Machine Learning Method for Estimating Heterogeneous Causal Effects of the World Bank Projects

Jianing Zhao
College of William Mary
P.O. Box 8795
Williamsburg, VA 23187-8795
jzhao@cs.wm.edu

Daniel M. Runfola
AidData
427 Scotland Street
Williamsburg, VA. 23185 USA
drunfola@aiddata.org

Peter Kemper
College of William Mary
P.O. Box 8795
Williamsburg, VA 23187-8795
kemper@cs.wm.edu

ABSTRACT

The world bank funds aid projects in third world countries all over the world each year. There is a natural interest in better understanding what makes a project succeed and what not. In technical terms, we are interested in estimating heterogeneity in causal effects and conduct inference about the magnitude of the differences in project effects across subsets of world bank projects. To do so, we analyze a data set with data on world bank projects from 1982 to 2014 that contains project characteristics such as temperature and precipitation as well as spatial and spatially related information. The key challenge for this analysis is that the observational data does not allow us to directly measure the difference between the result of performing a project and of not performing a project as reality only gives us the choice to do one of the two. Following recent research results by Athey and Imbens, we employ a combination of machine learning techniques such as regression trees with techniques from causal inference to measure the average treatment effect, i.e. the average effect of a project, for subsets of geographic locations.

Categories and Subject Descriptors

B.8.1 [**PERFORMANCE AND RELIABILITY**]: Reliability, Testing, and Fault-Tolerance

Keywords

ACM proceedings; LaTeX; text tagging

1. INTRODUCTION

Draw causal effects from data is one of the most interesting research problem across many disciplines. For example, people want to know the effects of a drug in medical studies, companies would like to know the effect of their advertisement on customers, government seeks to evaluate the effect of public policies, for our case, world bank wants to know the effect of the aid projects they investigate around the world over 30 years.

Instead of investigate the causal effects for the whole population, in this paper, we are interested in estimating heterogeneous causal effects for subpopulations by features or covariates. We can estimate heterogeneity by covariates on causal effects and then conducting inference for a distinct unit.

To avoid getting extreme treatment effects which lead to a spurious heterogeneous result, in disciplines such as clinical trial, they use pre-planed subgroup to analyze, for economic, they have pre-analysis plans for randomized experiments. With a data driven approach, the advantage is to discover some other causal effects instead of only the pre-planed subgroups.

To estimate heterogenous causal effects, there are several candidates, for example, classification and regression tree [2], random forest [3], LASSO [11], SVM [12] and so on. In this paper, we use the regression tree, the other methods such as random forest is also good candidate, but we focus on the regression tree in this paper.

In tradition, we can use decision trees to do prediction using the trained data or labeled data. We can build the regression tree to predicting the causal effects with the features as nodes in the tree. However, for the causal inference, the challenge is we do not have such data,in rubin causal model [7], we can only have the treated data or untreated data, but not both at the same time, hence we do not know the ground truth for prediction. We can't follow the traditional supervised machine learning method that we contract the tree with the trained the data and then use the the test data to do prediction based on the constructed tree. Follow the work of Athey and Imbens [1], we use causal tree to do heterogeneous causal effects estimation. However, in practice,for example in our case the world bank data set, within a node, there maybe only treated or untreated data, we will discuss in the paper how to explain such data and other issues

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

2. METHODOLOGY

2.1 conditional average treatment effects

Suppose we have a data set with n iid units with $i = 1, \cdots, n$, for each unit, it has a feature vector $X_i \in [0,1]^d$, a response $Y_i \in \mathbb{R}$ and treatment indicator $W_i \in \{0,1\}$.

For unit level causal effect, we can use Rubin causal model to estimate the average causal effect as shown in function 1,

$$\tau_i = Y_i(1) - Y_i(0) \tag{1}$$

In this paper, we are interested in heterogenous causal effect as 2, this estimator is proposed by [5],

$$\tau(x) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big] \tag{2}$$

The challenge is we know either $Y_i(0)$ or $Y_i(1)$, but not both at the same time. We need to make a unconfoundness assumptions to estimate $\tau(x)$.

$$W_i \perp (Y_i(1), Y_i(0)) \mid X_i \tag{3}$$

Under the unconfoundness assumption, we can get the causal effect as

$$\tau(x) = \mathbb{E}\big[Y^* \mid X_i = x\big] \tag{4}$$

where Y^* is function 5, $e(x)$ is function 6, to estimate the propensity score, there are several ways for calculation such as [9], [6], in this paper, we use logic regression to calculate the pscore.

$$Y_i^* = Y_i^{obs} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \tag{5}$$

$$e(x) = \mathbb{E}\big[W_i \mid X_i = x\big] \tag{6}$$

2.2 Causal Tree Model

We use regression tree to estimate the heterogeneous causal effects, the first step is to construct the tree. To construct the regression tree, we recursively partition the node until the size of the node is less than a threshold we set or the gain of split is negative.
In classic regression tree, mean square error (MSE) is often used to as the criterion for node splitting, the average value within the node is used as the estimator. Following Asthey and Imbens [1], we use 7 as the estimator and we calculate the error of the node by summing $Y_i - \hat{\tau}(X_i)$.

$$\begin{aligned}\hat{\tau}^{CT}(X_i) = &\frac{\sum_{i:X_i \in \mathbb{X}_l} Y_i^{obs} \cdot W_i/\hat{e}(X_i)}{\sum_{i:X_i \in \mathbb{X}_l} W_i/\hat{e}(X_i)} \\ &- \frac{\sum_{i:X_i \in \mathbb{X}_l} Y_i^{obs} \cdot (1 - W_i)/(1 - \hat{e}(X_i))}{\sum_{i:X_i \in \mathbb{X}_l} (1 - W_i)/(1 - \hat{e}(X_i))}\end{aligned} \tag{7}$$

2.3 Pruning the tree

To avoid overfitting of the tree, we need to prune the tree. We use the minimal cost complexity pruning and we define it as 8. α is the complexity parameter, with it we can construct the regression with the right size.

$$R_\alpha(T) = R(T) + \alpha \left|\tilde{T}\right| \tag{8}$$

where R(T) is the resubstitution error estimate of tree T, $\left|\tilde{T}\right|$ is defined as the complexity of the tree, which is the number of leaves in the tree,
To estimate the error of a node, we use function 9,

$$R(t) = \sum_{i=1}^{N} (Y_i - \hat{\tau}(X_i)) \tag{9}$$

where N is the total units in the nodes.

To get a sequence of α, we minimize function 10,

$$g(t, T) = \frac{R(t) - R(T_t)}{\left|\tilde{T}_t\right| - 1} \tag{10}$$

where T_t is a subtree of T rooted at node t.
We use the weakest link cutting to determine α and use it as the complexity parameter when we build the tree for the whole data set.

Because the tree structure is not stable, we use V-fold cross validation to estimate the errors for different α.

$$R(T(\alpha)) = \frac{1}{N} \sum_{1}^{N} Err(\alpha; Y_i, X_i) \tag{11}$$

where $Err(\alpha; Y_i, X_i)$ is the errors for construction the tree using Y_i, X_i data with parameter α.

3. DATA

3.1 data sources and collection

We leverage the following data sources in this analysis:

Variables	
ForestCover	NASA Long Ter
WorldBankProjectLocations	Double-blind geocoded inform
DistancetoRivers	The c
DistancetoCommercialRivers	Calculate
DistancetoRoads	
Elevation	Elevation data mea
Slope	Slope data calculat
AccesibilitytoUrbanAreas	European Commission
PopulationDensity	Center for International Earth Scie
AirTemperature	University of Delware Long term, glob
Precipitation	University of Delware Long term, glob

3.2 data pre-processing

This analysis uses three key types of data: satellite data to measure vegetation, data on the geospatial locations of World Bank projects, and covariate datasets (the sources of which are detailed above). Our primary variable of interest is the fluctuation of vegetation proximate to World Bank projects, which is derived from long-term satellite data (NASA 2015). There are many different approaches to using satellite data to approximate vegetation on a global scale, and satellites have been taking imagery that can be used for this purpose for over three decades. Of these approaches, the most frequently used is the Normalized Difference Vegetation Index (NDVI). The NDVI is a metric that has been used since the early 1970s, and is one of the simplest and most frequently used approaches to approximating vegetative biomass. NDVI measures the relative absorption and reflectance of red and near-infrared light from plants to quantify vegetation on a scale of -1 to 1, with vegetated areas falling between 0.2 and 1. The reflectance by chlorophyll is correlated with plant health, and multiple studies have illustrated that it is generally also correlated with plant biomass. In other words, healthy vegetation and high plant biomass tend to result in high NDVI values (Dunbar 2009). Using NDVI as an outcome measure has a number of other benefits, including the long and consistent time periods for which it has been calculated. While the NDVI does have a number of challenges - including a propensity to saturate over densely vegetated regions, the potential for atmospheric noise (including clouds) to incorrectly offset values, and reflectances from bright soils providing misleading estimates - the popularity of this measurement has led to a number of

improvements over time to offset many of these errors. This is especially true of measurements from longer-term satellite records, such as those used in this analysis, produced from the MODIS and AVHRR satellite platforms (NASA 2015).

The second primary dataset used in this analysis measures where - geographically - World Bank projects were located. This dataset was produced by AidData (2016), relying on a double-blind coding system where two experts employ a defined hierarchy of geographic terms and independently assign uniform latitude and longitude coordinates, precision codes, and standardized place names to each geographic feature. If the two code rounds disagree, the project is moved into an arbitration round where a geocoding project manager reconciles the codes to assign a master set of geocodes for all of the locations described in the available project documentation. This approach also captures geographic information at several levelsâĂŤcoordinate, city, and administrative divisionsâĂŤfor each location, thereby allowing the data to be visualized and analyzed in different ways depending upon the geographic unit of interest. Once geographic features are assigned coordinates, coders specify a precision code that varies from 1 (exact point) to 9 (national-level project or program). AidData performs many procedures to ensure data quality, including de-duplication of projects and locations, correcting logical inconsistencies (e.g. making sure project start and end dates are in proper order), finding and correcting field and data type mismatches, correcting and aligning geocodes and project locations within country and administrative boundaries, validating place names and correcting gazetteer inconsistencies, deflating financial values to constant dollars across projects and years (where appropriate), strict version control of intermediate and draft data products, semantic versioning to delineate major and minor versions of various geocoded datasets, and final review by a multidisciplinary working group.

4. EXPERIMENTS

4.1 Simulation Experiments

4.2 World Bank aid data

In this subsection, we examine an application of these methods to examine the research question: "How effective have the environmental safeguards of World Bank projects been in preventing deforestation?". These safeguards include environmental education and impact assessment programs, reforestation activities, and other environmental protection activities (Nielson and Tierney 2003; Ledec and Posas 2003; Quintero 2007). To examine the impact these activities have had on forest cover, we seek to isolate the causal impact of any given World Bank project on the forest cover - measured using satellite data - which it is proximate to.

In this paper, we specifically examine the benefits which causal trees have for examining causal impacts at a global scale. Because World Bank projects exist in highly heterogeneous environments (see figure X), traditional causal methods which identify a single effect across the entire sample - without identifying impacts unique to meaningful sub populations - are insufficient. In this approach, we illustrate that the causal tree enables an examination of impact on sub-populations within a given dataset, without the a-priori definition of those sub-populations. We pose that it is a more effective and accurate way to understand the geographically varied impacts of World Bank projects.

Our method is based on the R package rpart. Rpart support user defined split function, therefor, we can use the split criterion function 7. To improve the efficiency of the r program, we use rcpp and call C++ functions inside the split and evaluation function for each node in the tree. To further improve the c++ functions, we use openmp inside the C++ functions.

5. RELATED WORK

Causality [8] plays an important role in many area. In this paper, we focus on the heterogeneous causal effects. Some paper in the literature use tree based machine learning technique to estimate heterogeneous causal effects. In [10], they use statistical test as the criterion for node splitting. In [1], they use causal trees to estimate heterogeneous treatment effect. However, they do not show what if in some nodes, there is only treated units or only untreated units and then how to estimate the heterogeneous causal effects.
Some paper use forest based machine learning technique to estimate heterogeneous causal effects. In [13], they use casual forest to do heterogeneous causal effects estimation, and they share the same idea in paper [4] that they use difference data for the structure of the tree and the estimation value within each node.

6. CONCLUSIONS

7. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

8. REFERENCES

[1] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects, 2015.

[2] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.

[3] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[4] M. Denil, D. Matheson, and N. de Freitas. Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning (ICML)*, 2014.

[5] K. Hirano, G. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[6] D. Ho, K. Imai, G. King, and E. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236, 2007.

[7] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA, 2015.

[8] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000.

[9] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[10] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158, 2009.

[11] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[12] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[13] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests, 2015.