Optimal Data Analysis
Vol. 8 (March 21, 2019), 64-67

2155-0182/10/$3.00

Some Machine Learning Algorithms Find Relationships Between Variables When None Exist -- CTA Doesn't

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

Automated machine learning algorithms are widely promoted as the best approach for estimating propensity scores, because these methods detect patterns in the data which manual efforts fail to identify. If classification algorithms are indeed ideal for identifying relationships between treatment group participation and covariates which predict participation, then it stands to reason that these algorithms would also be unable to find relationships when none exist (i.e., covariates do *not* predict treatment group assignment). Accordingly, we compare the predictive accuracy of maximum-accuracy classification tree analysis (CTA) *vs*. classification algorithms most commonly used to obtain the propensity score (logistic regression, random forests, boosted regression, and support vector machines). However, here we use an artificial dataset in which ten continuous covariates are randomly generated and by design have no correlation with the binary dependent variable (i.e., treatment assignment). Among all of the algorithms tested, only CTA *correctly* failed to discriminate between treatment and control groups based on the covariates. These results lend further support to the use of CTA for generating propensity scores as an alternative to other common approaches which are currently in favor.

Recently, the use of classification algorithms has been promoted as an alternative to logistic regression for estimating the propensity score (i.e. the probability of being a participant in the treatment group in an observational study).[1-9] Machine learning classification algorithms find the best fitting model through automated processes which search through the data to detect patterns that may include interactions between variables, as well as interactions within subsets of variables. If indeed classification algorithms are ideal for identifying relationships between the dependent variable (e.g., treatment group status) and covariates which predict participation (e.g., demographic characteristics), then it stands to reason that these algorithms would also be unable to find relationships when none exist (i.e., covariates do *not* have any relationship with the dependent variable).

2155-0182/10/$3.00

In this paper we compare the predictive accuracy of CTA[9] to the accuracy achieved by some of the most commonly used classification algorithms for predicting the propensity score: logistic regression, random forests, boosted regression, and support vector machines.[1-9] However, as a novel twist, herein we use an artificial dataset in which ten continuous covariates are randomly generated and therefore designed to have no correlation with the binary dependent variable. In other words, we assess how well the alternative algorithms work when covariates should *not* predict treatment status.

Methods

Data

A simulated data set of 1000 observations was generated which included ten uniformly distributed random variates over the interval (0,1), and one binary variable representing the treatment group status (with 500 treated and 500 non-treated observations randomly sorted).

Analyses

Five classification algorithms were applied to the data (logistic regression, random forests, boosted regression, support vector machines, and CTA). In all models, the ten covariates (attributes) were used to predict treatment status (the class variable), using the default parameters of the respective algorithm.

A receiver operating characteristics (ROC) analysis was then conducted in which actual treatment status was set as the reference variable and the predicted probabilities from the respective models were set as the classification variable. The area under the ROC curve (AUC) can range from 0.50 to 1.0, where a model with perfect discriminatory ability will have an AUC of 1.0, while a model unable to distinguish between individuals with or without the outcome (treatment group status) will have an AUC of 0.50.[10]

Other than CTA, all machine learning models were implemented using Stata statistical software version 15.1 (StataCorp, College Station, Texas) with the following user-written packages; randomforest, svmachines, and boost. The CTA model was generated using the CTA software package.[11]

Results

CTA was the only algorithm to correctly identify that the data were random, that is, no model could be generated because no relationship exists between the covariates and the treatment indicator.

Conversely, all of the other algorithms generated models that were able to discriminate between treatment groups using the covariates. AUCs for the models are as follows (sorted from high to low): random forest = 1.00 (95% CI: 1.00, 1.00), boosted regression = 0.7953 (95% CI: 0.768, 0.823), support vector machines = 0.5714 (95% CI: 0.537, 0.605), and logistic regression 0.5665 (95% CI: 0.531, 0.602).

Discussion

This paper highlights an important limitation of popular machine learning algorithms used for generating propensity scores. That is, they are likely to find relationships between variables which really don't exist. In fact, in our artificial data, the random forest algorithm was able to perfectly discriminate between treatment and control groups, even though the data were completely random! While it is possible that tuning model parameters may improve the fit (or in this case, "no fit") to the data, in reality the investigator does not know the true nature of the relationship between variables. Of the algorithms used presently, only CTA was able to detect that there were no true relationships between variables.

In the context of propensity scoring, having a model which identifies erroneous

2155-0182/10/$3.00

relationships between covariates and treatment status suggests that the treatment groups are not truly balanced on observed characteristics (i.e. the groups are not comparable). In turn, by using erroneous propensity scores (either in matching or weighting), treatment effect estimates will remain as (or randomly more) biased than if no adjustment was made at all.

These results 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.[9-29] Perhaps the time has come to consider revising the guidelines for how health improvement interventions are evaluated.[30]

References

[1]Cook EF, Goldman L (1988). Asymmetric stratification. An outline for an efficient method for controlling confounding in cohort studies. *American Journal of Epidemiology*, *127*, 626-639.

[2]Barosi G, Ambrosetti A, Centra A, Falcone A, Finelli C, Foa P, et al. (1998). Splenectomy and risk of blast transformation in myelofibrosis with myeloid metaplasia. Italian Cooperative Study Group on Myeloid with Myeloid Metaplasia. *Blood*, *91*, 3630-3636.

[3]McCaffrey DF, Ridgeway G, Morral AR (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403-425.

[4]Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, *17*, 546-555.

[5]Lee BK, Lessler J, Stuart EA (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346.

[6]Westreich D, Lessler J, Funk MJ (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*, 826–833.

[7]Wyss R, Ellis AR, Brookhart MA, Girman CJ, Funk MJ, LoCasale R, Stürmer T (2014). The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology*, *180*, 645-55.

[8]Neugebauer R, Schmittdiel JA, van der Laan MJ (2016). A case study of the impact of data-adaptive versus model-based estimation of the propensity scores on causal inferences from three inverse probability weighting estimators. *The International Journal of Biostatistics*, *12*, 131-55.

[9]Linden A, Yarnold PR (2017). Using classification tree analysis to generate propensity score weights. *Journal of Evaluation in Clinical Practice*, *23*, 703-712.

[10]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.

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

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

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

2155-0182/10/$3.00

[14]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.

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

[16]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.

[17]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.

[18]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.

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

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

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

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

[23]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.

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

[25]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.

[26]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.

[27]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.

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

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

[30]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).

Author Notes

No conflict of interest was reported by either author.