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I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and I do not wish to place any further restrictions on access to this work.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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references given.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

LIST OF TABLES

LIST OF FIGURES

Multivariate Regression Trees for Cluster Analysis

Post Processing Regression Ensembles

Predictive Weighting for Tree-Based Cluster Ensembles