Optimal Data Analysis
2013, Vol. 2, Release 1 (September 27, 2013), 31-36

2155-0182/11/$3.00

Assessing Hold-Out Validity of CTA Models Using UniODA

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

The use of UniODA software to assess hold-out validity of a CTA model is demonstrated using both raw and standardized data.

A commonly-used, straightforward method of assessing the classification error rate of a CTA model is called the hold-out, one-sample jackknife, cross-generalizability, or cross-validation procedure, and it essentially involves attempting to replicate a finding using an independent random sample.[1] For applications with large samples investigators sometimes randomly split the total sample into halves, selecting one to use as the training sample (to develop the model) and the other for the hold-out validity sample. To estimate hold-out validity, first develop a CTA model using a training sample, and then use the CTA model to classify observations in one or more independent hold-out samples. The classification error rate for the hold-out sample(s) is used as the estimated hold-out classification error rate for the model.[2]

This methodology is illustrated with data from a study using information available prior to hospital admission to develop a staging system for categorizing in-hospital mortality risk of HIV-associated community-acquired pneumonia (CAP).[3] Data were acquired using retrospective review of medical records of 1,415 patients with HIV-associated CAP, hospitalized in 1995-1997 at 86 hospitals in seven metropolitan areas. The sample was randomly halved. One half-sample was randomly selected to be the *training sample* and used to develop the CTA model in Figure 1: classification performance met the criterion for a moderately strong result.[1] As seen in Table 1, the two-node model accurately predicted the mortality status of 68% of patients who lived and 80% who died, and it was accurate in 97% of the cases predicted to live, and in 20% of the cases predicted to die.

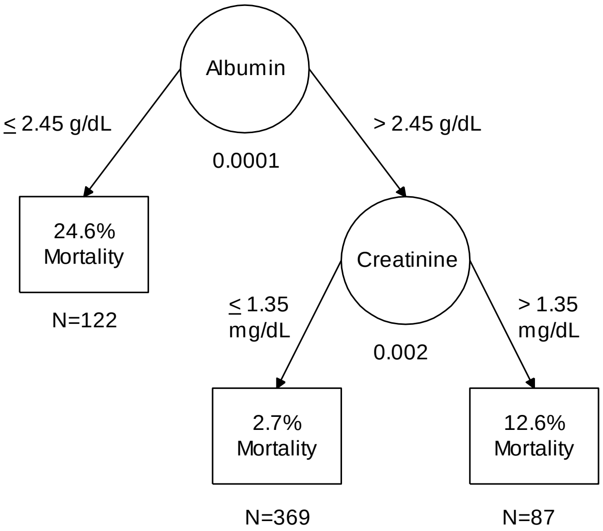

Figure 1: Training Sample CTA Model

The other half-sample is presently used as a *hold-out sample* and employed to assess the hold-out validity of the CTA model developed in training analysis.

Table 1: Confusion Table for Training Model

		Patient Predicted Status		
		Alive	Dead	
Patient Actual Status	Alive	359	168	68.1%
	Dead	10	41	80.4%
		97.3%	19.6%	

Validity analysis is conducted using the holdout functionality provided in UniODA software.[1] Using this procedure every model node is individually evaluated, starting at the root and working down model branches. Any discrepancies which may emerge between training and hold-out results will be discovered at their inception in the tree model.

Raw Score Method

Hold-out validity analyses are typically conducted using data recorded in their original metric (i.e., raw scores), which therefore is how this exposition begins. Table 2 gives descriptive statistics for the model attributes by sample.

Table 2: Descriptive Statistics for Model Attributes, Separately by Sample

Attribute	Sample	N	Mean	SD
Albumin	Training	580	3.02	0.82
	Hold-Out	541	3.09	0.79
Creatinine	Training	717	1.37	1.73
	Hold-Out	674	1.32	1.82

Note: Units are g/dL for albumin, and mg/dL for creatinine.

In the first step of the procedure the root node was evaluated via the following UniODA software[1] code: training.dat and holdout.dat are the training and hold-out datasets; “creat” is creatinine; 25,000 monte carlo experiments are run to estimate exact Type I error[4]; and software control commands are indicated in red.

```
open training.dat;
output holdout.out;
vars mortal albumin creat;
class mortal;
attr albumin;
mc iter 25000;
holdout holdout.dat;
go;
```

Program output for this analysis includes the UniODA model for the root attribute of the CTA model: if albumin≤2.45 g/dL then predict the patient died ($p<0.0001$, ESS= 41.4). Output also provides performance data for exactly this model applied to holdout data, and the results indicate that the root attribute was replicated: $p<$ 0.0001, ESS=40.5. Type I error for the hold-out results is obtained with one-tailed Fisher’s exact test (isomorphic with directional UniODA for a binary confusion table[1]): the *a priori* hypothesis is the training model will replicate when used to classify observations in the hold-out sample.

The second and final attribute was then evaluated by adding three lines of code:

```
in albumin>2.45;
attr creat;
go;
```

Program output for this analysis includes the UniODA model for the second attribute of the CTA model: if creatinine≤1.35 mg/dL then predict the patient lived. The output also gives performance data for exactly this model applied to holdout data, and results indicate this node was replicated: $p<0.004$, ESS=21.0. The overall hold-out validity performance (ESS=44.3) was 8.8% lower than was achieved for the training sample. Hold-out findings for the CTA model are illustrated in Figure 2.

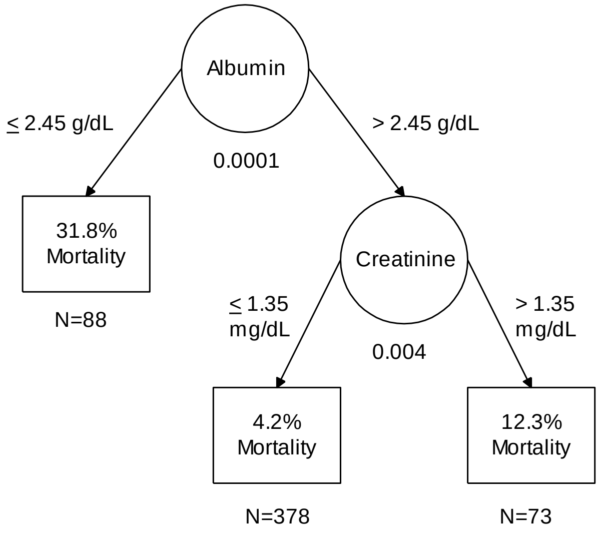

Figure 2: Classifying the Hold-Out Sample Using the Training Model and Raw Data

Summarized in Table 3, the CTA model accurately predicted mortality status of 74% of patients who lived (9% more accurate than the training model) and of 70% of patients who died (13% less accurate than training model). When the model predicted that a patient would live it was accurate in 96% of cases (2% less accurate than the training model), and in 23% of cases predicted to perish (17% more accurate than the training model).

Table 3: Confusion Table for Application of Training Model to Raw Hold-Out Data

		Patient Predicted Status		
		Alive	Dead	
Patient Actual Status	Alive	362	124	74.5%
	Dead	16	37	69.8%
		95.8%	23.0%	

It is natural to wonder what would occur in a split-half analysis if the training and validity data sets were switched, and data originally used in training were instead used in validation, and *vice versa*. This issue motivated development of a family of bootstrap methods which involve the iterative resampling of jackknife half-samples.[1] A variation of the one-sample jackknife known as *leave-one-out* is an efficient error estimation procedure available in UniODA and CTA software, in which each observation is a hold-out validity sample of size N=1.

A potentially serious problem involves a hold-out sample for which there is a significant mean difference in the attributes entering a CTA model as compared with the training data. For example, imagine that scores on an attribute are much higher in the training sample than in the hold-out sample. In this case, the cut-points obtained by CTA on an attribute for the training sample may similarly be much too high (or too low) to be *equivalently representative* in the hold-out sample. In this circumstance, transforming the original metric into a new sample-equivalent isometric, such as normative *z* scores, is necessary for a successful analysis.[1]

Between-sample mean differences noted presently were not extreme, suggesting that any potential gain in hold-out validity arising from separate standardization may be limited, in particular because the overall ESS for training and hold-out models based on raw data are relatively comparable. However, decline in classification performance of the model in hold-out analysis suggests the possibility of a gain in ESS, so this method is demonstrated next.

Standardized Score Method

Before beginning the validity analysis it is necessary to first normatively standardize[1] the attributes (here, albumin and creatinine):

$$z_i = (X_i - \text{Mean}) / \text{SD}, \qquad (1)$$

where z_i is the normative standard score (mean= 0, SD=1) and X_i is the raw score of the *i*th observation, and Mean and SD are computed for the sample from which the observation is drawn.

Copyright 2013 by Optimal Data Analysis, LLC
2155-0182/11/$3.00

This was done for observations in the training sample using parameters (Mean, SD) obtained for the training sample, and for observations in the hold-out sample using parameters obtained for the hold-out sample (see Table 2). In order to facilitate clarity, fewer significant digits are used herein than were used in software code, in which eight significant digits were employed to minimize round-off error. The minimally sufficient number of significant digits to use in code is determined when standardizing attributes, as the number of digits which is required to produce a sample having Mean=0, and SD=1.

In the first step of the procedure the root node was evaluated with the following UniODA code: ztrain.dat and zholdout.dat are training and hold-out datasets, zalbumin is z_i for albumin, and zcreat is z_i for creatinine.

```
open ztrain.dat;
output zexample.out;
vars mortal zalbumin zcreat;
class mortal;
attr zalbumin;
mc iter 25000;
holdout zholdout.dat;
go;
```

Program output for this analysis gives the UniODA model for the root attribute of the CTA model: if zalbumin≤ -0.695 g/dL then predict the patient died (p<0.0001, ESS=41.4). The standardized cut-point may also be computed by using (1): standard cutpoint=(2.45-3.02)/0.82.

Output also reports performance data for this model applied to holdout data, and results indicate that the root attribute was replicated (p< 0.0001). However, using an equally representative cut-point resulted in 11.5% greater overall accuracy at the root node (ESS= 45.2) compared with analysis using raw scores.

The second and final attribute was then evaluated by adding three lines of code:

```
in zalbumin>-0.695;
attr zcreat;
go;
```

Program output for this analysis includes the UniODA model for the second attribute of the CTA model: if zcreat ≤ -0.0091 mg/dL [i.e., (1.35-1.37)/1.73] then predict the patient lived. This is exactly the same finding as was derived for raw data for the training sample, because the raw and standardized cut-points have precisely the same relative value for training data.

The output also gives performance data for exactly this model applied to holdout data, and results indicate this node was replicated: p< 0.04, ESS=18.6: this is 11.3% lower than the corresponding effect for raw data at this node.

Overall training performance of the CTA model was identical with raw and standardized data. In contrast, hold-out validity performance obtained with standardized data (ESS=45.9) was 3.7% greater than achieved using raw data. The hold-out validity findings obtained for standard data are illustrated in Figure 3, and summarized in Table 4.

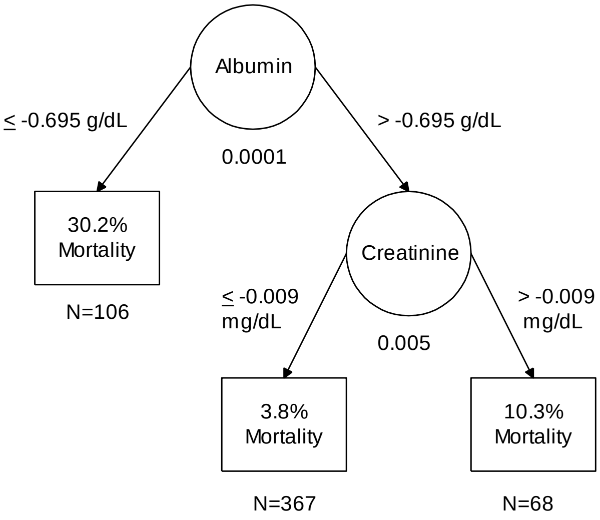

Figure 3: Classifying the Hold-Out Sample
With the Training Model and Standardized Data

Table 4: Confusion Table for Application of Training Model for Standardized Data to Hold-Out Sample

		Patient Predicted Status		
		Alive	Dead	
Patient Actual	Alive	353	135	72.3%
Status	Dead	14	39	73.6%
		96.2%	22.4%	

As seen, when compared to the hold-out results obtained using raw data, the CTA model for standardized data accurately predicted the mortality status of 72% of the patients who lived (3% less accurate) and 74% of patients who died (5% more accurate). When the model predicted that a patient would live it was accurate in 96% of cases (0.5% more accurate), and 22% of the cases predicted to perish (1% less accurate than the training model).

Discussion

Separate standardization of attributes in training and all hold-out samples is a recommended practice, and is a safeguard against unwitting induction of Simpson's paradox.[5] If model statistics are desired in their raw units of measure, standardized units may easily be reconverted. The use of standard scores provides information about relative magnitude of model cut-points. For example, for creatinine the cut-point was virtually zero, which indicates a value approximating the sample Mean. For raw data, the albumin cutpoint of -.70 falls beneath the sample Mean, representing the 24th percentile if the sample is normally-distributed.

What can be done if at some point in the hold-out validity assessment process, the finding for a model node fails to replicate? First, all nodes on any branches emanating from the non-replicating node should be evaluated to assess the full extent of replication failure.

Second, UniODA may be used in exploration of a cut-point value for the node (on the attribute), optimized for the hold-out sample, in an effort to retain integral training model geometry by relaxing parameter estimates (cut-point values). Similar parameter relaxation may be required in the tree from that point to the end of emanating branches.

Third, use of the gen (generalizability) command in UniODA software frees all of the samples to compromise optimal parameters in pursuit of a single model which achieves an operator-defined "minimally acceptable performance" level in the worst case. That is, a "middle-ground" model for which the worst performance observed for any sample exceeds the criterion for unacceptable weakness. When gen was applied to the present data the hold-out model for raw data was identified.

Finally, all else failing, exploratory CTA comparing the training and hold-out samples on all common attributes should be conducted so as to characterize inter-sample differences

References

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

[2]Stone M (1974). Cross-validatory choice and assessment of statistical problems. *Journal of the Royal Statistical Society*, *36*, 111-147.

[3]Arozullah AM, Parada J, Bennett CL, Deloria-Knoll M, Chmiel JS, Phan L, Yarnold PR (2003). A rapid staging system for predicting mortality from HIV-associated community-acquired pneumonia. *Chest*, *123*: 1151-1160.

[4]Yarnold PR, Soltysik RC (2010). Precision and convergence of Monte Carlo estimation of two-category UniODA two-tailed *p*. *Optimal Data Analysis*, *1*: 43-45.

[5] Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, *56*, 430-442.

Author Notes

Correspondence by mail should be sent to: Optimal Data Analysis LLC, 1220 Rosecrans St., #330, San Diego, CA 92106. Send e-mail to: Journal@OptimalDataAnalysis.com