Minority Class Oriented Active Learning for Imbalanced Datasets

Umang Aggarwal[1,2]; Adrian Popescu[1]; Celine Hudelot[2]
[1]CEA- LIST, [2]Centrale Supelec, France

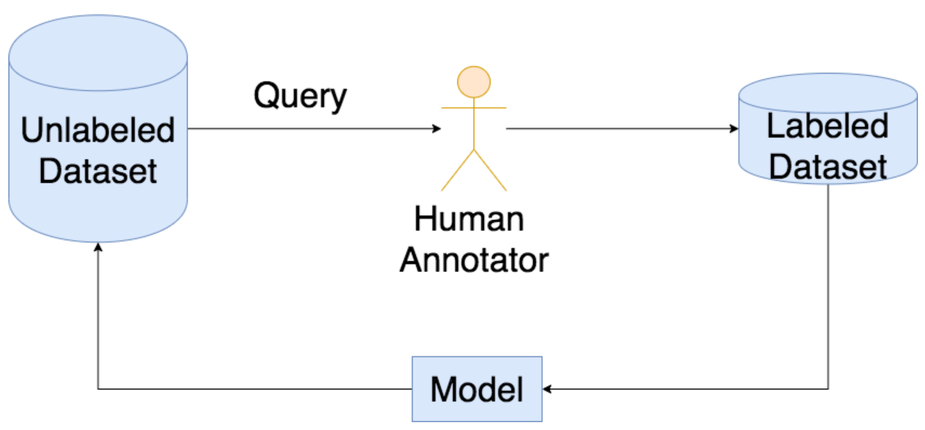
Introduction

Context
1. Iterative Active learning with small initial dataset
2. Unlabelled dataset can contain class imbalance

Motivation
1. Unstable model predictions at start of iterative active learning process
2. Imbalance in unlabelled dataset is propagated to labeled dataset

Solutions
1. Selecting samples that are predicted as minority class.
2. Learning shallow classifier over fixed representation as an alternate to classical fine tuning strategy.

Iterative Active Learning Pipeline



Proposed Method

Minority Class Oriented Sampling

1) Selecting samples predicted as minority class

Samples selected for a class:

$$\mathbb{D}_c^{U(k)} = \{\forall x \in \mathbb{D}_k^U, if\ P(c^1 = c|x)\}$$

Motivation:

if the sample is annotated as minority class :

help to mitigate imbalance

else if annotated as majority class :

help in decision boundary of minority class

2) Number of samples per class depends on imbalance and budget

For a given class (c), at iterative step (k):

Average number of class (μ_k) - Budget / number of classes.

Number of samples in class (c)

$$m_k^c = \begin{cases} \mu_k - s_k^c, & \text{if } s_k^c < \mu_k \\ 0, & \text{otherwise} \end{cases}$$

3) Allows use of any other AF if imbalance is mitigated or if not enough minority class samples for found

Three variants of the methods to select either certain, uncertain or most diverse samples belonging to the minority class

1) Certainty-oriented Minority Class Sampling

$$CMCS = arginvsort_{\forall x \in \mathbb{D}_c^{U(k)}} marg(x)$$

2) Uncertainty-oriented Minority Class Sampling

$$UMCS = argsort_{\forall x \in \mathbb{D}_c^{U(k)}} marg(x)$$

3)Diversity-oriented Minority Class Sampling

$$DMCS = core(\mathbb{D}_c^{U(k)}, \mathbb{D}_c^{L(k)})$$

Experiments

Dataset	Class	Images	Mean(μ)	Std(σ)	ir
FOOD-101	101	22956	227.28	180.31	0.793
CIFAR-100	100	17168	171.68	126.98	0.740
MIT-67	67	14281	213.15	168.16	0.789

TABLE I
DATASET STATISTICS. ir IS THE IMBALANCE RATIO.

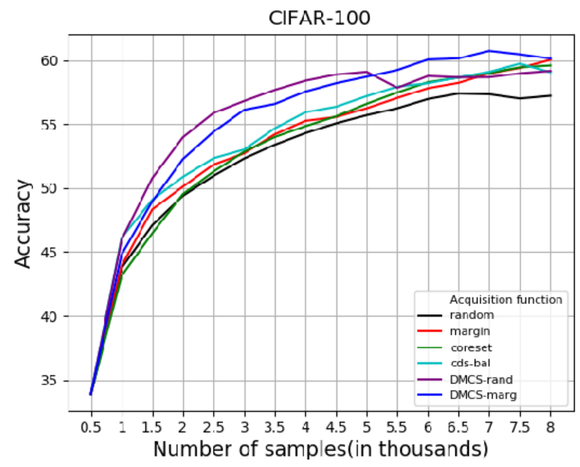
Initial Budget- 500, Iteration- 15, Total budget - 8000, Model- ResNet18

Training schemes
1. Fine-tuning ResNet18 with thresholding
2. Cost-Sensitive SVM over pre-trained ResNet18 features

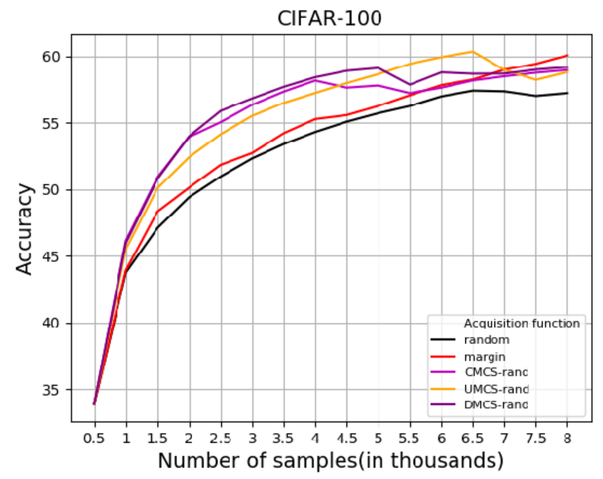
Results

Baselines:
1. Random Sampling
2. Uncertainty based- margin sampling
3. Core set

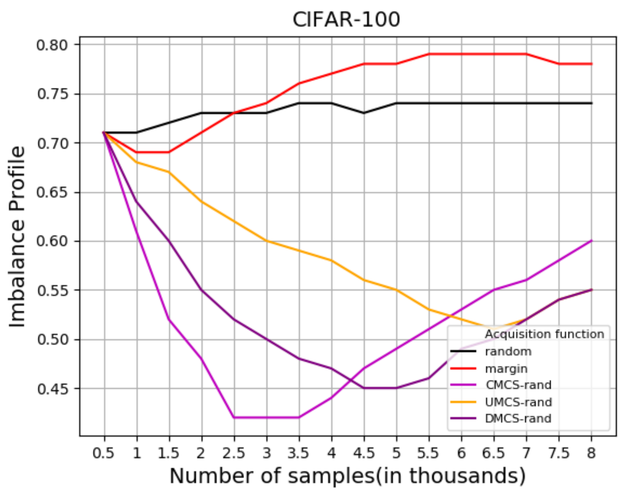
Iterative active learning performance for baselines and the proposed method DMCS



Iterative active learning performance for baselines and the three variants of proposed method.



Imbalance profiles of different methods



Conclusions

- Imbalance needs to be treated at the time of sample selection
- Cost-Sensitive SVM over fixed representation acts as a good alternative to CNN-FT
- Certainty-oriented Minority Class Sampling provides best mitigation to imbalance, while diversity-oriented minority class sampling performs best overall

Contact Email: umang.aggarwal@cea.fr