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