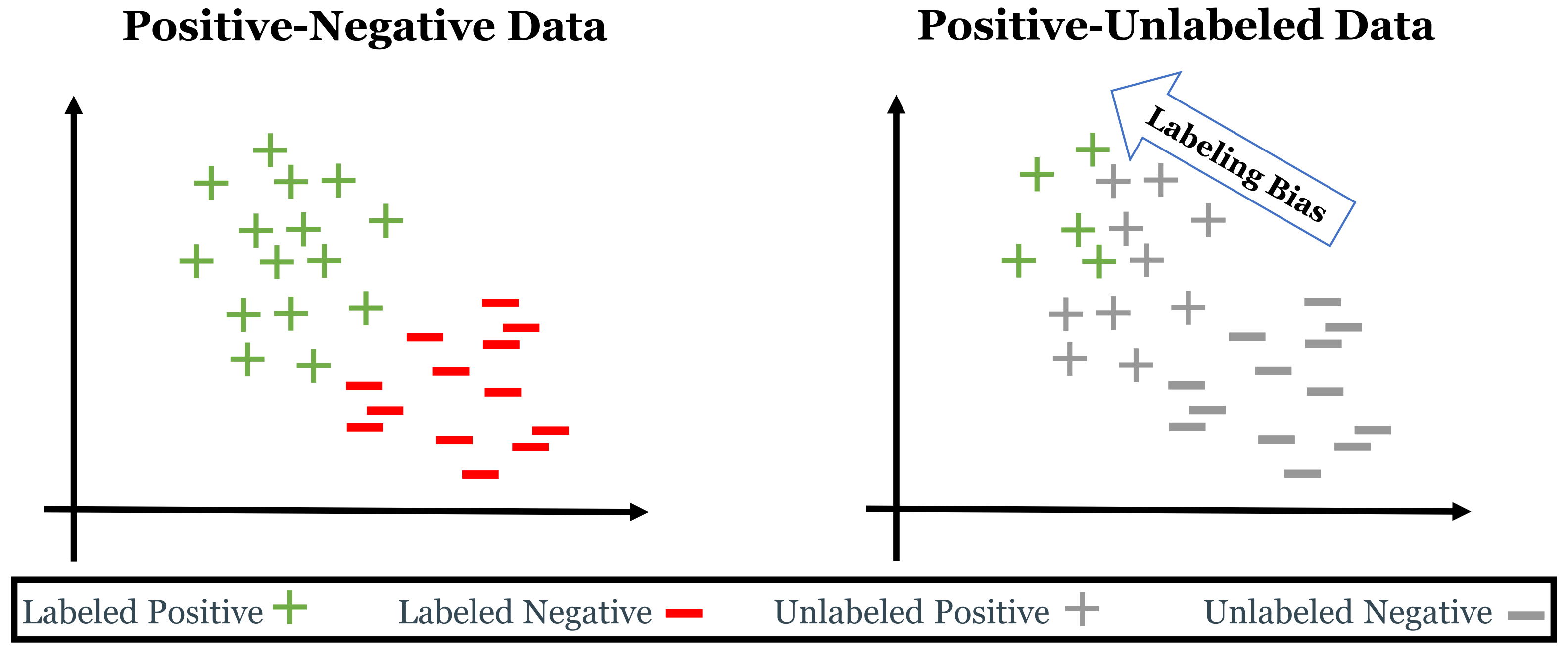
Recovering The Propensity Score From Biased Positive Unlabeled Data

Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, Elke Rundensteiner

Background: Positive-Unlabeled Data



PU Learning Definition:

$(x, \ell), x \rightarrow$ features, $\ell \rightarrow$ label, $\ell = 1 \rightarrow$ positive label, $\ell = 0 \rightarrow$ unlabeled

$p(\ell = 1 | y = -1) = 0$, $y \rightarrow$ ground truth, $y = 1 \rightarrow$ positive, $y = -1 \rightarrow$ negative

Goal: find $f(x)$ such that $f(x) = p(y = 1|x)$ given only Positive Unlabeled data

Motivating Example:

Instance

Possible Classes

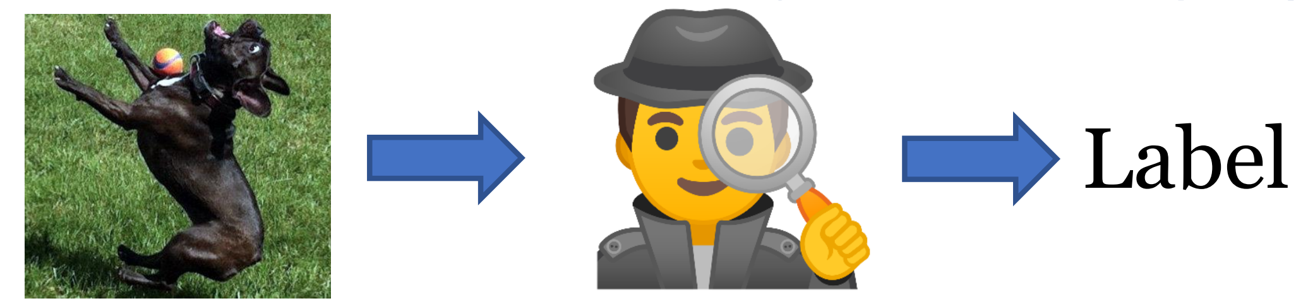
Person
Dog
Cat
Tree
Sky
⋮
Ball
Grass
Cow

- Too expensive/time consuming to label every class
- Apply labels only if positive instance of class

If annotator can miss classes, class not labeled => Unknown

Labels applied: + Dog, + Ball, + Grass
Missing label: + Grass

Key PU idea: Model The Labeling Mechanism (Propensity Score)



$e(x)$

$e(x) = p(l = 1|y = 1)$

Liu 2003
Elkan 2008
Kiryo 2017
Jain 2017
Guo 2020

Labeling is unbiased (Majority of PU methods)

$e(x) = p(l = 1|y = 1, x) \neq p(l = 1|y = 1)$

Labeling can be biased (Recent works)

Bekker 2019
Kato 2019
Jain 2020
Hammoudeh 2020

- Propensity score e(x): Probability that a true positive is labeled
- Knowledge of the propensity score lets us perform unbiased PN classification (Bekker 2019)
- Despite its importance, no prior work to determine when propensity score is **identifiable**
  - Identifiable: able to be uniquely recovered given sufficient data

Recovering The Propensity Score

Our Goal

1. Determine when the propensity score is **identifiable**
2. Recover true the propensity score when identifiable

Positive Unlabeled Assumptions

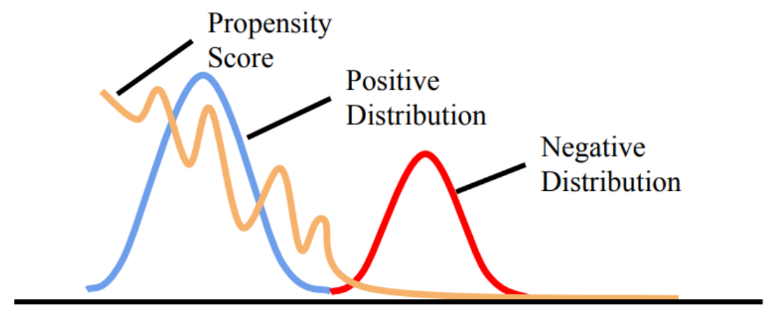
- Local Certainty/Separable Classes
  - Bayes Error of 0 between positive and negative distributions
- Positive Subdomain
  - There is some region A of the feature space determined by partial attribute assignment such that the Bayes error is 0
- Positive Function
  - There is some region A of the feature space determined by an arbitrary function for which the Bayes error is 0
- Irreducibility
  - The negative distribution is not a mixture containing the positive distribution

Theorem 1 *Let propensity score e be an arbitrary function of x, $e : \mathcal{X} \rightarrow (0, 1]$. Let the PU assumption hold (y is unobserved, ℓ and x are observed). Then, e is non-identifiable under the Positive Subdomain, Positive Function, and Irreducibly scenarios.*

- We show that in general the propensity score is **not** identifiable in the positive subdomain, positive function, and irreducibility scenarios

Identifiability Under Local Certainty

- Holds if the positive and negative distributions are separable



There is a **100%** probability of a "dog" in this picture

There is a **0%** probability of a "dog" in this picture

$$e^*(x) = \begin{cases} \frac{p(\ell = 1)p(x|\ell = 1)}{p(x)} & e^*(x) \neq 0 \\ 0 & e^*(x) = 0 \end{cases}$$

- We show that this is equivalent to propensity score under Local Certainty
- Easy to estimate from nonstandard classifier or density ratio estimation

Identifiability Under Probabilistic Gap

Recent biased PU methods utilize positive function + invariance of order (IOO)

- Invariance of order: $p(y = 1|x_1) > p(y = 1|x_2) \rightarrow p(\ell = 1|x_1) > p(\ell = 1|x_2)$

Theorem 2 *Let the Positive Function scenario and invariance of order assumption hold. Then, the propensity score is not identifiable*

- We show that in general the propensity score is **not** identifiable even with IOO

Scaled propensity: Strengthening IOO

- $e(x) := k * p(y = 1|x)$
- "Probabilistic Gap"

$$e^*(x) = \sqrt{Sup_{x \sim \chi}\,[p(\ell = 1|x)] * p(\ell = 1|x)}$$

- We show that this is equivalent to propensity score under Probabilistic Gap

Estimation approach:

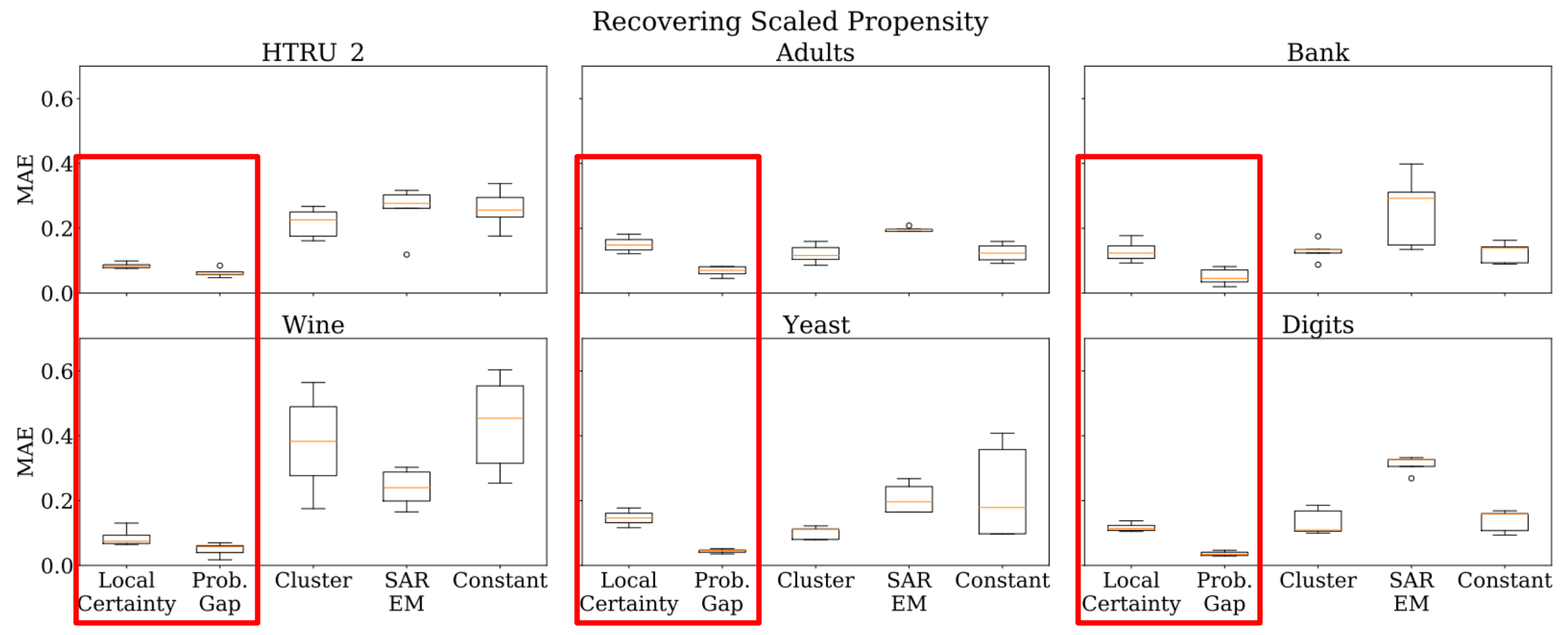
1. Train a probabilistic model $f_\ell(x)$ to predict labeling
2. Find k, maximum value of f_ℓ, on PU dataset
3. Set e*(x) = $\sqrt{k * f_\ell(x)}$

Experimental Evaluation

Compared Methods

- TiCE/Constant (Bekker 2018). Assumes propensity score is equal for all instances. Baseline.
- SAR-EM (Bekker 2019): Expectation-maximization algorithm for finding the propensity score
- Cluster (Jain 2020): Assumes that propensity score is constant within each cluster

Recovering Propensity Score



- Our probabilistic gap outperforms the state-of-the-art
- Our local certainty almost always performs equal or better to state-of-the-art
- Additional results in main manuscript

Using Recovered Propensity For Unbiased Classification

Dataset	HTRU 2	Adult	Bank	Wine	Yeast	Digits
LC (Ours)	**0.04**+/-0.00	**0.35**+/-0.02	**0.06**+/-0.01	**0.24**+/-0.02	**0.44**+/-0.00	**0.21**+/-0.01
PG (Ours)	0.10+/-0.00	0.40+/-0.00	0.22+/-0.01	0.36+/-0.02	0.47+/-0.00	0.33+/-0.01
Cluster	0.05+/-0.00	0.37+/-0.00	0.09+/-0.00	0.48+/-0.01	0.46+/-0.00	0.23+/-0.00
SE	0.10+/-0.05	0.37+/-0.01	0.09+/-0.01	0.47+/-0.05	0.46+/-0.01	0.33+/-0.03
Constant	0.43+/-0.00	0.70+/-0.01	0.17+/-0.01	0.34+/-0.02	0.46+/-0.00	0.29+/-0.01

Table 2: Classification error for arbitrary propensity score scenario

Dataset	HTRU 2	Adult	Bank	Wine	Yeast	Digits
LC (Ours)	0.14+/-0.01	0.75+/-0.01	0.38+/-0.03	0.26+/-0.01	0.45+/-0.01	0.51+/-0.02
PG (Ours)	0.05+/-0.00	**0.25**+/-0.03	**0.30**+/-0.01	**0.25**+/-0.02	**0.41**+/-0.01	**0.27**+/-0.01
Cluster	**0.04**+/-0.00	0.27+/-0.00	0.33+/-0.00	0.78+/-0.01	0.45+/-0.00	0.51+/-0.01
SE	0.14+/-0.08	0.26+/-0.01	0.35+/-0.01	0.51+/-0.06	0.44+/-0.01	0.49+/-0.03
Constant	0.43+/-0.00	0.46+/-0.03	0.39+/-0.01	0.34+/-0.03	0.46+/-0.00	0.54+/-0.01

Table 3: Classification error for scaled propensity score scenario

- Our approaches almost always lead to more accurate classifiers

Conclusion

In this work, we:

- Laid the groundwork for identifiability of the labeling mechanism for biased PU setting
- Proved that the propensity score is not identifiable for most common PU settings
- Identified two scenarios for which the propensity score is identifiable
  - One with strong distribution assumptions but weak assumptions on propensity function
  - One with weak distribution assumptions but strong assumptions on propensity function
- Provided a methods to recover the propensity score in those two settings

Acknowledgements

- WPI WASH Research group
- WPI DAISY Lab
- DARPA WASH Grant #FA8750-18-2-0077

References

Liu, B. et al. 2003. Building text classifiers using positive and unlabeled examples. ICDM.
Elkan, C. et al. 2008. Learning classifiers from only positive and unlabeled data. KDD.
Kiryo, R. et al. 2017. Positive-unlabeled learning with non-negative risk estimator. NeurIPS.
Jain, S. et al. 2017. Recovering True Classifier Performance in Positive-Unlabeled Learning. AAAI.
Guo, T. et al. 2020. On positive-unlabeled classification in GAN. CVPR.
Bekker, J. et al. 2019. Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data. ECML PKDD.
Kato, M. et al. 2019. Learning from Positive and Unlabeled Data with a Selection Bias. ICLR.
Jain, S. et al. 2020. Class Prior Estimation with Biased Positives and Unlabeled Examples. AAAI.
Hammoudeh, Z. et al. 2020. Learning from Positive and Unlabeled Data with Arbitrary Positive Shift. NeurIPS.