Empirical Analysis of Model Selection for Heterogeneous Causal Effect Estimation

Divyat Mahajan, Ioannis Mitliagkas, Brady Neal* & Vasilis Syrgkanis*

International Conference on Learning Representations (ICLR) 2024 (Spotlight)

*Equal Advising
The causal effect of exercise on cholesterol will be different for the group of young people vs old people.

Need to estimate conditional average treatment effect (CATE) rather than the average effect (ATE) for better decision making!
CATE Estimation

- CATE: $\tau(x) = \mathbb{E}[Y(1) - Y(0) | X = x]$
- Meta-Learners estimate $\tau(x)$ as a function of nuisance models $\hat{\eta} = (\hat{\mu}, \hat{\pi})$
  - Potential Outcome Model: $\hat{\mu}_w(x) = \mathbb{E}[Y | W = w, X = x]$
  - Propensity Model: $\hat{\pi}_w(x) = \mathbb{P}(W = w | X = x)$

$X$: Covariates
$W$: Binary Treatments
$Y(0), Y(1)$: Potential Outcomes

Diagram:
- $X$ connected to $W$ and $Y$
- $W$ connected to $Y$
CATE Estimation

- Indirect Meta-Learner:
  - T-Learner: $\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

- Direct Meta-Learner:
  - DR-Learner: $\hat{\tau}_{DR} := \hat{f}_{DR} = \arg \min_{f \in F} \sum_{x, w, y} (y^{DR}(\hat{\eta}) - f(x))^2$

$X$: Covariates
$W$: Binary Treatments
$Y(0), Y(1)$: Potential Outcomes

X: Covariates
W: Binary Treatments
Y(0), Y(1): Potential Outcomes
How to select between CATE Estimators?

Precision of Heterogeneous Effects (PEHE): \( L(\hat{t}) = \mathbb{E}_X[(\hat{t}(X) - \tau(X))^2] \)

- True CATE \( \tau(X) \) is not known as we don’t observe both potential outcomes
- Cannot perform cross-validation unlike machine learning!
How to select between CATE Estimators?

Surrogate PEHE: \[ L(\hat{\tau}) = \mathbb{E}_X[(\hat{\tau}(X) - \tilde{\tau}(X))^2] \]

- Surrogate Metrics: Estimate true CATE on the validation set \( \tilde{\tau}(X) \) in PEHE
- Different strategies for estimating \( \tilde{\tau}(x) \) lead to different surrogate metrics

We have a poor understanding about the relative advantages/disadvantages of surrogate metrics!
Contribution

We perform a comprehensive empirical study over 78 datasets to benchmark 34 surrogate metrics for CATE model selection, where model selection task is made challenging by training 415 CATE estimators per dataset.
CATE Estimators in our study

We allow for diverse collection of estimators for each direct meta-learner to make the task of CATE model selection more challenging.
We use AutoML to have low bias in estimating the nuisance parameters ($\hat{\eta}$) of surrogate metrics, which enhances their model selection ability.
Proposed Evaluation Framework

CATE Estimation

Dataset

List of CATE Estimators

Meta-Learners $E_A$

Meta-Learners $E_B$

Select using $M^A(\hat{\tau})$

Select using $M^B(\hat{\tau})$

List of optimal Meta-Learners

Meta-Learners $E^*_A$

Meta-Learners $E^*_B$
Proposed Evaluation Framework

Dataset → List of CATE Estimators → Meta-Learners $E_A$ → Select using $M^A(\hat{\tau})$ → List of optimal Meta-Learners → Select using $M(\hat{\tau})$ → Ensemble $E_M^*$

List of CATE Estimators

Meta-Learners $E_A$

Meta-Learners $E_B$

Select using $M^B(\hat{\tau})$ → Meta-Learners $E_A^*$

Meta-Learners $E_B^*$

Ensemble of optimal Meta-Learners
Proposed Evaluation Framework

Dataset

CATE Estimation

Meta-Learners $E_A$

Meta-Learners $E_B$

List of CATE Estimators

Select using $M^A(\hat{\tau})$

Select using $M^B(\hat{\tau})$

List of optimal Meta-Learners

Select using $M(\hat{\tau})$

Ensemble of optimal Meta-Learners

Counterfactual Data

Potential Outcomes $(Y(0), Y(1))$

PEHE to judge $M(\hat{\tau})$

Ensemble $E^*_M$
Main Findings

• Plug-in Surrogate Metrics are optimal as well!
  • Implication of well-tuned nuisance models via AutoML for surrogate metrics

• Two-level selection strategy provides strict improvement over single-level selection strategy!
  • Better performance in 28.7% cases, otherwise statistically indistinguishable.

• Ensemble selection provides further improvement!
  • Better performance in 5.8% cases, otherwise statistically indistinguishable.
Chat with us during the poster session!