Using forest-based models to personalise ventilation treatment in the ICU

Optimising positive end-expiratory pressure assignment based on the MIMIC-IV dataset

Responsible Professor Jesse Krijthe

Author Hubert Nowak
H.D.Nowak@student.tudelft.nl

Supervisors Rickard Karlsson
Jim Smit

1. BACKGROUND

Positive end-expiratory pressure (PEEP) Setting

- Plays a key role in mechanical ventilation treatment of patients with lung failure [1], has considerable influence on patient's chances of survival.
- It is not know if high or low PEEP setting is more beneficial for patients [2], perhaps it should be assigned on a case-to-case basis.

Conditional average treatment effect (CATE) estimation

- Tries to estimate the difference in outcomes when different treatment value is assigned.

$$\tau(X) = E[Y_1 - Y_0|X]$$

Where, in our case:

- Y_1 - outcome with high PEEP value
- Y_0 - outcome with low PEEP value
- X - given patient's characteristics
- $Y_w = 1$ - patient survived
- $Y_w = 0$ - patient died

2. RESEARCH GOAL

The main goal is to investigate whether one can use forest-based models to estimate CATE for PEEP assignment based on patient's characteristics. Specifically, we look at Causal Forest [3], as well as S- and T-learners [4] with Random Forest as the base model. We analyse and evaluate their performance using MIMIC-IV dataset [5] to determine if they can be used for the given task. Furthermore, we verify our claims using data from a randomised controlled trial.

3. METHOD

- Determine models' efectiveness in different scenarios using various types of simulated data, comparing the mean squared error of predicted treatment effect.
- Pre-process MIMIC-IV database and select appropriate features.
- Train models on pre-processed MIMIC-IV data, evaluate their performance and analyse the results using Qini curve and associated metrics.
- Evaluate the outcomes using data from randomised controlled trial (RCT).

4. RESULTS

SIMULATIONS

We firstly evaluated the models' performance in artificial settings. In total, seven experiments were conducted, with the data generated by specifying number of features, propensity score and the response functions.

- **S-learner** performed best when there was no treatment effect.
- **T-learner** outperformed other models when the response functions were independent of each other.
- **Causal forest** performed overall well, achieving low error rates in almost all tests.

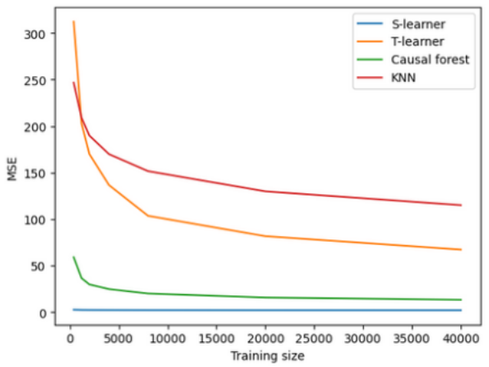

Figure 1: Evolution of MSE of the predicted treatment effect on the test set, as a function of the size of training set, in one of the simulations with treatment effect equal to 0. Note that already at the beginning S-learner is able to correctly estimate CATE.

MIMIC-IV DATABASE

- The models were strongly overfitting and the performance was rather poor.
- Hyperparameter tuning helped in fixing these issues, but did not alleviate them completely, with Qini AUC scores on training set being 4-5 times greater than on test set.
- Confidence intervals for all results were broad, indicating that:
  - The outcomes might be inaccurate.
  - The performance heavily depended on the chosen train/test split.
- When trained on some of the data splits, the models showcased moderate ability to correctly distinguish within patients those who benefit and suffer from high PEEP (see Figure 2).
- All models identified features *age*, *platelets*, *urea* and *pco2* as most impactful for the CATE estimates.

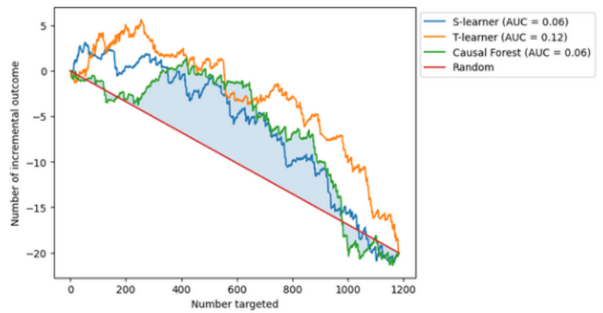

Figure 2: Qini curves for the models with optimal hyperparameters, found as best ones among 10 random data splits. We can see that all three models outperform the random baseline.

RCT DATASET

Not all features from MIMIC-IV were available in this dataset, thus we had to re-train the models (using the best found hyperparameters) omitting three of the selected variables.

- The real average treatment effect in this dataset was positive, while all our models estimated it to be negative.
- Qini AUC scores for all models were close to 0, indicating that the models performed only marginally better than a random baseline.

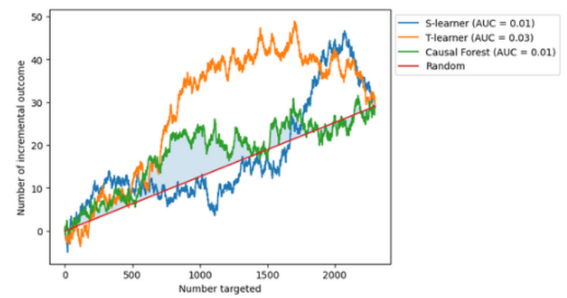

Figure 3: Qini curves for the RCT data. We can see that the AUC is much lower than on the MIMIC-IV data.

5. CONCLUSIONS & FUTURE WORK

- When trained with correct parameters and on correct data split, the models, to some extent, make valuable predictions.
- However, the results from RCT data suggest that the models offer little improvement over random CATE estimates.
- Increasing amount of data, as well as including other variables/parameters in our experiments could improve the models' performance and boost reliability of the results.

REFERENCES

[1] S. K. Sahetya and R. G. Brower, "Lung Recruitment and Titrated PEEP in Moderate to Severe ARDS: Is the Door Closing on the Open Lung?," JAMA, vol. 318, pp. 1327–1329, 10 2017
[2] A. Walkey et al. "Higher PEEP versus Lower PEEP Strategies for Patients with Acute Respiratory Distress Syndrome: A Systematic Review and Meta-Analysis," Annals of the American Thoracic Society, vol. 14, 10 2017
[3] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using randomforests," Journal of the American Statistical Association, vol. 113, no. 523, pp. 1228–1242, 2018
[4] S. R. Kunzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," Proceedings of the national academy of sciences, vol. 116, no. 10, pp. 4156–4165, 2019
[5] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-iv (version 2.2)." PhysioNet (2023)

CSE3000 Research Poster

TU Delft