References

Alsing, S. G., Bauer Jr., K. W. and Miller, J.O., (2002) A multinomial selection procedure for evaluating pattern recognition algorithms, Pattern Recognition 35, pp 2397 – 2412.

Aronszajn, N., (1950) Theory of reproducing kernels, Trans. Amer. Math. Soc. 686, pp 337- 404.

Augustyn, G. L., Batko, W. and Wierzbicki, J., (2003) Contex Filtering, Tenth International Congress on Sound and Vibration 7-10 July. Stockholm, Sweden, pp 4213-4219.

Baydar, N. and Ball, A., (2000) Detection of gear deterioration under varying load conditions by using the Instantaneous Power spectrum, Mechanical Systems and Signal Processing 14(6), pp 907-921.

Baydar, N. and Ball, A., (2001) A comparative study of acoustic and vibration signals in detection of gear failure using Wigner-Ville Distribution, Mechanical Systems and Signal Processing 15(6), pp 1091-1107.

Bishop, C. M. and Nabney., I.T., (1996) Netlab neural network software, http:/www.ncrg.aston.ac.uk/netlab.

Bishop, C.M., (1995) Neural networks for pattern recognition, Oxford University Press, Oxford, UK.

Braun S. and Seth, B., (1980) Analysis of repetitive mechanism signatures, Journal of Sound and Vibration 70(4), pp 513-526.

Braun, S., (1975) The extraction of periodic waveforms by time domain averaging, Acustica 32, pp 69-77.

Bringham, E.O., (1974) The Fast Fourier Transform. Englewood Cliffs, NJ: Prentice-Hall.

Broomhead, D.S. and Lowe, D., (1988) Multivariable functional interpolation and adaptive networks, Complex Systems 2, pp 321-355.

Burges, C. J. C., (1998) A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery (2), pp 121-167.

Davel, J. G., (2003) Correlation between vibration levels and expected life of cylindrical gears, BEng (Mech). Final year project, University of Pretoria.

Decker, H.J., (2002)[a] Crack detection for aerospace quality spur gears, *NASA/TM*, 2002-211492.

Decker, H.J., (2002)[b] Gear crack detection using tooth analysis, *NASA/TM*, 2002-211491,

Dempsey, P.J. and Afjeh, A.A., (2002) Integrating oil debris and vibration gear damage detection technologies using fuzzy logic, *NASA/TM*, 2002-211126.

Dempsey, P.J., Handschuh, R.F. and Afjeh, A.A., (2002) Spiral bevel gear damage detection using decision fusion analysis, *NASA/TM*, 2002-211814.

Fidêncio, P. H., Poppi, R.J. and de Andrade, J. C., (2002) Determination of organic matter in soils using radial basis function networks and near infrared spectroscopy, Analytica Chimica Acta 453, pp125-134.

Gardner, M.W. and Dorling, S.R., (1999) Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London Atmospheric Environment 33, pp 709-719.

Gaudart, J., Giusiano, B. and Huiart, L., (2002) Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data, Computational Statistics & Data Analysis.

Geman, S., Bienenstock, E., and Doursat, R., (1992) Neural networks and the bias/variance dilemma, Neural Computation 4, pp 1-58.

Gunn, S. R., (1998) Support vector machines for classification and regression, Technical Report, Department of Electronics and Computer Science, University of Southampton.

Gunn, S. R., Brown, M. and Bossley, K. M., (1997). Network performance assessment for neurofuzzy data modelling. In Liu, X., Cohen, P., and Berthold, M., editors, Intelligent Data Analysis, volume 1208 of Lecture Notes in Computer Science, pp 313-323.

Gunn, S., (1998) Matlab SVM Toolbox, http:/www.isis.ecs.soton.ac.uk/ resources/ svminfo

Haykin, S., (1999) Neural networks, 2nd edition, Prentice-Hall Inc, New Jersey, USA.

Heyns, P.S., (2002) Mechanical vibrations measurement and analysis, MEV 732 course notes, University of Pretoria.

Hinton, G.E., (1987) Learning translation invariant recognition in massively parallel networks, In J.W. de Bakker, A.J. Nijman, and P.C. Treleaven (Eds.), Proceedings PARLE Conference on Parallel Architectures and Languages Europe, pp. 1-13. Berlin: Springer-Verlag.

Hopfield, J.J., (1987) Learning algorithms and probability distributions in feed-forward and feed-back networks, Proceedings of the National Academy of Science, Vol. 84, 8429-8433.

MacKay D. J. C., (1994) Bayesian non-linear modelling for the energy prediction competition, ASHRAE Transactions 100(2), pp 1053-1062.

Marwala, T., (2001) Fault identification using neural networks and vibration data, Ph.D. Thesis, University of Cambridge.

McFadden, P.D. and Smith, J.D., (1985) A signal processing technique for detecting local defects in a gear from the signal average of the vibration, Proceeding of the Institute of Mechanical Engineers 199(C4), pp 287-292.

McFadden, P.D., (1986) Detecting fatigue cracks in gears by amplitude and phase nodulation of the meshing vibration, American Society of Mechanical Engineers, Journal of Vibration Acoustics Stress and Reliability in design 199(2), pp 165-170.

McFadden, P.D., (1987) A revised model for the extraction of periodic waveforms by time domain averaging, Mechanical Systems and Signal Processing 1, pp 83-95.

McFadden, P.D., (1987) Examination of a technique for the early detection of failure in gears by signal processing of the time domain average of the meshing vibration, Mechanical Systems and Signal Processing 1, pp 173-183.

McFadden, P.D., (1989), Interpolation techniques time domain averaging of gear vibration, Mechanical Systems and Signal processing 3(1), pp 87-97.

McFadden, P.D., Cook, J.G. and Forster, L.M, (1999) Decomposition of gear vibration signals by the Generalised S Transforms, Mechanical Systems and Signal Processing 13(5), pp 691-707.

Mdlazi, L., Marwala, T., Stander, C., Scheffer, C. and Heyns P.S. (2003) Principal component analysis and Automatic relevance determination for damage identification in structure, Proceedings of the 21st International Modal Analysis Conference, San Antonio, pp 37-42.

Moczulski, W., (1987) The digital synchronous filtering technique, Mechanical Systems and Signal Processing 1, pp 197-210.

Møller, M., (1993) A scaled conjugate gradient algorithm for fast supervised learning, Neural Networks, Vol 6, pp 525-533.

Moody, J. and Darken, C.J., (1989) Fast learning of networks of locally-tuned processing units, Neural Computation 1 (2), pp 281-294.

Neal, R. M., (1996) Bayesian Learning for Neural Networks. New York, NY: Springer Verlag.

Neal, R. M., (1998) Assessing relevance determination methods using DELVE, Neural Networks and Machine Learning. New York, NY: Springer-Verlag.

Norton, M. P., (1989) Fundamentals of noise and vibration analysis for engineers, New York: Cambridge University Press.

Paya, B.A., Esat, I.I. and Badi, M.N.M., (1997) Artificial Neural Networks based fault diagnostics of rotating machinery using Wavelet Transforms as a pre-processor, Mechanical Systems and Signal Processing 11(5), pp 751-765.

Powell, M. J. D., (1987) Radial basis functions for multivariable interpolation: a review. In. Mason, J. C. and Cox M. G., Algorithms for approximation, pp 143-167. Oxford: Clarendon Press.

Raath, A.D., (1992) Structural dynamic response reconstruction in the time domain, PhD thesis, Department of Mechanical and Aeronautical Engineering, University of Pretoria.

Ramesh, R., Mannan, M.A., Poo, A.N. and Keerthi, S.S., (2003) Thermal error measurement and modelling in machine tools. Part II. Hybrid Bayesian Network—support vector machine model, International Journal of Machine Tools & Manufacture.

Stander, C.J. and Heyns, P.S., (2001) Fault detection on gearboxes operating under fluctuating load conditions, Proceeding of the 14thInternational Congress on Condition Monitoring and Diagnostic Engineering Management Manchester UK 4-6 September, pp 457-464.

Stander, C.J. and Heyns, P.S., (2002)[a] Instantaneous Shaft Speed monitoring of gearboxes under fluctuating load conditions, Proceeding of the 15th International Congress on Condition Monitoring and Diagnostic Engineering Management, Birmingham UK 2-4 September 2002, pp 220-230.

Stander, C.J., Heyns, P.S. and Schoombie, W., (2002)[b] Using vibration monitoring for local fault detection on gears operating under fluctuating load conditions, Mechanical Systems and Signal Processing 16(6), pp 1005-1024.

Stander, C.J. and Heyns, P.S., (2003) Condition monitoring of gearboxes under cyclic and non-cyclic loading conditions, Proceeding of the 16th International Congress on Condition Monitoring and Diagnostic Engineering Management, Sweden 27-29 August, pp 601-610.

Staszewski, W.J. and Tomlinson, G.R., (1994) Application of the Wavelet Transform to fault detection in a spur gear, Mechanical Systems and Signal Processing 8(3), pp 289-307.

Staszewski, W.J., Worden, K. and Tomlinson, G.R., (1997) Time-Frequency analysis in gear fault detection using the Wigner-Ville and Pattern recognition, Mechanical Systems and Signal Processing 11(5), pp 673-692.

Stewart, R.M., (1977) Some useful data analysis techniques for gear diagnostics, Institute of Sound and Vibration Research, Paper MHM/R/10/77.

Taurino A.M., Distante, C., Siciliano, P. and Vasanelli, L., (2003) Quantitative and qualitative analysis of VOCs mixtures by means of a micro sensors array and different evaluation methods, Sensors and Actuators B 93, pp 117-125.

Trimble C.R., (1968) What is signal averaging?, Hewlett-Packard Journal 19(8), pp. 2-7.

Vapnik, V., Golowich, S. and Smola, A., (1997) Support vector method for function approximation, regression estimation, and signal processing. In Mozer, M., Jordan, M. and Petsche, T., editors, Advances in Neural Information Processing Systems 9, pp 281-287, Cambridge, MIT Press.

Vapnik, V.N., (1995) The nature of statistical learning theory, Springer-Verlag, New York, USA.

Walde, J.F., Tappeiner, G., Tappeiner, U., Tasser, E. and Holub, H.W., (2003) Statistical aspects of multilayer perceptrons under data limitations, Computational Statistics & Data Analysis.

Wang, W.J. and McFadden, P.D, (1993) Early detection of gear failure by vibration analysis-II. Interpretation of the time-frequency distribution using image processing techniques, Mechanical Systems and Signal Processing 7(3), pp 205-215.

Wang, W.J. and McFadden, P.D., (1995) Application of Orthogonal Wavelets to early gear damage detection, Mechanical Systems and Signal Processing 9(5), pp 497-507.

Wang, W. and Wong, A.K., (2000) Linear prediction and gear fault diagnosis, Proceeding of the 13th International Congress on Condition Monitoring and Diagnostic Engineering Management Houston Texas 3-8 December, pp 707-807.

White, G., (1991) Amplitude demodulation- a new tool for Predictive maintenance, Sound and Vibration, pp 14-19.

Yang, H., Chan, L. and King, I., (2002) Support Vector Machine Regression for volatile stock market prediction, IDEAL 2002, LNCS 2412, pp 391-396.

Zacksenhouse, M., Braun, S., Feldman, M. and Sidahmed, M., (2000) Towards helicopter diagnostics from a small number of examples, Mechanical Systems and Signal Processing 14(5), pp 523-543.

Zhong, M., Ding, S.X., Lam, J. and Wang, H., (2003) An LMI approach to design robust fault detection filter for uncertain LTI systems, Automatica (39), pp 543-550.

Appendix A

A.1 Experimental set-up

A schematic diagram of the accelerated gear life test rig used in this work is presented in Figure A.1. Figure A.2 shows a diagram of the accelerated gear life test rig.

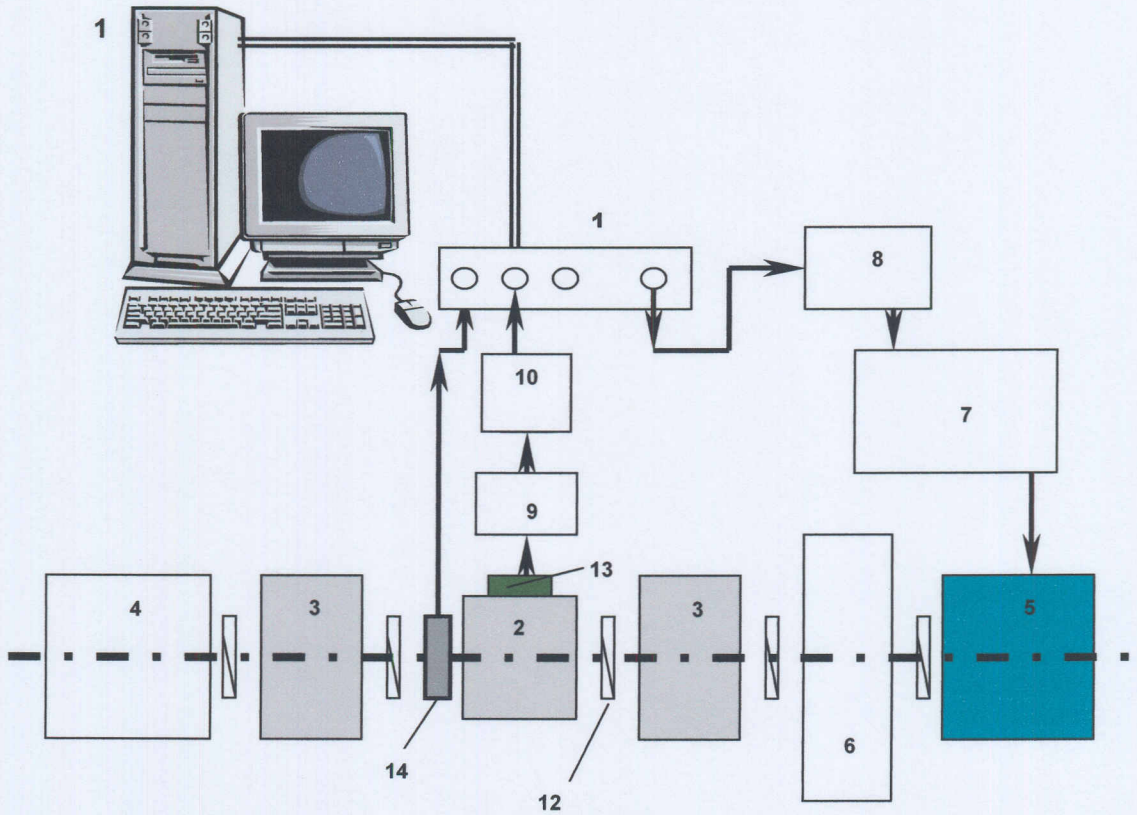

Figure A.1 Schematic diagram of experimental set-up

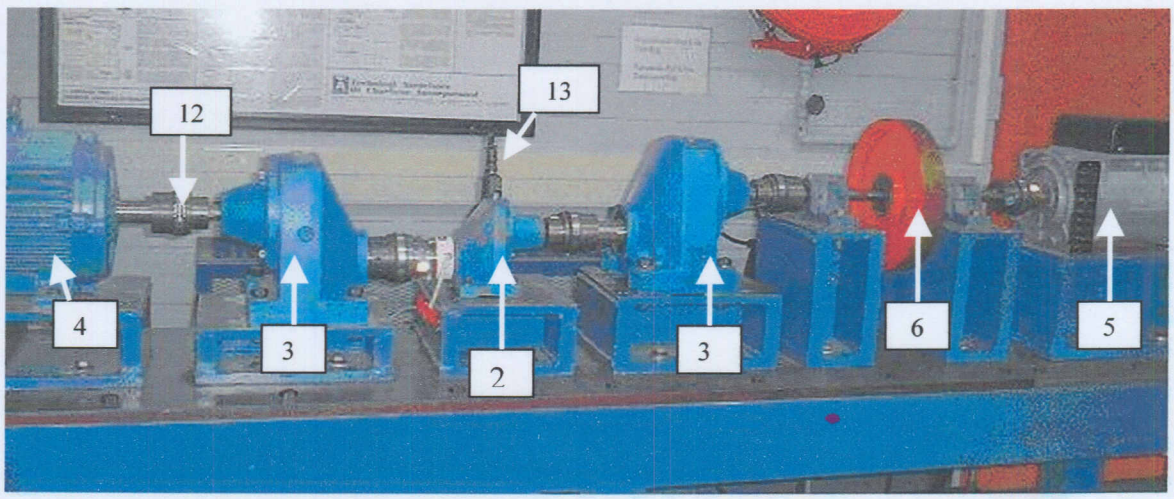

Figure A.2 Accelerated gear life test rig.

Table A1 give the specifications of the accelerated gear life test rig.

Table A.1 Accelerated life gear test rig Specifications

Item Number	Item	Description
1	PC	
2	Gearbox	Flender Himel Type E20A Ratio 1:96:1
3	Gearbox	Flender Himel Type E60A Ratio 4:72:1
4	Motor	WEG 380V / 50 Hz, three-phase
5	Alternator	Mecc alte 5.5 kVA, three-phase
6	Flywheel	Fenner 2517-25
7	Current controller	JEC current controller
8	DC Power supply	0-5V
9	Anti-aliasing low pass filter	4^{th} order low pass Butterworth filter with 300Hz cut-off
10	Signal conditioner	PCB ICP Model 482A22
11	Siglab analyser	Siglab model 20-42
12	Flexible couplings	
13	PCB accelerometer	5 V/g
14	Shaft encoder	Hengstler Himel type 0053 163 /10-30V DC /30mA

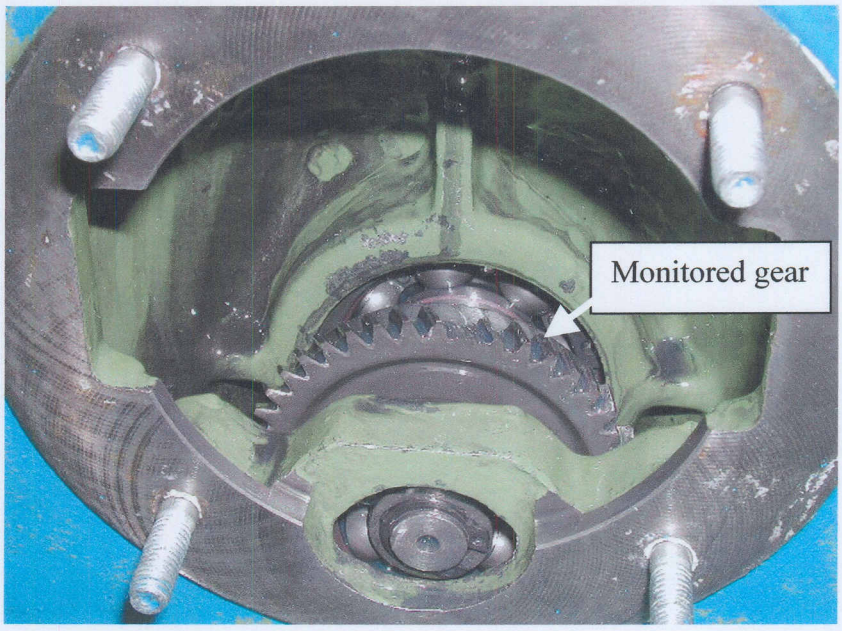

Figure A.3 Monitored gear inside the gearbox.

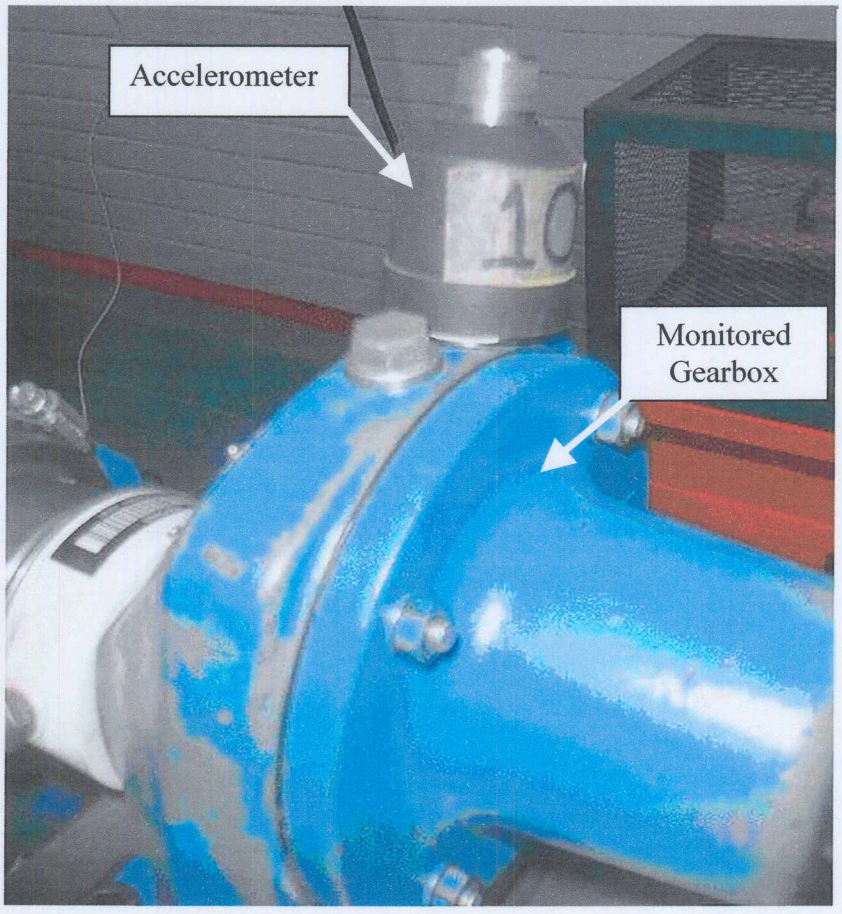

Figure A.4 Measurement point and mounting of accelerometer .

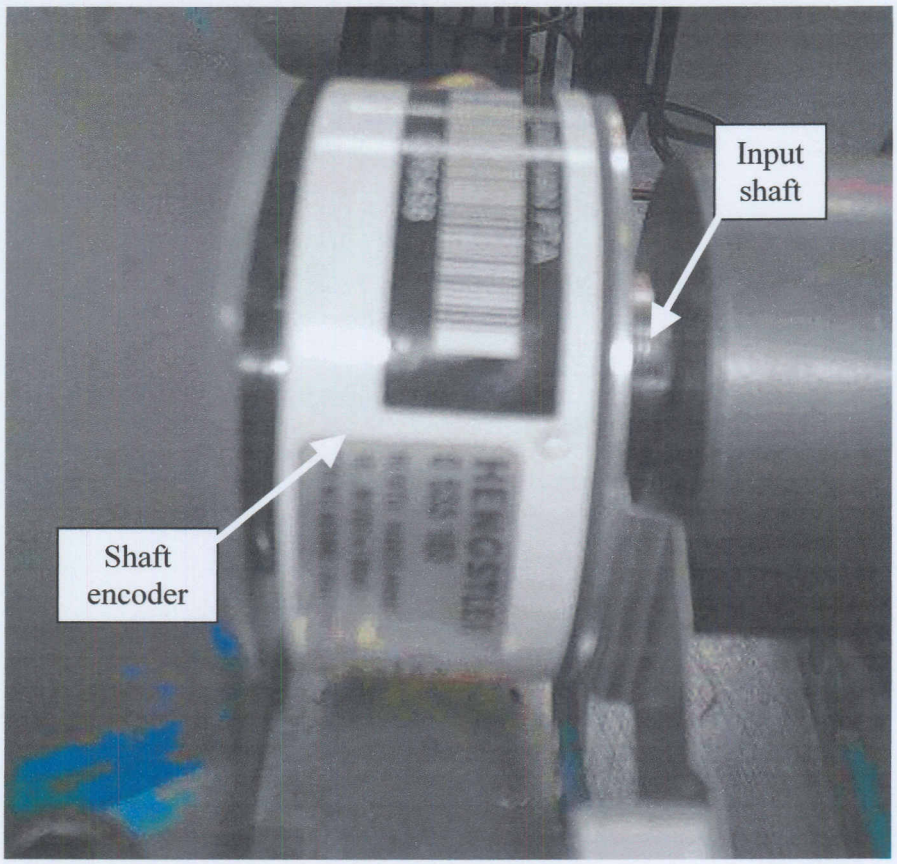

Figure A.5 Shaft encoder mounted on input to shaft to the monitored gearbox.

Appendix B

B.1 Back-Propagation method

In this study, the output is the time domain average of the rotation synchronised gearbox vibration data. In Figure 3.1 the output of the j^{th} hidden unit is obtained by calculating the weighted linear combination of the d input values.

$$a_j = \sum_{i=1}^{d} W_{ji}^{(1)} X_i \tag{B.1}$$

Where $W_{ji}^{(1)}$ indicates weights in the first layer, going from input i to hidden unit j while $W_{j0}^{(1)}$ indicates the bias for the j^{th} hidden unit. The activation of the j^{th} hidden unit is obtained by transforming the output a_j in equation (B.1) into z_j, which is shown in Figure 3.1, is

$$z_j = f_{inner}(a_j) \tag{B.2}$$

The output of the second layer is obtained by transforming the activation of the second hidden layer using the second layer weights. Given the output of the hidden layer z_j in equation (B.2), the output of unit k is given by

$$a_k = \sum_{j=0}^{M} W_{kj}^{(2)} y_j \tag{B.3}$$

Similarly equation (B.3) may be transformed into the output units by using some activation function as follows:

$$y_k = f_{outer}(a_k) \tag{B.4}$$

Combining equations (B.1), (B.2), (B.3) and (B.4) the input x to the output y can be related by a two-layered non-linear mathematical expression, which may be written as follows:

$$y_k = f_{outer}\left(\sum_{j=0}^{M} w_{kj}^{(2)} f_{inner}\left(\sum_{i=0}^{d} w_{ji}^{(1)}\right) + w_{k0}^{(2)}\right) \tag{B.5}$$

Where d is the number of input units, M is the number of hidden units, w_{ij} is the weight-vector, the function f_{outer} is linear and f_{inner} is a hyperbolic tangent function. These functions are defined as:

$$f_{outer}(v) = v \tag{B.6}$$

and

$$f_{inner}(v) = \tanh(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \tag{B.7}$$

The weights w_i and in the hidden layers are varied until the error between the network prediction and the output from the training data is minimised.

Given the training set $D = \{X_k, t_k\}_{k=1}^{N}$ and assuming that the targets t_k are sampled independently given the inputs x_k and the weight parameters w_{kj} the sum of square of error cost function E is given by

$$E = \frac{1}{2}\sum_{n}\sum_{k}(y_{nk} - t_{nk})^2 \tag{B.8}$$

Where n is the index for the training pattern and k is the index for the output units.

The minimisation of E is achieved by solving for the derivative of the error in equations (B.8) with respect to the weights. The derivative of the error is calculated with respect to the weights that connects the hidden layer to the output layer and may be written using the chain rule as follows:

$$\begin{aligned}\frac{\partial E}{\partial w_{kj}} &= \frac{\partial E}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}} \\ &= \frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}} \\ &= \sum_{n} f'_{outer}(a_k)\frac{\partial E}{\partial y_{nk}} z_j\end{aligned} \tag{B.9}$$

where z_j is given in equation (B.2). The derivative of the error with respect to the weights which connects the hidden layer to the output layer may be written using the chain rule is given by

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}}$$
$$\frac{\partial E}{\partial w_{kj}} = \sum_n f'_{inner}(a_j)\sum_k w_{kj} f'_{outer}(a_k)\frac{\partial E}{\partial y_{nk}} \tag{B.10}$$

The derivative of the sum of square cost function in equation (B.8) is written as

$$\frac{\partial E}{\partial y_{nk}} = t_{nk} - y_{nk} \tag{B.11}$$

The derivatives of the linear activation function in equation (B.6) is:

$$f'_{outer}(a_k) = c \tag{B.12}$$

while the derivative of the hyperbolic tangent function is:

$$f'_{inner}(a_j) = \text{sech}^2(a_j) \tag{B.13}$$

This appendix shows the derivatives of the errors with respect to weights. Equation (B.11) shows the derivative of the cost functions that could be incorporated into equations (B.9) and (B.10). Equations (B.12) and (B.13) show the derivatives of the two possible activation functions.

Appendix C

C.1 Gradient method

In this appendix the scaled conjugate optimisation method is described. Before introducing the scaled conjugate gradient method the conjugate gradient method is introduced. In supervised neural network training, the main goal is to identify weights that give the best prediction of the output whenever presented with the input. The scaled conjugate gradient method is used to sample through the weight space until the weight vector that minimises the distance between the neural network prediction and the target data is obtained.

C.2 Conjugate gradient method

The weight vector that gives the minimum error is obtained by taking successive steps through the weight space as follows:

$$w(n+1) = w(n) + \Delta w(n) \tag{C.1}$$

where n is the iteration step and Δ represents change. Different algorithms choose this step size differently. In this section, gradient descent method will be discussed, followed by how it is extended to the conjugate gradient method. For the gradient descent method, the step size in equation (C.1) is defined as:

$$\Delta w^n = -\eta \nabla E(w(n)) \tag{C.2}$$

where the parameter η is the learning rate and the gradient of the error is calculated using the back-propagation technique described in Appendix B. If the learning rate is sufficiently small, the value of error will decrease at each successive step until a minimum is obtained. The disadvantage with this approach is that it is computationally expensive compared to other techniques.

For the conjugate gradient method the quadratic function of error is minimised at each iteration over a progressively expanding linear vector space that includes the global minimum of the error. For the conjugate gradient procedure, the following steps are followed (Haykin, 1999; Marwala, 2001):

- Choose the initial weight $w(0)$.
- Calculate the gradient vector $\nabla E(w(0))$.

At each step n use the line search to find $\eta(n)$ that minimises $E(\eta)$ representing the cost function expressed in terms of η for fixed values of w and $-\nabla E(w(0))$.

- Check that the Euclidean norm of the vector $-\nabla E(w(n))$ is sufficiently less than that of $-\nabla E(w(0))$.
- Update the weight vector $w(n+1) = w(n) - \eta(n)\nabla E(w(n))$. For $w(n+1)$ compute the updated gradient $\nabla E(w(n+1))$.
- Use Polak-Ribiére method to calculate $\beta(n+1)$

$$\beta(n+1) = \frac{\nabla E(w(n+1))^T (\nabla E(w(n+1)) - \nabla E(w(N)))}{\nabla E(w(n))^T \nabla E(w(n))} \tag{C.3}$$

- Update the direction vector $\nabla E(w(n+1)) = \nabla E(w(n+1)) - \beta(n+1)\nabla E(w(n))$.
- Set $n = (n+1)$ and go back to step 3.
- Stop when the condition $\|\nabla E(w(n))\| = \varepsilon \|\nabla E(w(0))\|$ is satisfied in which ε is a small number.

C.3 Scaled conjugate gradient method

The scaled conjugate gradient method differs from conjugate gradient method in that it does not involve the line search described in step 3 in the previous section.
The step-size (see step 3) is calculated by using the following formula (Møller, 1993):

$$\bar{\eta}(n) = 2\left(\eta(n) - \frac{\nabla E(n)^T H(n)\nabla E + \eta(n)\|\nabla E(n)\|^2}{\|\nabla E(n)\|^2}\right) \tag{C.4}$$

where H is the Hessian matrix of the gradient.

Appendix D

D.1 Feature space

This appendix discusses the methods that can be used to construct a mapping into a high dimensional feature space by the use of reproducing kernels and implementation issues with regards to SVM. The idea of kernel functions is to enable operations to be performed in the input space rather than a potentially higher dimensional feature space. This provides a way of addressing the curse of dimensionality. The computation is however still critically dependent upon the number of training patterns and to provide a good data distribution for the high dimensional problem will generally require a large training set.

D.1 Kernel functions

The following theory is based on the Reproducing Kernel Hilbert Space (RKHS) (Aronszajn, 1950; Gunn , 1998). An inner product in the feature space has an equivalent kernel in input space,

$$K(x,x') = \langle \phi(x), \phi(x') \rangle, \tag{D.1}$$

provided certain conditions hold. If K is a symmetric positive definite function, which satisfies Mercer's condition,

$$K(x,x') = \sum_{m}^{\infty} a_m \phi_m(x) \phi_m(x'), \quad a_m \geq 0, \tag{D.2}$$

$$\iint K(x,x') g(x) g(x') dx dx' > 0, \quad g \in L_2, \tag{D.3}$$

Then the kernel represents a legitimate product in feature space. Valid functions satisfying the Mercer's conditions that were investigated in this study are given below. These functions are valid for all real x and x' unless otherwise stated.

D.2.1 Gaussian radial basis function

Radial basis functions have received significant attention, most commonly with a Gaussian of the form,

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right). \tag{D.4}$$

Classical techniques utilising radial basis functions employ some method of determining a subset of centres. Typically the method of clustering is employed to select the subset of centres. An attractive feature of the SVN is that this selection is implicit, with each support vector contributing one local Gaussian function, centred at that data point. By further consideration it is possible to select the global basis function width, s, using the SRM principle (Vapnik, 1995).

D.2.2 Exponential radial basis function

The form below defines an exponential radial basis function,

$$K(x,x') = \exp\left(-\frac{\|x-x'\|}{2\sigma^2}\right). \tag{D.5}$$

This form produces a piecewise solution that can be attractive when discontinuities are acceptable.

D.2.3 Splines

Splines are a popular choice for modelling because of their flexibility. A finite spline, of order κ, with N knots located at τ_s is given by,

$$K(x,x') = \sum_{r=0}^{\kappa} x^r x'^r + \sum_{s=1}^{N} (x-\tau_s)_+^{\kappa} (x'-\tau_s)_+^{\kappa}. \tag{D.6}$$

An infinite spline is defined on the interval [0, 1) by

$$K(x,x') = \sum_{r=0}^{\kappa} x^r x'^r + \int_0^1 (x-\tau_s)_+^{\kappa} (x'-\tau_s)_+^{\kappa} \, d\tau. \tag{D.7}$$

In the case where $\kappa = 1, \left(S_1^\infty\right)$, the kernel is defined by,

$$K(x,x') = 1 + \langle x,x' \rangle + \frac{1}{2}\langle x,x' \rangle \min(x,x') - \frac{1}{6}\min(x,x')^3, \tag{D.8}$$

where the solution is piecewise cubic.

D.2.4 B-splines

Bsplines are another popular spline formulation. The B-splines kernel is defined on the interval [-1, 1], by the attractive closed form

$$K(x,x') = B_{2N+1}(x - x'). \tag{D.9}$$

D.3 Loss functions

Using the quadratic loss function in Figure 3.6 (a),

$$L_{quad}(f(\mathrm{x}) - y) = (f(\mathrm{x}) - y)^2. \tag{D.10}$$

The solution is given by,

$$\begin{aligned} \max_{\alpha,\alpha^*} W(\alpha,\alpha^*) = \max_{\alpha,\alpha^*} & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j \rangle \\ & + \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)y_i - \frac{1}{2C}\sum_{i=1}^{l}\left(\alpha_i^2 + (\alpha_i^*)^2\right). \end{aligned} \tag{D.11}$$

The corresponding optimisation can be simplified by exploiting the KKT conditions, and noting that these imply $\beta_i^* = |\beta_i|$. The resultant optimisation problem is,

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_j\langle x_i, x_j \rangle - \sum_{i=1}^{l}\beta_i y_i + \frac{1}{2C}\sum_{i=1}^{l}\beta_i^2 \tag{D.12}$$

with constraints,

$$\sum_{i=1}^{l} \beta_i = 0. \tag{D.13}$$

For the Huber loss function in Figure 3.6 (c),

$$L_{Hubber}\left(f(\mathrm{x})-y\right)=\begin{cases}\frac{1}{2}\left(f(\mathrm{x})-y\right)^2 & \text{for} \quad |f(\mathrm{x})-y|<\mu \\ \mu|f(\mathrm{x})-y|-\frac{\mu^2}{2} & \text{Otherwise}\end{cases}, \tag{D.14}$$

the solution is given by,

$$\begin{aligned}\max_{\alpha,\alpha^*} W(\alpha,\alpha^*) = \max_{\alpha,\alpha^*} & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_i,x_j\rangle \\ & +\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i-\frac{1}{2C}\sum_{i=1}^{l}\left(\alpha_i^2+(\alpha_i^*)^2\right)\mu.\end{aligned} \tag{D.15}$$

The resultant optimisation is

$$\min_{\beta}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_j\langle x_i,x_j\rangle-\sum_{i=1}^{l}\beta_i y_i+\frac{1}{2C}\sum_{i=1}^{l}\beta_i^2\mu \tag{D.16}$$

with constraints,

$$\begin{aligned}&-C\le\beta_i\le C, \qquad i=1,\mathrm{K}\ ,l \\ &\sum_{i=1}^{l}\beta_i=0.\end{aligned} \tag{D.17}$$

D.4 Implementation issues

For SVM the resulting optimisation problems are dependent upon the number of training examples. For large data sets methods have been proposed for speeding up the algorithm by decomposing the problem into smaller ones. The optimisation problem for an ϵ-insensitive loss function can be expressed in matrix format as,

$$\min_{x}\frac{1}{2}x^T H x+c^T x \tag{E.18}$$

where

$$H = \begin{bmatrix} XX^T & -XX^T \\ -XX^T & XX^T \end{bmatrix}, \quad c = \begin{bmatrix} \varepsilon + Y \\ \varepsilon - Y \end{bmatrix}, \quad x = \begin{bmatrix} \alpha \\ \alpha^* \end{bmatrix} \tag{E.19}$$

with constraints

$$x \cdot (1,\mathrm{K},1,-1,\mathrm{K},-1) = 0, \quad \alpha_i, \alpha_i^* \geq 0, \quad i = 1,\mathrm{K},l. \tag{E.20}$$

where

$$X = \begin{bmatrix} \mathbf{x}_1 \\ \mathrm{M} \\ \mathbf{x}_l \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \mathrm{M} \\ y_l \end{bmatrix} \tag{E.21}$$