Automatic Interaction Detection for Longitudinal Data

Jacky Galpin
School of Statistics and Actuarial Science
University of the Witwatersrand
South Africa
jacky@galpin.co.za

Tree based methods for detecting variables predictive of a continuous or categorical dependent variable are well developed (see for example, Kass, 1980, Hawkins and Kass, 1982, Breiman, Olshen and Stone, 1984, among many others), and are available in several statistical packages for the case of a single dependent variable. The case of multiple dependent variables has also received some attention in the literature (see for example Zhang and Singer, 1999, and references there-in).

There have been a number of articles relating to the extension of these methods to longitudinal responses, but these generally use techniques to summarize the response profile over time for each respondent (see for example Zhang and Singer, 1999, and references in that text). These include the use of regression models (including splines), and latent class analysis. Little has appeared concerning the case of longitudinal data treated as such, rather than being summarized via some statistical model.

Looking at the methods for analyzing a single response variable, the most widely reported methods available are Chi-squared Automatic Interaction Detection (CHAID) and Classification and Regression Trees (CART), discussed in the references above.

The CART methods use techniques such as boosting and bagging to ensure the stability of the tree, essential to evaluating the usefulness of the identified model.

However, an issue that arises in the evaluation of market research data (among others) is that of problems with data collection, and/or sampling. An example of this is the checking of successive waves of the AMPS®, RAMS® and TAMS® (All Media and Products Survey, Radio Audience Measurement Survey, and Television Audience Monitoring Systems, all commissioned by the South African Advertising Research Foundation).

The aim of the analyses of these data sets is to compare successive surveys (for which the data is weighted to population totals) to detect changes over the waves. Some of these are readily explainable (such as increased listening, viewing and reading during sporting events). Others may be signs of problems with the surveys.

For other data sets, these techniques could be useful in examining the possible failure of model assumptions. One example would be the failure of the assumptions of a specific mean and/or covariance structure in the longitudinal generalized linear model. This could include the assumptions relating to the fixed and random effect parameters and distributions. It could also include the detection of aberrant observations which could result in an incorrect model being fitted. In other words, there may be some usefulness in doing the analyses without boosting and bagging (as well as with these techniques).

This paper addresses in particular the possible use of these methods in detecting possible mis-specification of the structure of the mixed model proposed for the data.

References:

Breiman, L, Friedman, JH, Olshen, RA and Stone, CJ (1984). *Classification and Regression Trees.* Wadsworth, California.

Hawkins, DM and Kass, GV (1982). Automatic Interaction Detection. In *Topics in Applied Multivariate Analysis,* Ed DM Hawkins, Cambridge University Press.

Kass, GV (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 29(2): 119-127.

Zhang, H and Singer, B (1999). *Recursive Partitioning in the Health Sciences*. Springer.