Positive-Unlabeled Learning using Random Forest via Recursive Greedy Risk Minimization

Jonathan Wilton[1], Abigail M. Y. Koay[2], Ryan K. L. Ko[2], Miao Xu[2,3], Nan Ye[1]

[1] School of Mathematics and Physics, The University of Queensland
[2] School of Information Technology and Electrical Engineering, The University of Queensland
[3] RIKEN, Japan 103-0027

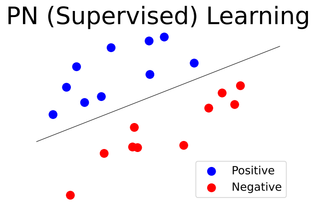
Motivation

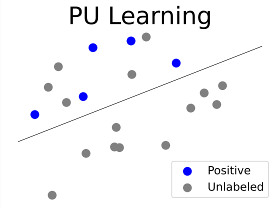
- Popular PU learning approaches use neural network classifiers
- RFs are promising but previously under-explored for PU learning tasks
- What we found from our novel PU RF algorithm:

	NN	RF
Predictive performance	✓	✓
Interpretability	✗	✓
Hyperparameter robustness	✗	✓

Background

Problem Setting





- Objective: learn a binary classifier g to minimize the expected risk

$$R(g) = \mathbb{E}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)}\, \ell(g(\boldsymbol{x}), y).$$

- Need to estimate the risk using only positive and unlabeled data

Background

Risk Estimators

- Unbiased (uPU) risk estimator[1]:

$$\widehat{R}_{\mathrm{uPU}}(g) := \sum_{\boldsymbol{x}\in P} w_{\mathrm{p}}\ell(g(\boldsymbol{x}),+1) + \overbrace{\sum_{\boldsymbol{x}\in U} w_{\mathrm{u}}\ell(g(\boldsymbol{x}),-1) - \sum_{\boldsymbol{x}\in P} w_{\mathrm{p}}\ell(g(\boldsymbol{x}),-1)}^{\text{may be negative} \,\to\, \text{overfitting}}$$

- Better: Nonnegative (nnPU) risk estimator[2]:

$$\widehat{R}_{\mathrm{nnPU}}(g) := \sum_{\boldsymbol{x}\in P} w_{\mathrm{p}}\ell(g(\boldsymbol{x}),+1) + \max\left\{0, \sum_{\boldsymbol{x}\in U} w_{\mathrm{u}}\ell(g(\boldsymbol{x}),-1) - \sum_{\boldsymbol{x}\in P} w_{\mathrm{p}}\ell(g(\boldsymbol{x}),-1)\right\}$$

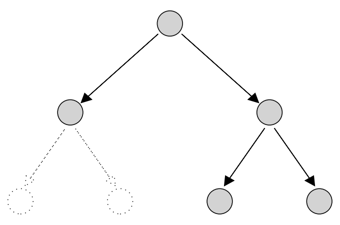
This paper: Construct decision tree to minimize PU risk estimator

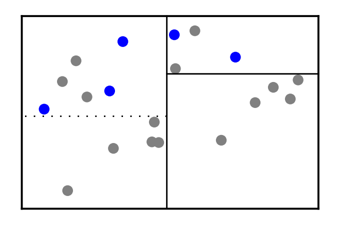
[1] Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. "Analysis of learning from positive and unlabeled data". In: *Advances in neural information processing systems* 27 (2014).

[2] Ryuichi Kiryo et al. "Positive-unlabeled learning with non-negative risk estimator". In: *Advances in neural information processing systems* 30 (2017).

PU Decision Tree Construction

Tree-growing by node splitting

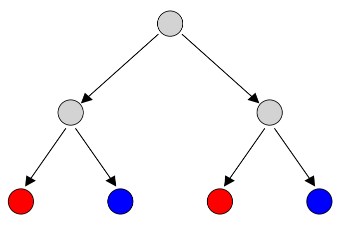


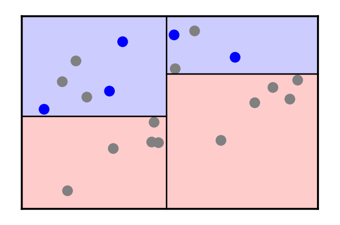


- Use binary splits to partition feature space in a recursive and greedy manner
- Split quality measured by:
  - **PN Learning:** decrease in label impurity based on PN data
  - **PU Learning:** decrease in risk estimate based on PU data
- **Special Cases**:
  - Quadratic loss $\rightarrow$ Gini impurity decrease
  - Logistic loss $\rightarrow$ entropy impurity decrease

PU Decision Tree Construction

Optimal Predictions





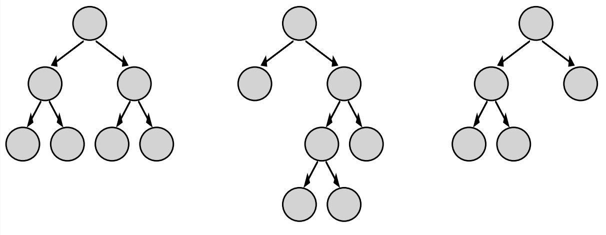
- Estimate proportion v^* of positive data at leaf node using weighted PU data
- Binary prediction that minimizes the risk estimator is

$$\begin{cases} +1, & v^* > 0.5 \\ -1, & v^* \leq 0.5 \end{cases}$$

- For uPU/nnPU risk estimators of many loss functions

Ensemble of PU Decision Trees

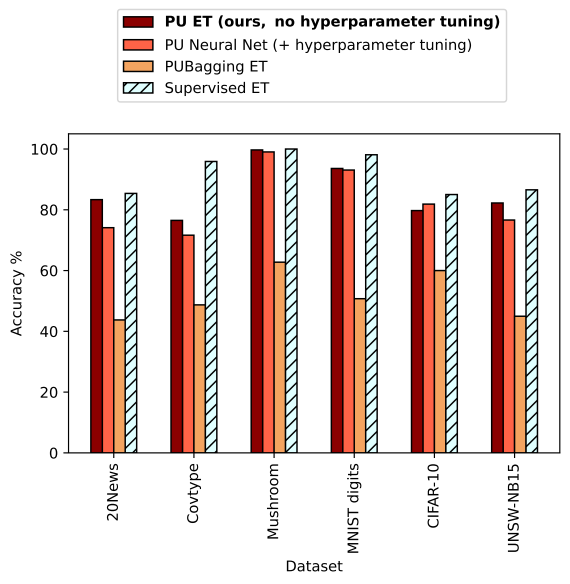
PU Extra Trees



- Combine predictions from many PU decision trees with majority vote
- Tree construction randomized for efficiency (based on Extra Trees[3])

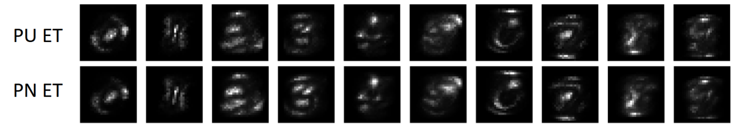
[3] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine learning* 63.1 (2006), pp. 3–42.

Experiments - Predictive Performance

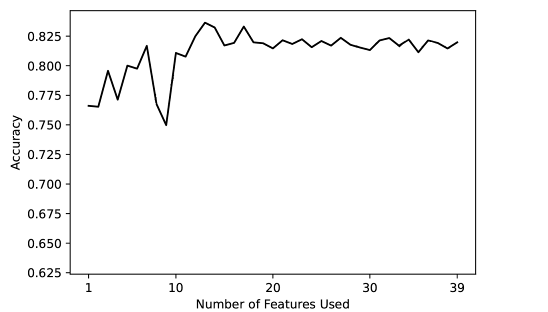


Experiments - Feature Importance

- PU feature importance score is the contribution to empirical risk reduction
- PU ET and supervised ET learn similar feature importances on MNIST.



- Our PU feature importance score is effective for selecting useful features on UNSW-NB15.



Conclusions

	NN	RF
Predictive performance	✓	✓
Interpretability	✗	✓
Hyperparameter robustness	✗	✓

- Additional experiments + theoretical results provided in our paper
- Code: *https://github.com/puetpaper/PUExtraTrees*