A Tree-based Model Averaging Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources

Xiaoqing Tan

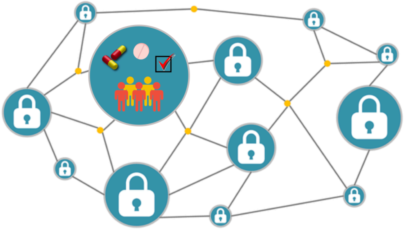
Department of Biostatistics
University of Pittsburgh

Joint work with Drs. Joyce Chang, Ling Zhou, and Lu Tang

Motivation: to improve CATE estimation in a target site

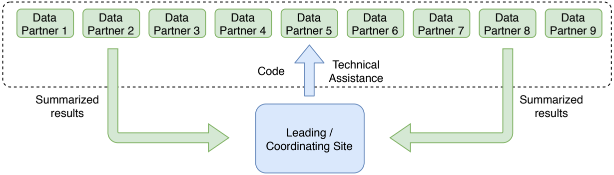
- Interested in causal effects conditional on subject characteristics, i.e. conditional average treatment effects (CATE)
- A site is under-powered, hope to borrow information across site

Challenges of information borrowing in distributed data networks



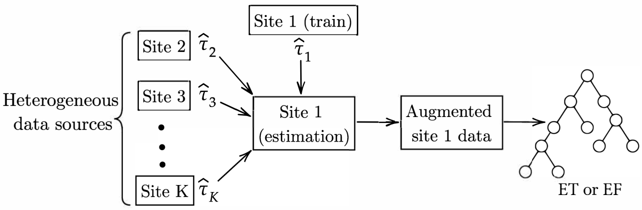
- Data are highly heterogeneous across sites
- Privacy concerns of subject-level data sharing across sites
- Challenging in causal inference settings: no “ground truth” outcome

Possible solution: leveraging models from different sites



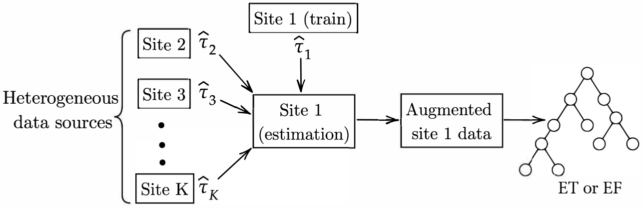
- Model averaging framework in distributed data networks
- Multiple sites collectively contribute to the tasks of statistical modeling without sharing sensitive subject-level data
- There is no established model averaging approach with the goal of improving the estimation of CATE

Sketch of the proposed framework



- Without loss of generality, let site 1 to be the target/coordinating site
- Local stage: each of the K sites estimate τ_k independently
    - can use any CATE estimator
    - in parallel
- Sites pass $\{\tau_k(\boldsymbol{x})\}_{k=1}^{K}$ to site 1

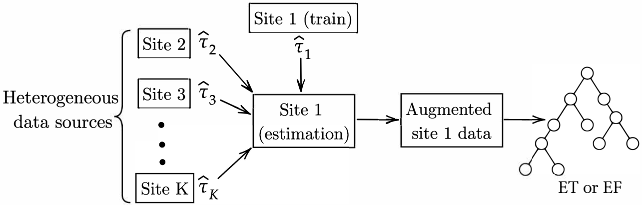
Sketch of the ensemble algorithm



- Create the augmented data in site 1

Subject	Site	$\boldsymbol{X}$	S	τ
1	1	$\boldsymbol{x}_1$	1	$\tau_1(\boldsymbol{x}_1)$
		$\vdots$		
1	1	$\boldsymbol{x}_1$	K	$\tau_K(\boldsymbol{x}_1)$
2	1	$\boldsymbol{x}_2$	1	$\tau_1(\boldsymbol{x}_2)$
		$\vdots$		

Sketch of the proposed framework



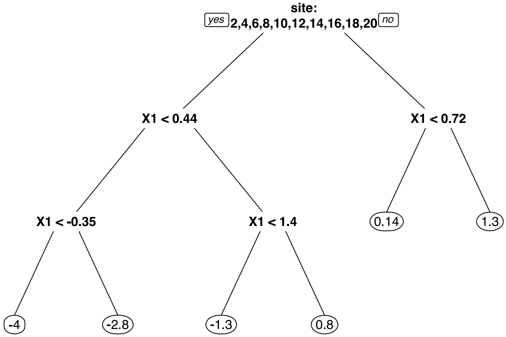
- Ensemble stage: either a **E**nsemble **T**ree (ET) or a **E**nsemble **F**orest (EF)

$$\text{causal estimand } \mathcal{T}(\boldsymbol{x}, s)$$

A site indicator of which individual model is used, S, along with the patient characteristics $\boldsymbol{X}$ are used as covariates

Visualization of the proposed tree-based estimator

- Suppose there are $K = 20$ sites in total
- CATE function $\tau(\boldsymbol{x}, k) = \mathbb{1}\{x_1 > 0\} \cdot x_1 + (x_1 - 3) \cdot c \cdot U_k$



Asymptotic properties

We provide the consistency guarantee of $\widehat{\mathcal{T}}_{\mathrm{EF}}$ for the true target τ_1

Theorem

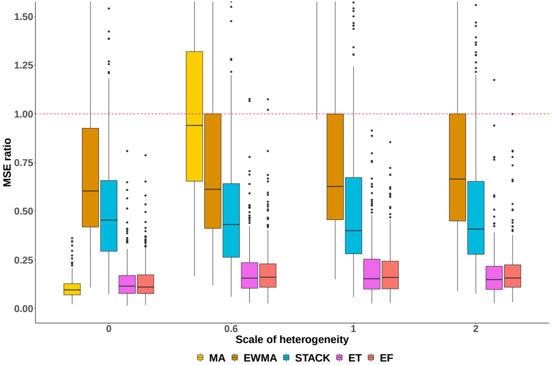
Suppose the subsamples used to build each tree in an ensemble forest are drawn from different subjects in the augmented site 1 data. Under the following conditions:

1. *Bounded covariates: Features $\boldsymbol{X}_i$ and the site indicator S_i are independent and have a density that is bounded away from 0 and infinity*
2. *Lipschitz response: the conditional mean function $\mathbb{E}[\mathcal{T}|\boldsymbol{X}=\boldsymbol{x}, S=k]$ is Lipschitz-continuous*
3. *Honest trees: trees in the random forest use different data for placing splits and estimating leaf-wise responses*

Then $\widehat{\mathcal{T}}_{EF}(\boldsymbol{x}, s) \xrightarrow{p} \tau_s(\boldsymbol{x})$, for all $\boldsymbol{x}$ and s, as $\min_k n_k \to \infty$
Hence, $\widehat{\tau}^*_{EF}(\boldsymbol{x}) \xrightarrow{p} \tau_1(\boldsymbol{x})$

Comparison among various methods

- Compare the proposed estimators with local estimator (LOC) and multiple existing model averaging or ensemble approaches

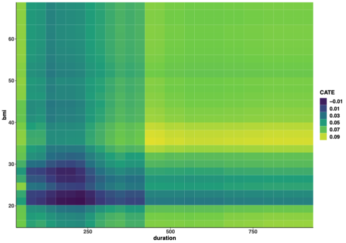


Application: oxygen therapy on survival

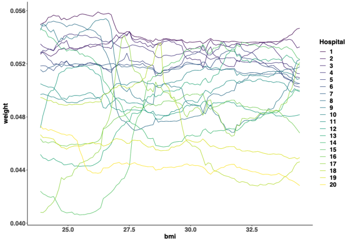
- **Our goal:** estimate CATE of oxygen saturation SpO_2 within 94-98% range on hospital mortality among critically ill patients with respiratory disease and with at least 48-hour of oxygen therapy
- eICU database: data from critical care units throughout the U.S.
- $K = 20$ hospitals, $N = 7,022$
- Earlier studies considered using random effects to model site heterogeneity (van den Boom et al., 2020)

Decision rule	Avg. survival rate
Baseline	0.762
LOC-based	0.772
EF-based	0.791

(a)



(b)



(c)

Discussion

- We have proposed an efficient and interpretable tree-based model averaging framework to enhance the estimation of CATE
- Our work facilitates practical collaboration within distributed research networks
- Can be extended beyond causal inference to a general $f(\boldsymbol{x})$
- R package `ifedtree` is built and available at GitHub (`https://github.com/ellenxtan/ifedtree`)

Thank you!