Optimal Data Analysis
Vol. 8 (April 10, 2019), 76-80

2155-0182/10/$3.00

Effect of Sample Size on Discovery of Relationships in Random Data by Classification Algorithms

Ariel Linden, Dr.P.H. and Paul R. Yarnold, Ph.D.
Linden Consulting Group, LLC Optimal Data Analysis, LLC

In a recent paper, we assessed the ability of several classification algorithms (logistic regression, random forests, boosted regression, support vector machines, and classification tree analysis [CTA]) to correctly *not* identify a relationship between the dependent variable and ten covariates generated completely at random. Only classification tree analysis correctly observed that no relationship existed. In this study, we examine whether various randomly derived subsets of the original N=1000 dataset change the ability of these models to correctly observe that no relationship exists. The randomly drawn samples were 250 and 500 observations. We further test the hold-out validity of these models by applying the generated model's logic onto the remaining sample and computing the area under the receiver operator's characteristics curve (AUC). Our results indicate that limiting the sample size has no effect on whether classification algorithms correctly determine that a relationship does not exist between variables in randomly generated data. Only CTA consistently identified that the data were random.

Predictive accuracy achieved using CTA was recently compared with accuracy obtained by other classification algorithms including logistic regression (LR), random forests (RF), boosted regression (BR), and support vector machines (SVM).[1] In that study, an artificial dataset was constructed with 500 "group 1" and 500 "group 2" observations as well as ten randomly generated continuous covariates (attributes)—which by design have no correlation with the binary dependent (class) variable. Of all the algorithms tested only CTA *correctly failed* to discriminate between groups using random covariates.

Table 1: Initial Results[1] for N=1000

Model	AUC	LCL	UCL
Logistic	0.567	0.531	0.602
BR	0.795	0.768	0.823
SVM	0.571	0.537	0.605
RF	1.000	1.000	1.000
CTA	---	---	---

Note: AUC=Area Under (ROC) Curve (0.50=chance, 1.0= perfect accuracy); LCL=Lower 95% Confidence Level; UCL=Upper 95% Confidence Level; dashes indicate no model was obtained.

2155-0182/10/$3.00

The present study examines whether this finding holds in smaller samples (and hold-out validity) for random samples of 50% (N= 500) and 25% (N=250), randomly drawn from the original sample of N=1000.[1]

The N=250 subsample, which is treated as a training sample and is used to create new models by each method, was created by randomly drawing 250 observations from the original sample (with roughly equal sizes per group). The remaining N=750 cohort was treated as a hold-out sample and used to assess the cross-generalizability of the training model when used to classify an independent random sample.[2]

The N=500 subsample, similarly treated as a training sample and used to identify new models by each method, was created by randomly drawing 500 observations from the original sample (with roughly equal sizes per group). The remaining random N=500 cohort was treated as a hold-out sample and used to assess cross-generalizability of the training model.

The LR, RF, BR, SVM and CTA classification algorithms were applied to the N=250 and N=500 samples. In all models, ten covariates (attributes) were used to predict class status using the default parameters of the respective algorithm. A receiver operating characteristics (ROC) analysis was then conducted with actual class status used as the reference variable and predicted probabilities from respective models used as the classification variable. Area under the ROC curve (AUC) ranges from 0.50 to 1.0: a model yielding perfect discrimination of the two groups has AUC=1.0, and a model unable to predict group (class) status has AUC=0.50.[3]

Statistically significant training models were used to classify the remaining sample (of 750 and 500 observations, in the N=250 and N=500 hold-out samples, respectively).[2]

All machine learning models but CTA[4] were implemented by Stata statistical software version 15.1 (StataCorp, College Station, TX) with the following user-written packages; randomforest, svmachines, and boost.

Table 2: Training and Hold-Out Analysis Results

	In sample (Training)			Out of sample (Hold-Out)		
Model	AUC	LCL	UCL	AUC	LCL	UCL
LR 250	0.623	0.553	0.692	0.498	0.457	0.539
LR 500	0.579	0.529	0.629	0.549	0.498	0.599
BR 250	0.831	0.779	0.882	0.484	0.443	0.526
BR 500	0.815	0.777	0.852	0.491	0.440	0.542
SVM 250	0.633	0.565	0.702	0.502	0.460	0.543
SVM 500	0.580	0.531	0.629	0.548	0.498	0.597
RF 250	1.000	1.000	1.000	0.502	0.461	0.544
RF 500	1.000	1.000	1.000	0.526	0.475	0.577
CTA 250	---	---	---	---	---	---
CTA 500	---	---	---	---	---	---

Note: AUC=Area Under (the ROC) Curve (0.50=chance, 1.0= perfect accuracy); LCL= Lower 95% Confidence Level (bound); UCL=Upper 95% Confidence Level. CTA was unable to identify a training model thus no hold-out analysis is possible.

2155-0182/10/$3.00

Results of training and hold-out analyses are summarized in Table 2. As occurred in the initial N=1,000 experiment[1], CTA was the only algorithm evaluated which *correctly* identified that the data were random: no model could be generated for the N=250 or N=500 samples because no relationship exists between the covariates and the class variable. If there is no training model, then no hold-out validity analysis is possible.

And, conversely, as also occurred in the initial N=1,000 study, all four other algorithms generated models which discriminated between the two class categories using the covariates, in both the N=250 and N=500 samples (Table 2). However, no training models were validated in hold-out reproducibility analysis, as the LCL of AUC dropped below 0.50. The latter finding reaffirms the importance of conducting cross-generalizability analysis to assess the potential reproducibility of the finding.[5-7] In light of the chaos that jackknife instability creates[8,9] in a classification tree model, and the objective to identify reproducible effects, CTA models are generally constrained to only retain attributes stable in LOO analysis in the model.[4]

This paper supports the results derived in our previous analysis[1], identifying an important limitation of popular machine learning algorithms used for predicting binary outcomes (e.g., propensity scores). That is, they are likely to find relationships between variables which really don't exist, regardless of sample size.

These results should be independently replicated, and the limits of this phenomenon probed. For example, future research should evaluate the influence of the number of random attributes made available to algorithms, number of significant digits (scientific precision) of the attributes, and number of class levels in the model, in affecting training AUC. Additionally, research should investigate applications involving randomized categorical attributes with different numbers of levels.

These findings also provide one more reason why we strongly advocate using ODA and CTA frameworks to draw causal inferences about treatment effects in both observational data and in data from randomized controlled trials.[2, 3-26] Now, more than ever, it is time to consider revising the guidelines for how health improvement interventions are evaluated.[27,28]

References

[1]Linden A, Yarnold PR (2019). Some machine learning algorithms find relationships between variables when none exist -- CTA doesn't. *Optimal Data Analysis, 8*, 64-67.

[2]Linden A, Yarnold PR (2017). Modeling time-to-event (survival) data using classification tree analysis. *Journal of Evaluation in Clinical Practice, 23*, 1299-1308.

[3]Linden A (2006). Measuring diagnostic and predictive accuracy in disease management: An introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice, 12*, 132-139.

[4]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160.

[5]Yarnold PR, Soltysik RC (2005). *Optimal Data Analysis: Guidebook with Software for Windows*. Washington, D.C.: APA Books.

[6]Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

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

[8]Yarnold PR (2016). Using UniODA to determine the ESS of a CTA model in LOO analysis. *Optimal Data Analysis, 5*, 3-10.

2155-0182/10/$3.00

[9]Yarnold PR (2016). Determining jackknife ESS for a CTA model with chaotic instability. *Optimal Data Analysis*, *5*, 11-14.

[10]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*, *22*, 839-847.

[11]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, *22*, 848-854.

[12]Linden A, Yarnold PR (2016). Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*, *22*, 855-859.

[13]Linden A, Yarnold PR, Nallomothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*, *22*, 860-867.

[14]Yarnold PR, Linden A (2016). Using machine learning to model dose-response relationships via ODA: Eliminating response variable baseline variation by ipsative standardization. *Optimal Data Analysis*, *5*, 41-52.

[15]Linden A, Yarnold PR (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, *22*, 868-874.

[16]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, *22*, 875-885.

[17]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis. *Optimal Data Analysis*, *22*, 65-73.

[18]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *22*, 171-174.

[19]Yarnold PR, Linden A (2017). Computing propensity score weights for CTA models involving perfectly predicted endpoints. *Optimal Data Analysis*, *6*, 43-46.

[20]Linden A, Yarnold PR (2017). Minimizing imbalances on patient characteristics between treatment groups in randomized trials using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *23*, 1309-1315.

[21]Linden A, Yarnold PR (2018). The Australian gun buy-back program and the rate of suicide by firearm. *Optimal Data Analysis*, *7*, 28-35.

[22]Linden A, Yarnold PR (2018). Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of Evaluation in Clinical Practice*, *24*, 353-361.

[23]Linden A, Yarnold PR (2018). Estimating causal effects for survival (time-to-event) outcomes by combining classification tree analysis and propensity score weighting. *Journal of Evaluation in Clinical Practice*, *24*, 380-387.

[24]Linden A, Yarnold PR (2018). Using machine learning to evaluate treatment effects in multiple-group interrupted time series analysis. *Journal of Evaluation in Clinical Practice*, *24*, 740-744.

[25]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials. *Optimal Data Analysis*, *7*, 46-49.

[26]Linden A, Yarnold PR (2018). Using ODA in the evaluation of randomized controlled trials: Application to survival outcomes. *Optimal Data Analysis*, *7*, 50-53.

2155-0182/10/$3.00

[27]Linden A, Adams J, Roberts N (October, 2003). *Evaluation methods in disease management: determining program effectiveness*. Position Paper for the Disease Management Association of America (DMAA).

[28]Linden A, Roberts N (2005). A Users guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care*, 11, 81-90.

Author Notes

No conflict of interest was reported.