

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

123
61349

A COMPARISON OF TREE-BASED AND TRADITIONAL CLASSIFICATION METHODS

A thesis presented in partial fulfilment of the requirements for the
Degree of PhD in Statistics at Massey University.

Robert D Lynn
1994

ABSTRACT

Tree-based discrimination methods provide a way of handling classification and discrimination problems by using decision trees to represent the classification rules. The principal aim of tree-based methods is the segmentation of a data set, in a recursive manner, such that the resulting subgroups are as homogeneous as possible with respect to the categorical response variable. Problems often arise in the real world involving cases with a number of measurements (variables) taken from them. Traditionally, in such circumstances involving two or more groups or populations, researchers have used parametric discrimination methods, namely, linear and quadratic discriminant analysis, as well as the well known non-parametric kernel density estimation and Kth nearest neighbour rules.

In this thesis, all the above types of methods are considered and presented from a methodological point of view. Tree-based methods are summarised in chronological order of introduction, beginning with the Automatic Interaction Detector (AID) method of Morgan and Sonquist (1963) through to the IND method of Buntine (1992).

Given a set of data, the proportion of observations incorrectly classified by a prediction rule is known as the apparent error rate. This error rate is known to underestimate the actual or true error rate associated with the discriminant rule applied to a set of data. Various methods for estimating this actual error rate are considered. Cross-validation is one such method which involves omitting each observation in turn from the data set, calculating a classification rule based on the remaining $(n-1)$ observations and classifying the observation that was omitted. This is carried out n times, that is for each observation in the data set and the total number of misclassified observations is used as the estimate of the error rate.

Simulated continuous explanatory data was used to compare the performance of two traditional discrimination methods, linear and quadratic discriminant analysis, with two tree-based methods, Classification and Regression Trees (CART) and Fast Algorithm for Classification Trees (FACT), using cross-validation error rates. The results showed that linear and/or quadratic discriminant analysis are preferred for normal, less complex data and parallel classification problems while CART is best suited for lognormal, highly complex data and sequential classification problems. Simulation studies using categorical explanatory data also showed linear discriminant analysis to work best for parallel problems and CART for sequential problems while CART was also preferred for smaller sample sizes. FACT was found to perform poorly for both continuous and categorical data. Simulation studies involving the CART method alone provided certain situations where the 0.632 error rate estimate is preferred to cross-validation and the one standard error rule over the zero standard error rule. Studies undertaken using real data sets showed that most of the conclusions drawn from the continuous and categorical simulation studies were valid. Some recommendations are made, both from the literature and personal findings as to what characteristics of tree-based methods are best in particular situations.

Final conclusions are given and some proposals for future research regarding the development of tree-based methods are also discussed. A question worth considering in any future research into this area is the use of non-parametric tests for determining the best splitting variable.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my three supervisors, Associate Professor Dick Brook, Mr Greg Arnold and Dr S Ganesalingam for their constant support and helpful advice throughout my PhD study. I would also like to thank Mum and Dad, Judith and Robin Lynn, for providing me with cheap board and lodgings over the years as well as encouraging me to persevere to the end. I am indebted to Massey University for the use of their computer facilities, and in particular, to the Department of Statistics for providing me with employment over the past five years. Last, and by no means least, I owe a great deal of thanks to Paula McMillan for her efforts in typing this thesis, without her skill in reading my often illegible script this thesis may never have been completed!

ADDITIONAL PUBLICATIONS

Ganesalingam, S and Lynn, R D (1991). Posterior probability based estimator for the overall error rate associated with a linear discriminant function. Occasional Publications in Mathematics and Statistics, **23**, Massey University.

Lynn, R D and Brook, R J (1991). Classification by decision trees and discriminant analysis. New Zealand Statistician, **26**, pp 18-26.

Lynn, R D, Brook, R J and Arnold, G C (1993). A comparison of four classification methods: linear and quadratic discriminant analysis, CART and FACT. Mathematical and Information Sciences Report, Series B: **1**, Massey University.

Table of Contents

1. INTRODUCTION	1
2. TRADITIONAL DATA DISCRIMINATION METHODS	5
2.1 INTRODUCTION	5
2.2 LINEAR DISCRIMINANT ANALYSIS.....	5
2.2.1 Stepwise Discriminant Analysis.....	11
2.3 QUADRATIC DISCRIMINANT ANALYSIS.....	12
2.4 THE ROBUSTNESS OF LINEAR AND QUADRATIC DISCRIMINANT ANALYSIS	12
2.4.1 Modifications to Linear Discriminant Analysis	14
2.5 KERNEL DENSITY ESTIMATION.....	14
2.6 Kth NEAREST NEIGHBOUR METHODS	17
2.7 CRITIQUES OF KERNEL DENSITY ESTIMATION AND KTH NEAREST NEIGHBOUR METHODS	18
3. A TABULAR COMPARISON ON TEN TREE-BASED METHODS.....	21
3.1 ORIGINS OF TREE-BASED METHODS.....	21
3.2 INTRODUCTION.....	21
3.3 AID.....	31
3.4 THAID	34
3.5 ID3.....	36
3.6 CHAID	39
3.7 CART	41
3.8 C4.5	47
3.9 FACT.....	50
3.10 KnowledgeSeeker.....	52
3.11 Splus Trees ()	55
3.12 IND.....	59

4. SIMULATION STUDIES INVOLVING CONTINUOUS DATA	65
4.1 INTRODUCTION	65
4.2 ERROR RATES	65
4.3 SIMULATION STUDY I.....	74
4.3.1 Study Plan.....	74
4.3.2 Results	76
4.3.3 Summary.....	82
4.4 SIMULATION STUDY II	83
4.4.1 Study Plan.....	83
4.4.2 Results	83
4.4.3 Summary and Discussion	91
4.5 THE EFFECTS OF PRIORS ON ERROR RATES	93
4.5.1 Introduction	93
4.5.2 Purpose of this study.....	93
4.5.3 Study Plan.....	94
4.5.4 Results	95
4.5.5 Summary.....	104
4.6 SIMULATION STUDY III.....	107
4.6.1 Introduction	107
4.6.2 Study Plan.....	107
4.6.3 Results	108
4.6.4 Summary.....	112
4.7 CONCLUSIONS	113
5. SIMULATION STUDIES INVOLVING CATEGORICAL DATA	115
5.1 INTRODUCTION	115
5.2 PREVIOUS STUDIES	116
5.3 SIMULATION STUDY I.....	117
5.3.1 Study Plan.....	117
5.3.2 Results	118
5.3.3 Summary.....	121

5.4	SIMULATION STUDY II.....	122
5.4.1	Introduction	122
5.4.2	Study Plan.....	122
5.4.3	Results	122
5.4.4	Summary.....	126
5.5	CONCLUSIONS.....	127
6.	CART SIMULATION STUDY	129
6.1	INTRODUCTION	129
6.2	ERROR RATE ESTIMATION FOR CONTINUOUS DATA IN CART	129
6.2.1	Previous Studies	129
6.2.2	Study Plan.....	130
6.2.3	Results	132
6.2.4	Summary.....	143
6.3	ERROR RATE ESTIMATION FOR CATEGORICAL DATA IN CART	144
6.3.1	Study Plan.....	144
6.3.2	Results	145
6.3.3	Summary.....	151
6.4	THE STANDARD ERROR RULE IN CART.....	151
6.4.1	Previous Studies	151
6.4.2	Study Plan.....	152
6.4.3	Results	152
6.4.4	Summary.....	158
6.5	TRANSFORMATIONS OF ERROR RATES.....	158
6.5.1	Study Plan.....	158
6.5.2	Results	159
6.5.3	Summary.....	161
6.6	CONCLUSIONS.....	161
7.	CASE STUDIES.....	165
7.1	INTRODUCTION.....	165
7.2	PREVIOUS STUDIES	165

7.3	COMPARATIVE STUDIES.....	166
7.3.1	Methods and Data Sets	166
7.3.2	Cross-Validation Error Rate Results	173
7.3.3	0.632 Error Rate Results.....	176
7.3.4	Individual Class Error Rates.....	176
7.3.5	The Standard Error Rule in CART	179
7.3.6	Splus Trees() versus CART.....	179
7.3.7	Summary.....	181
7.4	ILLUSTRATIVE CASE STUDY	183
7.4.1	Methods and Data.....	183
7.4.2	Linear Discriminant Analysis.....	184
7.4.3	CART	186
7.4.4	FACT.....	189
7.4.5	KnowledgeSeeker.....	192
7.4.6	Splus Trees()	203
7.4.7	Summary	208
8.	WHICH CHARACTERISTICS OF TREE-BASED METHODS ARE PREFERRED.....	209
8.1	INTRODUCTION.....	209
8.2	WHICH CHARACTERISTICS OF TREE-BASED METHODS ARE PREFERRED?.....	209
8.2.1	The Method of Splitting	209
8.2.2	Binary versus Multiway Splits	211
8.2.3	Univariate versus Linear Combination Splits.....	212
8.2.4	Costs and Priors.....	214
8.2.5	Stopping Rules and Tree Pruning.....	214
8.3	HUMAN COMPREHENSIBILITY AND USER-FRIENDLINESS OF.....	216
9.	CONCLUSIONS AND PROPOSALS FOR THE FUTURE.....	221
	NOTATION INDEX	235
	BIBLIOGRAPHY	239