JCU ePrints

This file is part of the following reference:

Hancock, Timothy Peter (2006) ***Multivariate consensus trees: tree-based clustering and profiling for mixed data types.*** **PhD thesis, James Cook University.**

Access to this file is available from:

http://eprints.jcu.edu.au/17497

JAMES COOK UNIVERSITY

Multivariate Consensus Trees:

Tree-based clustering and profiling for mixed data types

Thesis submitted by

Timothy Peter Hancock BSc(Hons)

in 2006

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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

…………………………………………….. ………………………
(Signature) (Date)

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 I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and;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 restriction on access to this work.

______________________ ________________

Signature Date

ELECTRONIC COPY

I, the undersigned, the 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 Contributions

I would like to thank the following people and organizations for the funding they gave to me over the last three and a half years.

1. The Australian government postgraduate award (APA) scholarship.
2. DRS grant from James Cook University.
3. The department of Mathematics, Statistics, Physics, Computer Science and IT at James Cook University.
4. The department of Analytical Chemistry and Pharmaceutical Technology at the Vrije Univeriteit Brussels.

Acknowledgements

Studying the same thing for three and a half years and then writing a thesis to prove it attains a PhD. With this in mind I would first like to thank the people who have distracted me and made me realise that there is far more to life than tree-based models for mixed type clustering and profiling. In particular I would like to thank all mu friends and colleagues that have had to bare the load of everything that went even just a little bit wrong over the last three and a half years. My family in particular must also be mentioned as over the course of my PhD they had to put up with even more. Thanks Dad for reading mine opus. I would like to thank my supervisors, Danny Coomans and Bruce Litow, for the time and effort they gave such that my PhD was. Also I would like to thank the people at FABI for food, shelter and support for my trip over to Belgium, Europe and the world.

In short this thesis is the culmination of a lot of people's insight and effort that has been condensed into the next 200 odd pages, thanks all …

Abstract

Multivariate profiling aims to find groups in a response dataset that are described by relationships with another. Profiling is not predicting each variable within the response set, but finding stable relationships between the two datasets that define common groups. Profiling styles of analysis arise commonly within the context of survey, experimental design and diagnosis type of studies. These studies produce complex multivariate datasets that contain mixed variables often with missing values that require analysis with a flexible, stable statistical technique.

The profiling model under consideration within this thesis is a Classification and Regression Tree (CART). A standard CART model finds groups within a univariate response by building a decision tree from a set of predictor variables. The flexible structure of a CART model allow it to be used for either discriminate or regression analysis whilst also catering for mixed types within the predictor set.

The goal of this thesis to develop methods that extend CART for a multivariate response dataset involving mixed data types. Multivariate regression for CART (MRT) has recently been shown to be a powerful profiling and clustering tool. However the same successes in extending CART for multivariate classification and multivariate mixed type analysis is yet to be realised. To begin with thesis explores simple extensions to CART for multivariate mixed type analysis. These are binary substitution of categorical variables within the response set and partitioning of a distance matrix using Db-MRT. These techniques use already existing extensions to

CART methods and are used as comparison methods to gauge the performance of the ensemble and consensus approaches that are the focus of this thesis.

Ensemble models using CART, such as random forests and treeboost, not only improve the overall accuracy of the model predictions but also introduce an ensemble proximity matrix as a measure of similarity between observations of the response set. In this thesis, through MRT, extensions to both random forests and treeboost are developed such that they predict a multivariate response. Furthermore, by binary substitution of the categorical variables within the response set these multivariate ensemble techniques are further extended to mixed type profiling. A result of this extension is that the ensemble proximity matrix now describes the groups found within the multivariate response. In this way multivariate tree-base ensembles can be interpreted as a cluster ensemble method, where the ensemble proximity matrices can be seen as cluster ensemble consensus matrices. In this thesis these proximity matrices are found to be powerful visualisation tools providing improved resolution of group structure found by a multivariate ensemble method. More so, as in cluster ensembles using these matrices as an input in to a clustering method improves the accuracy of the groups found.

The main work of this thesis is the development of the Multivariate Consensus Tree (MCT) framework for mixed type profiling. Motivating the MCT approach is the need to further understand which variables relate to the groups observed within the proximity matrix. To do this MCTs describe three methods to intelligently combine the ensemble proximity matrices of individual responses into one overall consensus matrix. This consensus matrix is a summary of the overall group structure within each individual proximity matrix. As MCTs work solely with proximity matrices they are independent of the data types within the variables of the response set. Furthermore as each response variable is explicitly predicted it is possible to assess the quality of each proximity matrix in terms of predictive accuracy of the corresponding ensemble.

The MCT consensus matrix is a visualisation tool for the groups present within both the response and predictor datasets. As a consensus matrix is a similarity matrix this thesis proposes five new splitting criteria for tree-based models that search for decision rules within variables of the predictor set that partition the consensus matrix into the observed groups. This tree provides a logical decision path that predicts each group. As the groups within the response are now defined by their relationships within the predictor set, the MCT profiling is complete. This thesis proposes two algorithms for building an MCT; global MCTs and local MCTs. Global MCTs construct an overall consensus matrix spanning all observations, and recursively partition on this matrix to build the tree. Local MCTs build a new consensus matrix at each terminal node to evaluate each new split.

As MCTs have the proximity matrices to summarise the group structure within each response variable methods to identify important subgroups within these variables are also proposed. This search for subgroups within the response can be done on two levels. Firstly to identify subgroups of response variables for overall analysis; and secondly to identify subsets of response variables within any specific group found by the MCT. By finding subsets of response variables that relate to specific group structure the understanding of structure within the dataset is greatly improved.

This thesis shows tree-based methods for profiling, in particular MCTs, to be a powerful tool for mixed type analysis. Firstly, the visualisation of the tree, combined with the proximity matrices, provide a unique view of the groups found and allow for their easy interpretation within the context of the analysis. Secondly, MCTs are shown to accurately estimate the number of groups and provide measures on their stability and accuracy. Furthermore, MCTs are found to be resistant to noise variables within the analysis. Finally they provide methods to find subgroups within the response variables and to identify unimportant variables from the analysis. Throughout this thesis these tree-based methods are compared with standard clustering techniques to provide an accurate benchmark for their performance.

Table of Contents

1. INTRODUCTION 1

2. BACKGROUND 8

2.1 Relating Objects 9

2.2 Relating Mixed Types 10

2.3 Cluster Analysis 12
2.3.1 Hierarchical agglomeration 13
2.3.2 K-means and medoids (PAM) 14
2.3.3 Clustering challenges 15
2.3.4 Determining the number of groups 16
2.3.5 Cluster ensembles 19
2.3.5.1 Cluster ensemble objective functions 19
2.3.5.2 Cluster ensemble consensus matrices 22
2.3.6 Cluster ensembles and predictive ensembles 24

2.4 Profiling Analysis 26
2.4.1 Multivariate regression and multivariate analysis of variance 29
2.4.2 Canonical correlation analysis (CCA) 29
2.4.3 Procrustes analysis 30
2.4.4 Decision and association rules 31

2.5 Tree-based Profiling 32

3. METHODS 35

3.1 Tree-based Models 36

3.2 Classification and Regression Trees (CART) 38
3.2.1 Determining CART tree size 40
3.2.2 Finding the best split 42
3.2.3 Classification trees 43
3.2.4 Univariate regression trees 44
3.2.5 Multivariate regression trees (MRT) 45
3.2.6 CART on a distance matrix (Db-MRT) 46
3.2.7 Auto associative multivariate regression trees (AA-MRT) 48

3.3 Ensembles of Trees 49
3.3.1 Bagging, random forests and multivariate random forests 50
3.3.2 Auto associative random forests and the random forest proximity matrix (RFP) 53
3.3.3 Boosting, treeboost and multivariate treeboost 57
3.3.4 Auto associative treeboost and the treeboost proximity matrix 59
3.3.5 Interpreting ensembles 61
3.3.6 Multidimensional Scaling (MDS) representation of ensemble proximity matrices. 63

3.4 Issues With CART For Mixed Type Responses 64

3.5 Combining Proximity Matrices 67
3.5.1 Combining RFPs by general procrutes analysis (GPA) 69
3.5.2 Combining RFPs by a hierarchical beta-binomial model (BB) 70
3.5.2.1 Estimating the overall beta distribution 72
3.5.2.2 Estimating the overall count 74

3.5.3 Combining RFPs by plaid models (PLAID) 75
3.5.3.1 Estimating the plaid parameters 78

4. MULTIVARIATE CONSENSUS TREES (MCT) 81

4.1 MCT Splitting Functions 83
4.1.1 Splitting using sums of squares (SSR) 83
4.1.2 Splitting using margin reduction (MR) 84
4.1.3 Splitting using an odds ratio (OR) 85
4.1.4 Combining splitting functions (MR-SSR & OR-SSR) 86

4.2 Growing An MCT 88

4.3 Global MCTs 89
4.3.1 Tree size selection for global MCTs 90

4.4 Local MCTs 92
4.4.1 The local MCT augmented consensus matrix (ACM) 93
4.4.2 Tree size selection for local MCTs 94

4.5 Understanding MCTs Output 97
4.5.1 Terminal node labelling 97
4.5.2 Terminal node quality 97
4.5.3 Assessing the quality of the consensus 99
4.5.4 Response variable importance (YVIP) 100
4.5.5 Plaid terminal node filtering 100

5. SOFTWARE 105

6. MCT SENSITIVITY ANALYSIS 107

6.1 Simulation Tests 109
6.1.1 Simulating RFPs 111
6.1.2 Simulation Test 1: Four blurred equal sized groups 113
6.1.3 Simulation Test 2: Ten uneven but clear groups 121
6.1.4 Simulation Test 3: Addition of pure randomness 131

6.2 Random Forest Sensitivity Analysis Using Vietnam Data 136

6.3 Summary 147

7. BENCHMARK EXAMPLES 148

7.1 Clustering Quantitative Variables: Thyroid dataset 150
7.1.1 AA-MRT 151
7.1.2 AA-RF 152
7.1.3 AA-Treeboost 155
7.1.4 Global MCT 157
7.1.5 Local MCT 159
7.1.6 Thyroid summary 161

7.2 Clustering Categorical Variables: Breast Cancer Dataset 163
7.2.1 Gower dissimilarity Db-MRT 164
7.2.2 Binary substituted MRT 165
7.2.3 Binary substituted RF 167
7.2.4 Binary substituted treeboost 170
7.2.5 Base method misclassification results 173

7.2.6 Global MCT 174
7.2.7 Local MCT 183
7.2.8 Breast cancer summary 186

7.3 Mixed Type Profiling: Horse Colic Dataset 188
7.3.1 MRT methods 190
7.3.1.1 Gower dissimilarity Db-MRT 190
7.3.1.2 Binary substituted MRT 191
7.3.1.3 MRT method summary 192
7.3.2 Tree-based ensemble methods 193
7.3.2.1 Binary substituted random forests 194
7.3.2.2 Binary substituted treeboost 197
7.3.3 MCT methods 199
7.3.3.1 Complete response set global MCT 200
7.3.3.2 Complete response set global MCT plaid terminal node filtering 206
7.3.3.3 Plaid response variable filtering algorithm 209
7.4 Horse Colic Summary 223

8. DISCUSSION 225

9. CONCLUSIONS 239

10. REFERENCES 242

List of Figures

Figure 1: Cluster analysis example. 12

Figure 2: Example of profiling a response set $Y = \{y_1, y_2\}$ by predictor variables $X = \{x_1, x_2, x_3, x_4\}$. 28

Figure 3: Example decision tree classifying the species of flowers in the iris dataset. 37

Figure 4: CART algorithm. 39

Figure 5: 10-fold cross-validated RE graph for the iris dataset. 41

Figure 6: Example distance matrix partition. 47

Figure 7: Random forests algorithm. 52

Figure 8: A comparison between the unsupervised random forest and AA-RF proximities. The proximity matrices have been re-ordered by the known iris groups: (1) Setosa, (2) Vericolor, (3) Virginica. Yellow represents a high count between the observations, red a low count. 55

Figure 9: AA-RF predictive performance with predictions on the y-axis and actual variables on the x-axis and a reference line running through y = x. The multivariate $R^2 = 0.97$ and the individual variable R^2s are printed in the titles of each plot. 56

Figure 10: Treeboost algorithm. 58

Figure 11: AA-Treeboost proximity results. The proximity matrices have been re-ordered by the known iris groups: (1) Setosa, (2) Vericolor, (3) Virginica. Yellow represents a high count between the observations, red a low count. 60

Figure 12: AA-Treeboost predictive performance on the iris dataset with predictions on the y-axis and actual variables on the x-axis and a reference line running through y = x. The multivariate R^2 is 0.996, and the individual variable R^2s are printed in the titles of each plot. 61

Figure 13: Binary substitution of an example dataset where the nominal type response variable ‘Sex’ becomes two binary variables *P(M)* and *P(F)* corresponding to the probability of a person being male and female respectively. 65

Figure 14: An illustration of combining RFPs into a consensus proximity matrix, $\bar{C}$, using the iris dataset. 67

Figure 15: GPA consensus matrix for the iris dataset: (1) setosa, (2) versicolour, (3) virginica. 70

Figure 16: Illustration of the hierarchical beta-binomial model. 72

Figure 17: BB consensus matrix for the iris dataset: (1) setosa, (2) versicolour, (3) virginica. 75

Figure 18: Plaid model illustration for a single layer. 77

Figure 19: PLAID consensus matrix for the iris dataset: (1) setosa, (2) versicolour, (3) virginica. 78

Figure 20: MCT split as an odds ratio. 85

Figure 21: Illustrating the performance of each individual MCT splitting function, on the case of perfect separation between the groups. 87

Figure 22: Global MCT algorithm. 89

Figure 23: Global MCT 10-Fold CV for the iris dataset. 90

Figure 24: Global MCT for the iris dataset. 91

Figure 25: Local MCT algorithm. 92

Figure 26: Illustration of the construction of an ACM for the iris dataset. 93

Figure 27: Local MCT RE and AIC curves for the iris dataset. 95

Figure 28: Local MCT on the iris dataset. 96

Figure 29: Terminal node numbering scheme. 97

Figure 30: RMSEs for global and local MCTs for the iris example. 99

Figure 31: Example of a simulated RFP. 111

Figure 32: Four group simulation base configuration. 114

Figure 33: Four group simulation blurred RFP MDS plots. 114

Figure 34: Four group simulation combination RMSE. 115

Figure 35: Four group simulation RE graphs. 116

Figure 36: Four group simulation consensus configurations. 119

Figure 37: Ten group simulation original configuration. 122

Figure 38: Ten group simulation blurred configurations. 123

Figure 39: Ten group simulation consensus RMSEs. 124

Figure 40: Ten group simulation consensus configurations. 125

Figure 41: Ten group simulation RE curves. 126

Figure 42: Pure randomness simulation original configuration. 132

Figure 43: Pure randomness simulation combination method RMSE. 132

Figure 44: Pure randomness simulation RMSE for the consensus matrices. 133

Figure 45: Pure randomness simulation RE curves. 134

Figure 46: Pure randomness simulation consensus configurations. 135

Figure 47: Random forest sensitivity analysis RE curves. 141

Figure 48: Random forest sensitivity analysis best global MCT, MR-SSR splitting with GPA consensus, and random forest tree size of 5 with a minimum terminal node size of 15. 145

Figure 49: Random forest sensitivity analysis best local MCT, MR-SSR splitting with GPA consensus, and random forest tree size of 5 with a minimum terminal node size of 15. 146

Figure 50: Thyroid analysis AA-MRT RE graph. 151

Figure 51: Thyroid analysis AA-MRT. 152

Figure 52: Thyroid analysis AA-RF error convergence plot. 153

Figure 53: Thyroid analysis AA-RF proximity images. 154

Figure 54: Thyroid analysis SSR partition on the AA-RF proximity matrix. 154

Figure 55: Thyroid analysis AA-Treeboost error convergence plot. 155

Figure 56: Thyroid analysis AA-Treeboost proximity images. 156

Figure 57: Thyroid analysis MR partition on the AA-Treeboost proximity matrix. 156

Figure 58: Thyroid analysis global MCT 10-Fold cross-validated RE curves. 157

Figure 59: Thyroid analysis global MCT proximity images. 158

Figure 60: Thyroid analysis global MCT, constructed with MR splitting on the GPA consensus. 158

Figure 61: Thyroid analysis local MCT RE and AIC plots. 159

Figure 62: Thyroid analysis local MCT ACM images and MDS plots. 160

Figure 63: Thyroid analysis local MCT with SSR splitting and GPA consensus combining. 160

Figure 64: Breast cancer analysis Gower distance Db-MRT RE curve. 164

Figure 65: Breast cancer analysis Gower distance Db-MRT. 165

Figure 66: Breast cancer analysis binary substituted MRT RE curve. 166

Figure 67: Breast cancer analysis binary substituted MRT and MDS plot. 166

Figure 68: Breast cancer analysis binary substituted AA-RF error convergence plot. 167

Figure 69: Breast cancer analysis binary substituted RF RE curves. 168

Figure 70: Breast cancer analysis binary substituted random forests MCT built with SSR splitting to 2 splits. 168

Figure 71: Breast cancer analysis binary substituted RF proximity images. 169

Figure 72: Breast cancer analysis binary substituted treeboost error convergence plot. 170

Figure 73: Breast cancer analysis binary substituted treeboost RE curves. 171

Figure 74: Breast cancer analysis binary substituted treeboost MCT built with SSR splitting to 3 splits. 171

Figure 75: Breast cancer analysis binary substituted treeboost proximity images. 172

Figure 76: Breast cancer analysis individual RFP MDS plots. 175

Figure 77: Breast cancer analysis consensus MDS plots and RMSEs. 176

Figure 78: Breast cancer analysis global MCT 10-fold CV RE curves. 177

Figure 79: Breast cancer analysis best global MCTs and terminal node location MDS plots. 180

Figure 80: Breast cancer analysis local MCT RE and AIC plots. 184

Figure 81: Breast cancer analysis GPA, BB and PLAID, OR-SSR local MCT. 185

Figure 82: Horse colic analysis Gower Db-MRT RE curve. 190

Figure 83: Horse colic analysis Gower Db-MRT and terminal node locations. 191

Figure 84: Horse colic analysis binary substituted MRT RE curve. 191

Figure 85: Horse colic analysis binary substituted MRT and terminal node locations. 192

Figure 86: Horse colic analysis binary substituted RF error convergence plot. 194

Figure 87: Horse colic analysis binary substituted RF RE curves. 195

Figure 88: Horse colic analysis RF tree grown to 3 splits using SSR splitting. 196

Figure 89: Horse colic analysis RF proximity images. 196

Figure 90: Horse colic analysis binary substituted treeboost error convergence plot. 197

Figure 91: Horse colic analysis binary substituted treeboost RE curves. 197

Figure 92: Horse colic analysis treeboost tree grown to 3 splits using SSR splitting. 198

Figure 93: Horse colic analysis treeboost proximity images. 198

Figure 94: Horse colic analysis individual RFP MDS plots. The MDS plots are coloured by the predictor variable LESION. 202

Figure 95: Horse colic analysis consensus MDS plots and consensus RMSE plots. 203

Figure 96: Horse colic analysis global MCT 10-fold CV RE curves. 204

Figure 97: Horse colic analysis complete response set global MCT and terminal node location MDS plot. 208

Figure 98: Plaid variable filtering algorithm. 210

Figure 99: Horse colic analysis plaid variable reduction error convergence for the first group. 211

Figure 100: Horse colic analysis response variable group profiles. 215

Figure 101: Horse colic analysis plaid filtered variable group one MCT results. 217

Figure 102: Horse colic analysis plaid filtered variable group two MCT results. 218

Figure 103: Horse colic analysis plaid filtered variable group three MCT results. 219

Figure 104: Horse colic analysis plaid filtered variable group four MCT results. 220

Figure 105: Horse colic analysis classification tree classifying the groups within predictor LESION by the entire response set. (Correct classification rate of 77.33 %). 221

Figure 106: Horse colic analysis MCT grown to 10 splits. 222

List of Tables

Table 1: Plaid terminal node filtering of the iris dataset. The values in the table are *(κ_m)* x *(β_m)*. A zero represents agreement with the mean representation. 101

Table 2: Table of important MCT parameters 108

Table 3: Four group simulation experiment base configuration group centroids. 113

Table 4: Four group simulation MCT misclassification performance (Min node size = 25) 119

Table 5: Four group simulation MCT misclassification performance (Min node size = 50). 120

Table 6: Four group simulation comparative method results. 120

Table 7: Ten group simulation base configuration group centroids. 121

Table 8: Ten group simulation residual sums. 125

Table 9: Ten group simulation MCT misclassification table (Min node size = 5). 129

Table 10: Ten group simulation MCT misclassification table (Min node size = 10). 130

Table 11: Ten group simulation comparative method results. 130

Table 12: Pure randomness simulation base configuration group centroids. 131

Table 13: Pure randomness simulation MCT misclassification results (Min node size = 25). 136

Table 14: Pure randomness simulation comparative method results. 136

Table 15: Random forest sensitivity analysis Vietnam dataset description. 140

Table 16: Random forest sensitivity analysis global MCT misclassification tables. The number of observations misclassified for each group is presented in brackets. The optimal performance for each combination method is emboldened. 143

Table 17: Random forest sensitivity analysis local MCT misclassification table. The number of observations misclassified for each group is presented in brackets. The optimal performance for each combination is emboldened. 144

Table 18: Thyroid analysis AA-MRT misclassification table 152

Table 19: Thyroid analysis AA-RF misclassification table 154

Table 20: Thyroid analysis AA-Treeboost misclassification table. 156

Table 21: Thyroid analysis global MCT misclassification table. 158

Table 22: Thyroid analysis local MCT misclassification table. 160

Table 23: Breast cancer analysis dataset description. 163

Table 24: Breast cancer analysis misclassification performances of base methods. 173

Table 25: Breast cancer analysis global MCT misclassification performances. 182

Table 26: Breast cancer analysis local MCT misclassification performances. 186

Table 27: Horse colic analysis dataset description. 189

Table 28: Horse colic analysis plaid terminal node plaid filtering results. 207

Table 29: Horse colic analysis plaid response variable filtering results. 214

Table 30: Horse colic analysis global MCT terminal missing value distribution. 224