through at least one point), 2 or 3. Indeed, there is a vanishing probability that a generic data set will have 4 (or more) co-circular points (points on the same circle). In this case we can rewrite the Schölkopf inequality (35) as:

$$n_{out} \leq 1/C \leq n_{out} + 3$$

For values $C < 1/3$ it then follows that

$$3 < 1/C \leq n_{out} + 3 \quad \Rightarrow \quad n_{out} > 0.$$

So we arrive at the somewhat surprising conclusion that if the unit cost is less than 1/3, we are *guaranteed to have at least one outlier,* no matter what the data set looks like (as long as it is generic). This is somewhat counter-intuitive as far as the usual concept of an outlier is concerned!

This concludes the proof. **QED**

3 Linear Slacks and Linear Loss

3.1 Basic analysis

As announced earlier, this section busies itself with minimizing the linear function

$$\zeta_1(\mathbf{a}, R) := R + C \sum_i \varphi_i \quad \text{subject to} \quad \forall i : \, d_i \equiv \|\mathbf{x}_i - \mathbf{a}\| \leq R + \varphi_i, \quad \varphi_i \geq 0. \tag{36}$$

Again, we absorb the constraints into the function by introducing the ramp function:

$$\zeta_1(\mathbf{a}, R) = R + C \sum_i (d_i - R)_+ \tag{37}$$

Taking subgradients with respect to $\mathbf{a}$ and R yields:

$$\frac{\partial \zeta_1}{\partial R} = 1 - C \sum H(d_i - R)$$
$$\nabla_{\mathbf{a}} \zeta_1 = -C \sum H(d_i - R) \frac{(\mathbf{x}_i - \mathbf{a})}{\|\mathbf{x}_i - \mathbf{a}\|}$$

since it is straightforward to check that:

$$\nabla_{\mathbf{a}}(d_i) = \nabla_{\mathbf{a}} \sqrt{(\|\mathbf{x}_i - \mathbf{a}\|^2)} = -\frac{(\mathbf{x}_i - \mathbf{a})}{\|\mathbf{x}_i - \mathbf{a}\|}.$$

Equating the gradient to zero and re-introducing the notation $h_i = H(d_i - R)$ we find that the optimum is characterized by:

$$\frac{\partial \zeta_1}{\partial R} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} h_i = 1/C \tag{38}$$

$$\nabla_a \zeta_1 = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} h_i \frac{(\mathbf{x}_i - \mathbf{a})}{\|\mathbf{x}_i - \mathbf{a}\|} = 0 \tag{39}$$

Notice how eq. (38) is identical to eq. (28) whereas eq. (39) is similar but subtly different from eq.(29). In more detail:

1. Once again we can make the distinction between the n_{in} points that reside inside the sphere, the n_{on} points that lie on the sphere and the n_{out} points that are outside the sphere. The latter two categories constitute the *support vectors*: $\#SV = n_{on} + n_{out}$. Hence,

$$1/C = \sum_i h_i = \sum_{d_i<R} h_i + \sum_{d_i=R} h_i + \sum_{d_i>R} h_i$$
$$= \sum_{d_i=R} \theta_i + n_{out}.$$

   So also in this case we get (cf. eq. (35)):

$$n_{out} \leq \frac{1}{C} \leq \#SV. \tag{40}$$

2. Comparing eqs. (39) and (29) we conclude that we can expect the solution corresponding to linear loss function (36) to be *more robust with respect to outliers.* Indeed, in Section 2 we've already argued that eq. (29) implies that the sphere's centre is the (weighted) mean of the support vectors. Noticing that in eq. (39) the vectors have been substituted by the corresponding *unit vectors* reveals that in the case of a linear loss function the centre can be thought of as the *weighted median* of the support vectors. Indeed, for a set of 1-dimensional points $x_1, \ldots, x_n$ the median m is defined by the fact that it separates the data set into two equal parts. Noticing that $(x_i - m)/|x_i - m| =$

$\mathrm{sgn}(x_i - m)$ equals -1, 0 or 1 depending on whether $x_i < m$, $x_i = m$ or $x_i > m$ respectively, we see that the median can indeed be defined implicitly by:

$$\sum_i \frac{(x_i - m)}{\|x_i - m\|} = 0.$$

This characterization of the median has the obvious advantage that the generalization to higher dimensions is straightforward [7]. The improved robustness of the solution of the linear cost function (36) with respect to the original one (7) is nicely illustrated in Fig. 2.

3.2 Further properties

To gain further insight in the behaviour of solutions we once again assume that the centre of the sphere has already been located, so that the cost function depends solely on R. We also assume that the points have been labeled to produce an increasing sequence of distances $d_i = \|\mathbf{x}_i - \mathbf{a}\|$. Hence:

$$\zeta_1(R) = R + C\sum_i \varrho(d_i - R) = R + C\sum_i (d_i - R)H(d_i - R) = R + C\sum_i b_i(d_i - R),$$

where we have once again re-introduced the binary auxiliary variables b_i defined in eq.(16) Rearranging the terms we arrive at:

$$\zeta_1(R) = \Big(1 - C\sum b_i(R)\Big)\, R + C\sum_i b_i(R) d_i, \tag{41}$$

which elucidates that the function is piecewise linear, with a piecewise constant slope equal to $1 - C\sum b_i$. For notational convenience, we define

$$\beta(R) = 1 - C\sum_i b_i(R) \quad \text{and} \quad \delta(R) = \sum b_i(R) d_i,$$

resulting in $\zeta_1(R) = \beta(R)\, R + C\delta(R)$. Furthermore, $\beta(0) = 1 - nC$ and increases by jumps of size (multiples of) C to reach 1 when $R = d_n$. Hence the minimum R^* is located at the distance d_i for which β jumps above zero.

These considerations allow us to mirror the conclusions we obtained for the original cost function:

1. The optimal value of R^* coincides with one of the distances d_i which means that the optimal circle passes through at least one of the data points.
2. The optimal value R^* changes discontinuously whenever the unit cost takes on a value $C = 1/k$ (for $k = 2, \ldots, n$).
3. Non-trivial solutions exist only within the range:

$$\frac{1}{n} \leq C \leq \frac{1}{2}.$$

   For other values of C either all or no points are outliers.
4. The Schölkopf bounds (35) (and the ensuing conclusions) prevail.

4 Conclusions

In this paper we re-examined the *support vector data descriptor* (SVDD) (introduced by [8] and [5]) for one-class classification. Our investigation was prompted by the observation that the definition of slack variables as specified in the SVDD approach, lacks a clear geometric interpretation. We therefore re-analyzed the SVDD constrained optimization problem, focussing on the primal formulation, as this allowed us to gain further insight into the behaviour of the solutions. We applied the same analysis to two natural alternatives for the SVDD function. The first one turned out to suffer from unacceptable limitations, but the second one produces results that are very similar to the original formulation, but enjoys enhanced robustness with respect to outliers. We therefore think it could serve as an alternative for the original.

Acknowledgement This research is partially supported by the Specific Targeted Research Project (STReP) FireSense *Fire Detection and Management through a Multi-Sensor Network for the Protection of Cultural Heritage Areas from the Risk of Fire and Extreme Weather Conditions* (FP7-ENV-2009-1244088-FIRESENSE) of the European Union's 7th Framework Programme Environment (including Climate Change).

References

1. Ben-Hur A., Horn D., Siegelmann H. T., and Vapnik V.: Support vector clustering. Journal of Machine Learning Research, 2:125137 (2001)
2. Lee J. and Lee D.: An improved cluster labeling method for support vector clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27:461464, (2005)
3. Lee S. and Daniels K.: Cone cluster labeling for support vector clustering. In Proceedings, 2006 SIAM Conference on Data Mining, pages 484488, (2006)
4. Schölkopf, B., Williamson, R.C., Shrinking the tube: A new support vector regression algorithm. Advances in Neural Information Processing Systems, (1999)
5. Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J., Support vector method for novelty detection. In Advances in Neural Information Processing Systems, 12:582-588, MIT Press, (2000)
6. Shioda R. and Tuncel L.: Clustering via Minimum Volume Ellipsoids. Journal of Comp. Optimization and App. vol.37(3) (2007)
7. Small, C.G., A survey of multidimensional medians. International Statistical Review, 58(3):263-277, (1990)
8. Tax, D.M.J., Duin R.P.W., Support vector domain description. Pattern Recognition Letters, 20(11-13):1191-1199, December (1999)
9. Tax, D.M.J.L: One-class classification: concept learning in the absence of counter example. PhD Thesis, TU Delft, 2001.
10. Ypma, A., Duin, R., Support objects for domain approximation. In ICANN, Skovde, Sweden (1998)

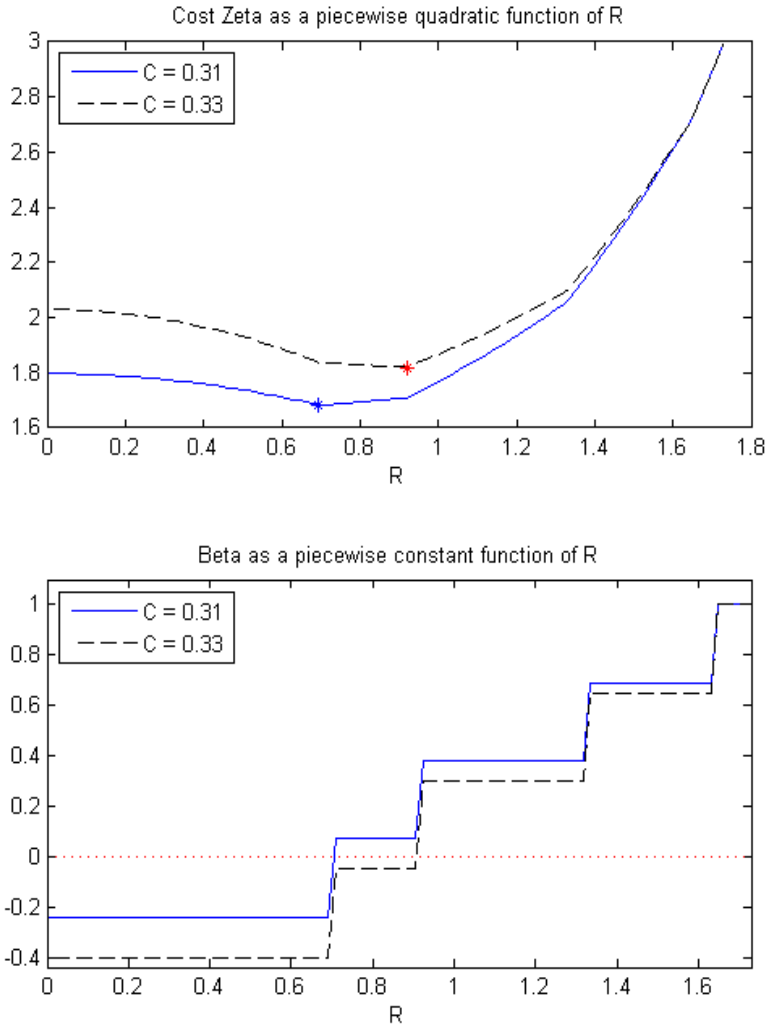

Fig. 1. *Top:* Total cost ζ (for two slightly different values of the unit cost C) as a function of the radius R for a simple data set comprising four points. This continuous function is composed of quadratic segments $\beta(R)R^2 + C\gamma(R)$. The piecewise constant behaviour of the β coefficient (which determines whether the segment is increasing or decreasing) is plotted in the bottom figure. *Bottom:* The quadratic coefficient $\beta(R) = 1 - C\sum_i b_i(R)$ is a piecewise constant function for which the jumps occur whenever R equals one of the distances $d_i = \|\mathbf{x}_i - \mathbf{a}\|$. For $C = 0.31$ this jump occurs around 0.7 resulting in a ζ-minimum at that same value. Increasing C slightly to $C = 0.33$ pushes the 2nd β-segment below zero, resulting in a ζ-minimum equal to $d_2 \approx 0.92$.

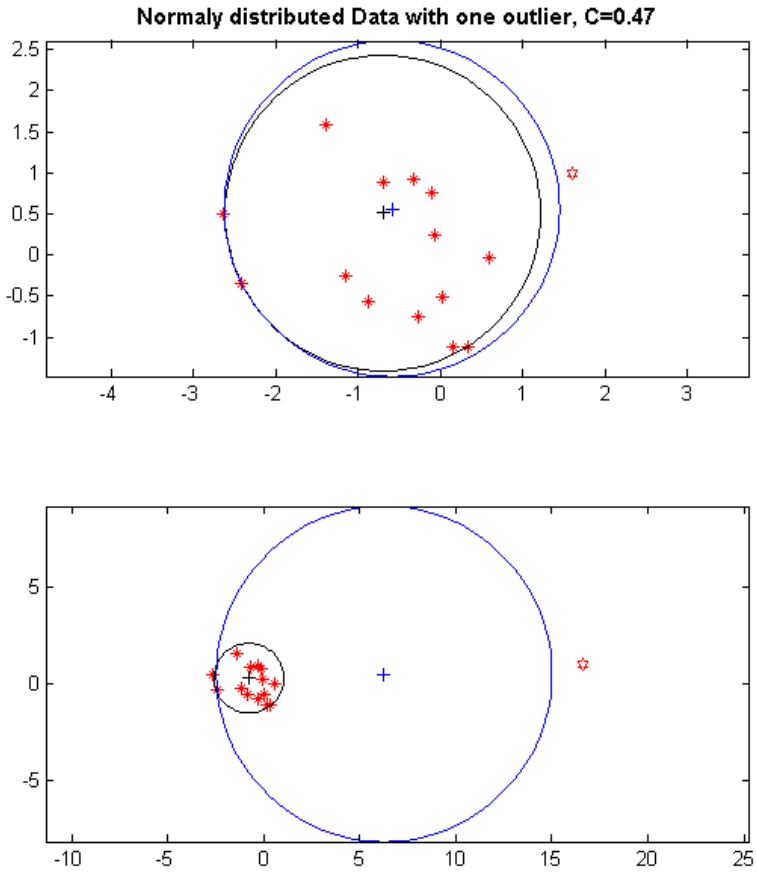

Fig. 2. Comparison of the optimal sphere for the original SVDD-function (in blue, cf. eq. (9), and the linear alternative (in black, cf. eq. (36)). The data sets in the top and bottom figures are identical except for the starred point on the right which, in the bottom figure (different scale!), has been moved far away from the rest of the cluster. Clearly, the optimal circle based on the linear function is essentially unaffected whereas the SVDD solution is dramatically inflated by this outlier.