Data Domain Description using Support Vectors

David M.J. Tax*, Robert P.W.Duin

Pattern Recognition Group, Faculty of Applied Science,
Delft University of Technology
Lorentzweg 1, 2628 CJ Delft, The Netherlands

Abstract. This paper introduces a new method for data domain description, inspired by the Support Vector Machine by V.Vapnik, called the Support Vector Domain Description (SVDD). This method computes a sphere shaped decision boundary with minimal volume around a set of objects. This data description can be used for novelty or outlier detection. It contains support vectors describing the sphere boundary and it has the possibility of obtaining higher order boundary descriptions without much extra computational cost. By using the different kernels this SVDD can obtain more flexible and more accurate data descriptions. The error of the first kind, the fraction of the training objects which will be rejected, can be estimated immediately from the description.

1. Introduction

In this paper we address an extension of the classification problem, the data domain description problem (also called One-class Classification). In domain description the task is not to distinguish between classes of objects like in classification problems or to produce a desired outcome for each input object like in regression problems, but to give a *description* of a set of objects. This description should be able to distinguish between the class of objects represented by the training set, and *all* other possible objects in the object space. The data domain description is used for outlier detection or novelty detection, the detection of objects which differ significantly from the given data set.

Different methods for data domain description or outlier detection have been developed. When an underlying statistical law for the outlying patterns is assumed, this underlying distribution should be estimated (see for instance [1]). When nothing about the outlier distribution can be assumed (or if an insufficient number of outlier examples is available), only a description of (the boundary of) the target class can be made. Most often a probability density of the available data is estimated and new test objects which are under some density threshold will be rejected (see [3]).

*This work was partly supported by the Foundation for Applied Sciences (STW) and the Dutch Organization for Scientific Research (NWO)

In this paper a new method for data domain description is presented. The method is inspired by the Support Vector Machines by V. Vapnik ([4]). Here the sphere with minimal volume (or minimal radius) containing all objects is found. This idea was already used in [2], but there it was just used to find an estimate of the VC-dimension of a hyperplane and not a description for a data set. We give the theoretical background of the method in section 2. In sections 3. and 4. we focus on choices for the free parameters and look at some characteristics of the methods. We give conclusions in section 5.

2. Theory

Of a data set containing N data objects, $\{x_i, i = 1, .., N\}$, a description is required. We try to find the sphere, described by center a and radius R, with minimum radius R and which contains all (or most of) the data objects. Because this description may become very sensitive to the most outlying objects in the target data set, we allow for some data points outside the sphere. Analogous to [4] we introduce slack variables ξ_i and we obtain the constraints:

$$\begin{aligned} (x_i - a)(x_i - a)^T &\leq R^2 + \xi_i \\ \xi_i &\geq 0 \end{aligned} \tag{1}$$

We minimize the radius R of the sphere and the size of the slack variables:

$$F(R, a, \xi_i) = R^2 + C \sum_i \xi_i \tag{2}$$

for given constant C, which gives the trade-off between the two error terms: volume of the sphere and the number of target objects rejected.

Incorporating the constraints in (2), we construct the Lagrangian:

$$L(R, a, \alpha_i, \xi_i) = R^2 + C \sum_i \xi_i - \sum_i \alpha_i \{R^2 + \xi_i - (x_i^2 - 2ax_i + a^2)\} - \sum_i \gamma_i \xi_i \tag{3}$$

with Lagrange multipliers $\alpha_i \geq 0$ and $\gamma_i \geq 0$. Setting the partial derivatives to zero, new constraints are obtained:

$$\begin{aligned} \sum_i \alpha_i &= 1 \\ a &= \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i} = \sum_i \alpha_i x_i \\ 0 &\leq \alpha_i \leq C \end{aligned} \tag{4}$$

Resubstituting gives to maximize with respect to α_i:

$$L = \sum_i \alpha_i (x_i \cdot x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \tag{5}$$

with constraints (4).

To test if a new object z is within the sphere, the distance to the center of the sphere has to be calculated. A test object z is accepted when this distance is smaller than the radius:

$$(z-a)(z-a)^T \quad = \quad (z \cdot z) - 2\sum_i \alpha_i (z \cdot x_i) + \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \leq R^2 \quad (6)$$

For a small set of objects in the training set the equality in equation (6) will be satisfied. These objects have $\alpha_i \neq 0$ and are called the support objects. When $C < 1$ support objects with $\alpha_i = C$ will occur. These objects are outside the sphere and are considered outliers. The rest of the training data is within the description. This method of making a data domain description will be called the Support Vector Domain Description (SVDD).

3. Generalizing to other kernels

Normally, data is not spherical distributed, even when the most outlying objects are ignored. To make a more flexible method, the object vectors x can be transformed to a higher dimensional feature space. As explained in [4] the inner products in equations 6 and 5 can be substituted by a kernel function $K(x_i, x_j)$, when this kernel satisfies Mercer theorem. The problem of finding a data domain description is now given by:

$$L = \sum_i \alpha_i K(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \quad (7)$$

with constraints given by equation (4). A test object z is accepted when:

$$K(z,z) - 2\sum_i \alpha_i K(z, x_i) + \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \leq R^2 \quad (8)$$

Taking different kernel functions K's result in other types of feature spaces and thus on differently shaped domain descriptions. This can make the description more flexible and more accurate than the very rigid spherical shape.

3.1. Kernel choices

When a polynomial kernel is used, $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$, with d is the degree of the polynomial, distances between objects will be enlarged for larger degrees d. The SVDD is completely adjusted to the objects which are most remote. This kernel results in a very large and sparse descriptions.

To suppress the growing distances for larger feature spaces, a Gaussian kernel is more appropriate:

$$K(x_i, x_j) = \exp(-(x_i - x_j)^2 / s^2) \quad (9)$$

The free parameter s gives the width, or spread, of the kernel.

For this Gaussian kernel is $K(x_i, x_i) \equiv 1$ and equation (7) boils down to:

$$L = 1 - \sum_i {\alpha_i}^2 - \sum_{i \neq j} \alpha_i \alpha_j K(x_i, x_j) \tag{10}$$

To find a maximum for equation (10) with constraints (4) two conflicting objectives have to be met:

1. the more α_i's are unequal to zero, the smaller each individual α_i can get and the smaller the factor $\sum_i {\alpha_i}^2$. This means an increase of the number of support vectors.
2. the more α_i's are equal to zero, the less cross products $\alpha_i \alpha_j$ contribute in the factor $\sum_{i \neq j} \alpha_i \alpha_j K(x_i, x_j)$ which means a decrease of the number of support vectors.

This tradeoff between the sizes of α_i and the cross products depends on the weighting factors $K(x_i, x_j)$ and thus on the choice of the free parameter s.

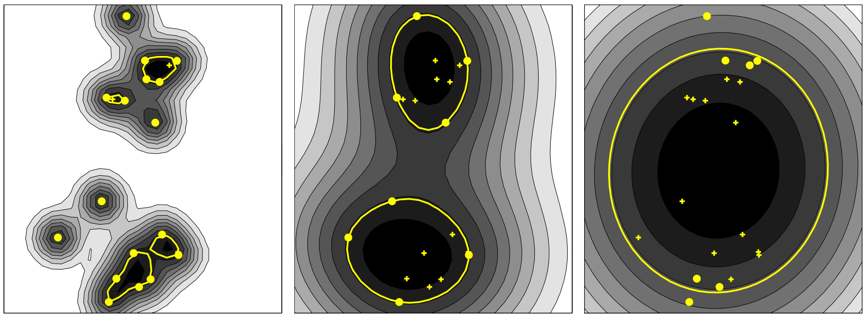

Figure 1: Distance to the center of the hypersphere, mapped back on the input space for three different values for s using a Gaussian kernel. The darker the color, the smaller the distance. The white dashed line indicates the surface of the hypersphere. The small circles indicate the support objects.

In figure 1 an artificial data set is shown with a Support vector domain description using a Gaussian kernel for three values of s. The data set contains 20 objects from a two dimensional mixture of two Gaussians. The gray value indicates the distance to the center of the sphere in the feature space. The white line crossing the support vectors (indicated by the small circles), is the boundary of the description. The width parameter s ranges from very small ($s = 0.5$ in the left figure) to large ($s = 12.5$ in the right figure). Note that the number of support vectors is decreasing and that the description becomes more sphere-like.

We can derive explicit solutions for equation (10) for the two different extreme situations, one for very small values and one for very large values for s.

For very small s is $K(x_i, x_j) \simeq 0, i \neq j$ and L is maximized when $\alpha_i = 1/N$. This is similar to the Parzen density estimation, where each object supports a kernel. For very large s is $K(x_i, x_j) \simeq 1$ and L is maximized when all $\alpha_i = 0$ except for one $\alpha_j = 1$. This solution is almost reached in the right subplot of figure 1. When a Taylor expansion of the Gaussian kernel is made, it can easily be shown that equation (7) becomes equation (5) (up to a scaling factor and an offset). In the case of moderate values of s (middle plot in figure 1) all values between 0 and 1 do exist for the $K(x_i, x_j)$ and an edited Parzen density estimator is obtained.

The parameter C gives the upper boundary for the parameters α_i (see equation (4)) and indicates how severely outlying objects are punished. Because $\sum_i \alpha_i = 1$ and $0 \leq \alpha_i \leq 1$, C should be $1/N \leq C \leq 1$. For $C < 1/N$ no solution can be found because then the constraint $\sum_i \alpha_i = 1$ can never be met, while for $C > 1$ one can always find the solution (α_i's are always less or equal to 1) which includes all training data.

Objects x_i which have a maximal α_i, $\alpha_i = C$, are special support objects. They are not lying on the boundary of the sphere but are outside the description. By restricting the parameter C, the cost of being outside the sphere is not very large and a greater fraction of the objects is allowed to be outside the sphere. This makes the sphere description less sensitive to noise in the remote objects but may also remove valuable objects. Two examples of these objects are shown in the right figure of figure 1.

4. Error estimation

To study the generalization or the overfitting characteristics of the SVDD, we have to get an indication of (1) the number of target patterns that will be rejected (errors of the first kind) and (2) of the number of outlying patterns that will be accepted (errors of the second kind).

We can estimate the error of the first kind by applying the leave-one-out method on the training set containing the target class (see also [4]). When leaving out an object from the training set which is no support object, the original description is found. When a support object is left out, the optimal sphere description can be made smaller and this left-out object will then be rejected. Thus the error can be estimated by:

$$E[P(\text{error})] = \frac{\#SV}{N} \tag{11}$$

where $\#SV$ is the number of support vectors.

Using a Gaussian kernel, we can regulate the number of support vectors by changing the width parameter s and therefore also the error of the first kind. When the number of support vectors is too large, we have to increase s, while when the number is too low, we have to decrease s. This guarantees that the width parameter in the SVDD is adapted for the problem at hand given the error.

The chance that outlying objects will be accepted by the description, the error of the second kind, cannot be estimated by this measure. We assumed only a good description of the target class in the form of a training set is available. **All** other patterns are outliers per definition. For an estimate of the error of the second kind assumptions about the outlying objects should be made.

5. Conclusions

In this paper we propose a data description which only depends on a few target objects, the support objects. In this Support Vector Data Description, SVDD, test objects only have to be compared with the support objects by an inner product or a more general kernel function. By adapting the kernel function, this method becomes more flexible than just a sphere in the input space. We focussed on a Gaussian kernel, which seems to work very well.

To become robust against outlying objects in the training set, the SVDD allows for some target objects not included in the sphere description. An extra parameter C gives the trade-off between the number of errors made on the training set and the size of the sphere description. In practice the size of this parameter is not very crucial for finding a good solution.

The Gaussian kernel in the SVDD has another free parameter, the width of the kernel s. Different extremes for this width parameter result in more or less flexible descriptions. For very small values for s a Parzen density estimation is obtained where all target objects become support objects. For very large values of s just one prototype for the complete data set is used and almost the complete training set can be disregarded. Applying a moderate value for the width parameter, an edited Parzen estimation is obtained.

In the SVDD method the error on the target class can be estimated immediately by calculating the fraction of target objects which become support objects. Setting the error on the training set beforehand, the value of s can be set to match the allowed accuracy on the data set at hand. This makes the SVDD both a flexible and a fast method for describing a data set.

References

[1] G. Ritter and M.T. Gallegos. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, 18:525–539, April 1997.

[2] B. Schoelkopf. *Support Vector Learning.* PhD thesis, Technischen Universitat Berlin, 1997.

[3] L. Tarassenko, P. Hayton, and M Brady. Novelty detection for the identification of masses in mammograms.

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