Computational Statistics and Data Analysis 106 (2017) 90–102

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

An alternative pruning based approach to unbiased recursive partitioning

Alberto Alvarez-Iglesias [a,*], John Hinde [b], John Ferguson [a], John Newell [a,b]

[a] HRB Clinical Research Facility, National University of Ireland Galway, University Road, Galway, Ireland

[b] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Ireland

ARTICLE INFO

Article history:
Received 6 March 2015
Received in revised form 12 August 2016
Accepted 13 August 2016
Available online 29 August 2016

Keywords:
Tree-based methods
Interactions
Pruning
False discovery rate

ABSTRACT

Tree-based methods are a non-parametric modelling strategy that can be used in combination with generalized linear models or Cox proportional hazards models, mostly at an exploratory stage. Their popularity is mainly due to the simplicity of the technique along with the ease in which the resulting model can be interpreted. Variable selection bias from variables with many possible splits or missing values has been identified as one of the problems associated with tree-based methods. A number of unbiased recursive partitioning algorithms have been proposed that avoid this bias by using *p*-values in the splitting procedure of the algorithm. The final tree is obtained using direct stopping rules (pre-pruning strategy) or by growing a large tree first and pruning it afterwards (post-pruning). Some of the drawbacks of pre-pruned trees based on *p*-values in the presence of interaction effects and a large number of explanatory variables are discussed, and a simple alternative post-pruning solution is presented that allows the identification of such interactions. The proposed method includes a novel pruning algorithm that uses a false discovery rate (FDR) controlling procedure for the determination of splits corresponding to significant tests. The new approach is demonstrated with simulated and real-life examples.

1. Introduction

The history of tree-based models goes back to the seminal work of Morgan and Sonquist (1963) and the development of the AID algorithm (Automatic Interaction Detection). In the presence of interaction effects, classical regression analysis needs to account for such effects by adding interaction terms to the model. In general, those effects are unknown a priori and their identification requires fitting many candidate models. The same is true in the case of traditional variable selection procedures like forward selection, backwards elimination or stepwise variable selection (or more recently lasso or ridge regression), where interaction effects have to be pre-specified in the model. Morgan and Sonquist (1963) proposed an algorithm that did not suffer from such constraints, making the identification of interaction effects an automatic process. The algorithm consists of an exhaustive search over all the possible partitions of the sample space generated by each one of the predictors, selecting the partition that leads to the greatest reduction in the variance of the response. Such an algorithm can be repeated for each one of the subgroups generated and the process can be iterated many times, i.e. recursive partitioning (RP).

One of the early problems of this method was the determination of the optimal size of the tree, which was usually selected based on some previously specified stopping criterion. A solution to this problem was given by Breiman et al. (1984) with

* Corresponding author.
E-mail address: alberto.alvarez-iglesias@nuigalway.ie (A. Alvarez-Iglesias).

http://dx.doi.org/10.1016/j.csda.2016.08.011
0167-9473/

the CART (Classification and Regression Trees) algorithm, which introduced the idea of pruning a large tree based on a cross-validated measure of the cost-complexity of a tree (a measure that evaluates the error of the tree taking into account the complexity of the model). Although the method was originally proposed for continuous and categorical responses, soon extensions were proposed for count (Chambers and Hastie, 1992; Therneau and Atkinson, 1997) and survival responses (Gordon and Olshen, 1985; Segal, 1988; Davis and Anderson, 1989; LeBlanc and Crowley, 1993, among others).

One of the drawbacks of the RP algorithm (a form of greedy search algorithm) is the problem of variable selection bias where predictors with many possible splits are more likely to be chosen as splitting variables. This problem has been studied by several authors including White and Liu (1994), Shih and Tsai (2004), and Kim and Loh (2001) who showed that the variable selection bias can also be due to the number of missing values of the predictor (predictors with many missing values have a higher probability of being selected).

Several unbiased recursive partitioning (URP) algorithms have been proposed that avoid this bias by, at each split, separating variable selection (usually based on hypothesis tests) and splitting point selection (see QUEST, (Loh and Shih, 1997), GUIDE, (Loh, 2002), *CTree*, (Hothorn et al., 2006), and MOB, (Zeileis et al., 2008), among others). In order to obtain the final tree, these methods use direct stopping rules (pre-pruning), generally based on p-values, or CART-style post-pruning strategies. However, as will be shown, stopping rules based on significance tests may generate trees where important effects are missing (such as interaction effects). Other algorithms such as evtree, Grubinger et al. (2014), use global optimization methods, including evolutionary algorithms, for learning globally optimal classification and regression trees. Here, only trees with binary splits and constant models in the terminal nodes are considered.

To begin, a review of the pre-pruning strategy based on Bonferroni-adjusted p-values proposed by Hothorn et al. (2006) (*CTree* algorithm) is presented. This approach will be used as an example of direct stopping rules based on significance and to introduce the interaction detection problem mentioned above. In Section 2, a description of the *CTree* algorithm is given. In Section 3, a synthetic example is shown that explains the difficulty of finding interactions using pre-pruning approaches based on p-values. In Section 4, a simple modification of the *CTree* algorithm is presented that allows the identification of such interactions using a post-pruning strategy based on false discovery rate (FDR) controlling procedures. In Section 5, the novel pruning procedure is compared to *CTree* and CART using a simulation study. Finally, in Section 6, the proposed method is applied to data from a breast cancer study.

2. Review of the *CTree* algorithm

The URP algorithm proposed by Hothorn et al. (2006) aims to overcome the problem of "selection bias towards predictors with many possible splits or missing values". Solutions to this problem had been suggested before in the literature (Loh and Vanichsetakul, 1988; Loh and Shih, 1997; Kim and Loh, 2001; Loh, 2002; Kim and Loh, 2003), by performing hypothesis tests in the splitting procedure, and selecting factors and covariates according to the obtained p-values. Unlike these methods, which use a great variety of tests, *CTree* uses conditional inference procedures, based on a general theory of permutation tests developed by Strasser and Weber (1999). The main strength of this form of test is its flexibility, as it can handle any type of response data within the same unified framework. Continuous, categorical, count, and survival outcomes and predictors can easily be accommodated with minor modifications to the algorithm.

At each node a hypothesis test is performed and the p-values obtained for each split determine whether tree growing should stop or continue. The following section describes the *CTree* algorithm.

2.1. The CTree algorithm

Let Y be the response (continuous, categorical, count, or time to event outcome) and $\{X_j, j = 1, \ldots, J\}$ a set of predictors (continuous or categorical). The method to grow the tree can be summarized in the following steps:

1. At any node (starting with the parent node), test the null hypothesis of independence between Y and any of the predictors X_j using the data belonging to that node. If $J > 1$ adjustments will be necessary to maintain the global significance level. If the adjusted p-value is not significant, the data are not split any further. If it is significant, choose the predictor with the strongest association (smallest p-value).
2. Using the predictor selected in step 1, find the cut-point that maximizes the separation between the two children nodes. This can be done by searching over all the possible cut-points and obtaining the p-values of the test of independence between the left and the right part. The cut-point leading to the smallest p-value is selected for the split.
3. Repeat steps 1 and 2 until no further splits are possible.

The separation of steps 1 (variable selection at each split) and 2 (selection of the splitting point) is key to avoid the variable selection bias (see Hothorn et al., 2006 for full details). Other algorithms, like CART and many of its variants, base the splitting mechanism on the selection of the predictor that contains the splitting point that maximizes separation, potentially leading to bias. This is one of the strengths of *CTree*, along with the flexibility to accommodate all types of response and explanatory variables.

The pre-pruning strategy based on p-values is used to protect locally against the discovery of false positives (splits on noise variables — that is splits on variables that are unrelated to the response) at a significance level α. It also works as a closed testing control procedure where subsequent hypotheses are only assessed if all previous ones were significant,