# Pump Failure Detection using Support Vector Data Descriptions

David M.J. Tax[1], Alexander Ypma[1], and Robert P.W. Duin[1]

Pattern Recognition Group
Dept. of Applied Physics, Faculty of Applied Sciences, Delft University of Technology
Lorentzweg 1, 2628 CJ Delft, The Netherlands
{davidt,ypma}@ph.tn.tudelft.nl

**Abstract.** For good classification preprocessing is a key step. Good preprocessing reduces the noise in the data and retains most information needed for classification. Poor preprocessing on the other hand can make classification almost impossible. In this paper we evaluate several feature extraction methods in a special type of outlier detection problem, machine fault detection. We will consider measurements on water pumps under both normal and abnormal conditions. We use a novel data description method, called the Support Vector Data Description, to get an indication of the complexity of the normal class in this data set and how well it is expected to be distinguishable from the abnormal data.

## 1 Introduction

For good classification the preprocessing of the data is a important step. Good preprocessing reduces the noise in the data and retains as much of the information as possible [1]. When the number of objects in the training set is too small for the number of features used (the feature space is under sampled), most classification procedures cannot find good classification boundaries. This is called the **curse of dimensionality** (see for an extended explanation [3]). By good preprocessing the number of features per object can be reduced such that the classification problem can be solved.

A particular type of preprocessing is feature selection. In feature selection one tries to find the optimal feature set from a already given set of features [5]. In general this set is very large. To compare different feature sets, a criterion has to be defined. Often very simple criteria are used for judging the quality of the feature set or the difficulty of the data set. See [2] for a list of different measures.

Sometimes we encounter a special type of classification problems, so-called outlier detection or data domain description problems. In data domain description the goal is to accurately describe **one** class of objects, the target class, as opposed to a wide range of other objects which are not of interest or are considered outliers [7]. This last class is therefore called the outlier class. Many standard pattern recognition methods are not well equipped to handle this type of problem; they require complete descriptions for both classes. Especially when

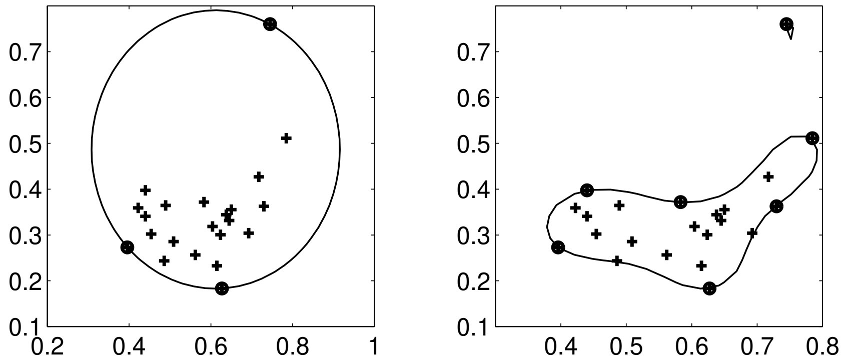


**Fig. 1.** Data description of a small data set, (left) normal spherical description, (right) description using a Gaussian kernel.

the outlier class is very diverse and ill-sampled, normal (two class) classifiers obtain very bad generalizations for this class.

In this paper we will introduce a new method for data domain description, the Support Vector Data Description (SVDD). This method is inspired on the Support Vector Classifier by V. Vapnik [9] and it defines a spherically shaped boundary with minimal volume around the target data set. Under some restrictions, the spherically shaped data description can be made more flexible by replacing normal inner products by some kernel functions. This will be explained in more detail in section 2.

In this paper we try to find the best representation of a data set such that the target class is optimally clustered and can be distinguished as best as possible from the outlier class. The data set here considered is vibration data recorded from a water pump. The target class contains recordings from the normal behavior of the pump, while erroneous behaviour is placed in the outlier class. Different preprocessing methods will be applied to the recorded signals in order to find the optimal set of features.

We will start with an explanation of the Support Vector Data Description in section 2. In section 3 the origins of the vibration data will be explained and in section 4 we will discuss the different types of features extracted from this data set. In section 5 the results of the experiments are shown and we will conclude with conclusions in section 6.

## 2 Support Vector Data Description

The Support Vector Data Description (SVDD) is the method which we will use to describe our data. It is inspired on the Support Vector Classifier of V. Vapnik ([9], or for a more simple introduction [6]). The SVDD is explained in more detail in [8], here we will just give a quick impression of the method.

The idea of the method is to find the sphere with minimal volume which contains all data. Assume we have data set containing $N$ data objects, $\{x_i, i = 1, .., N\}$ and the sphere is described by center $a$ and radius $R$. We now try to minimize an error function containing the volume of the sphere. The constraints that objects are within the sphere are imposed by applying Lagrange multipliers:

$$L(R, a, \alpha_i) = R^2 - \sum_i \alpha_i \{R^2 - (x_i^2 - 2ax_i + a^2)\} \tag{1}$$

with Lagrange multipliers $\alpha_i \geq 0$. This function has to be minimized with respect to $R$ and $a$ and maximized with respect to $\alpha_i$.

Setting the partial derivatives of $L$ to $R$ and $a$ to zero, gives:

$$\sum_i \alpha_i = 1 \qquad \text{and} \qquad a = \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i} = \sum_i \alpha_i x_i \tag{2}$$

This shows that the center of the sphere $a$ is a linear combination of the data objects $x_i$.

Resubstituting these values in the Lagrangian gives to maximize with respect to $\alpha_i$:

$$L = \sum_i \alpha_i (x_i \cdot x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \tag{3}$$

with $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$.

This function should be maximized with respect to $\alpha_i$. In practice this means that a large fraction of the $\alpha_i$ become zero. For a small fraction $\alpha_i > 0$ and these objects are called Support Objects. We see that the center of the sphere depends just on the few support objects, objects with $\alpha_i = 0$ can be disregarded.

Object $z$ is accepted when:

$$\begin{aligned} (z - a)(z - a)^T &= (z - \sum_i \alpha_i x_i)(z - \sum_i \alpha_i x_i) \\ &= (z \cdot z) - 2 \sum_i \alpha_i (z \cdot x_i) + \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \leq R^2 \end{aligned} \tag{4}$$

In general this does not give a very tight description. Analogous to the method of Vapnik [9], we can replace the inner products $(x \cdot y)$ in equations (3) and in (4) by kernel functions $K(x, y)$ which gives a much more flexible method. When we replace the inner products by Gaussian kernels for instance, we obtain:

$$(x \cdot y) \rightarrow K(x, y) = \exp(-(x - y)^2 / s^2) \tag{5}$$

Equation (3) now changes into:

$$L = 1 - \sum_i {\alpha_i}^2 - \sum_{i \neq j} \alpha_i \alpha_j K(x_i, x_j) \tag{6}$$

and and the formula to check if a new object $z$ is within the sphere (equation (4)) becomes:

$$1 - 2\sum_i \alpha_i K(z, x_i) + \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \le R^2 \tag{7}$$

We obtain a more flexible description than the rigid sphere description. In figure 1 both methods are shown applied on the same two dimensional data set. The sphere description on the left includes all objects, but is by no means very tight. It includes large areas of the feature space where no target patterns are present. In the right figure the data description using Gaussian kernels is shown, and it clearly gives a superior description. No large empty areas are included, what minimizes the change of accepting outlier patterns. To obtain this tighter description, one training object in the right upper corner is rejected from the description.

This Gaussian kernel contains one extra free parameter, the width parameter $s$ in the kernel (equation (5)). As shown in [8] this parameter can be determined by setting a priori the maximal allowed rejection rate of the target set, i.e. the error on the target set. Applying leave-one-out estimation on the training set shows that non-support objects will be accepted by the SVDD when they are left out of the training set, while support object will be rejected. Therefore the error on the target set can be estimated by the fraction of training objects that become support objects in the data description:

$$E[P(\text{error})] = \frac{\#SV}{N} \tag{8}$$

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

In [8] it is shown that equation 8 is a good estimate of the error on the target class. The fact that the fraction support objects can immediately be used for the estimate of the error on the target class, makes this data description method a very efficient one with respect to the number of objects needed. Because independent test data is not necessary, all available data can immediately used for estimating the SVDD.

Note we cannot set a priori restrictions on the error on the outlier class. In general we only have a good representation of the target class and the outlier class is per definition everything else.

## 3 Machine vibration analysis

Vibration was measured on a small pump in an experimental setting and on two identical pump sets in pumping station “Buma” at Lemmer. One of the pumps in the pumping station showed severe gear damage (pitting, i.e. surface cracking due to unequal load and wear) whereas the other pump showed no significant damage. Both pumps of the pumping station have similar power consumption, age and amount of running hours.

The Delft test rig comprises a small submersible pump, which can be made to run at several speeds (from 46 to 54 Hz) and several loads (by closing a membrane controlling the water flow). A number of faults were induced to this pump: loose foundation, imbalance, failure in the outer race of the uppermost ball bearing. Both normal and faulty behaviour was measured at several speeds and loads.

In both set-ups accelerometers were used to measure the vibration near different structural elements of the machine (shaft, gears, bearings). Features from several channels were collected as seperate samples in the the same feature space, i.e. inclusion of several channels increases the sample size, not the feature dimensionality. By putting the measurements of the different sensors into one data set, the data set increases in size, but information on the exact position of an individual measurement is lost.

For the Lemmer measurements three feature sets were constructed by joining different sensor measurements into one set:

1. one radial channel near the place of heavy pitting (expected to be a good feature),
2. two radial channels near both heavy and moderate pitting along with an (unbalance sensitive) axial channel, and
3. inclusion of all channels (except for the sensor near the outgoing shaft which might be too sensitive to non-fault related vibration).

As a reference dataset, we constructed a high-resolution logarithmic power spectrum estimation (512 bins), normalized w.r.t. mean and standard deviation and its linear projection (using Principal Components Analysis) on a 10-dimensional subspace. Three channels were included, expected to be roughly comparable to the second configuration previously described.

For the Delft dataset the same procedure was followed: the first set contains data from one channel near a fault location, the second set contains three channels near fault bearings and the third set contains all five channels.

## 4 Features for machine diagnostics

We compared several methods for feature extraction from vibration data. It is well known that faults in rotating machines will be visible in the acceleration spectrum as increased harmonics of running speed or presence of sidebands around characteristic (structure-related) frequencies. Due to overlap in series of harmonic components and noise, high spectral resolution may be required for adequate fault identification. This may lead to difficulties because of the curse of dimensionality: one needs large sample sizes in high-dimensional spaces in order to avoid overfitting of the train set. Hence we focused on relatively low feature dimensionality (64) and compared the following features:

**power spectrum:** standard power spectrum estimation, using Welch's averaged periodogram method. Data is normalized to the mean prior to spectrum estimation, and feature vectors (consisting of spectral amplitudes) are normalized w.r.t. mean and standard deviation (in order to retain only sensitivity to the spectrum shape).

**envelope spectrum:** a measurement time series was demodulated using the Hilbert transform, and from this cleaned signal (supposedly containing information on periodic impulsive behavior) a spectrum was determined using the above method. Prior to demodulation a bandpass-filtering in the interval 125 - 250 Hz (using a wavelet decomposition with Daubechies-wavelets of order 4) was performed: gear mesh frequencies will be present in this band and impulses due to pitting are expected to be present as sidebands. For comparison, this pre-filtering step was left out in another data set.

**autoregressive modelling:** another way to use second-order correlation information as a feature is to model the timeseries with an autoregressive model (AR-model). For comparison with other features, an AR(64)-model was used (which seemed sufficient to extract all information) and model coefficients were used as features.

**MUSIC spectrum estimation:** if a time series can be modeled as a model of sinusoids plus noise, we can use a MUSIC frequency estimator to focus on the important spectral components [4]. A statistic can be computed that tends to infinity when a signal vector belongs to the so-called signal subspace.
When one expects amplitudes at a finite number of discrete frequencies to be a discriminant indicator, MUSIC features may enable good separability while keeping feature size (relatively) small.

**some classical indicators:** three typical indicators for machine wear are

- rms-value of the power spectrum
- kurtosis of the signal distribution
- crest-factor of the vibration signal

The first feature is just the average amount of energy in the vibration signal (square root of mean of squared amplitudes). Kurtosis is the $4^{th}$ central moment of a distribution, that measures the 'peakedness' of a distribution. Gaussian distributions will have kurtosis near 0 whereas distributions with heavy tails (e.g. in the presence of impulses in the time signal) will show larger values. The crest-factor of a vibration signal is defined as the peak amplitude value divided by the root-mean-square amplitude value (both from the envelope detected time signal). This feature will be sensitive to sudden defect bursts, while the mean (or: rms-) value of the signal has not changed significantly.

## 5 Experiments

To compare the different feature sets the SVDD is applied to all target data sets. Because also test objects from the outlier class are available (i.e. the fault class defined by the pump exhibiting pitting, see section 3), the rejection performance on the outlier set can also be measured.

In all experiments we have used the SVDD with a Gaussian kernel. For each of the feature sets we have optimized the width parameter $s$ in the SVDD such that 1%, 5%, 10%, 25% and 50% of the target objects will be rejected, so for each data set and each target error another width parameter $s$ is obtained. For each feature set this gives an acceptance-rejection curve for the target and the outlier class.

We will start with considering the Lemmer data set with the third sensor combination (see section 3) which contains all sensor measurements. In this case we do not use prior knowledge about where the sensors are placed and which sensor might contain most useful information.

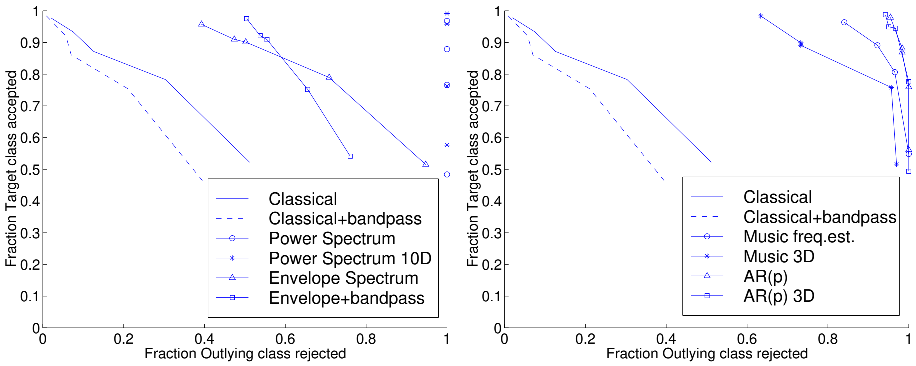


**Fig. 2.** Acceptance/rejection performance of the SVDD on the features on the Lemmer data set, with all sensor measurements collected.

In figure 2 the characteristic of the SVDD on this data is shown. If we look at the results for the power spectrum using 512 bins (see left figure 2) we see that for all target acception levels we can reject 100% of the outlier class. This is the ideal behavior we are looking for in a data description method and it shows that in principle the target class can be distinguished from the outlier class very well. Drawback in this representation though is that each object contains 512 power spectrum bins, it is both expensive to calculate this large a Fourier spectrum and expensive in storage costs. That is why we try other, smaller representations.

Reducing this 512 bin spectrum to just 10 features by applying a Principal Component Analysis (PCA) and retaining the ten directions with the largest variations, we see that still we can perfectly reject the outlier class.

Looking at the results of the classical method and the classical method using bandpass filtering, we see that the target class and the outlier class overlap significantly. When we try to accept 95% of the target class only 10% or less is rejected by the SVDD. Also considerable overlap between the target and the outlier class is present when envelope spectra are used. When 5-10% of the target class is rejected, still about 50% of the outlier class is accepted. Here also bandpass filtering does not improve the performance very much, only for large target acceptance rates, the bandpass filtering is useful.

Finally in the right figure 2 the results for the MUSIC estimator and the AR-model features are shown. The results on the AR-model feature set and the MUSIC frequency estimation feature are superior to all other methods, with the AR model somewhat better than the MUSIC estimator. The AR model approximates almost the 512 bin power spectrum, only for very large acceptance rates of the target class, we see some patterns from the outlier class being accepted. Applying the SVDD on the first three principal components deteriorates the performance of the MUSIC estimator. The AR model still performs almost optimal.

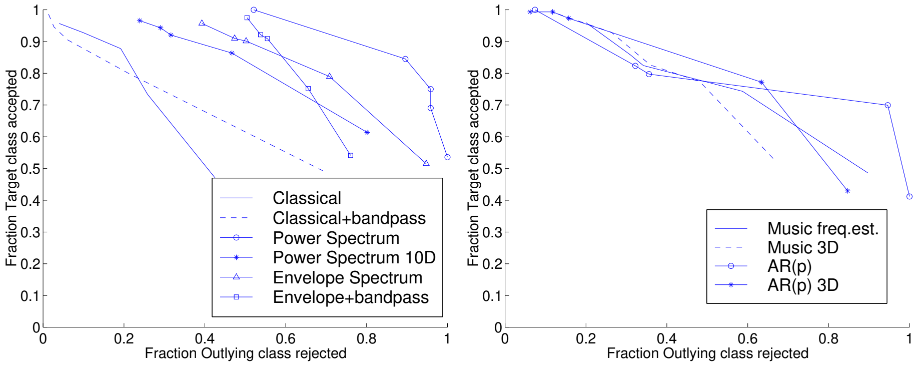


**Fig. 3.** Acceptance/rejection performance of the SVDD on the features on the Delft data set, with all sensor measurements collected.

In figure 3 similar figures are shown for the Delft measurements. Looking at the performance of the 512 bin power spectrum, we see that here already considerable overlap between the target and the outlier class exist. This indicates that this problem is more difficult than the Lemmer data set. The performances by the different features do not vary much, the MUSIC estimator, AR-model and the envelope spectrum perform about equal. In all cases there is considerable overlap between target and outlier class. This might indicate that for one (or more) of the outlier classes the characteristics are almost equal to the target class characteristics (and thus it is hard to speak of an outlier class).

The analysis was done on a data set in which all sensor information was used. Next we look at the performance of the first and the second combination of sensors in the Lemmer data set. In figure 4 the performance of the SVDD is shown on all feature sets applied on sensor combination (1) (on the left) and combination (2) (on the right). Here also the classical features perform poorly. The envelope spectrum works reasonably well, but both the MUSIC frequency estimator and the AR-model features perform perfectly. The data from sensor combination (1) is clearly better clustered than sensor combination (3). Only the AR model features and the Envelope detection with bandpass filtering on the single sensor data set shows reasonable performance.

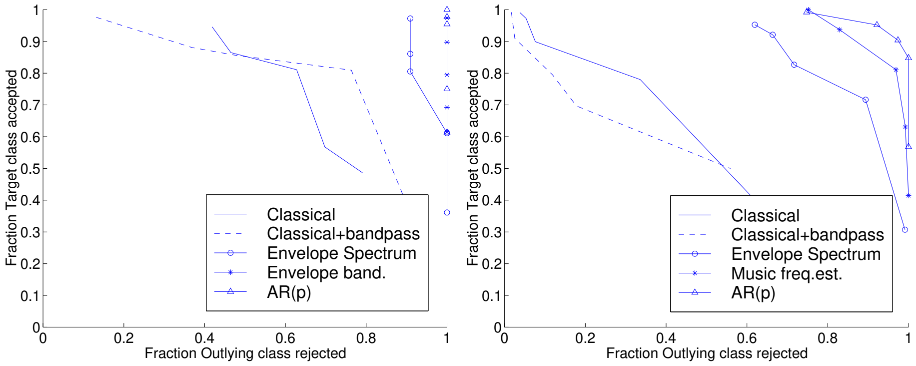


**Fig. 4.** Acceptance/rejection performance of the SVDD on the different features for sensor combination (1) and (2) in the Lemmer data set.

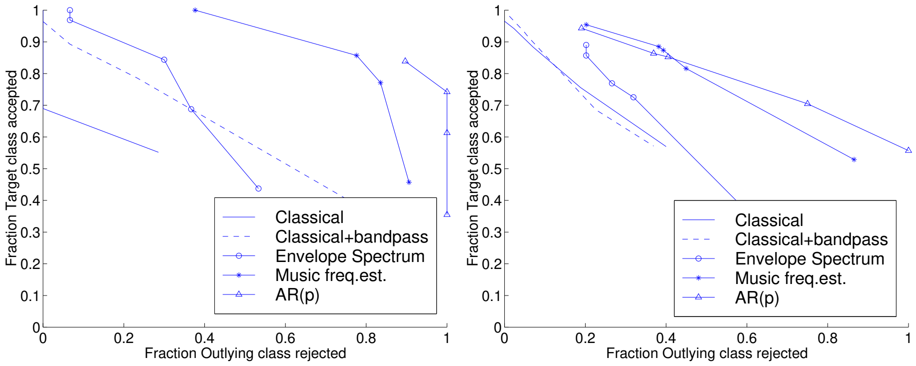


**Fig. 5.** Acceptance/rejection performance of the SVDD on the different features for sensor combination (1) and (2) in the Delft data set.

We can observe the same trend in figure 5, where the performances are plotted for sensor combination (1) and (2) in the Delft data set. Here also the MUSIC estimator and the AR model outperform the other types of features, but there are large errors, which can be expected considering the complexity of this problem.

## 6 Conclusion

In this paper we tried to find the best representation of a data set such that the target class can best be distinguished from the outlier class. This is done by applying the Support Vector Data Description, a method which finds the smallest sphere containing all target data. We applied the SVDD in a machine diagnostics problem, where the normal working situation of a pump in a pumping station (Lemmer data) and a pump in an experimental setting (Delft data) should be distinguished from abnormal behavior.

In this application data was recorded from several vibration sensors on a rotating machine. Three different subsets of the sensor channels were put together to create new data sets and several features were calculated from the time signals. Although the three sensor combinations show somewhat different results, a clear trend is visible.

As a reference a very high resolution power spectrum (512 bins) is used. In the case of the Lemmer measurements with this spectrum it is possible to perfectly distinguish between normal and abnormal situations, which means that in principle perfect classification is possible. In the case of Delft measurements, the distinction between target and outlier class becomes more difficult. This is probably caused by the fact that the Delft measurements contain more outlier situations which are very hard to distinguish from the normal class.

Performance of both MUSIC- and AR-features was usually very good in all three configuration data sets, but worst in the second configuration and best in the third configuration. This can be understood as follows: the sensors underlying configuration 2 are a subset of the sensors in configuration 3. Since the performance curves are based on percentages accepted and rejected, this performance may be enhanced by adding new points to a dataset (e.g. in going from configuration 2 to 3) that would be correctly classified according to the existing description. The sensor underlying the first configuration was close to the main source of vibration (the gear with heavy pitting), which explains the good performance on that dataset. From the results it is clear that there is quite some variation in discrimination power of different channels, but also that in this specific application inclusion of all available channels as separate samples can be used to enhance robustness of the method.

In the three dimensional classical feature set both classes severely overlap and can hardly be distinguished. This can be caused by the fact that one of the classical features is kurtosis, whose estimate shows large variance. Increasing the time signal over which the kurtosis is estimated might improve performance, but this would require long measurement runs.

When all sensor combinations and both Lemmer and Delft data sets are considered, the AR-model allows for the tightest description of the normal class when compared with all other features. We can conclude that if we want to use shorter representations of vibration data to overcome the curse of dimensionality, the AR-model is best choice.

## 7 Acknowledgments

This work was partly supported by the Foundation for Applied Sciences (STW) and the Dutch Organisation for Scientific Research (NWO). We would like to thank TechnoFysica B.V. and pumping station "Buma" at Lemmer, The Netherlands (Waterschap Noord-Oost Polder) for providing support with the measurements.

## References


1. C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Walton Street, Oxford OX2 6DP, 1995.
2. P.A. Devijver and J. Kittler. *Pattern Recognition, A statistical approach.* Prentice-Hall International, London, 1982.
3. R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.
4. J.G. Proakis and D.G. Manolakis. *Digital signal processing - principles, algorithms and applications, 2nd ed.* MacMillan Publ., New York, 1992.
5. P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.
6. D.M.J. Tax, D. de Ridder, and R.P.W. Duin. Support vector classifiers: a first look. In *Proceedings ASCI'97.* ASCI, 1997.
7. D.M.J. Tax and R.P.W Duin. Outlier detection using classifier instability. In Amin, A., Dori, D., Pudil, P., and Freeman, H., editors, *Advances in Pattern Recognition, Lecture notes in Computer Science*, volume 1451, pages 593–601, Berlin, August 1998. Proc. Joint IAPR Int. Workshops SSPR'98 and SPR'98 Sydney, Australia, Springer.
8. D.M.J. Tax and R.P.W Duin. Data domain description using support vectors. In Verleysen M., editor, *Proceedings of the European Symposium on Artificial Neural Networks 1999*, pages 251–256. D.Facto, Brussel, April 1999.
9. V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag New York, Inc., 1995.