The University of Texas at Austin
McCombs School of Business

Mitigating Label Bias via Decoupled Confident Learning

Label Bias Paper

Yunyi Li, Maria De-Arteaga, Maytal Saar-Tsechansky
University of Texas at Austin

Yunyi Li Maria Maytal

Abstract

Bias in labels is pervasive across important domains. However, there is a lack of methodologies to address this problem. We propose a pruning method — **De**coupled **Co**nfident **Le**arning (DeCoLe) — to **mitigate label bias**. After illustrating its performance on a synthetic dataset, we apply DeCoLe in the context of hate speech detection, and show that it successfully identifies biased labels and outperforms competing approaches.

Highlights

- ***Label bias*** refers to a systematic disparity between the ground truth labels intended to train an AI system and the observed labels, such that **the relationship underlying the mismatch differs across groups**. Label bias is very common in **human generated labels**.
- DeCoLe use decoupled classifiers to estimate label confidence and perform **group-specific pruning** to **reduce label bias**.
- DeCoLe is a **model-agnostic, data-centric** algorithm.

Algorithm

Notations: $\tilde{y} \rightarrow$ Observed Label;
$y^* \rightarrow$ Laten Ground Truth Label;
$g \rightarrow$ Group Indicator.

Assumptions: Suppose there exists a **group and class conditional noisy labeling process** that results in bias in observed labels $\tilde{y}$. For each group g_i, where i refers to a specific value of g, we have:

False Negative Rate of $g_i \rightarrow \pi_{0g_i} = P(\tilde{y} = 0 | y^* = 1, g = i)$

False Positive Rate of $g_i \rightarrow \pi_{1g_i} = P(\tilde{y} = 1 | y^* = 0, g = i)$

Algorithm 1 Decoupled Confident Learning

Input: Noisy dataset $\boldsymbol{D} := (\boldsymbol{x}, \tilde{y})^n$, group indicator g, initialize a set of classifiers $\{\text{clf}_{g_1}, \ldots, \text{clf}_{g_k}\}$

for $i = 1$ **to** k **do**

Part 1: Estimating $p(\boldsymbol{x})$ and thresholds

$\text{clf}_{g_i}.\text{fit}(\boldsymbol{x}_{g_i}, \tilde{y})$ where $\boldsymbol{x} \in g_i$

$\hat{p}(\boldsymbol{x}_{gi}) \leftarrow \text{clf}_{g_i}.\text{predict\_crossval\_prob}\ (\tilde{y} = 1 | \boldsymbol{x}_{g_i})$

$\text{LB}_{g_i} = \text{LB}(y^* = 1, g = i) = E_{\boldsymbol{x} \in \tilde{y}=1, g=i}[\hat{p}(x)]$

$\text{UB}_{g_i} = \text{UB}(y^* = 0, g = i) = E_{\boldsymbol{x} \in \tilde{y}=0, g=i}[\hat{p}(x)]$

Part 2: Pruning

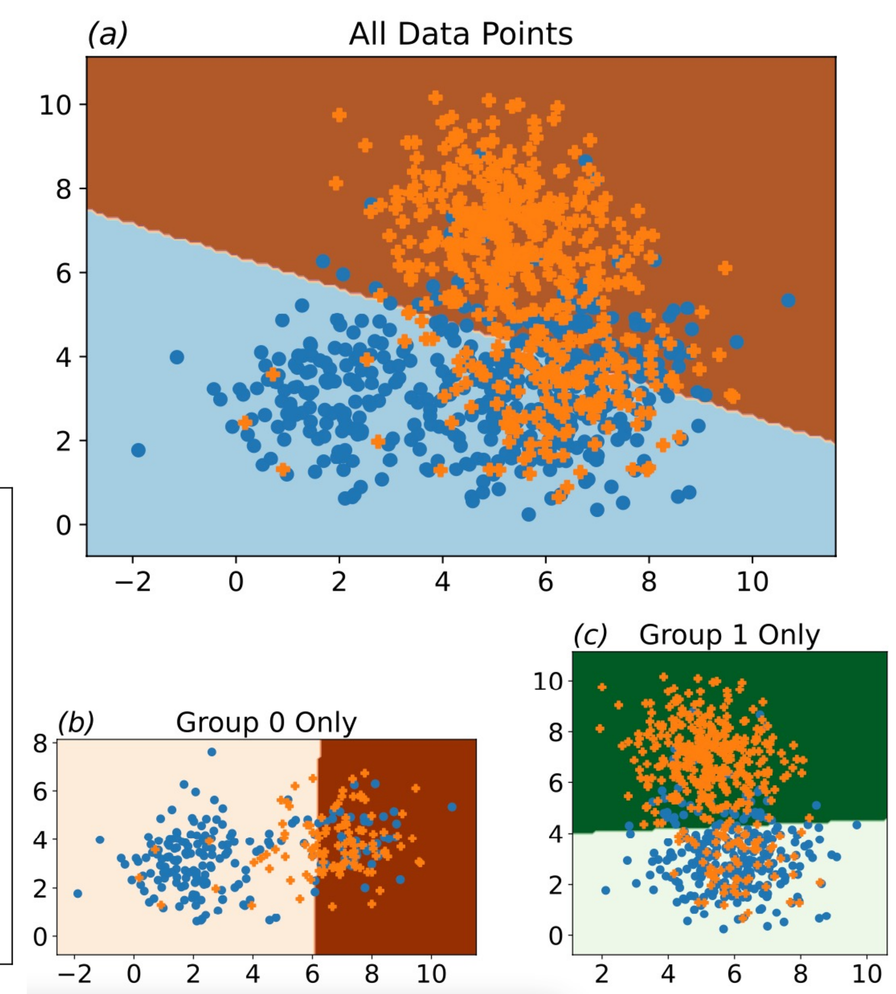
Remove $(\boldsymbol{x}_{g_i}, \tilde{y}) \in \boldsymbol{D}$ where $\tilde{y} = 1, \hat{p}(\boldsymbol{x}_{g_i}) < \text{UB}_{g_i}$

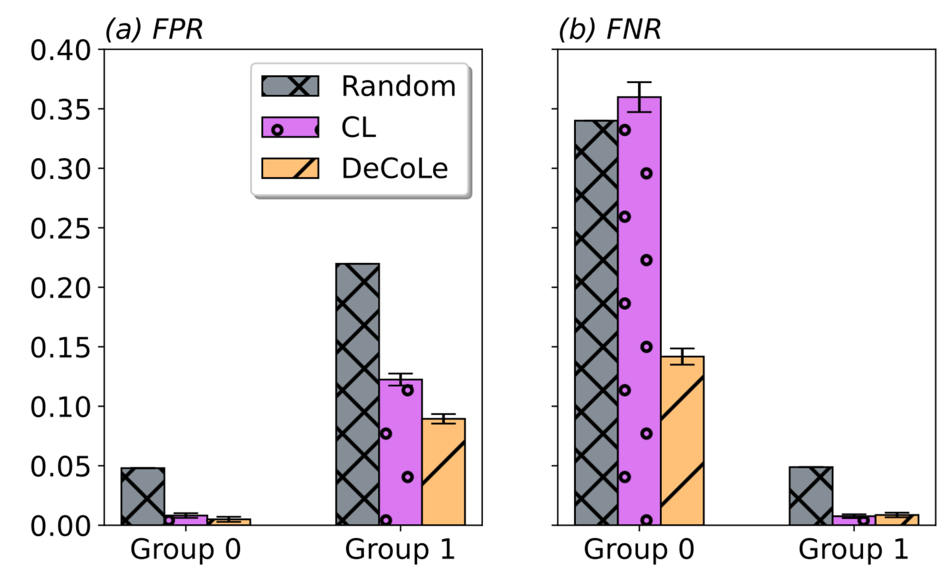
Remove $(\boldsymbol{x}_{g_i}, \tilde{y}) \in \boldsymbol{D}$ where $\tilde{y} = 0, \hat{p}(\boldsymbol{x}_{g_i}) > \text{LB}_{g_i}$

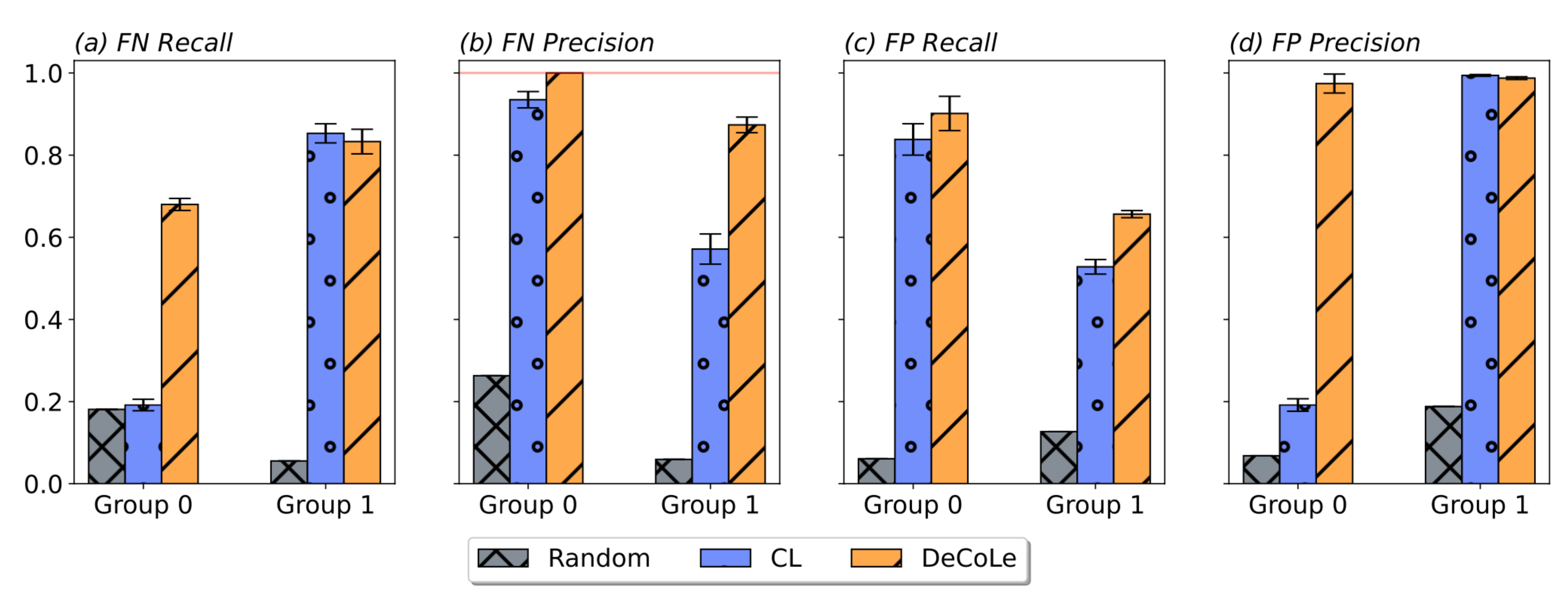
end for

Synthetic Experiments

We create a dataset with **group and class-conditional noise rates**. This allows us to have control of the relationship between $\tilde{y}$ and y^*. We also consider **group imbalance** (70% majority), and **differential sub-group validity**.



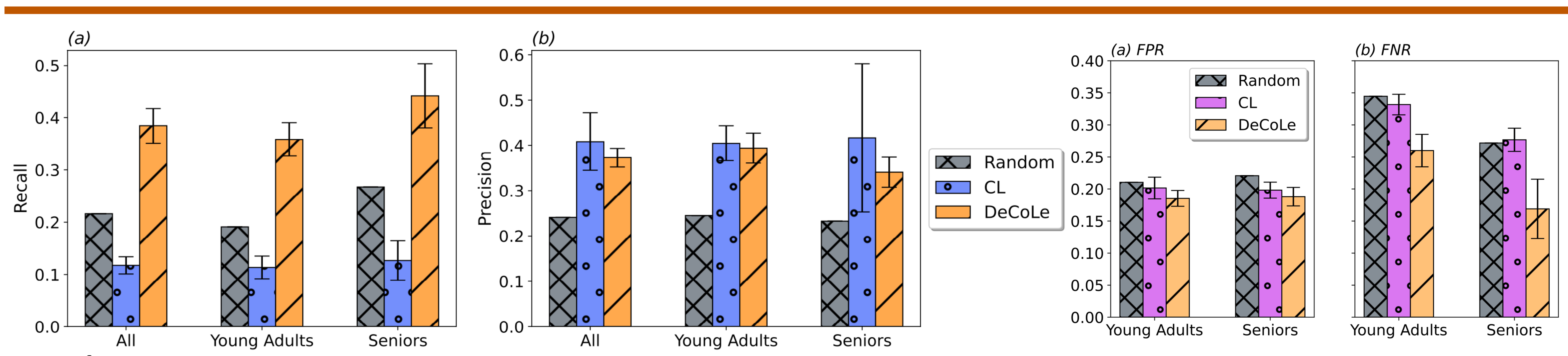




Results: 1. DeCoLe significantly outperforms CL in all scenarios, with **particularly remarkable higher accuracy** in correctly **identifying erroneous labels** of group g0, the **disadvantaged group**.

2. DeCoLe is substantially **more capable of mitigating** the two **most prominent error types** (false positives for group g1 and false negatives for group g0) and thus is more suitable for **preventing systematic bias** present **in labels**.

Hate Speech Label Bias Mitigation



Backgrounds:

- Hate speech **causes significant harm**. It is used to radicalize and recruit within extremist groups, incite violence.
- **Labeling hate speech is challenging** given that judgments of offensiveness depend on societal context.
- **Hate speech labels**, typically generated by crowdsourcing annotators, are **bias prone**.

Results:
1. DeCoLe demonstrates significant **improvement in recall** of erroneous labels compared to other methods.
2. DeCoLe significantly **reduces false negatives** for **both** posts **targeting young adults** and posts **targeting seniors**, surpassing the performance of CL algorithm.

Conclusion & Future Research

- We propose a novel approach called **De**coupled **Co**nfident **Le**arning (DeCoLe), a pruning method that mitigates label bias.
- DeCoLe **improves** upon existing noise-mitigation alternatives by **accounting for** the fact that noise may be **group and class conditioned**.
- Our **experimental results**, which focus on the **hate speech** domain, **validate the effectiveness of DeCoLe** in pruning erroneous instances and mitigating group-specific false negatives associated with hate speech labels.
- **Future research endeavors** should focus on the development of methodologies capable of handling **other forms of label bias structures**.