A study on the tree model grown by hybrid splitting criterion

Daewoo Choi
Dept. of Statistics, Hankuk Univ. of FS
Yongin, 449-791, Korea
dachoi@freechal.com

Joongsik Yoon
Dept. of Statistics, Hankuk Univ. of FS
Yongin, 449-791, Korea
lunch93@stat.hufs.ac.kr

1. Introduction

Tree model is the most popular classification algorithm in data mining due to easy interpretation of the result. In CART (Breiman *et al.*, 1984) and C4.5 (Quinlan, 1993) which are representative of tree algorithms, the split for classification proceeds to attain the homogeneous terminal nodes with respect to the composition of classes in response variable. But, for instance, in churn prediction modeling for CRM (Customer Relationship Management), the rate of churn is generally very low although we are interested in mining the churners. Thus it is difficult to get accurate prediction model using tree model based on the traditional split rule, such as gini or entropy.

Buja and Lee (1999) introduced a new split rule, one-sided purity. One-sided purity is very attractive for searching a small interesting group, but the accuracy of the model is not higher than that of the tree model based on gini or entropy index.

In this study we propose a new splitting criterion, hybridizing gini and one-sided purity to improve the classification accuracy for binary response variable.

2. Some properties of splitting criteria

Tree model is growing as splitting the data recursively based on minimizing criterion. Gini index as in (1) is the popular criterion with entropy index.

$$f(g) = \sum_j \hat{p}_j(g)(1 - \hat{p}_j(g)) \qquad (1)$$

In (1), $\hat{p}_j(g)$ is a proportion of j-th class of response variable in a node g. In case of binary response variable, $f(g)$ is the variance of Bernoulli distribution as a measurement of impurity in a node g.

In CART proposed by Breiman *et al.* (1984), binary-split proceeds for reducing gini indices of both left- and right-node. But, for the unbalanced data with respect to the proportion for each class among response variable, equal-balanced splitting criteria (EBSC) such as gini or entropy index

may not be efficient to search the cluster with the lower proportion.

Buja and Lee (1999) proposed a new splitting criterion, one-sided purity (OSP). The OSP splitting criterion for binary response level is as follows:

$$\min(\hat{p}_L(1-\hat{p}_L), \hat{p}_R(1-\hat{p}_R)) \quad (2)$$

In (2), $\hat{p}_L$ ($\hat{p}_R$) is a proportion of a class in left- (right-) node. From (2), we know that OSP let the tree growing up as peeling a pure cluster. Friedman and Fisher (1997) proposed a rule discovery algorithm, PRIM peeling off a fixed rate of data.

For data mining problem such as churn prediction in mobile telecommunication co., OSP is useful because the rate of churner among whole customer is very low. But we are not able to improve the accuracy of a tree model only with OSP splitting criterion. The reason is that the proportion for both classes gets balanced as the split by OSP goes on. That is, OSP has a limit to reach high accuracy.

We generate a data with binary response depicted in *Figure* 1. Restricting the depth of tree model less than equal to 30, we have a poor tree model grown only by OSP. (see *Figure* 2) But, comparing with OSP, gini works better as in *Figure* 3.

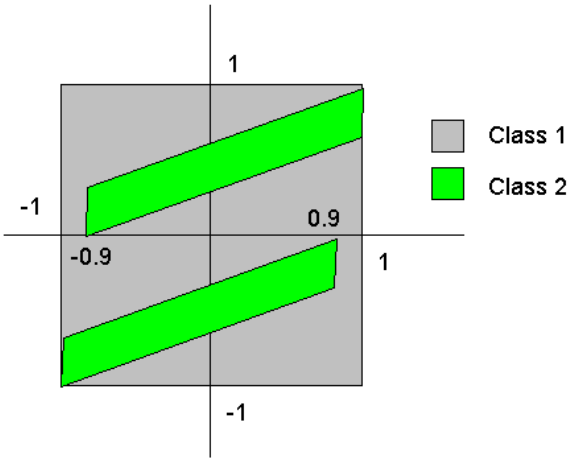

Figure 1. data set with binary response defined in [-1,1]x[-1,1]

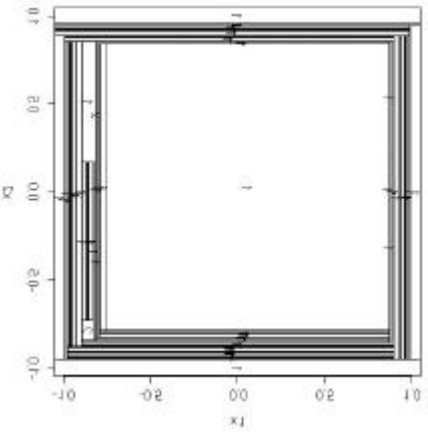

Figure 2. tree model by OSP

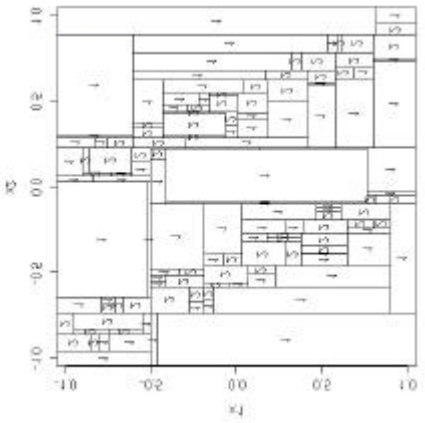

Figure 3. tree model by gini index

3. Exploration clusters by hybrid splitting criterion

Since OSP only concerns the smaller value among two gini indices of child nodes, we loose the information from the node with high impurity. In *Figure* 4, we plot the value of gini index tracking only the nodes with bigger gini index value at each split. At the node 51 where the jump of impurity happens, we change the splitting criterion from OSP to EBSC. That is, employing OSP for peeling out pure cluster, we have found impure cluster. Then, gini criterion is applied to the rest of the cluster.

Figure 5 is the result of tree model grown by hybrid splitting rule, Gini-after-OSP. Misclassification error rate of Gini-after-OSP (0.039) is slightly lower than that (0.042) of gini only, but clearly much lower than 0.187 of OSP only.

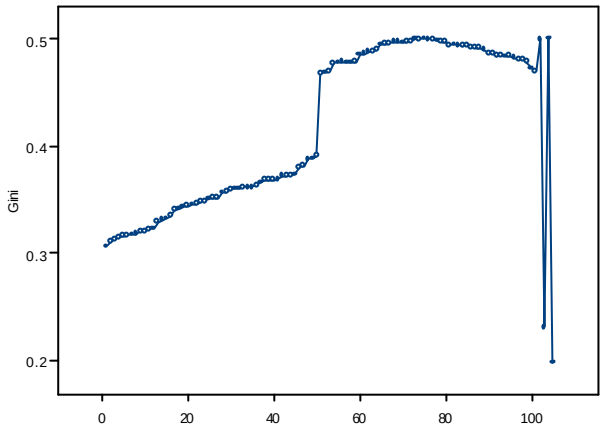

Figure 4. plot of gini index tracking only bigger index

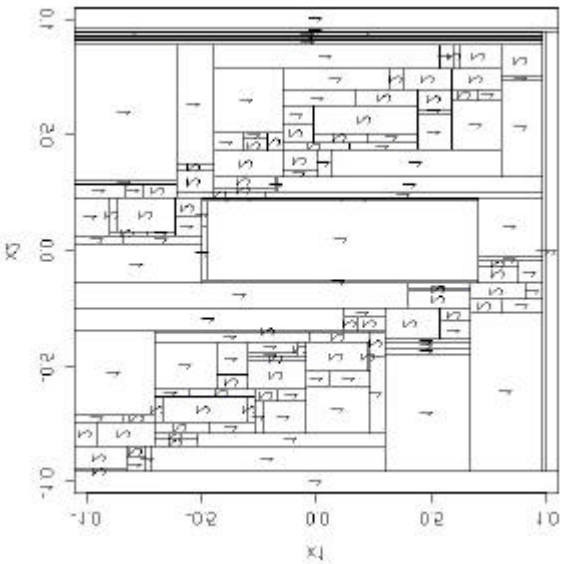

Figure 5. tree model by hybrid splitting criterion

REFERENCE

[1] Breiman, L., J. H. Friedman, R. A. Olshen, C. J. Stone (1984), *Classification and regression trees*, Wadsworth International

[2] Buja, A., and Y.-S. Lee (1999), "Data mining criteria for tree-based regression and classification", http://www.research.att.com/~andreas/papers/trees.ps.gz

[3] Friedman, J. H., and N. I. Fisher (1997), "Bump hunting in high-dimensional data", *Tech. report, Dept. of Staitsitcs, Stanford University.*

[4] Quilan, J. R., *C4.5: programs for machine learning* (1993), Morgan Kaufmann

RESUME

Daewoo Choi
Associate Professor, Dept. of Statistics, Hankuk Univ. of FS, Korea
Ph.D., Dept. of Statistics, Rutgers Univ., U.S.A.
M.S., Dept. of Statistics, Seoul National Univ., Korea
B.S., Dept. of Statistics, Seoul National Univ., Korea

Joongsik Yoon
Graduate Student, Dept. of Statistics, Hankuk Univ. of FS, Korea
B.S., Dept. of Statistics, Hankuk Univ. of FS, Korea