Support vector data description applied to machine vibration analysis

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

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

Keywords: pattern recognition, one-class problems, outlier detection, Support Vector Machines, Support Vector Data Description, machine diagnostics

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 makes classification almost impossible. In this paper we try to find good preprocessing for a special type of outlier detection problem, machine diagnostics in which we try to distinguish between a target class and all other possible patterns. We will consider measurements on a water pump under both, normal and abnormal conditions. We use a novel data domain description method 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 an important step. Good preprocessing reduces the noise in the data and retains as much of the information as possible (see [Bis95]). 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 [DH73]). By good preprocessing the number of features per object can be reduced such that the classification problem can be solved.

A special type of preprocessing is feature selection. In feature selection one tries to find the optimal feature set from a already given set of features (see [PNK94]). In general this set is very large. To compare different feature sets, a criterion has to be defined. Most often very simple criteria are used for judging the quality of the feature set or the difficulty of the data set. See [DK82] for a list of different measures. Most important measures are the Mahanalobis distance between two or more classes and the nearest neighbour measure.

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 as outliers[TD98]. 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 the outlier class is very diverse and ill-sampled, normal (two class) classifiers obtain very bad generalizations for this class.

Several data description methods exist. Moya [MH96] trained a neural network with the restriction that the network forms closed decision surfaces. Also several vector quantization methods have been constructed [CGR91, Koh95]. Unfortunately these methods are focussed more on the representation of the target class and not on the option to reject outliers. Therefore in these methods it is difficult to find a boundary around the data which can reliably reject non-target objects.

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 [Vap95] and it defines a spherically shaped

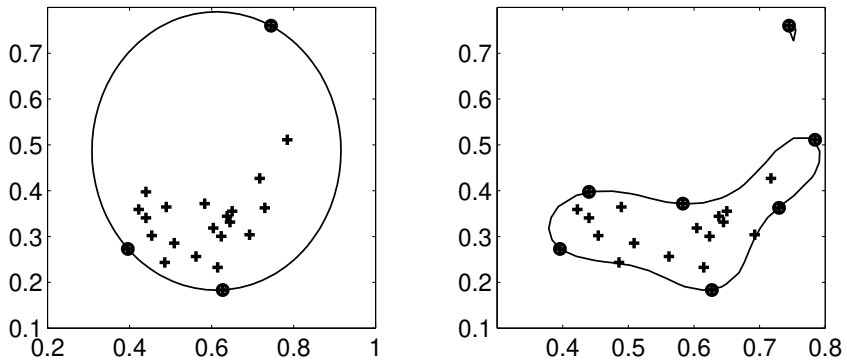

Figure 1: Data description of a small data set, (left) normal spherical description, (right) description using a Gaussian kernel.

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 which will be 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 (see [Vap95], or for a more simple introduction [TdRD97]). The SVDD is explained in more detail in [TD], 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 a 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 α_i.

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

$$\begin{aligned} \sum_i \alpha_i &= 1 \\ a &= \frac{\sum_i \alpha_i x_i}{\sum_i \alpha_i} = \sum_i \alpha_i x_i \end{aligned} \tag{2}$$

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 α_i:

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

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

This function should be maximized with respect to α_i. In practice this means that a large fraction of the α_i become zero. For a small fraction $\alpha_i > 0$ and the corresponding 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:

$$(z - a)(z - a)^T = (z - \sum_i \alpha_i x_i)(z - \sum_i \alpha_i x_i)$$

$$= (z \cdot z) - 2\sum_i \alpha_i (z \cdot x_i) + \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j)$$
$$\leq R^2 \qquad (4)$$

In general this does not give a very tight description. Analogous to the method of Vapnik [Vap95], 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:

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

Equation (3) now changes into:

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

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

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

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 empty areas are included, what minimizes the change of accepting outlier patterns.

This Gaussian kernel contains one extra free parameter, the width parameter s in the kernel (equation (5)). As shown in [TD] this parameter can be set by setting a priori the maximal allowed rejection rate of the target set, i.e. the error on the target set. This error can be estimated by the number of support vectors:

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

where $\#SV$ is the number of support vectors. We can regulate the number of support vectors by changing the width parameter s and therefore we can also set the error on the target set on the prespecified value.

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

The condition of rotating mechanical machinery can be monitored by measuring the vibration on the machine casing. After determining a suitable method for feature extraction from the measurement time series, a signature may be obtained that is unique for each machine. Significant deviations from this signature (*novelty*) will usually indicate faults or wear. However, since a machine will be used in several operating modes (differing loads, speeds, environmental conditions), the admissible ("normal") domain will consist of a set of signatures, hopefully clustered in feature space. We will use the previously described method for domain description to quantify the compactness of the normal class along with the amount of overlap with fault classes.

Vibration was measured on two identical pump sets in pumping station "Buma" at Lemmer, The Netherlands. This station is one of the two stations responsible for controlling the amount of water in the "Noord-Oost Polder". One pump showed severe gear damage (pitting, i.e. surface cracking due to unequal load and wear, see figure 2 adapted from [Tow91]), whereas the other showed no significant damage. Both pumps have similar power consumption, age and amount of running hours. The load of both pumps can be influenced by lowering or lifting a sliding door (which determines the amount of water that can be put through). Seven accelerometers were used to measure the vibration near different structural elements of the machine (shaft, gears, bearings).

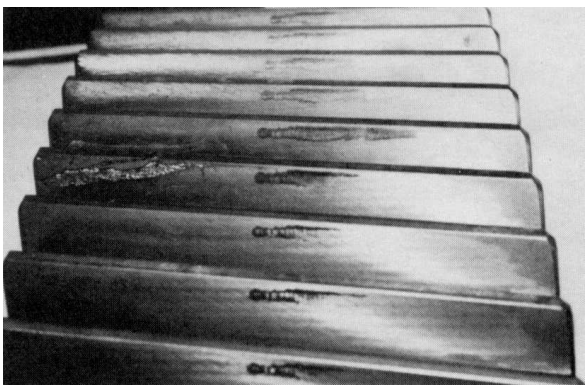

Figure 2: Severe case of pitting in a gear wheel

Vibration was measured with 7 accelerometers, placed near the driving shaft (in three directions), and upper and lower bearings supporting shafts of both gearboxes (that perform a two-step reduction of the running speed of the driving shaft to the outgoing shaft, to which the impeller is

attached). Measurements from several channels were combined in the following manner: from each channel a separate feature vector was extracted and added as a new sample to the dataset. Hence, inclusion of more than one channel gives rise to several data samples, each giving some information on the current measurement setting (as opposed to one sample if only one channel would be selected). If faults are adequately measurable by all sensors, we expect the amount of class overlap in data from a certain sensor to be roughly the same for all sensors. However, since the machine under investigation is quite large and measurement directions are not always the same, this assumption may not hold. Incorporation of multiple channels in the above manner might improve robustness (less dependence on particular sensor to be selected), but on the other hand might introduce class overlap, because uninformative channels are treated equally important as informative channels. In this paper, we try to quantify this effect. An alternative approach would be to combine channels in the time domain to overcome this dependence on sensor informativity [YP99].

Three feature sets were constructed by joining different sensor measurements into one set:

1. one radial channel near the place of heavy pitting,
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.

In all datasets we included measurements at various machine loads, e.g. samples corresponding to measurements from the healthy machine operating at maximum load were added to samples corresponding to smaller loads to form the total normal (target) dataset (and the same for data from the worn machine).

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 (figure 3) and noise, high spectral resolution may be required for adequate fault identification.

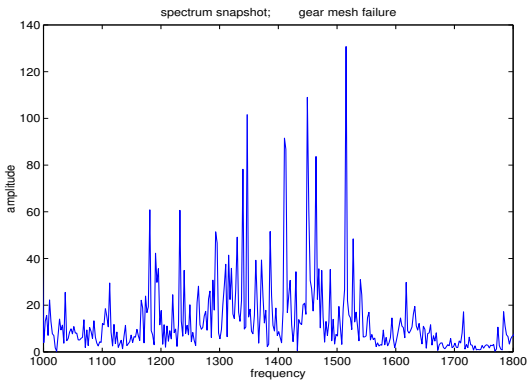

Figure 3: Overlapping harmonic series, visible in high resolution spectrum

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 fea-

ture 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

$$y(n) = x(n) + w(n) = \sum_{i=1}^{p} e^{j(2\pi n + \phi_i)} + w(n) \tag{9}$$

i.e. a model of sinusoids plus noise, we can use a MUSIC (MUltiple SIgnal Classification) frequency estimator to focus on the important spectral components ([PM92]).

A statistic can be computed that tends to infinity when a signal vector $\mathbf{e}_f$ (sinusoid with discrete frequency) belongs to the so-called signal subspace

$$P(f) = \frac{1}{L - \sum_{i=1}^{p} |\mathbf{e}_f^H \mathbf{u}_i|^2} \tag{10}$$

where $\mathbf{u}_i$ is the i^{th} eigenvector of R_{yy}, the correlation matrix of signal y having rank L.

In short, 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} order central moment of a distribution, measuring the 'peakedness' of a distribution. Gaussian distributions will have a normalized 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 $\frac{A_{peak}}{A_{rms}}$, i.e. 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 a acceptance-rejection curve for the target and the outlier class.

We will start with considering 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.

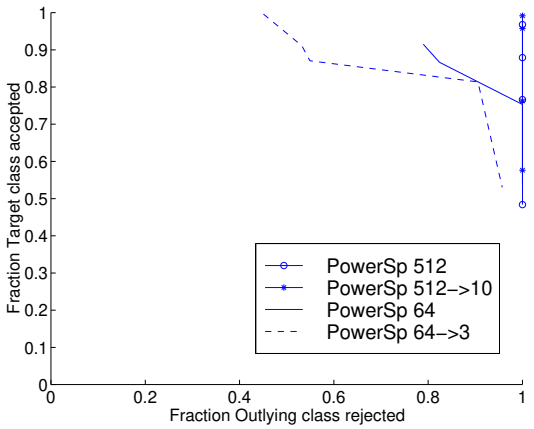

Figure 4: Acceptance/rejection performance of the SVDD on the 512-bin and 64-bin power spectrum features on sensor combination (3).

In figure 4 the characteristic of the SVDD is shown for the first data set using four different types of features. If we look at the results for the power spectrum using 512 bins we see that for all target acception levels we can always 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 we can still perfectly reject the outlier class. Using a less well sampled power spectrum of just 64 bins results in a decrease of performance. Only when 50% of the target class can be rejected, more than 95% of the outlier class is rejected. Finally, when just the three largest principal components are used for the SVDD, performance is comparable, only for small target rejection rates it is somewhat poorer.

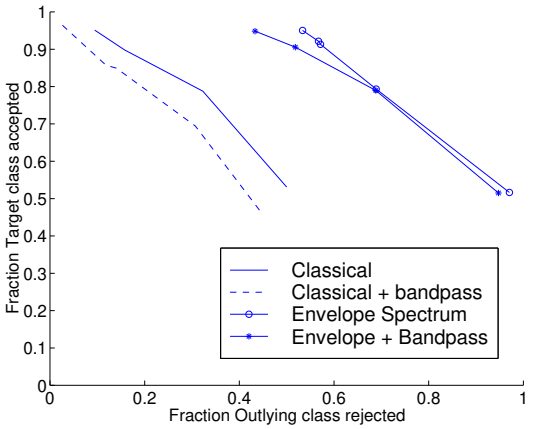

Figure 5: Acceptance/rejection performance of the SVDD on the Classical features and the Envelope spectrum on sensor combination (3).

In figure 5 the envelope spectrum feature set is compared with the classical features. Both the envelope spectrum and the bandpass filtered envelope spectrum features clearly outperform the classical method. Differences between the bandpassed and the original envelope spectrum features is very small. 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.

Finally in figure 6 the results on the AR-model feature set and the MUSIC frequency estimation

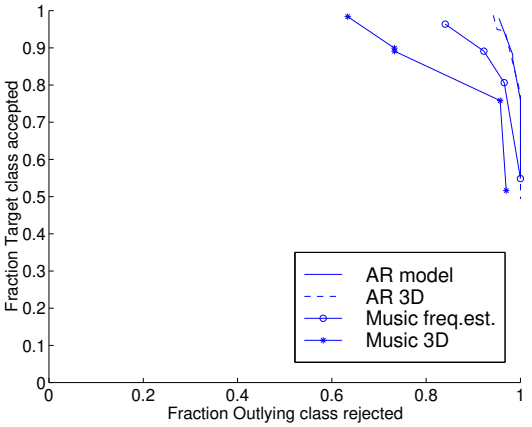

Figure 6: Acceptance/rejection performance of the SVDD on the AR-model and the MUSIC frequency estimation on sensor combination (3).

feature set are shown. The MUSIC estimator performs better than the already shown classical features, the envelope spectra with and without bandpass filtering. Taking the 3D PCA severely hurts the performance, especially for smaller target rejection rates. The AR model outperforms all other method, except for the 512 bin power spectrum. Even taking the 3D PCA does not deteriorate the performance. Only for very small rejection rates of the target class, we see some patterns from the outlier class being accepted.

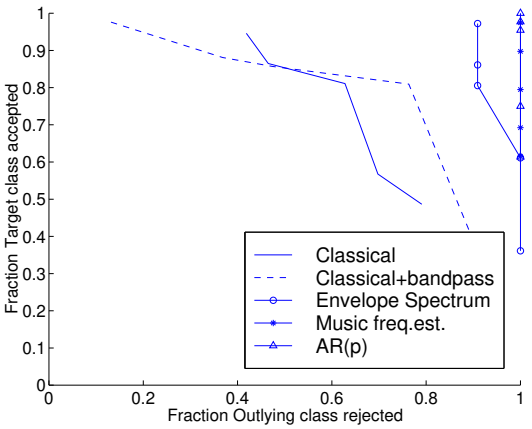

Figure 7: Acceptance/rejection performance of the SVDD on the different features for sensor combination (1).

This analysis was done on a data set in which all sensor information was used. We now look at the performance of the first and the second combination of sensors. In figure 7 the performance

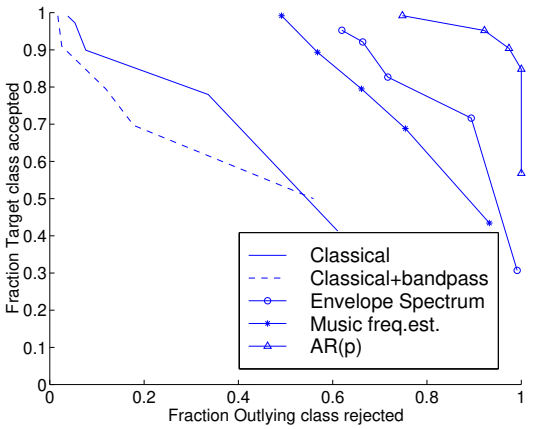

Figure 8: Acceptance/rejection performance of the SVDD on the different features for sensor combination (2).

of the SVDD is shown on all feature sets applied on sensor combination (1). 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).

We can observe the same trend in figure 8, where the performances are plotted for sensor combination (2). Here also the MUSIC estimator and the AR model outperform the other types of features, but here there are some errors. Total performance is worse than that of sensor combination (1) and (3).

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 should be distinguished from erroneous behavior.

From 7 sensors vibration data was recorded. Three subsets of the measurements of the 7 sensors were put together to create new data sets and several features are calculated from the recorded time signals. Although the three sensor combinations show somewhat different results, a clear trend is visible.

Performance of both MUSIC- and AR-features was usually very good in all three configuration datasets (see section 3). However, in comparison, the second configuration performed poorest and the third configuration performed best. 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 samples, 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 a certain variation in discrimination power of measurement 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 is probably due to 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 measuring runs.

The 64 dimensional envelope spectra with and without bandpass filtering perform better. Here a large fraction of the outlier class can be distinguished from the target class. Remarkable is that the bandpass filtering does not improve the description. Moreover, the normal power spectrum using 64 Fourier bins has comparable performance as the envelope spectrum.

Best performance is shown by the MUSIC frequency estimator and the AR-model. Even when just the first three principal components of the AR-model is used, the outlier class can be distinguished almost perfectly from the target class.

As a reference a very high resolution power spectrum (512 bins) is used. With this spectrum it is possible to perfectly distinguish between normal and abnormal situations, which means that in principle perfect classification is possible. If we want to use smaller representations in the vibration data analysis to overcome the curse of dimensionality, the AR-model is a good choice. Not only the AR-model gives the tightest description of the normal class when compared with other features, it is even comparable with the reference method when just 3 features are used.

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

[Bis95] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Walton Street, Oxford OX2 6DP, 1995.

[CGR91] G.A. Carpenter, S. Grossberg, and D.B. Rosen. ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition. *Neural Networks*, 4(4):493–504, 1991.

[DH73] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.

[DK82] P.A. Devijver and J. Kittler. *Pattern Recognition, A statistical approach*. Prentice-Hall International, London, 1982.

[Koh95] T. Kohonen. *Self-organizing maps*. Springer-Verlag, Heidelberg, Germany, 1995.

[MH96] M.M. Moya and D.R. Hush. Network contraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[PM92] J.G. Proakis and D.G. Manolakis. *Digital signal processing - principles, algorithms and applications, 2nd ed.* MacMillan Publ., New York, 1992.

[PNK94] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[TD] D.M.J. Tax and R.P.W Duin. Data domain description using support vectors. To appear in the special issue of Pattern Recognition Letters, proceedings of Pattern Recognition in Practice VI 1999.

[TD98] 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.

[TdRD97] D.M.J. Tax, D. de Ridder, and R.P.W. Duin. Support vector classifiers: a first look. In *Proceedings ASCI'97*. ASCI, 1997.

[Tow91] D. P. Townsend. *Dudley's gear handbook*. McGraw-Hill, Inc., 1991.

[Vap95] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995.

[YP99] A. Ypma and P. Pajunen. Rotating machine vibration analysis with second-order independent component analysis. In *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation, ICA'99*, pages 37 – 42, January 1999.