

# JCU ePrints

This file is part of the following reference:

**Smyth, Christine Wendy (2007) *Weighted tree-based cluster ensembles for high dimensional data*. PhD thesis, James Cook University.**

Access to this file is available from:

<http://eprints.jcu.edu.au/17524>



**Weighted Tree-Based Cluster Ensembles  
for High Dimensional Data**

**Thesis submitted by**

**Christine Wendy SMYTH BSc (Hons) / BA**

**in March 2007**

**for the degree of Doctor of Philosophy  
in the School of Mathematics, Physics, and Information Technology  
James Cook University**

## STATEMENT OF ACCESS

I, the undersigned, author of this work, understand that James Cook University will make this thesis available for use within the University Library and, via the Australian Digital Theses network, for use elsewhere.

I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and I do not wish to place any further restrictions on access to this work.

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

## **ELECTRONIC COPY**

I, the undersigned, author of this work, declare that the electronic copy of this thesis provided to the James Cook University Library is an accurate copy of the print thesis submitted, within the limits of the technology available.

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

## **STATEMENT OF 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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references given.

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

## **STATEMENT ON THE CONTRIBUTION OF OTHERS**

The main body of this thesis is comprised of six publications, completed during the period of candidature. I, the undersigned, conducted the relevant research, designed the experiments, wrote all the necessary computer code, and drafted and edited the manuscripts. The contributions of others included knowledgeable discussions and proofreading of the manuscripts.

Stipend support was also received in the form of a James Cook University Prestige Postgraduate Research Scholarship for the period of candidature.

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

## ACKNOWLEDGEMENTS

I don't think that I really lost the plot too many times over the last three years. This is in no doubt due to the concerted efforts of many. Firstly, I'd like to thank my parents and grandmother. They kept me grounded (Mum), entertained (Dad), and well-fed (Goo). My sister, Helen, prevented me from becoming a boring statistician for which I will be eternally grateful. I also thank my friends for keeping me up to date with the outside world.

I'd also like to thank my academic colleagues, particularly my knowledgeable supervisors, Professor Danny Coomans and Dr Yvette Everingham, for expert advice and guidance.

To all these people, I say "thank you".

## **ABSTRACT**

The increasing size of datasets is particularly evident in the field of bioinformatics. It is unlikely that analyzing these large datasets with a single model will produce an accurate solution. This has led to the ensemble approach, where many models are averaged to give a consensus representation of the data. Taking a weighted average of the individual models has improved the accuracy of both classification and regression ensembles. However, weighting models within a cluster ensemble has remained relatively undeveloped because there is no gold standard available for comparison.

This thesis explores a technique of weighting cluster ensembles. A regression technique, multivariate regression trees, is shown to produce an accurate clustering solution. Each solution (tree) is then weighted purely in terms of its predictive accuracy. Various weighting strategies are trialed to determine the superior technique. After each individual tree is assigned a weight, the trees' co-occurrence matrices are obtained. The co-occurrence matrices are then aggregated together, weighted according to the trees' predictive weights. The final result is a single weighted co-occurrence matrix.

A new technique, similarity-based k-means, is developed in order to partition the weighted co-occurrence matrix. Similarity-based k-means is demonstrated to produce accurate partitions of similarity matrices. The resulting clusters agree with the known groups in the investigated datasets.

Furthermore, this thesis develops two other techniques so that maximal information can be obtained in conjunction with the weighted cluster ensemble. The first method

suggests an estimate of the natural number of clusters in a dataset, by assessing the predictive performance and variability of similarity-based k-means for various numbers of clusters. The estimates agree with the known numbers of groups within the investigated datasets. The second method elucidates the variables that define the clusters. These variables have high classification power within the studied datasets.

Therefore, this thesis presents a holistic cluster analysis: clusters are accurately unearthed within large datasets; an estimate of the natural number of clusters is obtained; and the variables important in defining the clusters are also established. The weighted cluster ensemble technique is applied to a variety of small and large datasets. All results demonstrate the power of weighting the individual models within the ensemble: the developed weighted cluster ensemble technique consistently outperforms the other techniques. The results of analyzing two DNA microarray datasets are particularly promising. The discovered clusters overlap with the known diagnoses in the datasets, and the variables deemed important in defining the clusters have previously been suggested as biomarkers.

Whilst the size of contemporary datasets presents unique statistical challenges, the potential information within them is immense. Statistical techniques must be developed in order to accurately analyze these datasets. Motivated by the success of weighted regression and classification ensembles applied to large datasets, this thesis suggests a technique of weighting models within a cluster ensemble. The results highlight the potential of weighted cluster ensembles in high dimensional settings, such as the analysis of DNA microarrays.

# TABLE OF CONTENTS

<b>Background</b>	<b>1</b>
<b>Multivariate Regression Trees for Cluster Analysis</b>	<b>24</b>
Overview	25
Auto-Associative Multivariate Regression Trees for Cluster Analysis	26
Clustering Noisy Data in a Reduced Dimension Space via Multivariate Regression Trees	53
Synopsis	80
<b>Post Processing Regression Ensembles</b>	<b>81</b>
Overview	82
Parsimonious Ensembles for Regression	83
Post processing regression ensembles: imposing parsimony to improve predictions	102
Synopsis	127
<b>Predictive Weighting for Tree-Based Cluster Ensembles</b>	<b>131</b>
Overview	132
Predictive Weighting for Cluster Ensembles	134
Clustering Microarrays with Predictive Weighted Ensembles	171
Synopsis	199
<b>Concluding Summary</b>	<b>206</b>
<b>Future Work</b>	<b>211</b>
<b>Supporting Theory</b>	<b>217</b>
Regression Trees	218
Regression Tree Splitting	218
Predicting an Observational Unit Using a Regression Tree	220
Auto-Associative Multivariate Regression Trees	220
Multivariate Regression Trees with Principal Component Scores and Multivariate Regression Trees with Factor Scores	221
Dimension Reduction Techniques	221
Principal Components Analysis	221
Factor Analysis	223
Creating Individual Models for Ensembles	225
Bootstrapping and Random Features	225
Bagging	225
Random Forests	226
Bayesian Linear Regression	226

Determining a Suitable Prior Distribution	227
Calculating the Likelihood Function of the Data	228
Determining the Posterior Distribution of the Parameter	229
Sampling from the (Approximated) Posterior Distribution	229
Gibbs Sampler	230
Using Samples to Summarize the Posterior Density Function	231
Using Bayesian Linear Regression to Calculate Weight Coefficients	231
Distributions	232
Multivariate Normal Distribution	232
Inverse Gamma Distribution	232
Multivariate T Distribution	232
Weibull Distribution	233
Double Exponential Distribution	233
Posterior Distributions	233
Posterior Distribution – Multivariate Normal Prior	233
Posterior Distribution – Multivariate T Prior	235
Posterior Distribution – Weibull Prior	236
Posterior Distribution – Double Exponential Prior (Lasso)	238
Evolutionary Algorithms	239
Calculating Fitness Values	241
Fitness Scaling	241
Breeding Chromosomes According to Fitness	241
Crossover – Genetic Algorithm	242
Recombination – Evolution Strategy	242
Mutation	242
Mutation – Genetic Algorithm	243
Mutation – Evolution Strategy	243
Creation of a New Generation	243
Using Genetic Algorithms and Evolution Strategies to Calculate	
Weight Coefficients	244
Genetic Algorithm Approach	244
Evolution Strategy Approach	245
Quadratic Programming	246
Active Set Methods	250
Active Set Methods for Convex Quadratic Programs	251
Active Set Methods for Non-Convex Quadratic Programs	254
Using Quadratic Programming to Calculate Weight Coefficients	258
K-Means	258
Similarity-Based K-Means	260
Determining the Natural Number of Clusters	261
Figures of Merit	262
Extending Figures of Merit	263
<b>References</b>	<b>265</b>

# LIST OF TABLES

## Multivariate Regression Trees for Cluster Analysis

Auto-Associative Multivariate Regression Trees for Cluster Analysis	
<i>Results using the Vietnam dataset</i>	41
<i>Results using the Thyroid dataset</i>	45
Clustering Noisy Data in a Reduced Dimension Space via Multivariate Regression Trees	
<i>Adjusted rand index values at k=6 clusters using the Vietnam Dataset</i>	69
<i>Estimates of cluster number for the Vietnam dataset</i>	70
<i>Adjusted rand index values at k=3 clusters using the Thyroid dataset</i>	74
<i>Estimates of cluster number for the Thyroid dataset</i>	75

## Post Processing Regression Ensembles

Parsimonious Ensembles for Regression	
<i>Description of the Datasets</i>	96
<i>Results of the various weighting schemes over the five datasets</i>	100
Post processing regression ensembles: imposing parsimony to improve predictions	
<i>Description of the Datasets</i>	116
<i>R<sup>2</sup> values for the four datasets using each post processing technique</i>	118
<i>Number of models selected as non-zero by each post processing technique</i>	118
Synopsis	
<i>Comparison of the best regression models with the stepwise linear regression models</i>	129

## Predictive Weighting for Tree-Based Cluster Ensembles

Predictive Weighting for Cluster Ensembles	
<i>Explanation of the three datasets used in the analyses</i>	148
<i>Results of the Iris dataset split to three clusters</i>	152
<i>Results of the Iris dataset split to four clusters</i>	154
<i>Results of the Thyroid dataset split to three clusters</i>	158
<i>Results of the Thyroid dataset split to four clusters</i>	160
<i>Results of the Vietnam dataset split to six clusters</i>	162
<i>Sample variable importance list for the Vietnam dataset</i>	167
Clustering Microarrays with Predictive Weighted Ensembles	
<i>Number of Misclassifications for the Alon Dataset – Two Clusters</i>	188
<i>Number of Misclassifications for the Alon Dataset – Three Clusters</i>	188
<i>Important Variables for the Alon Dataset</i>	190
<i>Number of Misclassifications for the Golub Dataset – Two Clusters</i>	191
<i>Number of Misclassifications for the Golub Dataset – Three Clusters</i>	192

<i>Important Variables for the Golub Dataset</i>	193
Synopsis	
<i>Number of misclassifications using SBK on the co-occurrence matrices created by ensembles of AAMRTs with varying numbers of variables</i>	200
<i>Number of misclassifications using SBK on the co-occurrence matrices created by ensembles of cross-validated AAMRTs with varying numbers of variables</i>	201
<i>Number of misclassifications using SBK on the co-occurrence matrices created by ensembles of cross-validated AAMRTs weighted with evolution strategies</i>	202

# LIST OF FIGURES

## Multivariate Regression Trees for Cluster Analysis

Auto-Associative Multivariate Regression Trees for Cluster Analysis	
<i>Classification Tree of the Vietnam dataset</i>	38
<i>Classification Tree of the Thyroid dataset</i>	39
<i>Relative error curve for the Auto-Associative Multivariate Regression Tree grown on the Vietnam dataset</i>	42
<i>AFOM graphs for the Vietnam dataset</i>	43
<i>Auto-Associative Multivariate Regression Tree of the Vietnam dataset grown to six terminal nodes</i>	44
<i>Relative error curve for the Auto-Associative Multivariate Regression Tree grown on the Thyroid dataset</i>	46
<i>AFOM graphs for the Thyroid dataset</i>	47
<i>Auto-Associative Multivariate Regression Tree of the Thyroid dataset grown to three terminal nodes</i>	48

## Clustering Noisy Data in a Reduced Dimension Space via Multivariate Regression Trees

<i>Adjusted rand index values at <math>k=6</math> clusters for the Vietnam dataset perturbed by type I noise variables</i>	66
<i>Adjusted rand index values at <math>k=6</math> clusters for the Vietnam dataset perturbed by type II noise variables</i>	68
<i>Adjusted rand index values at <math>k=3</math> clusters for the Thyroid dataset perturbed by type I noise variables</i>	72
<i>Adjusted rand index values at <math>k=3</math> clusters for the Thyroid dataset perturbed by type II noise variables</i>	73

## Post Processing Regression Ensembles

### Parsimonious Ensembles for Regression

<i><math>R^2</math> values of the different weighting strategies applied to each dataset</i>	99
<i>Number of models required by each weighting strategy</i>	99

### Post processing regression ensembles: imposing parsimony to improve predictions

<i>Models selected by the post processing techniques for the Ozone dataset</i>	119
<i>Models selected by the post processing techniques for the Fat dataset</i>	120
<i>Models selected by the post processing techniques for the Boston Housing dataset</i>	121
<i>Models selected by the post processing techniques for the Friedman dataset</i>	122

### Synopsis

<i>Models selected as non-zero for the Friedman dataset using Bayesian linear regression with genetic algorithms</i>	128
--	-----

## Predictive Weighting for Tree-Based Cluster Ensembles

Predictive Weighting for Cluster Ensembles	
<i>Weighted co-occurrence matrix for the Iris dataset</i>	150
<i>Average co-occurrence matrix for the Iris dataset</i>	150
<i>AFOM graphs obtained by splitting the Iris dataset's weighted co-occurrence matrix using SBK to various numbers of clusters</i>	153
<i>Three known groups of the Iris dataset projected onto the discriminant plane</i>	155
<i>Two clusters found by SBK on the weighted co-occurrence matrix of the Iris dataset</i>	155
<i>Three clusters found by SBK on the weighted co-occurrence matrix of the Iris dataset</i>	155
<i>Four clusters found by SBK on the weighted co-occurrence matrix of the Iris dataset</i>	155
<i>Weighted co-occurrence matrix for the Iris dataset partitioned with SBK</i>	156
<i>Average co-occurrence matrix for the Iris dataset partitioned with SBK</i>	156
<i>AFOM graphs obtained by splitting the Thyroid dataset's weighted co-occurrence matrix using SBK to various numbers of clusters</i>	159
<i>Three known groups of the Thyroid dataset projected onto the discriminant plane</i>	161
<i>Two clusters found by SBK on the weighted co-occurrence matrix of the Thyroid dataset</i>	161
<i>Three clusters found by SBK on the weighted co-occurrence matrix of the Thyroid dataset</i>	161
<i>Four clusters found by SBK on the weighted co-occurrence matrix of the Thyroid dataset</i>	161
<i>AFOM graphs obtained by splitting the Vietnam dataset's weighted co-occurrence matrix using SBK to various numbers of clusters</i>	163
<i>Classification tree of the Vietnam dataset</i>	167
Clustering Microarrays with Predictive Weighted Ensembles	
<i>AFOM graph for the Alon dataset</i>	189
<i>AFOM graph for the Golub dataset</i>	192