A New Splitting Approach for Regression Trees

Yung-Seop Lee
Dongguk University, Department of Statistics
3ga 26, Phil-Dong,Chung-Gu
Seoul, Korea
yung@dongguk.edu

1. Introduction

There are several techniques for classification in data mining. One of them is a decision tree method. Two kinds of decision tree methods are regression trees and classification trees (Breiman, 1984). Regression trees are used to predict a continuous response, while classification trees are used to predict a class label. The goal of decision tree methods is to partition data into small nodes such that a response value (regression tree) or a class label (classification tree) can be well predicted in each node. Key components of a tree construction are the selection of splitting criteria, stopping criteria when a tree grows and the assignment of terminal nodes to a class. Splitting criteria in a tree is specified by a measure of impurity and by selection the best splitting variables and its threshold for the further split. If there is no improvement in the node impurity from parent node to child nodes, the split is worthless. In contrast, if a split results in pure child nodes, then the split is definitely best. This is what we called the impurity reduction. For regression trees, impurity can be measured by variants of Residual Sum of Squares, while for classification trees, impurity can be measured for example by misclassification rate, entropy, Gini Index. Let $i(\cdot)$ be some measure of within-node impurity at parent node ($i(\cdot)$), the left child node ($i(L)$) and right child node ($i(R)$). The best splitting variable and threshold having the most reduced impurity (the biggest Δi) in a node are selected when the tree is grown. Also, the formula for the conventional impurity reduction to select the splitting variable and threshold can be expressed as follows:

$$\Delta i = i(o) - \left(\frac{n_L}{n} i(L) + \frac{n_R}{n} i(R) \right) \qquad (1)$$

That is, the left and right child nodes are combined by a weighted sum of two sides. In effect, one compromises between left and right buckets by taking their sizes into account.

2. Alternative splitting criteria

There are two aspects of tree assessment: 1) global accuracy of trees for prediction, and 2) interpretability of trees. Most of tree algorithms, for example, CART (Breiman, 1984), Bagging(Breiman, 1996), AdaBoost(Freund and Schapire, 1996) and Bumping (Tibshirani and Knight, 1996) are concentrated on improving the accuracy. But in this paper, we are more concern

to the interpretable trees. That is, we attempt to find methods that search for meaningful nodes as close to the top of the tree as possible. By emphasizing the interpretability of nodes near top, we de-emphasize the precise calibration of the tree depth for prediction: Stopping criteria for growing and pruning trees are of low priority for our purposes because the interpreting data analyst can always ignore the lower nodes of a tree that was grown too deeply. This is harmless as long as the analyst does not interpret statistical flukes in the lower nodes. By de-emphasizing the lower parts of trees, we may sacrifice some global accuracy of prediction. A remark about simplicity and interpretability of trees is in order: Although the literature has a bias in favor of balanced trees, balance and interpretability are very different concepts: There exists a type of extremely unbalanced trees that are highly interpretable. The simplicity of this tree stems from the fact that all nodes can be described by one or two conditions, regardless of tree depth. Our new criteria for splitting sometimes generate trees that are less balanced and yet more interpretable than conventional balanced trees. Instead of the conventional weighted average of right and left child nodes for splitting, we propose the following two criteria:

$$1)\ \min\left(\frac{\hat{m}_L}{\hat{s}_L^2}, \frac{\hat{m}_R}{\hat{s}_R^2}\right) \qquad 2)\ \min\left(\frac{\hat{s}_L^2}{\hat{m}_L}, \frac{\hat{s}_R^2}{\hat{m}_R}\right)$$

In fact, these criteria are the extension of Buja and Lee (2001). They are just concentrated on pure nodes (based on $\hat{s}_L^2$ and $\hat{s}_R^2$) or extreme nodes (based on $\hat{m}_L$ and $\hat{m}_R$) separately in advance. But these extended criteria are the mixture of the previous criteria, that is, we can select the nodes by considering both pure and extreme in the same tree. The shape of trees are also very unbalanced but interpretable as expected. These criteria are not the generalized one of the previous criteria but another option depending on the circumstance. For example, in case your data have many outliers or extreme values, these criteria are more adequate and interpretable. The result trees will be shown in the on-site presentation by the short of space.

REFERENCES

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

Breiman,L., Friedman,J.H., Olshen,R.A., and Stone,C.J. (1984), Classification and Regression Trees, Pacific Grove, CA: Wadsworth.

Buja,A. and Lee, Y-S. (2001), Data Mining Criteria for Tree-Based Regression and Classification, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA (to appear).

Freund, Y. and Schapire,R.E. (1996), A Decision-Theoretic Generalization of On-Line Learning and Application to Boosting, Journal of Computer and System Science 55, 119-139.

Tibshirani,R. and Knight,K. (1996), Model Search and Inference via Bootstrap Bumping, Technical Report, Dept. of Statistics, U. of Toronto.