Optimal Data Analysis
Vol. 5 (June 2, 2016), 53-57

2155-0182/10/$3.00

Identifying the Descendant Family of HO-CTA Models by using the Minimum Denominator Selection Algorithm: Maximizing ESS versus PAC

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

Usually it is possible to identify numerous different hierarchically-optimal classification tree analysis (HO-CTA) models in applications having an adequate sample size and involving multiple attributes. The models differ in complexity—defined as the number of endpoints representing distinct sample strata: the fewer the number of strata, the more parsimonious the model. The different models also vary in normed predictive accuracy—defined as effect strength for sensitivity (ESS): 0 represents the predictive accuracy expected by chance for the application; 100 represents errorless prediction. The distance of each model from a theoretically ideal solution for the application—defined as a model having perfect accuracy and minimum complexity, is computed as a *D* statistic. The underlying descendant family of models (including the globally-optimal model with the lowest *D* possible by an HO-CTA model for the application) is identified by first obtaining a model without specifying minimum endpoint sample size, and then applying the minimum denominator selection algorithm (MDSA).[1] These methods are illustrated in an application seeking to identify aspects of nursing care delivered to patients that predict satisfaction among 1,045 strongly satisfied and 671 moderately satisfied Emergency Department (ED) patients.[2] HO-CTA models that explicitly maximize ESS versus the overall percentage of accurate classification (PAC) are contrasted.

Discovered in 1996, hierarchically optimal classification tree analysis[3] (HO-CTA) proved its merit by identifying the most accurate models in various fields of application.[1,4] The accuracy and generalizability over time of early HO-CTA models is particularly impressive considering that many were published prior to discovery of pruning to explicitly maximize the ESS achieved by the model: thus the models were likely neither as accurate nor as parsimonious as would have been the case had these features been explicitly maximized.[1,5]

As is the style across disciplines and analytic methods[6,7], applied research using HO-CTA typically reports a single model that meets the methodological criteria of such modeling, as well as the theoretical criteria of substantive aspects of the research. Furthermore, manual construction of an HO-CTA model using either UniODA[8] or MultiODA[9-11] software is an exacting process involving multiple analytic operations.[12] Using programmable (automated) CTA software[13] however, HO-CTA models are easily obtained, as is demonstrated herein.

However, newly-discovered novometric theory asserts that a *descendant family* consisting of multiple optimal models may exist for a given application.[1] Models within this family vary in terms of their comparative parsimony and accuracy. The different members (models) in the family are discovered using MDSA.[1] Final selection of the primary (featured) model is then based on analytic (e.g., D) as well as substantive (e.g., qualitative findings) features of the family of models.

Data for this exposition were obtained from patients who received care in the ED of a private Midwestern hospital, and were subsequently mailed and returned a completed satisfaction survey.[2] Items were answered using the identical five-point Likert-type scale: 1=very poor, 2=poor, 3=fair, 4=good, 5=very good. The class variable ("sat") was overall satisfaction with care received in the ED: data from 671 patients rating their overall satisfaction as good were compared with data from 1,045 patients rating overall satisfaction as very good. Attributes were patient ratings of six aspects of nurse care behavior: V1=courtesy; V2=took patient's problem seriously; V3=attention paid to patient; V4=concern to keep patient informed; V5=concern for patient privacy; V6= technical skill.

Maximizing ESS

The first step of the MDSA is accomplished by using CTA software[1,13] command syntax below (to explicitly maximize ESS, weighting by prior odds is turned on[1]):

```
OUTPUT nurse.out;
OPEN nurse.dat;
VARS sat v1 v2 v3 v4 v5 v6;
CLASS sat;
ATTR v1 v2 v3 v4 v5 v6;
MISSING all (-9);
PRIORS ON;
MC ITER 10000 CUTOFF .05 STOP 99.9;
PRUNE .05;
GO;
```

CTA software output lists: (a) the fully-grown HO-CTA model after pruning via a sequentially-rejective Sidak Bonferroni-type multiple comparisons procedure criterion[1,8] to ensure the desired experimentwise Type I error rate ($p < 0.05$); and (b) the CTA model in (a) after additional pruning to explicitly maximize ESS.[1,5] Summarized in Table 1, every model that *wasn't* pruned to maximize ESS had lower ESS, efficiency, and PAC than the corresponding model that *was* pruned to maximize ESS.

Only HO-CTA models that are pruned for experimentwise p as well as for maximum ESS are used in the MDSA.

In the first step the smallest N in any of the model endpoints was 26 observations (Table 1), so the second model in the descendant family is identified by forcing the CTA algorithm to increase the minimum endpoint N by at least one observation:

```
MINDENOM 27;
```

In the third step the smallest N in any model endpoint was 36 observations, so the third and in the present case final model in the descendant family[1] is identified by forcing the CTA algorithm to increase the minimum endpoint N by at least one observation:

```
MINDENOM 37;
```

Table 1: Summary of MDSA Procedure for Discriminating Patients who are Very *vs*. Moderately Satisfied: Maximize ESS

Step	Pruned	Strata	MinD	ESS	Eff	PAC	*D*
1	No	12	10	27.1	2.3	70.2	32.3
	Yes	5	26	58.1	11.6	77.2	3.6
2	No	8	36	21.6	2.7	68.8	29.1
	Yes	4	36	58.0	14.5	77.5	2.9
3	No	7	44	20.0	2.1	66.7	28.1
	Yes	2	769	56.4	20.2	77.3	1.5

Note: There were three steps in this MDSA. Models that were pruned explicitly maximized ESS. Strata is the number of endpoints identified by the CTA model. MinD is the smallest number of observations (patients) in any of the strata (i.e., the smallest model endpoint N). ESS is a normed index of classification accuracy on which 0 represents the level of accuracy expected by chance and 100 represents perfect (errorless) classification. By rule-of-thumb: ESS<25 is a relatively weak effect; ESS<50 is a moderate effect; ESS<75 is a relatively strong effect; and ESS>75 is a strong effect.[8] Efficiency, an index of parsimony, is ESS/number of strata. PAC is the overall proportion of the sample correctly predicted by the model. *D* is the number of additional equivalent effects needed to achieve a theoretically ideal model for this application.[1] Exact discrete 95% CIs for the model and chance aren't provided in this exposition—the focus here is on method used to identify the descendant family.[1,14]

Maximizing PAC

The first step of the MDSA is accomplished by using CTA software command syntax below (to explicitly maximize PAC, weighting by prior odds is turned off[1]):

```
OUTPUT nurse.out;
OPEN nurse.dat;
VARS sat v1 v2 v3 v4 v5 v6;
CLASS sat;
ATTR v1 v2 v3 v4 v5 v6;
MISSING all (-9);
PRIORS OFF;
MC ITER 10000 CUTOFF .05 STOP 99.9;
PRUNE .05;
GO;
```

In the first step the smallest N in any of the model endpoints was 2 observations (Table 2), so the second model in the descendant family is identified by forcing the CTA algorithm to increase the minimum endpoint N by at least one observation:

```
MINDENOM 3;
```

In the third and final step the smallest N in any model endpoint was 35 observations, so the third and in the present case final model in the descendant family is identified by forcing the CTA algorithm to increase the minimum endpoint N by at least one observation:

```
MINDENOM 36;
```

Table 2: Summary of MDSA Procedure for Discriminating Patients who are Very *vs*. Moderately Satisfied: Maximize PAC

Step	Pruned	Strata	MinD	ESS	Eff	PAC	*D*
1	No	5	2	55.5	11.1	77.8	4.0
	Yes	4	2	57.0	14.2	77.1	3.0
2	No	6	3	55.1	9.2	77.1	4.9
	Yes	3	35	56.9	19.0	77.1	2.3
3	No	3	99	54.6	18.2	78.0	2.5
	Yes	2	769	56.4	28.2	77.3	1.5

Note: See notes to Table 1.

Comparing HO-CTA Models Maximizing ESS *vs*. PAC

Considering only the models pruned for maximum accuracy, in Step 1 of the MDSA the PAC and ESS for the maximum ESS model (77.2, 58.1, respectively) was greater than PAC and ESS for the maximum PAC model (77.1, 57.0). However, whereas the former model uses five strata (*D*=3.6) the latter model uses two strata (*D*=3.0). Maximum ESS and maximum PAC models differed because they optimized different objective functions (ESS versus PAC).

In Step 2 of the MDSA the PAC and ESS for the maximum ESS model (77.5, 58.0)

Copyright 2016 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

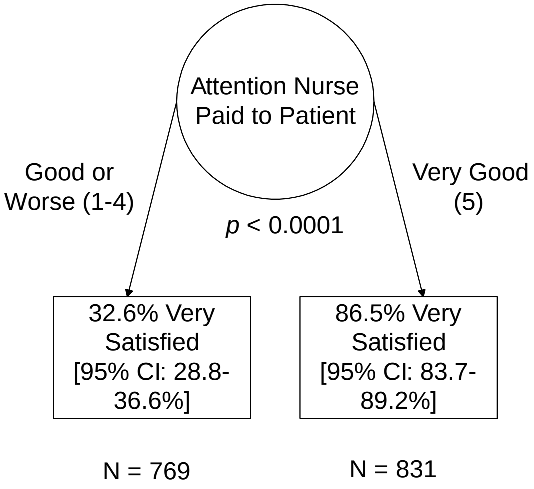
was greater than PAC and ESS for the maximum PAC model (77.1, 56.9): the former model uses four strata (*D*=2.9) the latter model uses two strata (*D*=2.3).

In Step 3 of the MSDA the maximum ESS and maximum PAC models converged to an elemental GO-CTA model with two strata: *D*= 1.5, PAC=77.3, ESS=56.4.

Maximum ESS and PAC GO-CTA Model

Figure 1 presents the elemental (two-strata) model identified on the basis of having the smallest *D* statistic for analyses maximizing ESS and for analyses maximizing PAC. This research suggests that the critical attribute in discriminating very satisfied versus moderately satisfied ED patients is the amount of attention that the nurse paid to the patient.

Figure 1: Two-Strata Model for Discriminating Patients who are Strongly versus Moderately Satisfied with Care Received in the ED


Attention Nurse Paid to Patient
Good or Worse (1-4)
Very Good (5)
p < 0.0001
32.6% Very Satisfied [95% CI: 28.8-36.6%]
86.5% Very Satisfied [95% CI: 83.7-89.2%]
N = 769
N = 831

To increase the number of very satisfied patients, and reduce the number of moderately satisfied patients, the model indicates that the nurses should maximize the number of patients who rate nurse attention as very good, and minimize the number of patients who rate nurse attention lower. The right-hand endpoint achieves homogeneity of almost 90%, which approaches the limits of measurement reliability for survey ratings if for no other reason than that of a ceiling effect. What is needed to improve this model is one additional currently unmeasured attribute loading on the left-hand branch of the CTA model, rather than (as now) having the branch end in a terminal endpoint. If this additional attribute results in one endpoint with 90% correct classification of very satisfied patients, and the other endpoint with 90% correct classification of less satisfied patients, then the resulting two-attribute, three-strata model would approach the rarified territory of a theoretical statistically ideal model—close to the absolute limits of the precision and validity of the survey methodology used presently.[1,15]

References

[1] Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago: ODA Books.

[2] Yarnold PR (2014). What most satisfies Emergency Department patients? *Optimal Data Analysis*, *3*, 98-101.

[3] 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.

[4] Yarnold PR (2013). Initial use of hierarchically optimal classification tree analysis in medical research. *Optimal Data Analysis*, *2*, 7-18.

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

[6] Grimm LG, Yarnold PR (Eds.). *Reading and Understanding More Multivariate Statistics*. Washington, DC: APA Books, 2000.

2155-0182/10/$3.00

[7] Grimm LG, Yarnold PR (Eds.). *Reading and Understanding Multivariate Statistics*. Washington, DC: APA Books, 1995.

[8] Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

[9] Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, *2*, 194-197.

[10] Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the wheat. *Optimal Data Analysis*, *2*, 202-205.

[11] Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, *2*, 220-221.

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

[13] Soltysik RC, Yarnold PR (2010). Introduction to automated CTA software. *Optimal Data Analysis*, *1*, 144-160.

[14] Yarnold PR, Soltysik RC (2014). Discrete 95% confidence intervals for ODA model- and chance-based classifications. *Optimal Data Analysis*, *3*, 110-112.

[15] Yarnold PR (2014). Triage algorithm for chest radiography for community-acquired pneumonia of Emergency Department patients: Missing data cripples research. *Optimal Data Analysis*, *3*, 102-106.

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
USA