Wavelet Based Feature Extraction Methods for the Discrimination and Regression of Spectral Data

Thesis submitted by
Yvette Lelia MALLET Bsc(Hons) *Qld*
in October 1997

for the degree of Doctor of Philosophy
in the School of Computer Science, Mathematics and Physics
James Cook University of North Queensland

In Memory of Tes Everingham

Statement of Access

I, the undersigned, the author of this thesis, understand that James Cook University of North Queensland will make it available for use within the University Library and, by microfilm or other means, allow access to users in other approved libraries. All users consulting this thesis will have to sign the following statement:

> In consulting this thesis I agree not to copy or closely paraphrase it in whole or in part without the written consent of the author; and to make proper public written acknowledgement for any assistance which I have obtained from it.

Beyond this, I do not wish to place any restriction on access to this thesis.

.... 3/3/98.....................
(Date)

STATEMENT ON SOURCES
DECLARATION

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

.... 3/3/98 ....
(Date)

Acknowledgements

Foremostly, I thank my supervisors A/Prof Danny Coomans and A/Prof Olivier de Vel, for their professionalism and encouragement in the development of this thesis. I am also grateful for the many hours that they have spent revising this and other manuscripts throughout the course of my research work.

I extend my appreciation to Professor Massart for allowing me to visit his laboratory at the Free University of Brussels and providing me with the opportunity to become more familiar with NIR data. Also from the Free University, I thank Delphine Jouan-Rimbaud, Paula Fernandez, Eric Bouveresse, Wu Wen, Beate Walczak and Wim Penninckx, for their assistance.

Sincere thanks is expressed to Dr Jaroslav Kautsky at Flinders University in Adelaide, who gave up his valuable time to introduce me to wavelets. I would also like to thank Dr Bill Moran, Radka Turcajová and Pavel Turcaj for their assistance and cooperation whilst I was visiting Flinders University.

I extend my appreciation, to Professor Trevor Hastie for supplying his S-plus code and providing valuable input. Many people have provided data which I have used as part of my thesis, these people are acknowledged in Sections 7.2 and 8.2 of this thesis.

Thanks to my colleagues in the School of Computer Science, Mathematics and Physics at James Cook University for their encouragement during the final stages of this thesis. In particular to Dr Wayne Read for revising parts of this thesis. The encouragement provided by A/Prof Bob Staudte was also much appreciated.

Whilst pursuing the research summarized in this thesis, I was primarily supported by the Australian Government in the form of an Australian Postgraduate Research Award.

I wish to thank my fiancé, Andrew Everingham for his continual support and encouragement. I would also like to thank my team-mates for their patience and understanding. Finally, I thank my mother for always thinking of the little things that mean so much.

Abstract

This thesis is concerned with the application of statistical methods to spectral data. A major concern which arises from spectral data is that the number of variables or dimensionality usually exceeds the number of available spectra. This leads to a degradation in performance of traditional statistical methods. There are basically two strategies which can be implemented for overcoming such situations. It is common practice to first reduce the dimensionality of the data by some feature extraction preprocessing method, and then use an appropriate low dimensional statistical procedure. An alternative procedure is to use a high dimensional statistical procedure which is capable of handling a large number of variables. This thesis considers both approaches, and investigates the applicability of wavelets as features for statistical analyses, as well as other feature extraction procedures. The particular statistical analyses investigated are discriminant and regression analysis.

It is shown that, the wavelet based methods, particularly wavelets which have been designed to suit a particular task, perform quite adequately when compared to traditional approaches.

Contents

List of Figures

List of Tables

List of Symbols

Non-bold Lower Case Letters

- $a(\mathrm{x})$ appreciation score of x
- $a_{\mathrm{ccr}}(\mathrm{x})$ appreciation score equal to 1 if $P(r \mid \mathrm{x}_{i(r)}) \geq P(r \mid \mathrm{x}_i)$and zero otherwise
- $a_A(\mathrm{x})$ appreciation score equivalent to $P(r \mid \mathrm{x}_{i(r)})$
- $a_Q(\mathrm{x})$ quadratic appreciation score of x
- a_{il} lth element in the ith principal component vector
- a parameter used in RDA which weights the pooled covariance matrix
- b parameter used in RDA which controls shrinkage of the weighted pooled covariance matrix
- b_i ith element in the vector of estimated regression coefficients $\mathbf{b}$
- band(j, t) τth band $\tau \in \{0, 1, \ldots, m-1\}$ at the jth level $j \in \{J, J-1, \ldots, J - \mathrm{max}_{\mathrm{lev}} + 1\}$ of the DWT
- $c_{j,k}$ scaling coefficients
- $d_{j,k}$ wavelet coefficients
- f_{V_j} orthogonal projection of $f(t)$ onto V_j
- $g(\mathrm{x}, r)$ classification score
- $g_{\mathrm{blda}}(\mathrm{x}, r)$ BLDA classification score
- $g_{\mathrm{bqda}}(\mathrm{x}, r)$ BQDA classification score
- h_k high pass filter coefficients
- $\hbar_{ii}$ is the element along the ith diagonal of the hat matrix $\mathcal{H}$
- j_* complex number $\sqrt{-1}$

- j parameter controlling the dilation of the wavelet basis functions
- k parameter controlling the translation of the wavelet basis functions
- ℓ_k low pass filter coefficients
- m number of bands in the DWT; downsampling rate
- $\max_{lev}$ maximum number of levels in the DWT.
- n number of observational units in the training data set
- n' number of observational units in the testing data set
- n_r number of observational units from class r; rth element in the vector n
- $n_{[l]}$ number of objects in node l of CART model
- n_{levels} number of levels that an object has been transformed, in the DWT
- p dimensionality of the data set
- p_* dimensionality of the reduced data set $p_* \ll p$
- p_o number of parameters to be estimated (including the intercept) in a MLR model
- $p(\mathrm{x})$ is the class probability density of x
- q the number of sub-matrices in the filter coefficient matrix A is $q+1$
- r index for class categories
- s_o minimum of one less than the total number of classes $(R-1)$, or the dimensionality (p).
- s_* number of discriminant variables used for assigning an object to a class; $s_* \leq s_o$
- x_i ith element in the data vector x
- $x_{i[l]}$ ith object in node l of CART model
- y_i ith element in the response vector $\mathbf{y}$
- y'_i ith element in the test response vector $\mathbf{y}'$

- $y_{i[l]}$ response value of ith object in node l of CART model
- $\hat{y}_{-i}$ predicted value of $\mathbf{x}_i$, obtained when $\mathbf{x}_i$ is deleted from the model building process
- $\hat{y}_i$ predicted response value for object $\mathbf{x}_i$
- $\hat{y}'_i$ predicted response value for object $\mathbf{x}'_i$
- y_{ij} element in row i and column j of $\mathbf{Y}$
- $\hat{y}_{ij}$ estimate of y_{ij}
- z index for wavelet filter $z = 1, \ldots, m-1$

Non-bold Upper Case Letters

- AIC Akaike's information criterion
- CCR correct classification rate
- CCR′ correct classification rate of test set
- C_p Mallows C_p
- CVCCR cross-validated correct classification rate
- DF degrees of freedom
- DEV deviation
- $\mathcal{D}(\mathbf{x}, r)$ distance between $\mathbf{x}$ and $\bar{\mathbf{x}}_r$ in the discriminant coordinate system
- E_{cross} cross entropy measure
- E_{sym} symmetric entropy measure
- E_{tot} total symmetric entropy measure
- $\mathcal{F}_{\text{CWT}}$ continuous wavelet transform
- $\mathcal{F}_{\text{DWT}}$ discrete wavelet transform
- $\mathcal{F}_{\text{FT}}$ Fourier transform

- $\mathcal{F}_{\mathrm{WFT}}$ windowed Fourier transform
- J highest level in the DWT ; $J = \text{ceiling}(\log p / \log m)$
- $\mathcal{J}$ criterion function applied in the adaptive wavelet or LDB algorithm
- $\mathcal{J}_\Lambda$ Wilk's lambda discriminatory criterion function
- $\mathcal{J}_E$ entropy discriminatory criterion function
- $\mathcal{J}_{\mathrm{cvqpm}}$ discriminatory criterion function based on the cross-validated quadratic probability measure
- $\mathcal{J}_{\mathrm{cvrsq}}$ regression criterion function based on the cross-validated r-squared measure
- $L^2(\mathbb{R})$ space of square integrable functions
- M_{ij} i, jth element in the Lawton matrix
- MCR misclassfication rate
- MSE mean square error
- N_i node identity in CART model
- N_f number of filter coefficients with nonnegative indices
- P_A average probability that an object is assigned to the correct class
- $\mathrm{P}_{\mathrm{QPM}}$ quadratic probability measure
- $\mathrm{P}_{\mathrm{CCR}}$ probability of correctly classifying objects
- $P(r)$ prior probability for class r
- $P(r \mid \mathbf{x})$ posterior probability that given some vector $\mathbf{x}$ it is from class r
- $P(r \mid \mathbf{x}_{i(r)})$ posterior probability for the true class of $\mathbf{x}_i$
- $P(\mathbf{x} \mid r)$ class probability density function
- $P(r \mid l)$ proportion of objects in node N_l of a CART model which are from class r

- $P_{-i}(r \mid \mathbf{x}_i)$ posterior probability for $\mathbf{x}_i$ when the covariance matrices and mean vectors in the probability density function have been calculated in the absence of $\mathbf{x}_i$
- PRESS predicted residual sum of squares
- RSS residual sum of squares
- RSS_{p_o} residual sum of squares of a MLR model with complexity p_o
- R^2 coefficient of variation (r-squared)
- R total number of class categories in a set of data
- R^* integer value less than or equal to $R - 1$
- TSS total sum of squares
- V number of testing groups used in a cross-validation routine
- V_j subspace containing all the possible approximations of functions in $L^2(\mathbb{R})$ at resolution 2^j
- W_j orthogonal complement of V_j

Bold Lower Case Letters

- $\mathbf{a}_i$ ith vector of principal component coefficients with dimension $p \times 1$
- $\mathbf{b}$ estimated vector of regression coefficients
- $\mathbf{b}_{r_{\text{os}}}$ rth column of the matrix of regression coefficients for the optimal scoring problem, $\mathbf{B}_{\text{os}}$
- $\mathbf{b}_{\text{pls}}$ estimated vector of regression coefficients from a PLS model
- $\mathbf{c}_j$ scaling coefficients at resolution (or level) j
- $\mathbf{d}_j$ wavelet coefficients at resolution (or level) j
- $\mathbf{d}_j^{(z)}$ wavelet coefficients at resolution (or level) j produced from the filter matrix $\mathbf{D}_{j+1}^{(z)}$

- $\mathrm{e}_{(r)}^{[j]}(\tau)$ class energy vector of wavelet (or wavelet packet) coefficients
- ℓ vector of low pass filter coefficients
- n $R \times 1$ vector of class sample sizes
- p $n \times 1$ vector containing principal component scores
- r vector of residuals in the PLS algorithm
- s output from low pass filtering operation
- t latent variables from PLS model
- u_i normalized vectors which are used to construct the wavelet matrix A
- v normalized vector which is used to construct the wavelet matrix A
- v $p \times 1$ vector of discriminant coefficients
- w_1 sums of squares and cross product between $\mathbf{X}$ and $\mathbf{y}$
- w output from high pass filtering operation
- x $p \times 1$ training data vector
- $\bar{\mathrm{x}}$ $p \times 1$ mean vector of the training data set
- x' $p \times 1$ testing data vector
- x^* $p \times 1$ column object vector from X^*
- $\mathrm{x}^{[j]}(\tau)$ column vector containing the coefficients in band(j, τ) of the DWT
- $\mathrm{x}_{i(r)}$ $p \times 1$ data vector from class r
- $\mathrm{x}_{i(r)}^*$ ith data object from X^* which belongs to class r
- $\bar{\mathrm{x}}_r^*$ mean class vector from X^*
- ${}^{o}\mathrm{x}^{[j]}(\tau)$ wavelet packet coefficients which occur at the jth level in the τth band of the wavelet packet transform

- $\mathbf{y}$ $n \times 1$ vector of training response values (regression) or class labels (discriminant analysis)
- $\hat{\mathbf{y}}$ $n \times 1$ predicted vector of response values (regression) or class labels (discriminant analysis)
- $\mathbf{y}'$ $n' \times 1$ vector of test response values (regression) or class labels (discriminant analysis)
- $\hat{\mathbf{y}}'$ $n' \times 1$ predicted vector of test response values (regression) or class labels (discriminant analysis)
- $\mathbf{z}$ $n \times 1$ discriminant variable

Bold Upper Case Letters

- $\boldsymbol{A}$ wavelet matrix
- $\boldsymbol{A}_i$ sub-matrix of the wavelet matrix $\boldsymbol{A}$
- $\mathbf{B}$ matrix of multivariate regression coefficients
- $\mathbf{B}_{\mathrm{os}}$ optimal scoring matrix of regression coefficients
- $\mathbf{C}_j$ low pass filtering matrix at level j in the DWT
- $\mathbf{D}_j$ high pass filtering matrix at level j in the DWT
- $\mathbf{D}_j^{(z)}$ high pass filtering matrix at level j in the DWT which contains the zth set of highpass filter coefficients
- $\mathbf{D}$ diagonal matrix whose ith diagonal element is equal to $D_{ii} = 1/\sqrt{\lambda^2_{i_{\mathrm{fda}}}(1 - \lambda^2_{i_{\mathrm{fda}}})}$
- $\boldsymbol{F}_i$ ith factor in the wavelet matrix $\boldsymbol{A}$
- $\mathbf{L}$ low pass convolution matrix
- $\mathcal{H}$ hat matrix $\mathcal{H} = \mathbf{X}^{\mathbf{T}}(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{X}$
- $\mathbf{H}$ high pass convolution matrix

- $\mathbf{P}$ matrix whose ith column contains the principal component scores vector $\mathbf{p}_i$
- $\mathbf{P}_1$ is a matrix which augments $\mathbf{1}_n$ to the first column of $\mathbf{P}$
- $\mathbf{P}_X, \mathbf{P}_{X^*}$ linear projector matrices
- Q orthogonal matrix used in contruction of the wavelet matrix A
- R projection matrix used in contruction of the wavelet matrix A
- S_B between covariance matrix
- S_W within covariance matrix
- $\mathrm{S}_{\mathrm{pooled}}$ pooled covariance matrix
- S_r covariance matrix of class r
- T matrix whose ith column contains the ith latent vector from PLS
- V_{s_o} matrix whose ith column is v_i for $i = 1, \ldots, s_o$.
- X $p \times n$ training data matrix
- X_1 training data matrix whose first row is equal to $\mathbf{1}_n^T$
- X_c $p \times n$ centered training data matrix
- X' $p \times n'$ testing data matrix
- X^* $p \times n$ data matrix which results from some feature selection/transformation procedure based on X.
- $\mathrm{X}^{[j]}(\tau)$ matrix containing the coefficients for the objects which would lie in band(j, τ)
- Y $n \times R$ class indicator matrix
- Z_{s_o} matrix whose ith column is z_i for $i = 1, \ldots, s_o$

Greek Letters

- β_i ith component in the vector of regression coefficients β
- $\delta(t)$ delta function
- δ_{ij} indicator variable; $\delta_{ij} = 1$ if $i = j$, zero otherwise
- ϵ_i ith component in the vector of regression residuals ϵ
- γ_i eigenvalue corresponding to the ith principal component
- λ is a measure of the discriminant criterion $\lambda = \mathbf{v}^T \mathbf{S}_B \mathbf{v}$
- $\lambda_{i_{\text{fda}}}$ ith element of λ_{fda}
- Λ Wilk's Lambda
- $\Lambda^{(i)}$ Wilk's Lambda at the ith iteration of a stepwise routine
- $\mho(j, \tau)$ discriminatory measure of band(j, τ) in the wavelet packet transform
- ν_i ith element in $\boldsymbol{\nu}$
- ω frequency
- $\phi(t)$ scaling function
- $\phi_{j,k}(t)$ scaling basis function; $\phi_{j,k}(t) = m^{j/2}\phi(m^j t - k)$
- $\psi(t)$ mother wavelet function
- $\psi_{j,k}(t)$ wavelet basis function; children wavelets; $\psi_{j,k}(t) = m^{j/2}\psi(m^j t - k)$
- $\hat{\rho}_{ij}$ correlation between the ith principal component and the jth variable
- $\hat{\sigma}_{\mathrm{x}_i}$ sample standard deviation of x_i
- τ band label for the DWT; $\tau \in 0, 1, \ldots, m-1$
- ϱ rank of a matrix
- $\boldsymbol{\beta}$ vector of regression coefficients

- β_{pcr} vector of regression coefficients from a PCR model
- β_{pls} vector of regression coefficients from a PLS model
- λ_{fda} vector whose elements are the eigenvalues of $\Psi^{*T}\Psi^{*}/n$
- Λ_{fda} diagonal matrix whose ith element is equal to $\lambda_{i_{\mathrm{fda}}}$
- $\eta(\mathrm{x}^{*})$ vector of fitted values for x^{*}
- $\bar{\eta}_r$ fitted centroid of all x^{*} objects belonging to class r
- ν vector of wavelengths
- Ψ^{*} class indicator matrix used in FDA and PDA
- $\hat{\Psi}^{*}$ estimate of the class indicator matrix Ψ^{*}
- Θ matrix whose columns are the eigenvectors of $\Psi^{*T}\Psi^{*}/n$

Miscellaneous Characters

- 1_i $i \times 1$ column vector whose elements are all equal to 1
- $\downarrow m$ downsample by a factor of m

List of Algorithms