

A Variable Kernel Classifier for Bearing Fault Diagnosis using Simple Statistical Features

Barend J. van Wyk¹, Michaël A. van Wyk¹, Johannes J. Naudé², Frederique Perrier¹

¹French South-African Technical Institute in Electronics
at the

Tshwane University of Technology
Staatsartillerie Road, Pretoria, South Africa

² Denel Aerospace Systems, Centurion, South Africa

vanwykb@tut.ac.za vanwykmal@tut.ac.za hannes.naude@kentron.co.za

ABSTRACT

The Variable kernel Similarity Metric (VSM) learning procedure was generalized by Naudé and Van Wyk to produce the generalized variable kernel similarity metric (GVSM) learning algorithm [1]. In this paper we will outline the application of this data pre-processing technique to boost the classification of a nearest neighbours classifier on a bearing data classification task using only simple statistical features such as the variance and various central moments. It is shown that by preprocessing the bearing data using the GVSM technique the performance of a nearest neighbour classifier can be boosted to be comparable to that of a more complex neural network.

KEY WORDS

Metric learning, variable kernel similarity metric, generalized variable kernel similarity metric, bearing data classification.

1. Introduction

Vibration signal analysis plays an important role in the condition monitoring of bearings in rotating machinery. The presence of noise and the wide variety of possible faults complicate diagnostic procedures. Very often fault diagnosis rely on expert experience, statistical analysis, or the use of classical time and frequency domain analysis techniques.

During the past decade various wavelet and machine intelligence approaches were added to the arsenal of available diagnostics tools: Nikolaou and Antoniadis [6] introduced an effective demodulation

method based on the use of complex shifted Morlet wavelets, Chen and Mo [7] used wavelet transform techniques in combination with a function approximation approach to extract fault features which were used with a neural network, Lou and Loparo [8] introduced a scheme based on the wavelet transform and a neuro-fuzzy classification strategy, Altman and Mathew [9] used discrete wavelet packet analysis to enhance the detection and diagnostics of low-speed rolling element bearing faults, Zhang *et. al* [17] introduced an approach based on localized wavelet packet bases of vibration signals, Sun and Tang [11] applied the wavelet transform to detect abrupt changes in vibration signals, and Prabhakar *et. al* [12] also showed that the discrete wavelet transform can be used for improved detection of bearing faults.

Subrahmanyam and Sujatha [13] demonstrated that a multilayered feed forward network and an ART-2 network can be used for the automatic detection and diagnosis of localised ball bearing defects, Kowalski *et. al* [14] showed that Kohonen networks can be used as an introductory step before a neural detector for initial classification, Sporre [15] applied the cascade correlation algorithm to bearing fault classification problems, Gelle and Colas [16] used blind source separation as a pre-processing step to rotating machinery fault detection and diagnosis, Zhang *et. al* used a genetic programming approach and Samanta *et. al* used a support vector machine in conjunction with a genetic algorithm.

In this paper we derive a variable kernel classifier for bearing fault diagnosis using simple statistical

features derived from the unfiltered time domain data. The metric used to measure the distance between exemplars has a huge influence on the performance of the trained classifier and indicate the need for a fully automated metric learning procedure. One such procedure is the Variable kernel Similarity Metric (VSM) learning procedure introduced by Lowe in [2]. This procedure was generalized by the Naudé and Van Wyk to produce the generalized variable kernel similarity metric (GVSM) learning algorithm [1]. In this paper we will outline the application of this data pre-processing technique to boost the classification of a nearest neighbours classifier on a bearing data classification task using only simple non-filtered time domain features such as the variance and various central moments. It is shown that by preprocessing the bearing data using the GVSM technique the performance of a nearest neighbour classifier can be boosted to be comparable to that of a neural network similar to the one used by Samanta and Al-Balushi [19].

Notation and preliminaries are discussed in section 2 and section 3 outlines our nearest neighbours strategy. GVSM optimization is discussed in section 4 and section 5 contains a description of the dataset used and the features extracted. Simulation results are discussed in section 6 and section 7 concludes our presentation.

2. Preliminaries

For the rest of the paper s_{ti} will denote the known probability (i.e. 1 or 0) that sample number $t \in 1, 2, \dots, T$ falls in class $i \in 1, 2, \dots, I$ and p_{ti} will denote the estimated probability that sample number t falls in class i based on the training set excluding sample t . Similarly s_{tki} will denote the known probability that the k -th nearest neighbour of sample number t from the training set falls in class i . \bar{x}_t and \bar{c}_{tk} will denote the feature vectors of the t -th sample and it's k -th nearest neighbour respectively.

3. K -nearest neighbour classification

The K -nearest neighbour technique uses the following expression to approximate the probability that a sample belongs to class i

$$p_i = \frac{\sum_{k=1}^K n_k s_{ki}}{\sum_{k=1}^K n_k}.$$

In the most basic form of the method all n_k coefficients are set to 1 and it becomes a simple vote. A slightly more sophisticated method attaches more importance to closer neighbours by determining the weight n_{tk} assigned to each neighbour using a kernel centred at x_t . In this case we will use a Gaussian kernel:

$$n_k = e^{-\frac{d_k^2}{2\sigma^2}}.$$

The width of this kernel is determined by σ . If σ is too small, the truly nearest neighbour will dominate the decision and generalization will be poor. If it is too large, the method will fail to capture significant changes in the output. In general σ may be chosen smaller, the more densely the data is sampled. Since the density of data varies over the input space, fixed values of σ will not normally perform well. In order to make the width of the window vary with the density of available training samples, σ is set to some multiple of the average distance to the M nearest neighbours. It is better if only a fraction (e.g. $M = \frac{K}{2}$) of the nearest neighbours are used so the kernel becomes small even when only a few neighbours are close to the input,

$$\sigma = \frac{r}{M} \sum_{m=1}^M d_k.$$

The difference between VSM and GVSM lies in the definition of distance. While VSM uses the expression

$$d_k^2 = \sum_{d=1}^D w_d^2 (x_d - c_{kd})^2$$

to define the distance between a sample x and it's k -th nearest neighbour c_k , GVSM uses the more general matrix norm

$$d_k^2 = (\bar{x} - \bar{c}_k)^T A (\bar{x} - \bar{c}_k) \quad (1)$$

where A is a positive definite symmetric matrix. For the case where A is a diagonal matrix, this reduces to VSM.

4. GVSM optimization

The first complication that arises when attempting to numerically optimize a matrix norm is the fact that the matrix must be constrained to be symmetric

and positive definite.

A necessary condition for a matrix A to be a symmetric positive definite matrix, is that it can be expressed as $A = L^T L$ where L is upper triangular with positive diagonal elements. A sufficient condition for A to be symmetric positive definite, is that it can be written as $A = L^T L$ where L can be any non-singular matrix. Therefore if L is a non-singular upper triangular matrix the factorization

$$A = L^T L, \quad (2)$$

is a necessary and sufficient condition for A to be positive definite.

Expressing L as

$$L = \begin{bmatrix} L_{11} & L_{12} & L_{13} & \dots & L_{1d} \\ 0 & L_{22} & L_{23} & \dots & L_{2d} \\ 0 & 0 & L_{33} & \dots & L_{3d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & L_{dd} \end{bmatrix}$$

allows us to optimize the various elements L_{uv} and yet assure that A will be a symmetric positive definite matrix.

The cross validation error is defined as

$$E = \sum_t \sum_i (s_{ti} - p_{ti})^2.$$

Obtaining the optimal elements L_{uv} via gradient descent requires

$$\frac{\partial E}{\partial L_{uv}} = -2 \sum_t \sum_i (s_{ti} - p_{ti}) \frac{\partial p_{ti}}{\partial L_{uv}}$$

where

$$\frac{\partial p_{ti}}{\partial L_{uv}} = \frac{\sum_k (s_{tki} - p_{ti}) \partial n_{tk} / \partial L_{uv}}{\sum_k n_{tk}}$$

and

$$\frac{\partial n_{tk}}{\partial L_{uv}} = \frac{n_{tk}}{2\sigma^2} \left(\frac{d_{tk}^2 r}{M\sigma} \sum_{m=1}^M \frac{1}{d_{tm}} \frac{\partial d_{tm}^2}{\partial L_{uv}} - \frac{\partial d_{tk}^2}{\partial L_{uv}} \right), \quad (3)$$

$$\frac{\partial d_{tk}^2}{\partial L_{uv}} = \sum_o \sum_p (x_{to} - c_{tko})(x_{tp} - c_{tkp}) \frac{\partial A_{op}}{\partial L_{uv}},$$

$$\frac{\partial A_{op}}{\partial L_{uv}} = \begin{cases} 2L_{uv} & \text{if } o = p = v \\ L_{uo} & \text{if } o \neq p = v \\ L_{up} & \text{if } p \neq o = v \\ 0 & \text{if } p \neq o \neq v \end{cases}.$$

In order to optimize the parameter r we simply use the derivative of n_{tk} with respect to r

$$\frac{\partial n_{tk}}{\partial r} = \frac{n_{tk} d_{tk}^2}{r\sigma^2}$$

in place of equation 3. For a d -dimensional input space GVSM optimizes $\frac{d(d+1)}{2} + 1$ parameters as opposed to the $d+1$ parameters optimized by VSM, which implies more power to select an appropriate metric, but also more potential for overfitting the data set. Thus VSM would be more appropriate if the size of the data set is small relative to the dimensionality of the input. On the other hand GVSM optimizes very few parameters compared to an equivalent neural net, which seems to indicate that the overtraining problem should not be excessive.

5. Dataset Description and Feature Extraction

The data used for our work are measurements from accelerometers on a submersible pump driven by an electric motor acquired by the *Delft Machine Diagnostics by Neural Networks* project with help from Landustrie B.V, The Netherlands and can be downloaded freely at <http://www.ph.tn.tudelft.nl/ypma/mechanical.html>.

Refer to [5] for more details regarding data acquisition and experimental setup. Separate measurements were obtained for a normal bearing and a bearing with an outer race defect at the upper end. The sensors were placed at five different positions and sampled at 51.2 kHz while the pump was rotating at 1123 rpm. A total of 20480 samples were recorded for each channel.

For our work the signals were divided into 590 overlapping bins of 1024 samples. Each bin was processed to extract the following five features as suggested in [19]: root mean square (rms), variance ($\bar{\sigma}^2$), normalized third central moment (γ_3), normalized fourth central moment (γ_4) and the normalized sixth central moment (γ_6). Since there is a normal and a faulty recording available for each sensor there were 1180 feature vectors available per sensor for testing and training. The features are given by

$$rms = \sqrt{\frac{\sum \bar{y}_i^2}{n}} \quad (4)$$

$$\bar{\sigma}^2 = E\{\bar{y}_i^2\} \quad (5)$$

$$\gamma_3 = \frac{E\{\bar{y}_i^3\}}{\sigma^3} \quad (6)$$

$$\gamma_4 = \frac{E\{\bar{y}_i^4\}}{\sigma^4} \quad (7)$$

and

$$\gamma_6 = \frac{E\{\bar{y}_i^6\}}{\sigma^6} \quad (8)$$

where $\bar{y}_i = y_i - \mu$ and $\mu = E\{y_i\}$.

6. Simulation Results

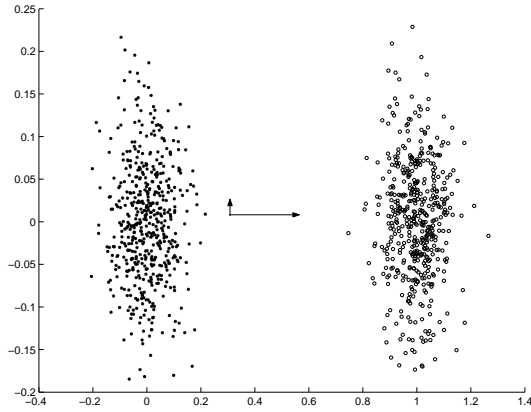


Figure 1: VSM learns a metric that accentuates differences along the x axis and suppresses those along the y axis. The arrows represent the kind of metric that VSM will typically learn for this case. (Classes are widely separated for illustrative purposes.)

While VSM learning performs well for cases where the input features are uncorrelated (such as depicted in figure 1) simple feature scaling is not powerful enough once the noise added to different features becomes cross-correlated (as depicted in figure 2). From figure 3 it is clear that fourth and sixth central moments from sensor 1 are ideal candidates for GVSM learning as for example opposed to the variance and the third central moment from sensor 5 depicted in figure 4. Cross correlation are observed for all fourth and sixth central moments from all sensors except for sensor 4. Visually the class separation between the fourth and sixth central

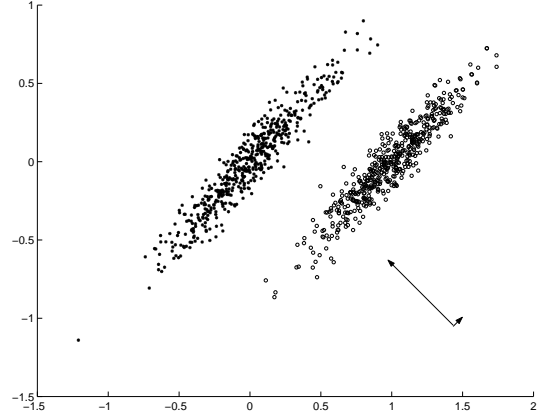


Figure 2: VSM is unable to learn an appropriate distance measure for cases where the noise affecting different features is correlated. The arrows represent the kind of metric that GVSM will typically learn for this case.

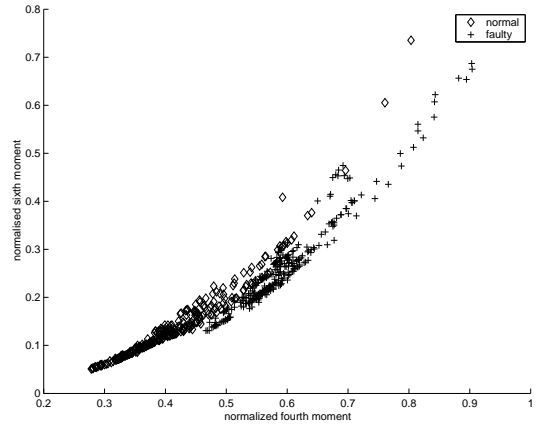


Figure 3: Plot of the fourth and sixth central moments from sensor 1.

moments are more distinct than between any other feature combination for the given dataset.

Figure 5 depicts the training set given in figure 3 after GVSM preprocessing. For this particular example the gradient descent procedure was performed with a stepsize of 0.1, $\sigma = 50$ and $K = 30$.

Table 2 shows the results obtained using the nearest neighbours procedure described in section 3 with $\sigma = 10$ and $K = 10$ where the input data was pre-processed as described in section 4. It was found that improved results are obtained for this particular dataset if the gradient descent procedure is performed with values for σ and K bigger than used for validation and testing. To obtain the A matrix gradient descent was performed with $K = 30$. Al-

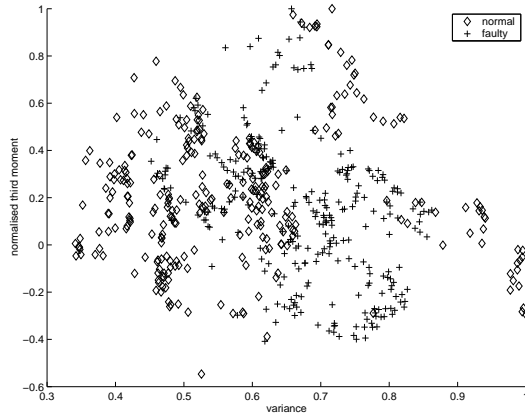


Figure 4: Plot of the variance and third central moment from sensor 5.

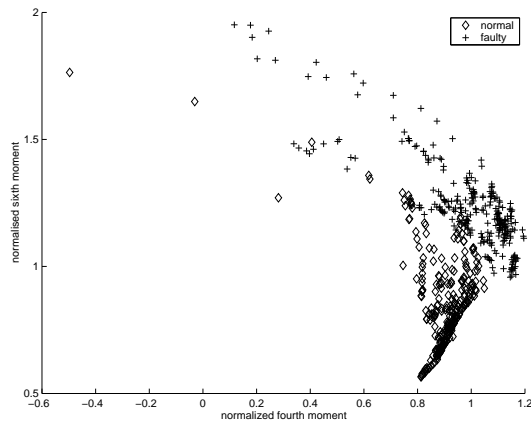


Figure 5: Plot of the fourth and sixth central moments from sensor 1 after GVSM pre-processing.

though VSM learning uses conjugate-gradient descent to minimize the cross validation error over the training set we have only implemented a gradient descent method with a primitive form of line search for use with GVSM. However convergence is still attained within a reasonable time for most problems.

For each sensor we had 590 feature vectors available for training 590 feature vectors available for testing. Due to reasons given earlier only the fourth and sixth central moments were used for GVSM nearest neighbour classification. From table 2 it is clear that the average training success is more than 90 percent and the average test success more than 85 percent if the results from sensor four are excluded. The best test success of 92 percent was obtained for sensor 5. It will be seen in the rest of this section that the results obtained using **only** the fourth and sixth central moments are comparable to the results

obtained by a feedforward neural network using the features proposed by Samanta and Al-Balushi [19].

| Sensor Signal | Training Success | Test Success |
|---------------|------------------|--------------|
| 1 | 97.5 | 84.5 |
| 2 | 91.0 | 86.5 |
| 3 | 90.0 | 86.5 |
| 4 | 74.5 | 63.0 |
| 5 | 88.6 | 92.0 |

Table 1: Percentage test success with GVSM pre-processing. A total of 590 training and test samples were processed for each sensor.

| Sensor | Features | ANN Settings | Test Success |
|--------|----------|--------------|--------------|
| 1 | 1 2 4 5 | 30/10 | 79 |
| 1 | 1 2 4 5 | 16/10 | 80 |
| 1 | 2 3 4 5 | 16/10 | 86 |
| 2 | 1 2 4 5 | 16/10 | 94 |
| 2 | 2 3 4 5 | 16/10 | 86 |
| 3 | 1 2 4 5 | 16/10 | 90 |
| 3 | 2 3 4 5 | 16/10 | 89 |
| 4 | 2 3 4 5 | 16/10 | 71 |
| 5 | 2 3 4 5 | 16/10 | 86 |
| 5 | 1 2 4 5 | 16/10 | 83 |
| 5 | 1 3 4 5 | 20/10 | 75 |

Table 2: Percentage ANN test success. A total of 590 test samples were processed for each sensor.

An effort to reproduce the results reported in Samanta and Al-Balushi [19] (where the same data set was used) was unsuccessful. In [19] the data was divided into 20 bins of 1024 non-overlapping samples each. This implies that they had at their disposal only 20 feature vectors per sensor for training and testing. A test success of 100 percent was reported when using features 2,3,4 and 5 from sensors two, three and four for the case where the sensor data was not filtered.

Neural network results obtained using 590 feature vectors for training 590 feature vectors for testing (as described in section 5) are reported in table 2. For our neural network experiment all sensors were treated separately. Feature combinations were chosen based on the arguments in [19], visual inspection and class separability measures described in [20]. The feedforward network with two hidden layers was implemented using the MATLAB neural network toolbox. The number of neurons in the first

hidden layer was varied between 10 and 30 and the number of neurons in the second hidden layer was varied between 5 and 10. As can be seen from table 2 the best test success of 94 percent was obtained for sensor 2.

7. Conclusion

This paper introduced a variable kernel classifier for bearing fault diagnosis using simple statistical features derived from unfiltered time domain data. It is shown that by preprocessing the bearing data using the GVSM technique the performance of a nearest neighbour classifier can be boosted to be comparable to that of a more complex neural network. It is expected that performance can be improved by preprocessing the time domain sensor data using band-pass filtering or wavelet processing.

8. References

- [1] Naudé, J.J., van Wyk, M.A., Generalized Similarity Metric Learning for a Variable-Kernel Classifier., Proceedings of the Thirteenth annual symposium of the Pattern Recognition Association of South Africa, November 2002.
- [2] Lowe, D. G., Similarity Metric Learning for a Variable-Kernel Classifier, *Neural Computation*, 7(1), 72-85, 1995.
- [3] Cover, T.M., and Hart, P.E., Nearest neighbour pattern classification., *IEEE Transactions on Information Theory*, IT13,1,21-27, 1967.
- [4] Nayar S. K., Nene S. A., Murase H., Subspace methods for robot vision, *IEEE Trans. on Robotics and Automation*, 12(5), 750-758, 1996
- [5] Ligteringen, R., Duin, R.P.W., Fritman, E.E.E., Ypma, A. - Machine diagnostics by neural networks: experimental setup, Proceedings of ASCI97, Heijen (The Netherlands), June 2 - 4, 1997
- [6] Nikolaou N.G., and Antoniadis I.A., Demodulation of vibration signals generated by defects in rolling element bearings using complex shifted Morlet wavelets, *Mechanical Systems and Signal Processing*, 16(4), 677-694, 2002.
- [7] Chen C. and Mo C., A method for intelligent fault diagnosis of rotating machinery, *Digital Signal Processing*, 14, 203-217, 2004.
- [8] Lou X. and Loparo K.A., Bearing fault diagnosis based on wavelet transform and fuzzy inference, *Mechanical Systems and Signal Processing*, 18, 1077-1095, 2004.
- [9] Altman J. and Mathew J., Multiple band-pass autoregressive demodulation for rolling-element bearing fault diagnosis, *Mechanical Systems and Signal Processing*, 15(5), 963-977, 2001.
- [10] Zhang S., Matyhew J., Ma L., Sun Y., Best basis-based intelligent machine fault diagnostics, *Mechanical Systems and Signal Processing*, 19, 357-370, 2005.
- [11] Sun Q. and Tang Y., Singularity analysis using continuous wavelet transform for bearing fault diagnosis, *Mechanical Systems and Signal Processing*, 16(6), 1025-1041, 2002.
- [12] Prabhakar S., Mohanty A.R., Sekhar A.S., Application of discrete wavelet transform for detection of ball bearing race faults, *Tribology International*, 35, 793-800, 2002.
- [13] Subrahmanyam M. and Sujatha C., Using neural networks for the diagnosis of localized defects in ball bearings, *Tribology International*, 30(10), 739-752, 1997.
- [14] Kowalski C.T. and Orłowska-Kowalska T., Neural networks application for induction motor fault diagnosis, *Mathematics and Computers in Simulation*, 63, 435-448, 2003.
- [15] Spoerre J.K., Application of the cascade correlation algorithm (CCA) to bearing fault classification problems, *Computers in Industry*, 32, 295-304, 1997.
- [16] Gelle G. and Colas M., Blind source separation: A tool for rotating machine monitoring by vibration analysis, *Journal of Sound and Vibration*, 248(5), 865-885, 2001.
- [17] Zhang L., Jack. L.B., and Nandi A.K., Fault detection using genetic programming, *Mechanical Systems and Signal Processing*, 19, 271-289, 2005.

- [18] Samanta B., Al-Balushi K.R. and Al-Araimi S.A., Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection, *Engineering Applications of Artificial Intelligence*, 16, 657-665, 2003.
- [19] Samanta B., and Al-Balushi K.R., Artificial neural network based fault diagnostics of rolling element bearings using time-domain features.
- [20] Theodoridis S. and Koutroumbas K., *Pattern Recognition*, Academic Press, London, 1999.