Optimal Data Analysis
Vol. 8 (January 22, 2019), 30-32

2155-0182/10/$3.00

Growing Classification Tree Models on the Basis of *a Priori* Performance Criteria

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

Growth of classification tree models based upon a secondary *a priori* performance criterion is demonstrated using a suboptimal classification tree model previously developed using the CHAID algorithm.

Growth of classification tree analysis (CTA) models[1] developed via hierarchically[2], enumerated[3] or globally-optimal[4] algorithms is usually governed by accuracy assessed using the Effect Strength for Sensitivity (ESS, norming accuracy *vs*. chance) or Percent of Accurate Classification (PAC, maximizing overall accuracy) objective functions.[5-8] CTA model growth may also be governed by a secondary objective function, for example achieving a criterion performance level which is specified *a priori*.

This paper reviews an example of a dual objective function methodology previously used to obtain a suboptimal classification tree model using the CHAID[9] (i.e., CHi-squared Automatic Interaction Detector) algorithm.[10] This early and innovative study proposed: "by employing two decision thresholds for identifying high- and low-risk cases—instead of the standard single threshold—the use of actuarial tools to make dichotomous risk classification decisions may be further enhanced" (p. 83).

CHAID assesses statistical significance of bivariate associations between every attribute and the class variable: once the "best" predictor (i.e., effect having lowest associated *P*-value) is identified, and the sample is partitioned accordingly, the procedure is repeated until no further partitioning is possible. Whereas CHAID does *not* offer a pruning algorithm to inhibit model over-fitting as well as to explicitly maximize classification accuracy (e.g., ESS, D, PAC), this may easily be accomplished vis-à-vis optimal pruning methodology.[5,6]

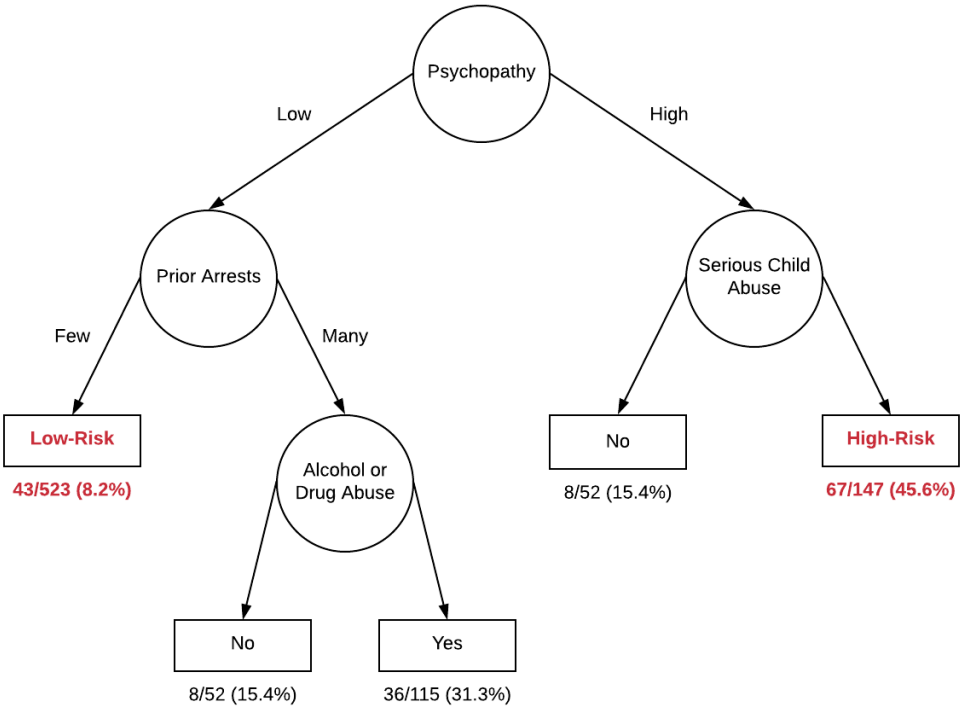
The secondary performance criterion was selected on the basis of the base rate of violence for the sample studied. Described by the authors: "The prevalence rate of violence during the first 20 weeks after hospital discharge for the full sample (939 discharged patients) was 18.7% (i.e., 18.7% of the patients committed at least one violent act during the first 20 weeks following hospital discharge). We defined any case assigned a predicted probability of violence (assessed using logistic regression analysis) that was greater than *twice* the base rate prevalence rate (>37%) as in the "high-risk" category, and any case whose predicted probability of violence was less than half the base prevalence rate (<9%) as in the "low-risk" category" (p. 89).

The original study first grew the full model possible using the CHAID algorithm: the final model employed a total of 13 attributes. Once the model was grown, the performance criterion was applied to identify the high- and low-risk cases.[10]

2155-0182/10/$3.00

The present study adjusts the original methodology: rather than first fully growing the model and then back-tracking to apply the secondary performance criterion, the performance criterion is instead evaluated at every step in the process of model development. This methodology yielded the model which is illustrated in Figure 1.

Figure 1: Modified CHAID Model Obtained by Applying the Secondary Performance Criterion at Every Step During Model Growth



Endpoints indicated using red font met the performance criteria for high- or low-risk strata—and thus application of the tree growth algorithm is terminated at these nodes. This methodology classifies 523+147=670 or 71.4% of the total of 939 patients—a 25% increase in number of classified patients when compared to the total of 57.1% of observations classified by the original methodology.

In the original article patients who were *not* classified in the first iteration formed a separate subsample which was subjected to the classification methodology a second time, using attributes which didn't enter the model in the first iteration. This procedure is reminiscent of optimal structural decomposition methodology which is used *vs*. log-linear modeling[11-13] and in novometric analysis.[14-16]

References

[1]Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, *56*, 656-667.

[2]Yarnold PR, Bryant FB (2015). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis*, *4*, 36-53.

2155-0182/10/$3.00

[3]Yarnold PR, Bryant FB (2015). Obtaining an enumerated CTA model via automated CTA software. *Optimal Data Analysis, 4*, 54-60.

[4]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, *3*, 78-84.

[5]Yarnold PR, Soltysik RC (2010). Maximizing the accuracy of classification trees by optimal pruning. *Optimal Data Analysis*, *1*, 23-29.

[6]Yarnold PR (2016). Pruning CTA models to maximize PAC. *Optimal Data Analysis, 5*, 58-61.

[7]Yarnold PR (2018). Objective functions optimized in ODA. *Optimal Data Analysis, 7*, 10-11.

[8]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis, 6*, 26-42.

[9]SPSS, Inc. (1993). *SPSS for Windows CHAID* (*Release 6.0*). Chicago, IL: SPSS, Inc.

[10]Steadman HJ, Silver E, Monahan J, Appelbaum PS, Robbins PC, Mulvey EP, Grisso T, Roth LH, Banks S (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior*, *24*, 83-100.

[11]Yarnold PR (2010). GenUniODA *vs*. log-linear model: Modeling discrimination in organizations. *Optimal Data Analysis*, *1*, 59-61.

[12]Yarnold PR (2015). UniODA-based structural decomposition *vs*. log-linear model: Statics and dynamics of intergenerational class mobility. *Optimal Data Analysis, 4*, 179-181.

[13]Yarnold PR (2016). Pairwise comparisons using UniODA *vs. not* log-linear model: Ethnic group and schooling in the 1980 Census. *Optimal Data Analysis*, *5*, 19-23.

[14]Yarnold PR (2016). Novometric *vs*. log-linear model: Intergenerational occupational mobility of white American men. *Optimal Data Analysis, 5*, 218-222.

[15]Yarnold PR (2016). Novometric *vs*. logit *vs*. probit analysis: Using gender and race to predict if adolescents ever had sexual intercourse. *Optimal Data Analysis, 5*, 223-224.

[16]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

Author Notes

No conflict of interest was reported.