Machine learning methods for propensity score estimation

Yeonho Jung[1] and Kyungseon Lee[1]

[1] Department of Statistics, Seoul National University, Seoul, Republic of Korea, {dusgh9514, ppleeqq}@ snu.ac.kr
[2] Corresponding author

Abstract

Treatment and control are often unbalanced and uncontrollable in the observational study of the causal interface. In this case, a propensity score[2] can be given to resolve the data imbalance. In the process of estimating Average Treatment Effect(ATE) by estimating propensity score, we use a machine learning-enhanced technique inspired by the Improving Propensity Score Weighting Using Machine Learning paper[4]. This method enhances the propensity score weighting in a way that is highly useful in real data.

In this paper[4], they create data from seven scenarios and run simulations, and experiment with increasing extensions from the data where the ideal additivity and linearity assumptions are established to the non-linearity and non-additivity datasets similar to the real data. In the most ideal dataset, evaluation is better in all models when estimating propensity, but in the most realistic datasets, performance is better in Random Forest[1] or reinforced Classification And Regression Tree[5] (pruned-CART, bagged-CART, boosted-CART) than logistic regression. We will apply this models to real data, not simulation data, to see if performance is really good in real data.

First, we starts with a famous and well-refined datasets, Chemical and expands to an unrefined datasets, such as national nutrition health survey dataset and oo dataset. We aims to verify the arguments of this paper using same performance metrics which are covariate balance, standard error, percent absolute bias, and 95 percent confidence interval (CI) coverage as shown in this paper[4].

Keywords— *Propensity Score Weighting, Random Forest, CART*

I. INTRODUCTION

Recently, researchers are widely applying to observational studies for estimating the effects of treatments Zi, exposures Yi using covariates. The methods of using propensity score are used frequently to estimate the treatment effects in terms of observational data. For covariates Xi, we consider the conditional probability of receiving treatment, e(x) and We call it the propensity score. The e(X) is a balancing score of treated and controlled units having same propensity score that have similar distributions of observed covariates[3].

Many estimation methods have been developed by using e(X) and the trait of variable reduction technique in confounding adjustment is appealing in recent days. For example, the propensity score is widely used in estimation methods such as covariate adjustment, stratification or subclassification on the propensity score, matching and weighting techniques. Rubin says that an advantage of using propensity score methods is that those things make observational studies possible to be designed similar to randomized experiments. In other words, conditioning on the estimated propensity score has reduced observed systematic differences between the treated and the controlled.

Generally, several steps exist using propensity scores. First, we generate a propensity score for each person or unit, and then decide how we want to use the scores such as matching, subclassification or weighting[3]. Finally, we run optimal regression model or according to the types of response variables. In parametric models, statistical functional form, the types of distributions of variables and some specification terms need to be assumed. Unless the assumptions are right, covariate balance may not be achieved by conditioning on the propesntiy score resulting in a biased effect estimate consequently.

Propensity score are normally estimated using logistic regression, but there are some limitations when using real data and in some constraints. For that reason, machine learning methods can be an alternative by using diverse techniques of classification and prediction algorithms. Compared to the regression that requires statistical method to modeling when setting parameters for modeling, machine learning is an process which tries to find and extract the relationship among variables. Because of the traits 'black box' in machine learning which means it is difficult for interpreting results and setting algorithms, propensity score using machine learning are mostly researched especially in medical field.

Among several methods in machine learning, this report treat 3 methods to compare the effects and the results with the regression model in terms of weighing using estimated propensity score. Those methods are 'Classification and Regression Tress(CART)', additionally 'Pruned CART' and 'Random Forest', a kind of ensemble methods. Despite the increasing interest and popularity of the machine learning method, relatively little is approaching and known about what situations the outcomes of those methods and skills are more efficient than traditional regression methods[6]. We already know that in Monte-Carlo simulation experiments which is based on some linear and additivity situations, the results of linear regression model are somewhat superior compared to the other machine learning techniqes in terms of ASAM(Average Stadardized absolute mean difference), Bias, SE, CI coverage and weights. However, if the model is simulated with complexity with quadratic terms and kinds of interactions similar to the realistic data set, its estimation ability is relatively low. For this reason in the real world, the basic concepts of Decision trees can be an alternative. It divides a dataset into nodes such that within each nodes, observations are as homogeneous as. A 'CART' covers both meanings of whether the type of predicted outcome is a class or numerical and observations have similar probabilities within each node of the tree. Although CART has some disadvantage of overfitting, the 'pruned' techniques exists for reducing excessive tree splits to be less sensitive to noise and generalize to new data called test dataset. In addition to 'CART' and 'pruned CART' which treat a single tree as a classifier, we use ensemblem methods, which are related to iterative and bootstrap steps, using multiple trees and samplings. The Random Forest is one of ensemble methods that utilize and combine several models of prediction to improve accuracy. Its main concept is 'diversity' that produces a number of train data set from the original data set and 'random subspace' which chooses variables randomly when constructing decision tree models.

In this report, due to the limitation of calculating the true treatment effects, we use estimated values in the context of propensity score weighting. Using simulated data and two kinds of real data, we compare 'LR', 'CART', 'Pruned CART' and 'Random Forest(RF)' by the performance metrics or standard such as 'ASAM', 'Bias', 'SE', 'CI coverage', 'Weights' using plots and tables to show the efficiency of machine learning methods as an alternative to LR.[4] Although we can not utilize all the performances compared to the simulated data set because of unknown true treatment effects, we can roughly compare the results with several methods.

II. Main Framework

Keyword

- Propensity score : $e\left(X_i\right) = Pr(Z_i = 1|X_i)$
- $\widehat{\Delta}_{IPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i \cdot Y_i}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-Z_i)\cdot Y_i}{1-\hat{e}_i}$
- $\widehat{\Delta}_{SIPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i \cdot Y_i}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-Z_i)\cdot Y_i}{1-\hat{e}_i}$
- Methods :
  - Logistic regression
  - CART and Pruned CART
  - Random Forest

  This report aims for examining the use of machine learning as an alternative to logistic regression not only simulated data but also real data.
- Performace metrics
  - ASAM = $\frac{\left|\frac{\sum_{i=1}^{N} W_{ij} Z_i C_{ij}}{N_t} - \frac{\sum_{i=1}^{N} W_{ij}(1-Z_i) C_{ij}}{N_C}\right|}{\sigma_j}$

    * $\sigma_j = \frac{\sum_{i=1}^{N}\left(W_{ij} Z_i C_{ij} - \sum_{i=1}^{N} W_{ij} Z_i C_{ij}/N_t\right)^2}{N_t}$
  - Bias, SE, CI coverage, weight

Propensity score methods can be a tool of controlling for confounding in observational studies. Conditioning on observed covariates, the probability of receiving a treatment called propensity score become a scalar value. Several methods such as regression adjustment, matching, stratification and weighting are introduced with respect to propensity score, but matching and weighting are being widely used and appealing in some instances among them. By studying the knowledge and the traits in the lecture, matching may be an effective tool for being designed to randomized experiments but in some cases, it can not use all of the data points(units) and sometimes should discard the unmatched data. Unlike matching, propensity score weighting is similar with survey sampling weighting, giving an description for more or less sampling by weighing the units to account for the population where the units are drawn. The most widely used value is inverse probability weighting(IPW), which uses propensity score in the context of probability. The treated and controlled units are weighted to be designed similar to randomized experiment. Additionally, the limitations of estimated IPW for being close to 0 and 1 in some cases, the weight, IPW, can be extremely large dominating the $\widehat{\Delta}_{IPW}$ and highly unstable. For an alternative, Stabilized inverse probability weighting is used for weighted average of Y additional weights forming bounds. The main keyword is 'weight' in terms of propensity score. However, as we do not know the true propensity score, the average treatment effect(ATE) in the real world within any observational data-set, they are needed to be estimated along with (S)IPW. The simulation framework setup cited by Setoguchi and colleagues consists of N=500, 1000 and 2000 observations with a binary exposure, continuous outcome and several covariates with 7 scenarios A to G. Those covariates are made up of standard normal random variables, some interaction and quadratic terms[7]. The properties of linearity or non-linearity and additivity or non-additivity would make the results comparable using several methods such as logistic regression, CART or pruned CART and some ensemble techniques. In the section 'Experiment', those values will be compared by several methods in addition to the real data-set. Propensity scores are generally estimated using logistic regression. However, owing to the parametric models requiring assumptions for the functional form, distribution of variables, if any of these assumptions are wrong, covariate balance, an important measure for an experiment, can not be achieved with respect to propensity score and the facts make a biased effect estimate as a result. This is the reason for studying and examining the application of machine learning methods versus logistic regression.

Machine learning is not yet widely introduced in causal inference field because of the 'black box' nature of algorithms which is difficult for interpreting results. Simply speaking, the concept of machine learning is simply for extracting the relationship between outcome and predictor through a learning algorithm without an a *priori* data model. To begin with, decision tree is a starting point for the simplest algorithms. It divides a data-set into regions that within each region called node, observations are as homogeneous as possible. Whether the outcome is class or numerical, the decision tree is called as classification tree or regression tree. Collectively speaking, 'Classification and Regression tree' called CART is a kind of domain as decision tree. Regardless of types of observations such as class or numerical, observations have similar probabilities in the same categorized node. But for traits of overfitting in CART, pruning methods are introduced to reduce the number of tree splits. Next step is for ensemble methods which predict outcomes with several classfiers and 'Random Forest(RF)' is one of models for ensembling. The 'RF' utilizes randomly sub-samplings from the data set with muptiple trees and results in the most voted outcomes as a result. In the simulated data, Setoguchi et al. did not consider ensemble methods which would perform well in classification with respect to prediction values.[4] This report utilized three kinds of data-sets such as one 'simulated data' and two real data-sets as 'Chemical data used in the lecture not' and 'KHN data-set'. In the real world(real data-set), the true treatment effects are unknown unfortunately, we can not utilize all the performance metrics with the real data such as Bias, SE and CI. Anyway, this report aims for checking and examining the performance of machine learning methods compared to the logistic regression. The main comparison will be shown by using 'logistic regression', 'CART', 'Random Forest' in terms of propensity score weighting. The last framework is performance metrics which describes specifically using propensity score and its weighting. The concepts of (S)IPW is trimming high weights downwards and vice versa. Then, we evaluate the performance of weight trimming by checking 'Average Stadardized Absolute Mean difference(ASAM)', 'Bias(the absolute percentage difference compared to the true treatment effect)', 'Confidence Interval(CI) coverage' and standard error of effect estimates. The experiment results will be proposed using box plot and the numerical table in R with the 3 data-sets. In conclusion, the main framework is based on estimating propensity score and its weghting with several methods such as logistic regression and machine learning models, so this report will compare the conclusion using the performance metrics one simulated data-set and two real data-set. In real data-set, it is realistically hard for us to compare all the observations using performance metrics. Anyway, we can make an conclusion that machine learning models such as CART, RF and another models using pruning, boosting, bagging techniques are showing consistently superior performance performance of propensity score estimation compared to the logistic regression which provided adequate covariate variance with only main effect terms(linear and additivity) in simulated data-set.

III. Experiment

Our experiments are conducted using R programming, which has well-implemented classification and regression model packages. The propensity score is estimated using the models and ATE is estimated to perform propensity score weighting. We measure the performance using the performance metrics of the paper[4] and visualize them as graphs and tables.

We used some actual Chemical Datasets and Korea National Health and Nutrition Examination Survey Dataset to verify that this paper is correct. From these datasets, we conducted an experiment by selecting columns under the subject of causal inference of death or disease, which is in high demand.

A. *Chemical Dataset*

This dataset is real observed dataset from a cohort study of 5000 subjects. The columns consist of age, sex, prior blood pressure, poisox, subsequent blood pressure and mortality. We estimate ATE and confirm if poisox is the cause of mortality. In the case of prior blood pressure column and subsequent blood pressure column, the difference was more important than each value, so a new column called blood difference was created and analyzed. Columns such as poisox are values that cannot be adjusted, so it is good to estimate ATE by applying propensity score weighting.

B. *Korea National Health and Nutrition Examination Survey Dataset*

This dataset is an real life dataset that surveyed the health and nutrition of the Korean people by the Korea Centers for Disease Control and Prevention using a survey in 2019-2020. The columns consist of basic variables, health survey variables, health checkup variables, and nutrition survey variables. To determine whether drinking is the cause of Hepatitis, we selected columns such as gender, age, drinking status, AST, ALT, BMI and Hepatitis. The treatment group was those who drank, and the control group was those who did not drink. Outcome is whether Hepatitis is diagnosed. Since drinking column is an unbalanced treatment, it can be said that it is an appropriate dataset for estimating ATE by assigning a propensity score.

C. *Results*

In the paper[4], to evaluate the performance of propensity score estimation models, they set the true treatment effect and create 1000 simulation datasets. It explains how good propensity score estimation using machine learning is by comparing the treatment effect estimation of the datasets with the true treatment effect value. However, since we cannot know the true treatment effect of the real life datasets, we evaluate the performance by obtaining only the ASAM, weight distribution, and propensity score distribution for each models among the performance metrics implemented in the paper.

	ASAM in Chemical	ASAM in KHN
LR	7.0444	24.387
RF	3.5544	19.517
CART	0.5460	20.665

Table 1. ASAM in the Chemical dataset and KHN dataset.

ASAM is short for Average Standardized Absolute Mean difference. So, a lower ASAM indicates that the treatment and comparison groups has more balance with respect to all covariates. This makes our unbalanced data balanced, allowing ATE estimation from observational data. In table 1, using machine learning models such as Random Forest and CART, ASAM is decreasing and we can get balance in those datasets.

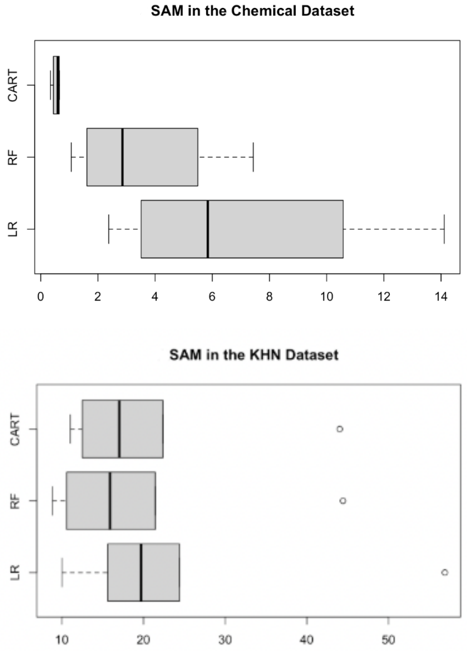

Fig. 1. SAM distribution in the Chemical dataset and KHN dataset.

In the case of logistic regression model, SAM of some covariates are small. But SAM of other covariates are larger than others. It means that balance is different by covariate. It is improved a little bit in the Random Forest model and SAM of most covariates are decreased in the CART model.

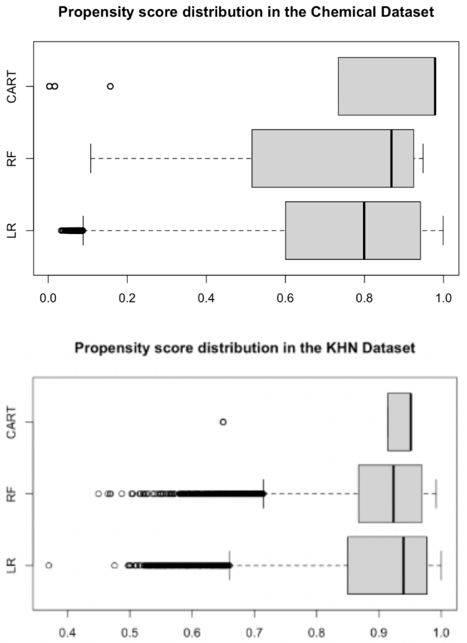

Fig. 2. Propensity score distribution in the Chemical dataset and KHN dataset.

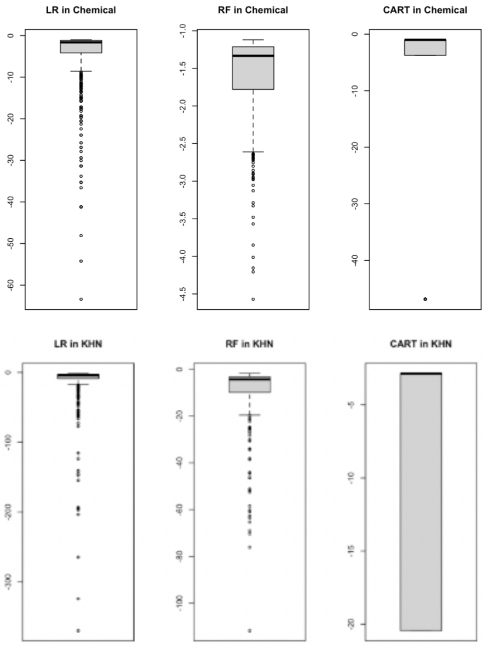

Fig. 3. Weight distribution in the Chemical dataset and KHN dataset.

Since the equation of ATE estimation has propensity score and 1-propensity score as denominator, the closer the propensity score is to 1 or 0, the worse the ATE is. In figure 2, it can be seen that the propensity score of the logistic regression model has many values closer to 1 than other models in both datasets.

The performance of weighting methods can be adversely affected if weights are extreme, as a result of estimated propensity scores that are close to 0 or to 1. We can see that the weights of the logistic regression are much more extreme. This can be seen by looking at the scale of the box plot of the weight distribution. In Chemical dataset LR has weight distribution of values below -60, whereas CART and RF have values above -10, except for one outlier in CART. As with the previous dataset, in KHN dataset, minimum weight value of logistic regression model is less than -300 and it also greatly differ to machine learning model's one.

IV. Limitation and Future Work

When estimating the propensity score, performance is very poor when propensity score is close to 0 or 1, but the logistic regression model rarely comes out with values close to 0 or 1. Therefore, even if other models can reduce the imbalance of treatment distribution in most datasets, ATE values could not be obtained for cases with extreme propensity score estimation. Therefore, it is necessary to find a model that improves performance with various machine learning techniques and at the same time does not cause performance degradation in any dataset.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[3] Ariel Linden DrPH. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Journal of Evaluation in Clinical Practice*, 34(1):3661–3679, 2015.

[4] Brian Lee, Justin Lessler, and Elizabeth Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29:337–46, 11 2009.

[5] Wei-Yin Loh. Classification and regression trees. *WIREs Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

[6] Soko Setoguchi, Sebastian Schneeweiss, M Brookhart, Robert Glynn, and E Cook. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and drug safety*, 17:546–55, 06 2008.

[7] Soko Setoguchi, Sebastian Schneeweiss, M Brookhart, Robert Glynn, and E Cook. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and drug safety*, 17:546–55, 06 2008.