Improving class and group imbalanced classification with uncertainty-based active learning

Alexandru Ţifrea*, John Hill*, Fanny Yang

ETH Zürich, Georgia Tech

Active learning

Goal: Get same accuracy as passive learning with less labeled data

- prior focus: average-case performance

One common strategy: Uncertainty-based AL (U-AL):

- Repeat until labeling budget is exhausted
  1. train classifier $\hat{f}$ on current labeled set
  2. label points of highest uncertainty wrt classifier $\hat{f}$ e.g. points close to the current decision boundary
- often **performs poorly** [Mussmann et al'18; Ţifrea et al'23] and cannot decide a priori if U-AL is suitable for a task

Alternative to U-AL: representativeness-based AL

- ϵ-greedy U-AL, BADGE, Coreset AL, TypiClust, etc

How does U-AL perform on imbalanced classification problems?

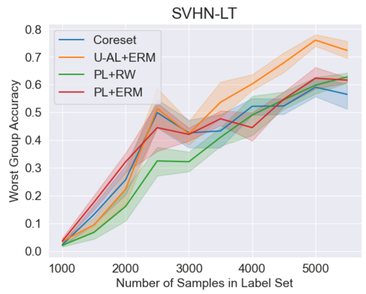
U-AL selects a more balanced dataset

Proposition for symmetric 2-GMM data

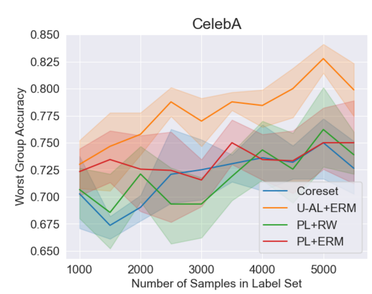
U-AL collects more balanced labeled set than passive learning.

- similar observation by [Ertekin et al'07] for Bayes classifier

U-AL improves worst-group accuracy



(a) Class imbalance



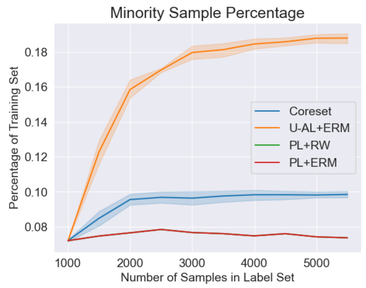
(b) Group imbalance

- U-AL improves accuracy for **both the worst-class and worst-group**

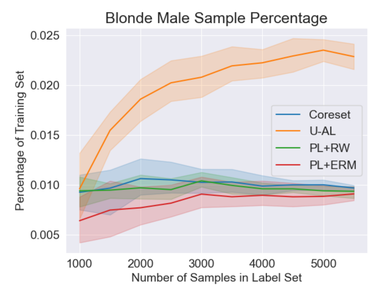
Note: U-AL can improve further with re-weighting (RW)

U-AL vs other AL strategies

- U-AL collects more balanced data for **both class- and group-imbalance**



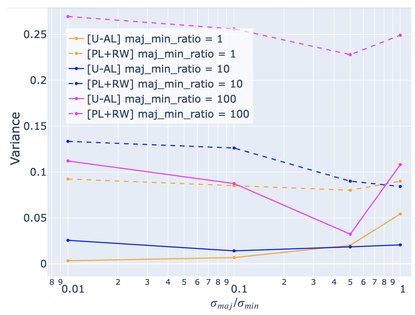
(a) SVHN-LT



(b) CelebA

U-AL vs PL mitigations

- U-AL does not hurt variance as much as PL mitigations (e.g. RW)

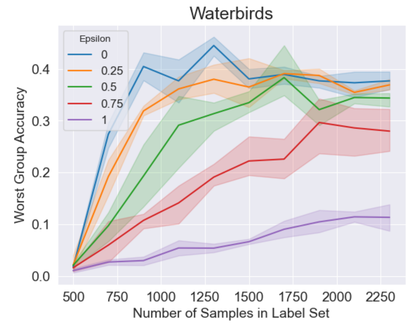


Representativeness-based AL

Idea: interpolate between U-AL and PL e.g. ϵ-greedy U-AL

- U-AL = informativeness
- PL = representativeness

PL with probability ϵ, U-AL with probability $1-\epsilon$



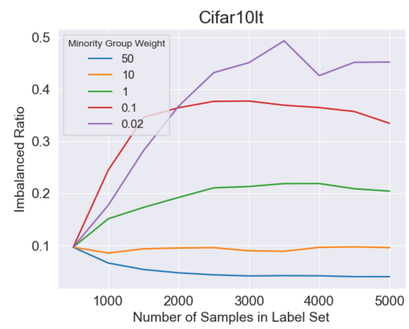
Takeaway: representativeness-based AL hurts worst-class and worst-group accuracy even when combined with U-AL

Biased classifiers sample balanced data

Idea: bias the classifier used for sampling new labeled data

- re-weighting with lower (rather than higher) weight for minority samples

  i.e. increase (rather than decrease) bias



Takeaway: increasing classifier bias leads to collecting a more balanced labeled dataset