The CoIL Challenge:
an Application of Classification Trees with Bootstrap Aggregation

A.P. White [*†] (University of Birmingham)
W.Z. Liu (Seagate Software)

Abstract

A classification approach involving data tuning, followed by the use of a binary classification tree, augmented by bootstrap aggregation, yielded a ‘hit’ score of 110.

Introduction

The original intention of the authors was to use a combination of classification procedure, as it has been shown by Henery (1997) that this can yield enhanced classification performance. Unfortunately, shortage of time constrained the authors to the use of a single technique, namely that of a binary classification tree. However, this was preceded by various forms of data tuning and augmented by bootstrap aggregation (‘bagging’). Each of these techniques is described in more detail in the following sections.

Software

The classification tree software (PREDICTOR) that was employed for this task was designed and written some years previously by the authors themselves. PREDICTOR operates by a recursive binary partitioning of the data space, under a form of statistical control that branches preferentially on the more important variables. In this respect, it is similar to CART, described by Breiman et al. (1984). A stopping rule based on significance testing principles guards against excessive branching and the specification of a minimum terminal node frequency provides additional protection against overfitting.

For ordered attributes, the operation of the algorithm is described in detail by White & Liu (1990). Briefly, the algorithm operates in the following way. At each node, each attribute is partitioned into two sub-ranges, thereby deriving a binary attribute whose values define the sub-ranges. This is done by considering each possible cutting point for the ordered attribute and selecting the one which gives the maximum association with class, for the derived binary attribute. Branching from the node under consideration is then performed using that derived binary attribute which has the strongest association with class. For both determination of the cutting points and selection of the attributes, the χ^2 measure of association is used.

[*] Dr. A.P. White is an Associate Member of the School of Mathematics and Statistics.
[†] All correspondence should be addressed to Dr. A.P. White, School of Computer Science, University of Birmingham, Edgbaston, Birmingham B15 2TT. Alternatively, send email to a.p.white@bham.ac.uk .

The growth of each path is subject to termination either by attainment of node purity, or by χ^2 failing to reach some preset threshold value (expressed in terms of a significance level), or by the number of cases at the node falling below a preset threshold value. (These last two terminating conditions are both mechanisms to prevent branching from proceeding too far, in order to protect against overfitting). At each terminal node, the probabilities of class membership are estimated in the obvious way from the relative class frequencies at that node.

PREDICTOR operates by dynamic path generation (White, 1987), in which only the path necessary to classify the particular case under consideration is generated, rather than the entire tree. This avoids much of the computational expense of cross-validation.

Data Tuning

Before PREDICTOR was used on the data, it was necessary to carry out recoding of two of the attributes. This was because the splitting mechanism for nominal (non-ordered) attributes is different from that for ordered attributes, involving a choice between many more possible splits. The two attributes in question were MOSTYPE (1) and MOSHOOFD (5). The 'trick' employed with these two attributes was to employ recoding in order to derive attributes with a quasi-ordering, based on monotonically ordered probabilities of membership of the target class, estimated from the data in the training set.

Each of the attributes was then subjected to cross-tabulation against the class variable (CARAVAN). Those that showed substantial evidence of a quadratic relationship (or higher-order polynomial) were then also recoded into derived variables having a monotonic relationship with class. Attributes modified at this stage were PBRAND (59), MKOOPKLA (43), MSKA (25), MGODPR (7), MGODGE (9) and MGODOV (8). The reasoning behind this move was to reduce the branching that PREDICTOR would have to do. This would, in turn, reduce the corresponding partitioning of the data space (White & Liu, 1997), leading to a smaller number of larger regions, which can be shown to give better classification results.

Bootstrap Aggregation

Breiman (1996) describes a technique based on the bootstrap (Efron & Tibshirani, 1993) which can be used to augment the performance of various discriminant algorithms. This is called bootstrap aggregation ('bagging'). It involves taking a number of bootstrap replicates of the training set and deriving from each one classification predictions for the entire test set and averaging these over the bootstrap replicates. Classification based on this average is usually superior to that from any of the replicates themselves, or indeed to that obtained by just using the classification obtained from the training set itself.

Method

The first step in the procedure was to find the parameter setting for PREDICTOR that gave optimal classification performance under n-fold (leaving-one-out) cross-validation for the entire training set. The minimum terminal node frequency was fixed at 5. The significance level for χ^2 was varied in approximately equal logarithmic steps of 0.1, 0.05, 0.02, 0.01 and so on. The best cross-validated classification performance was produced with a setting of 0.002.

Twelve bootstrap replicates of the training set were then generated. Each of these involved sampling with replacement to generate a sample of the same size as the training set itself. Each of these bootstrap replicates was then used (with the parameter settings as previously specified) to perform classification on the test set. This produced 12 probabilities of membership of the target set for each case in the test set. These were then averaged and the cases with the 800 largest mean probability values were selected as the solution.

Results and Discussion

The resulting classification score on the 800 cases selected from the test set was 110 'hits'. It is interesting to compare this with the figure of 105 which would have been obtained if the original training set had been used (with the same parameter settings), rather than bootstrap aggregation.

It is also instructive to examine the classification performance of each bootstrap replicate on the test set. Here, the scores ranged from 82 to 109, with a mean of 90.25. Taken together, these figures illustrate the power of bootstrap aggregation in that the final result is better than that derived from the best of the bootstrap replicates and also better than would have been obtained if just the training set had been used.

References

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and regression trees*. Belmont: Wadsworth.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123 – 140.

Efron, B.S. & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Henery, R.J. (1997). Combining classification procedures. In *Machine Learning and Statistics: the Interface,* edited by G. Nakhaeizadeh & C.C. Taylor, pp. 153 – 177. New York: John Wiley & Sons.

White, A.P. (1987). Probabilistic induction by dynamic path generation in virtual trees. In *Research and Development in Expert Systems **III**,* edited by M.A.Bramer, pp. 35 – 46. Cambridge: Cambridge University Press.

White, A.P. & Liu (1990). Probabilistic induction by dynamic path generation for continuous variables. In *Research and Development in Expert Systems **VII***, edited by T.R. Addis & R.M. Muir, pp. 285 – 296. Cambridge: Cambridge University Press.

White, A.P. & Liu (1997). Statistical properties of tree-based approaches to classification. In *Machine Learning and Statistics: the Interface,* edited by G. Nakhaeizadeh & C.C. Taylor, pp. 23 – 44. New York: John Wiley & Sons.