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). 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).
In this paper we compare the predictive accuracy of CTA\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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
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. Perhaps the time has come to consider revising the guidelines for how health improvement interventions are evaluated.

References


**Author Notes**

No conflict of interest was reported by either author.