Boosting subgroup performance without any group annotations

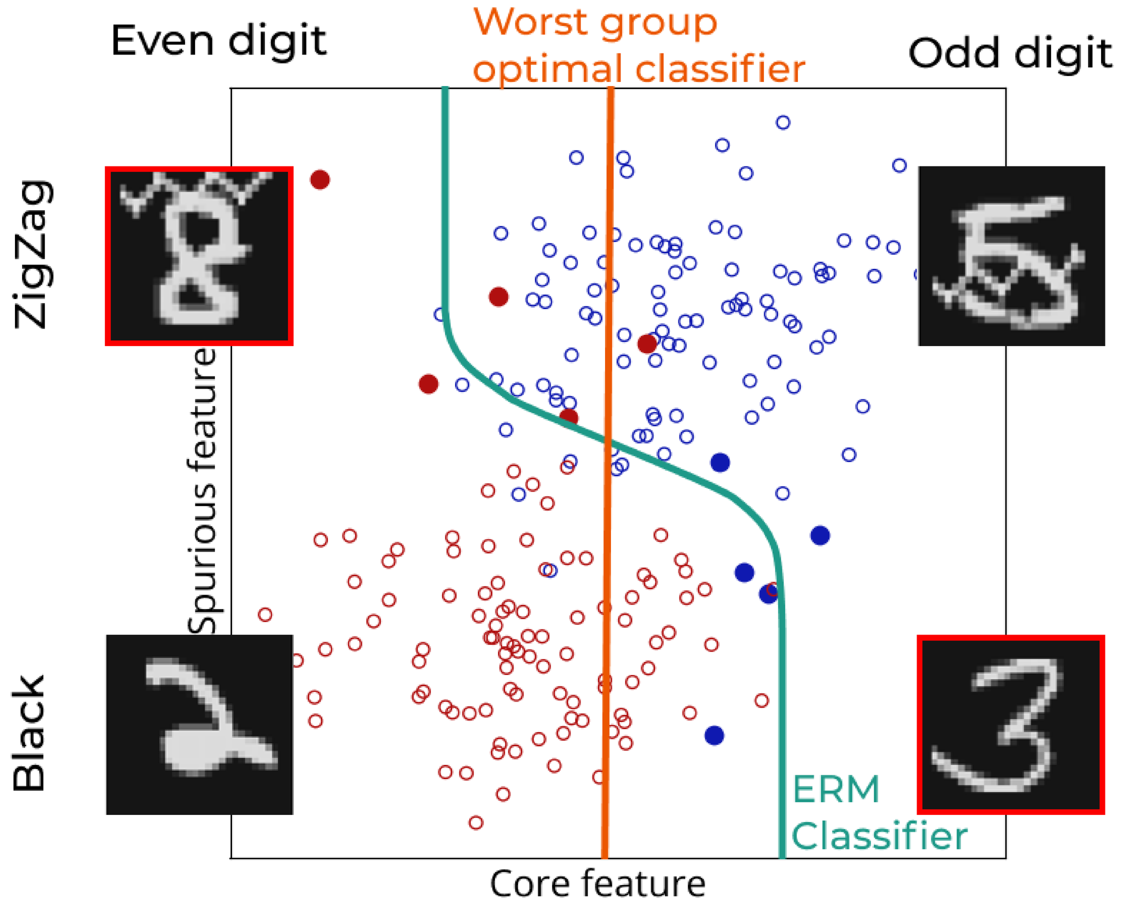
Vincent Bardenhagen, Alexandru Ţifrea, Fanny Yang

Department of Computer Science, ETH Zurich

Data with imbalanced subpopulations

Setting:

- Training data: Consists of imbalanced subgroups.
- Majority groups: Spurious features correlated with class labels.
- Minority groups: Only core features are predictive of class label.



Problem:

1. ERM classifier predictions are incorrect on underrepresented subgroups.
2. Problem can be mitigated given group labels. But they are difficult to collect.

Previous approaches

Use **train + validation** data with group annotations:

- **Drastically improve worst group accuracy**
- **Require large amounts of annotated data**

Use **validation** data with group annotations:

- **Similar worst-group accuracy, at a reduced labeling cost**
- **Sometimes impossible to collect group-annotations (e.g. ethnicity, sexual orientation etc.)**

Can we improve worst-group accuracy when **no group labels** are available?

YES! We perform model selection without group labels!

Debiased classifier in two simple steps

- **First stage:** Train biased predictor $\hat{f}_{t_1,\theta_1}$ using regularization t_1 and optimal hyperparameters θ_1.

  Annotate samples in the error set of $\hat{f}_{t_1,\theta_1}$ as minority ($g=1$).

  $\bar{S}(\hat{f}_{t_1,\theta_1}) = \{(x_i, y_i, g_i) : (x_i, y_i) \in S, g_i = \mathbb{1}[\hat{f}_{t_1,\theta_1}(x_i) \neq y_i]\}$
- **Second stage:** Train unbiased predictor using IW/GDRO:

$$\hat{f}_{t_1,\theta_1,t_2,\theta_2} = \underset{f \in \mathcal{F}(t_1,\theta_1,t_2,\theta_2)}{\arg\min} \mathcal{L}_{\mathrm{IW}}(f, \bar{S}(\hat{f}_{t_1,\theta_1}))$$

- **Prior work:** Hyperparameter tuning **requires group labels!**

  Select $t_1^*, \theta_1^*, t_2^*, \theta_2^*$ using WgAcc on group-annotated $\bar{V}_{\text{oracle}}$.

Model selection without group labels

1. Early-stopping after one epoch: $t_1^* = 1$
2. Optimize AvgAcc on V to increase bias:

$$\theta_1^* \in \underset{\theta_1 \in \Theta_1}{\arg\max} \, \mathbf{AvgAcc}(\hat{f}_{t_1,\theta_1}, V)$$

3. Optimize WgAcc wrt estimated group labels $\bar{V}(\hat{f}_i)$:

$$t_2^*, \theta_2^* \in \underset{t_2,\theta_2 \in \mathrm{T}_2 \times \Theta_2}{\arg\max} \underbrace{\frac{1}{K}\sum_{i=1}^{K} \mathbf{WgAcc}(\hat{f}_{t_1^*,\theta_1^*,t_2,\theta_2}, \bar{V}(\hat{f}_i))}_{\mathcal{C}(\hat{f}_{t_1^*,\theta_1^*,t_2,\theta_2}, \{\bar{V}(\hat{f}_i)\}_{i=1}^{K})}$$

Experimental results

		Corrupt-MNIST	Waterbirds	CelebA	Color-MNIST	Adult	Poverty
No group labels	ERM	71.2	74.9	60.7	82.6	41.6	**55.6**
	Ours	**96.5**	**78.5**	**78.9**	**96.6**	**68.0**	50.0
Val group labels	ERM WG	79.8	**86.7**	77.8	84.4	61.2	51.5
	JTT	**91.3**	**86.7**	**81.1**	**94.8**	**63.3**	**60.5**
Train & val group labels	GDRO	93.1	89.4	92.2	93.1	71.4	67.5

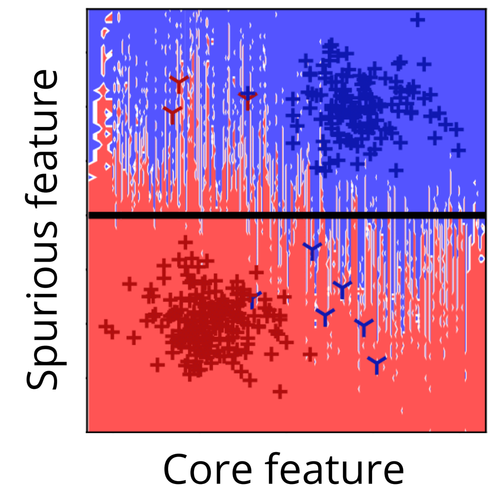
Our procedure significantly outperforms the ERM baseline

- similar performance to approaches that use group labels.

Empirical insights

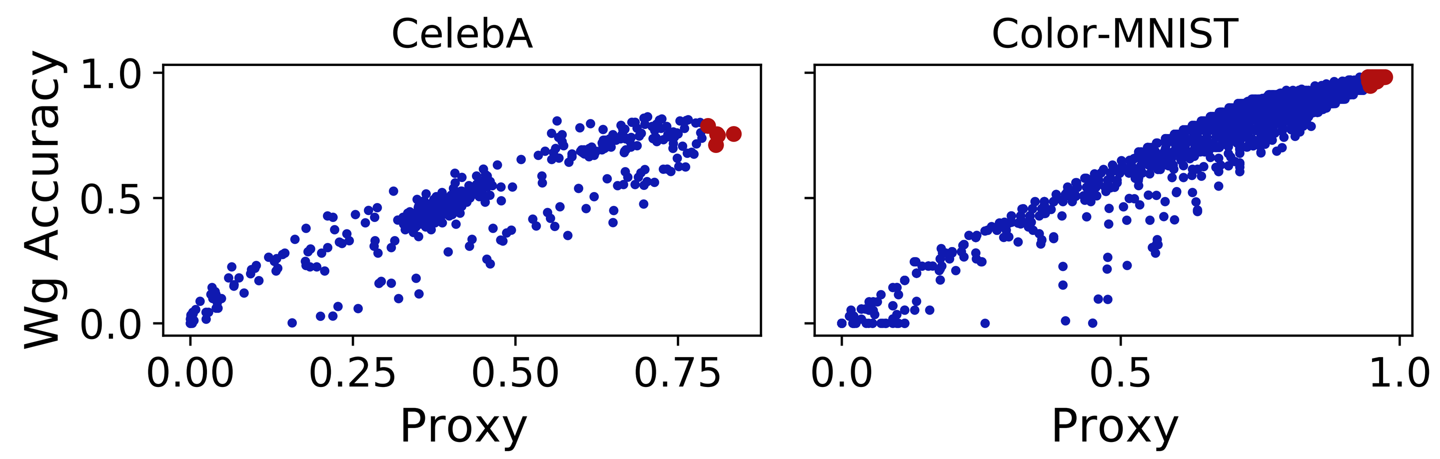
First stage:

Extreme regularization + optimizing AvgAcc leads to biased predictors



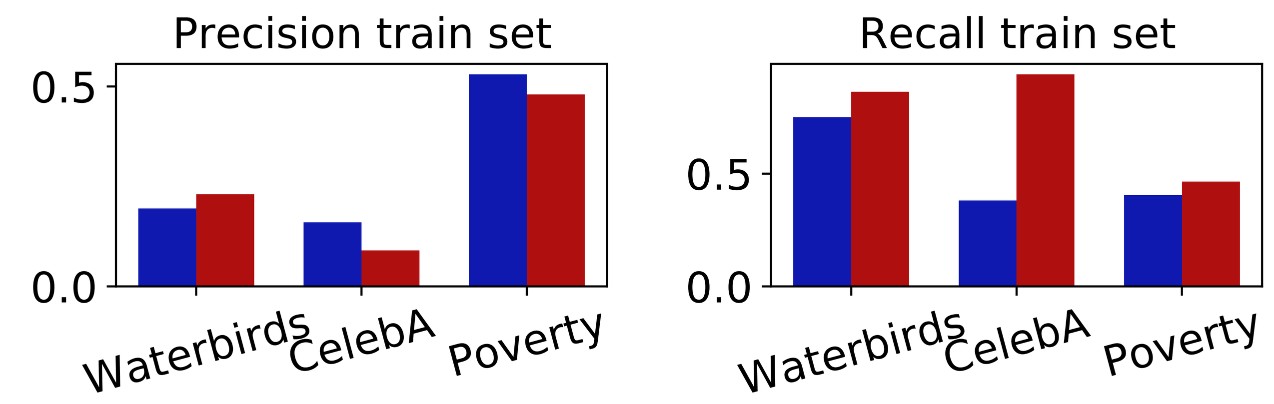
Second stage:

Criterion $\mathcal{C}(f, \{\bar{V}(\hat{f}_i)\}_{i=1}^{K})$ correlates well with WgAcc on $\bar{V}_{\text{oracle}}$



Ablation studies

1. Identifying minority samples: Access to group labels (JTT) can increase recall, but precision remains similar to ours.



2. Ensembling: sample consistently identified as minority have higher impact on model selection.
   - optimal early-stopping time is selected more reliably