Neuroevolutionary Feature Representations for Causal Inference*

Michael C. Burkhart[1][0000−0002−2772−5840] and
Gabriel Ruiz[2][0000−0003−3349−760X]

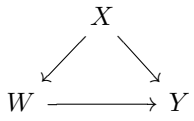
[1] University of Cambridge, Cambridge, U.K.
mcb93@cam.ac.uk
[2] UCLA, Los Angeles, CA, U.S.A.
ruizg@ucla.edu

Abstract Within the field of causal inference, we consider the problem of estimating heterogeneous treatment effects from data. We propose and validate a novel approach for learning feature representations to aid the estimation of the conditional average treatment effect or CATE. Our method focuses on an intermediate layer in a neural network trained to predict the outcome from the features. In contrast to previous approaches that encourage the distribution of representations to be treatment-invariant, we leverage a genetic algorithm to optimize over representations useful for predicting the outcome to select those less useful for predicting the treatment. This allows us to retain information within the features useful for predicting outcome even if that information may be related to treatment assignment. We validate our method on synthetic examples and illustrate its use on a real life dataset.

Keywords: causal inference, heterogeneous treatment effects, feature representations, neuroevolutionary algorithms, counterfactual inference

1 Introduction

In this note, we aim to engineer feature representations to aid in the estimation of heterogeneous treatment effects. We consider the following graphical model



(1)

where $X \in \mathbb{R}^d$ denotes a vector of features, $W \in \{0, 1\}$ represents a boolean treatment, and $Y \in \mathbb{R}$ denotes the outcome. Suppose (X_i, W_i, Y_i) for $i = 1, \ldots, n$ are i.i.d. samples from a distribution P respecting the graph (1). Within the potential outcomes framework [10], we let $Y_i(0)$ denote the potential outcome

*M.B. and G.R. were supported by Adobe Inc. (San José, Calif., U.S.A.).

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_1

if W_i were set to 0 and $Y_i(1)$ denote the potential outcome if W_i were set to 1. We wish to estimate the conditional average treatment effect (CATE) defined by $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$. We impose standard assumptions that the treatment assignment is unconfounded, meaning that $\{Y_i(0), Y_i(1)\} \perp W_i \mid X_i$, and random in the sense that $\epsilon < P(W_i = 1|X_i = x_i) < 1 - \epsilon$ for all i, some $\epsilon > 0$, and all x_i in the support of X_i. These assumptions jointly constitute *strong ignorability* [13] and prove sufficient for the CATE to be identifiable. Under them, there exist methods to estimate the CATE from observed data that then allow us to predict the expected individualized impact of an intervention for novel examples using only their features. Viewing these approaches as black box estimators, we seek a mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^m$ such that the estimate of the CATE learned on the transformed training data $(\Phi(X_i), W_i, Y_i)$ is more accurate than an estimate learned on the original samples (X_i, W_i, Y_i). In particular, we desire a function Φ yielding a corresponding representation $\Phi(X)$ such that (1) $\Phi(X)$ is as useful as X for estimating Y, and (2) among such representations, $\Phi(X)$ is least useful for estimating W. In this way, we hope to produce a new set of features $\Phi(X)$ that retain information relevant for predicting the outcome but are less related to treatment assignment. *We propose learning Φ as a hidden layer in a neural network estimating a functional relationship of Y given X. We apply a genetic algorithm to a population of such mappings to evolve and select the one for which the associated representation $\Phi(X)$ is least useful for approximating W.*

Feature representations have previously been used for causal modeling. Johansson, et al. [4,14] viewed counterfactual inference as a covariate shift problem and learned representations designed to produce similar empirical distributions among the treatment and control populations. Li & Fu [8] and Yao, et al. [16] developed representations designed to preserve local similarity. However, we generally agree with Zhang et al.'s [17] recent argument that domain invariance often removes too much information from the features for causal inference.[†] *In contrast to most previous approaches, we develop a feature representation that attempts to preserve information useful for predicting the treatment effect if it is also useful for predicting the outcome.*

Outline. The next section describes related work. In section 3, we outline our methodology. We validate our method on artificial data in section 4 and on a publicly available experimental dataset in section 5, before concluding in section 6.

2 Related work

In this section, we discuss meta-learning approaches for inferring the CATE and briefly introduce genetic algorithms.

Meta-learners. Meta-learning approaches leverage an arbitrary regression framework (e.g., random forests, neural networks, linear regression models) to estimate

[†]Zhao et al. [18] make this argument in a more general setting.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_1

the CATE from data. The **S-learner** (single-learner) uses a standard supervised learner to estimate $\mu(x,w) = \mathbb{E}[Y|X=x, W=w]$ and then predicts $\hat{\tau}_S(x) = \hat{\mu}(x,1) - \hat{\mu}(x,0)$. The **T-learner** (two-learner) estimates $\mu_1(x) = \mathbb{E}[Y(1)|X=x]$ from treatment data and $\mu_0(x) = \mathbb{E}[Y(0)|X=x]$ from control data and then predicts $\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$. The **X-learner** [6] estimates μ_1 and μ_0 as in the T-learner, and then predicts the contrapositive outcome for each training point. It then estimates $\tau_1(x) = \mathbb{E}[Y_i - \hat{\mu}_0(X_i) \mid X=x]$ and $\tau_0(x) = \mathbb{E}[\hat{\mu}_1(X_i) - Y_i \mid X=x]$ before predicting $\hat{\tau}_X(x) = g(x)\hat{\tau}_0(x) + (1-g(x))\hat{\tau}_1(x)$ where $g : \mathbb{R}^d \to [0,1]$ is a weight function.[‡] The **R-learner** [11] leverages Robinson's decomposition that led to Robin's reformulation of the CATE function as the solution to $\tau(\cdot) = \arg\min_\tau \{\mathbb{E}_{(X,W,Y)\sim P}[|(Y-m(X)) - (W-e(X))\tau(X)|^2]\}$ in terms of the treatment propensity e and conditional mean outcome $m(x) = \mathbb{E}[Y|X=x]$.

Neuroevolutionary algorithms Holland introduced genetic algorithms [2] as a nature-inspired approach to optimization. These algorithms produce successive generations of candidate solutions. New generations are formed by selecting the fittest members from the previous generation and performing cross-over and/or mutation operations to produce new offspring candidates. Evolutionary algorithms encompass extensions to and generalizations of this approach including memetic algorithms that perform local refinements, genetic programming that acts on programs represented as trees, and evolutionary programming and strategies that operate on more general representations. When such methods are applied specifically to the design and training of neural networks, they are commonly called neuroevolutionary algorithms. See Stanley et al. [15] for a review.

3 Methodology

We now describe how to create our feature mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^m$. Individual candidate solutions derive from a hidden layer in a network trained to predict Y from X. We then evolve cohorts of parameter sets for such maps to minimize the functional usefulness of candidate representations for predicting W.

Candidate solutions. We consider neural networks $f_\Theta : \mathbb{R}^d \to \mathbb{R}$ of the form $f_\Theta(x) = M_2 \cdot a(M_1 \cdot x + b_1) + b_2$ for a nonlinear activation function a where the parameter set Θ denotes $M_1 \in \mathbb{R}^{m\times d}$, $M_2 \in \mathbb{R}^{1\times m}$, $b_1 \in \mathbb{R}^m$, and $b_2 \in \mathbb{R}^1$. Though f_Θ is decidedly not a deep neural network, we note that, as a neural network with a single hidden layer, it remains a universal function approximator in the sense of Hornik et al. [3]. Optimizing the network f_Θ in order to best predict Y from X seeks the solution $\Theta_* = \arg\min_\Theta \mathbb{E}\,|Y - f_\Theta(X)|^2$. For fixed Θ, we let $\Phi_\Theta : \mathbb{R}^d \to \mathbb{R}^m$ given by $\Phi_\Theta(x) = a(M_1 \cdot x + b_1)$ denote the output of the hidden layer.

[‡]It is also possible to estimate τ from $\{(X_i, Y_i - \hat{\mu}_0(X_i))\}_{W_i=1} \cup \{(X_i, \hat{\mu}_1(X_i) - Y_i)\}_{W_i=0}$ or, using $\hat{\mu}(x,w)$ from the S-learner approach, with $\{(X_i, Y_i - \hat{\mu}(X_i,0))\}_{W_i=1} \cup \{(X_i, \hat{\mu}(X_i,1) - Y_i)\}_{W_i=0}$. We find that these alternate approaches work well in practice and obviate the need to estimate or fix g.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_1

Fitness function. For Θ near the optimum Θ_*, $\Phi_\Theta(X)$ should be approximately as useful as X for estimating Y. However, the mapped features $\Phi_\Theta(X)$ may also carry information useful for predicting W. To this end, we define $g_{\Psi,\Theta} : \mathbb{R}^d \to [0,1]$ by $g_{\Psi,\Theta}(x) = \sigma(M_4 \cdot a(M_3 \cdot \Phi_\Theta(x) + b_3) + b_4)$ for a nonlinear activation a, sigmoidal activation σ, and parameter set Ψ consisting of $M_3 \in \mathbb{R}^{k\times m}$, $M_4 \in \mathbb{R}^{1\times k}$, $b_3 \in \mathbb{R}^k$, and $b_4 \in \mathbb{R}$. We define the fitness of a parameter set Θ to be $\mu(\Theta) = \min_\Psi \mathbb{E}\,|W - g_{\Psi,\Theta}(X)|^2$. In this way, we express a preference for representations $\Phi_\Theta(X)$ that are less useful for predicting W.

Evolutionary algorithm. Given training and validation datasets, we form an initial cohort of c candidates independently as follows. For $1 \le j \le c$, we randomly instantiate Θ_j using Glorot normal initialization for the weights and apply the Adam optimizer on training data to seek Θ_*. We use Tikhonov regularization for the weights and apply dropout after the $a(x) = \tanh(x)$ activation function[§]to prevent overfitting. For each constituent Θ_j in the cohort, we then initialize and train a network g_{Ψ,Θ_j} to seek $\Psi_j = \arg\min_\Psi \mathbb{E}\left|W - g_{\Psi,\Theta_j}(X)\right|^2$ on the training set and then evaluate $\mathbb{E}\left|W - g_{\Psi_j,\Theta_j}(X)\right|^2$ empirically on the validation set to estimate $\mu(\Theta_j)$. For each of the $\binom{\ell}{2}$ pairs formed from the ℓ fittest members of the current cohort, we apply Montana and Davis's node-based crossover [9] method to the parameters M_1 and b_1 that we use to form Φ. The next generation then consists of the fittest candidate from the previous generation, the candidates formed from cross-over, and new candidates generated from scratch.

Remarks. Due to our choice of representation Φ, after training the network $f_\Theta(x)$, we expect the relationship between the learned features $\Phi(X)$ and the outcome Y to be approximately linear. In particular, $Y \approx M_2 \cdot \Phi(X)$. For this reason, the causal meta-learners trained using a linear regression base learner may benefit more extensively from using the transformed features instead of the original features, especially in cases where the relationship between the original features and outcomes is not well-approximated as linear.

In order to use the represented features $\Phi(X_i)$ in place of the original features X_i, we require that strong ignorability holds for the transformed dataset $(\Phi(X_i), W_i, Y_i)$, $i = 1, \ldots, n$. One sufficient, though generally not necessary, assumption that would imply strong ignorability is for Φ to be invertible on the support of X [14, assumption 1]. Unconfoundedness would also be guaranteed if $\Phi(X)$ satisfied the backdoor condition with respect to (W, Y) [12, section 3.3.1].

4 Ablation study on generated data

Due to the fundamental challenge of causal inference (namely, that the counterfactual outcome cannot be observed, even in controlled experiments), it is

[§]We tested rectified and exponential linear unit activation functions for a in Φ_Θ but noticed only minor differences in subsequent performance of the causal forest.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_1

common practice to compare approaches to CATE estimation on artificially generated datasets for which the CATE can be calculated. In this section, we perform experiments using Setups A and C from Nie & Wager's paper [11].‡

Comparison methodology. For both setups, we ran 100 independent trials. Within each trial, we randomly partitioned a simulated dataset it into training, validation, and testing subsets at a 70%-15%-15% rate. We trained causal inference methods on the training and validation sets, and predicted on the test dataset. We then developed a feature map using the training and validation data as described in the previous section, applied this map to all features, and repeated the training and testing process using the new features. To determine the impact of the fitness selection process, we also learned a feature transformation that did not make use of the fitness function at all. It simply generated a single candidate mapping and used it to transform all the features. This ablative method is referred to as "no-fitness" in Table 1. We compared the causal forest [1] with default options, and the S-, T-, and X-learners with two base learners: LightGBM [5] and cross-validated ridge regression.

Results. We report results in Table 1. For both setups, we consider a paired t-test for equal means against a two-sided alternative. For setup A, we find that the improvement in MSE from using the transformed features in place of the original features corresponds to a statistically significant difference for the following learners: the causal forest ($p < 0.001$), the S-learner with ridge regression ($p < 0.001$), the T-learner with both LightGBM ($p < 0.001$) and ridge regression ($p < 0.001$), and the X-learner with both LightGBM ($p < 0.001$) and ridge regression ($p < 0.001$). For setup C, we again find significant differences for the causal forest ($p = 0.023$), S-learner with LightGBM ($p = 0.003$), T-learner with ridge regression ($p < 0.001$) and X-learner with ridge regression ($p < 0.001$). In summary, we find that our feature transformation method improves the performance of multiple standard estimators for the CATE under two data generation models.

learner	features	Set. A	Set. C
Causal forest	initial	0.175	0.035
	no-fitness	0.114	0.029
	transformed	0.120	0.029

learner	features	Set. A	Set. C
S-L. LGBM	initial	0.149	0.226
	no-fitness	0.140	0.211
	transformed	0.135	0.204
S-L. Ridge	initial	0.093	0.015
	no-fitness	0.079	0.015
	transformed	0.081	0.015
T-L. LGBM	initial	0.666	0.567
	no-fitness	0.536	0.551
	transformed	0.512	0.544

learner	features	Set. A	Set. C
T-L. Ridge	initial	0.745	0.178
	no-fitness	0.333	0.125
	transformed	0.325	0.128
X-L. LGBM	initial	0.411	0.313
	no-fitness	0.335	0.313
	transformed	0.317	0.298
X-L. Ridge	initial	0.630	0.166
	no-fitness	0.289	0.109
	transformed	0.288	0.114

Table 1: Average Mean Squared Error (MSE) over 100 independent trials.

‡Nie & Wager's paper included four setups, namely A–D; however setup B modeled a controlled randomized trial and setup D had unrelated treatment and control arms.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_1

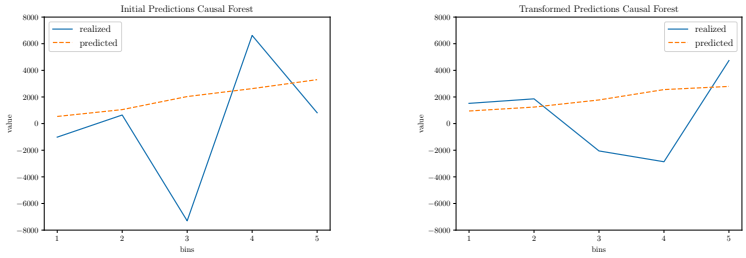


Figure 1: We plot estimated realized and predicted average treatment effects versus the quintiles of predicted treatment effect (5 bins) for a causal forest using (left) the initial features and (right) the features transformed using our method.

5 Application to econometric data

In this section, we apply our feature engineering method to the LaLonde dataset [7] chronicling the results of an experimental study on temporary employment opportunities. The dataset contains information from 445 participants who were randomly assigned to either an experimental group that received a temporary job and career counseling or to a control group that received no assistance. We consider the outcome of earnings in 1978 (in $, after treatment).

We cannot determine true average treatment effects based on individual-level characteristics (i.e. the true CATE values) for real life experimental data as we can with the synthetic examples of the previous section. Instead, we evaluate performance by comparing the average realized and predicted treatment effects within bins formed by sorting study participants according to predicted treatment effect as demonstrated in Figure 1. Applying the causal forest predictor to the original features results in a root mean square difference between the average predicted and realized treatment effects of 4729.51. Using the transformed features improves this discrepancy to 3114.82. From a practical perspective, one may learn the CATE in order to select a subset of people for whom a given intervention has an expected net benefit (and then deliver that intervention only to persons predicted to benefit from it). When we focus on the 20% of people predicted to benefit most from this treatment, we find that the estimated realized benefit for those chosen using the transformed features ($4732.89) is much greater than the benefit for those chosen using the original feature set ($816.92). This can be seen visually in Figure 1 by comparing bin #5 (the rightmost bin) in both plots.

6 Conclusions

Causal inference, especially on real life datasets, poses significant challenges but offers a crucial avenue for predicting the impact of potential interventions. Learned feature representations help us to better infer the CATE, improving our ability to individually tailor predictions and target subsets of the general

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_1

population. In this paper, we propose and validate a novel representation-based method that uses a neuroevolutionary approach to remove information from features irrelevant for predicting the outcome. We demonstrate that this method can yield improved estimates for the CATE on standard synthetic examples and illustrate its use on a real life dataset. We believe that representational learning is particularly well-suited for removing extraneous information in causal models and anticipate future research in this area.

Acknowledgements. We would like to thank Binjie Lai, Yi-Hong Kuo, Xiang Wu, and the anonymous reviewers for their insights and suggestions.

References

1. Athey, et al.: Generalized random forests. Ann. Statist. **47**(2), 1148–1178 (2019)
2. Holland: Adaptation in Natural and Artificial Systems. U. Michigan Press (1975)
3. Hornik, et al.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
4. Johansson, et al.: Learning representations for counterfactual inference. In: Int. Conf. Mach. Learn. pp. 3020–3029 (2016)
5. Ke, et al.: LightGBM. In: Adv. Neur. Inf. Proc. Sys. pp. 3146–3154 (2017)
6. Künzel, et al.: Metalearners for estimating heterogeneous treatment effects using machine learning. Proc. Natl. Acad. Sci. **116**(10), 4156–4165 (2019)
7. LaLonde: Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. **76**(4), 604–620 (1986)
8. Li & Fu: Matching on balanced nonlinear representations for treatment effects estimation. In: Adv. Neur. Inf. Proc. Sys. pp. 930–940 (2017)
9. Montana & Davis: Training feedforward neural networks using genetic algorithms. In: Int. Jt. Conf. Artif. Intell. pp. 762–767 (1989)
10. Neyman: Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Rocz. Nauk Rol. **10**, 1–51 (1923)
11. Nie & Wager: Quasi-oracle estimation of heterogeneous treatment effects. Biometrika **108**(2), 299–319 (2021)
12. Pearl, J.: Causality. Cambridge U. Press, 2nd edn. (2009)
13. Rosenbaum & Rubin: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
14. Shalit, et al.: Estimating individual treatment effect: generalization bounds and algorithms. In: Int. Conf. Mach. Learn. vol. 70, pp. 3076–3085 (2017)
15. Stanley, et al.: Designing neural networks through neuroevolution. Nat. Mach. Intell. **1**, 24–35 (2019)
16. Yao, et al.: Representation learning for treatment effect estimation from observational data. In: Adv. Neur. Inf. Proc. Sys. vol. 31, pp. 2633–2643 (2018)
17. Zhang, et al.: Learning overlapping representations for the estimation of individualized treatment effects. In: Int. Conf. Artif. Intell. Stats. (2020)
18. Zhao, et al.: On learning invariant representations for domain adaptation. In: Int. Conf. Mach. Learn. (2019)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_1