# Learning To Sample: an Active Learning Framework

Jingyu Shao, Qing Wang and Fangbing Liu

Research School of Computer Science
Australian National University

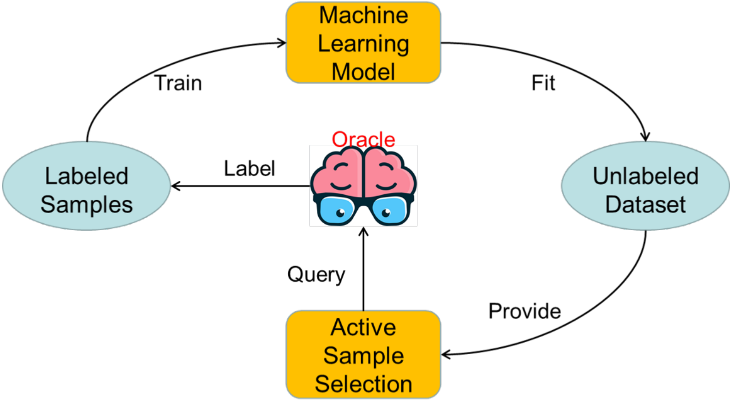
November 11, 2019

Acknowledgement: This work was partially funded by the Australian Research Council (ARC) under Discovery Project DP160101934

# Introduction: Active Learning

- Active learning seeks for the most representative and informative samples to be labeled by leveraging observations from previously labeled samples.

- Active learning seeks for the most representative and informative samples to be labeled by leveraging observations from previously labeled samples.
- A general active learning process:



- **Limitation**

  No one-fit-all solution for active learning, i.e., the “best” varies due to the variety of datasets and machine learning models.

- **Limitation**

  No one-fit-all solution for active learning, i.e., the "best" varies due to the variety of datasets and machine learning models.

- **Solution**

  Instead of using pre-defined strategies for active learning, we consider to learn the "best" active learning strategy based on the estimated performance of a model.

## Related Work: Learning based Active Learning

- *Active Learning by Learning* (ALBL) relates active learning with multi-armed bandit learner.

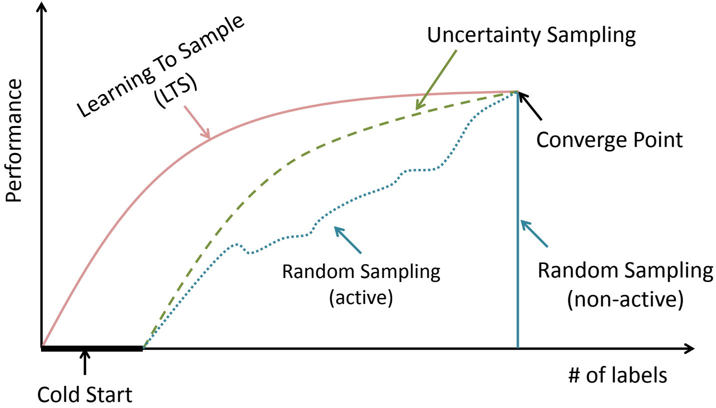
  Active learning by learning, Hsu and Lin, AAAI 2015

- *Learning Active Learning* (LAL) aims to train a regressor which can predict the generalization error reduction of each unlabelled instance and greedily select one with highest error reduction for labelling.

  Learning active learning from data, Konyushkova et al., NIPS 2017

# Our Objective

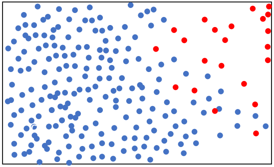
- To build a learning-based active learning framework:



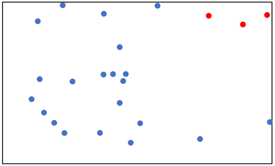
# Uncertainty and Diversity



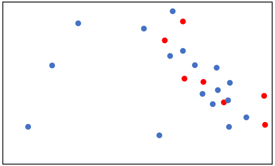
- Uncertainty sampling: using a function to measure uncertainty, e.g., probabilistic confidence, fisher information and entropy
- Diversity sampling: considering data distribution (e.g., samples with different feature values).



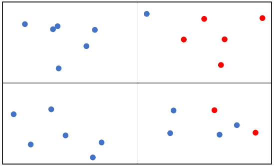
(a) Entire Data Distribution



(b) Random Sampling

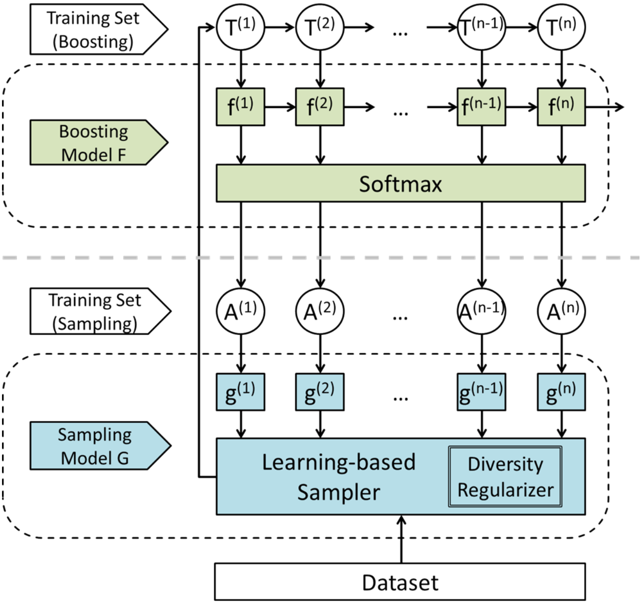


(c) Uncertainty Sampling



(d) Diversity Sampling
(4 groups)

# Learning to Sample (LTS)



# Learning to Sample (LTS)

- Two key components: a boosting model $F$ and a sampling model $G$, dynamically learn from each other in iterations for improving the performance of each other.

- Two key components: a boosting model $F$ and a sampling model $G$, dynamically learn from each other in iterations for improving the performance of each other.

- The sampling model $G$ incorporates uncertainty and diversity of samples into a unified process for optimization.

- Two key components: a boosting model $F$ and a sampling model $G$, dynamically learn from each other in iterations for improving the performance of each other.
- The sampling model $G$ incorporates uncertainty and diversity of samples into a unified process for optimization.
- We actively select samples based on the joint impacts of probabilities of being mis-classified by a boosting model and the distribution of samples in a sample space.

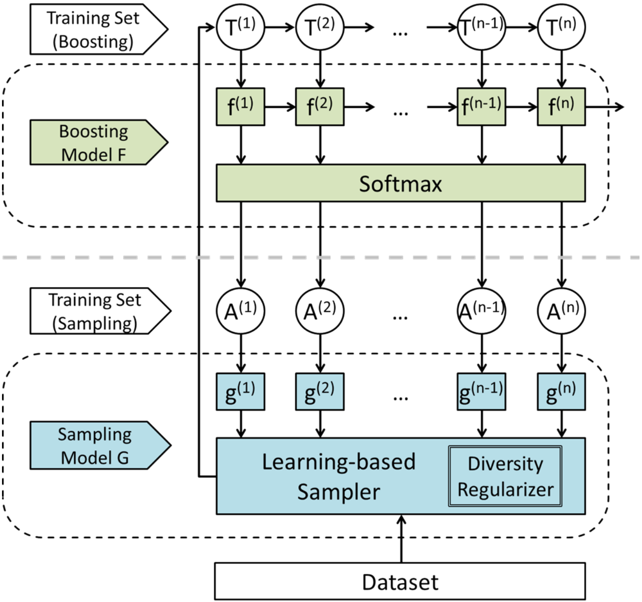
- Given a training set $T^{(t)}$, $f^{(t)} \in F$ in the $t$-th iteration is trained by minimizing:

$$\sum_{(x_i, y_i) \in T^{(t)}} \ell_1(\hat{y}_i^{(t-1)} + f^{(t)}(x_i), y_i) + \Omega_1(f^{(t)})$$

  - $\hat{y}_i^{(t-1)} = \sum_{k=1}^{t-1} f^{(k)}(x_i)$ is the prediction of $x_i$ in the ($t$-1)-th iteration;
  - $\ell_1$ is a differentiable loss function;
  - $\Omega_1(f^{(t)})$ is the penalty for the complexity of $f^{(t)}$.

- $F$ is a sequence of functions $\langle f^{(1)}, \ldots, f^{(n)} \rangle$.

# Learning based Framework (LTS)



# Sampling Model $G$

- A sampling model $G$ actively selects a set $\Delta^{(t)}$ of representative samples at the $t$-th iteration by:

$$\begin{aligned} \textbf{maximize} \quad & \sum_{i=1}^{k} v_i g^{(t)}(x_i) + \alpha \times \Gamma(\mathbf{v}) \\ \textbf{subject to} \quad & ||\mathbf{v}||_1 = |\Delta^{(t)}| \end{aligned}$$

where $k = |X_U^{(t)}|$, $\mathbf{v} = (v_1, ..., v_k)^T \in \{0, 1\}^k$ and $\alpha$ is a parameter.

- Two kinds of sampling strategies:
  - $g^{(t)}(x_i)$ as a regressor, learns the uncertainty of samples which are likely to be mis-classified by the boosting model;
  - $\Gamma(\mathbf{v})$ as a regularizer, controls the diversity of samples in terms of distribution.

- Given a training set $A^{(t)}$, the regressor for uncertainty sampling is trained by minimizing:

$$\sum_{(x_i, z_i^{(t)}) \in A^{(t)}} w_i^{(t)} \ell_2(g^{(t)}(x_i), z_i^{(t)}) + \Omega_2(g^{(t)})$$

where:

  - $A^{(t)} = \{(x_i, z_i^{(t)}) | (x_i, y_i) \in T^{(t)}, z_i^{(t)} \in [0, 1]\}$;
  - $z_i^{(t)}$ represents the uncertainty of a sample $x_i$ in $T^{(t)}$;
  - $w_i^{(t)}$ is a weight for $x_i$;
  - $\ell_2$ is a differentiable loss function;
  - $\Omega_2(g^{(t)})$ is the penalty for the complexity of $g^{(t)}$.

## Strategy: Diversity Sampling

- Given a number of partitioned groups $\{\mathbf{v}_1, \ldots, \mathbf{v}_b\}$ from the sample space $\mathbf{v}$, the diversity $\Gamma(\mathbf{v})$ is defined using a $l_{2,1}$-norm function:

$$\Gamma(\mathbf{v}) = ||\mathbf{v}||_{2,1} = \sum_{j=1}^{b} ||\mathbf{v}_j||_2$$

where:

  - $\sum_{j=1}^{b} |\mathbf{v}_j| = |\mathbf{v}|$;
  - $\mathbf{v}_j \in \{0, 1\}^m$;
  - $m = |X_j^{(t)}|$.

# Experimental Setup: Datasets

| Datasets | # of Attributes | # of Instances | # of Classes | Class Imbalance Ratio |
|---|---|---|---|---|
| *Image classification* | | | | |
| Mnist | $28 \times 28$ | 60,000 | 10 | N/A |
| *Salary level prediction* | | | | |
| Adult | 14 | 48,842 | 2 | 1 : 3 |
| *Entity resolution* | | | | |
| Cora | 12 | 837,865 | 2 | 1 : 49 |
| DBLP-Scholar | 4 | 168,112,008 | 2 | 1 : 71,233 |
| DBLP-ACM | 4 | 6,001,104 | 2 | 1 : 2,698 |
| NCVoter | 18 | 10M | 2 | 1:420 |

# Experimental Setup: Baselines

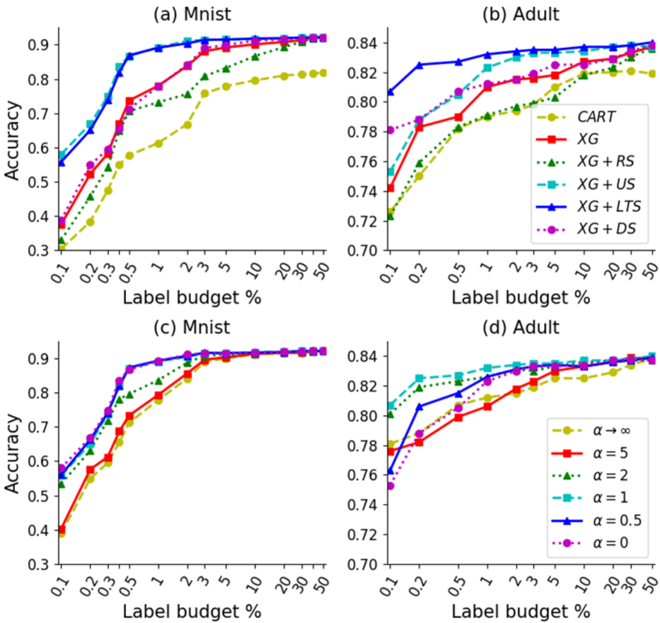
Australian National University

- CART: Classification And Regression Tree
- XG: eXtreme Gradient Boosting
  - + RS: Random sampling
  - + US: Uncertainty sampling
  - + DS: Diversity sampling
  - + LTS: Learning to sample with equal sampling distribution
  - + LTS(E): Learning to sample with exponentially decreasing sample budget

# Results: Different Label Budgets

| Dataset | Label Budget $\zeta$ (% of $\lvert X \rvert$) | CART | XG | XG+RS | XG + US $\alpha = 0$ | XG+LTS $\alpha = 1$ | XG + DS $\alpha \to \infty$ | XG + LTS(E) $\alpha = 1$ |
|---|---|---|---|---|---|---|---|---|
| Cora | 0.01 | 0 | 0 | 0 | 0 | 0.857 | **0.878** | 0.862 |
| | 0.05 | 0.741 | 0.763 | 0.750 | 0.827 | 0.864 | **0.885** | 0.867 |
| | 0.1 | 0.788 | 0.796 | 0.787 | 0.823 | 0.862 | 0.886 | 0.870 |
| | 0.5 | 0.848 | 0.835 | 0.835 | 0.873 | **0.900** | 0.893 | 0.890 |
| | 1 | 0.868 | 0.878 | 0.880 | 0.870 | **0.902** | 0.894 | 0.896 |
| | 5 | 0.878 | 0.897 | 0.892 | 0.907 | **0.915** | 0.898 | 0.904 |
| NCVoter | 0.01 | 0 | 0 | 0 | 0 | 0.324 | **0.875** | 0.571 |
| | 0.05 | 0 | 0 | 0 | 0 | 0.954 | 0.991 | 0.934 |
| | 0.1 | 0 | 0 | 0 | 0 | **0.994** | 0.993 | 0.993 |
| | 0.5 | 0 | 0 | 0 | 0 | **0.994** | 0.991 | **0.994** |
| | 1 | 0.334 | 0.379 | 0.398 | 0 | 0.993 | **0.994** | 0.993 |
| | 5 | 0.993 | 0.993 | 0.994 | 0.993 | **0.997** | 0.993 | 0.994 |
| DBLP-ACM | 0.1 | 0 | 0 | 0 | 0 | 0 | **0.397** | 0 |
| | 0.5 | 0 | 0 | 0 | 0 | 0.702 | 0.632 | 0.679 |
| | 1 | 0.348 | 0.347 | 0.279 | 0 | **0.878** | 0.721 | 0.793 |
| | 2 | 0.599 | 0.767 | 0.680 | 0.403 | **0.884** | 0.783 | 0.854 |
| | 5 | 0.870 | 0.850 | 0.803 | 0.874 | **0.931** | 0.833 | 0.891 |
| | 10 | 0.903 | 0.911 | 0.890 | 0.926 | **0.981** | 0.899 | 0.933 |
| DBLP-Scholar | 0.1 | 0 | 0 | 0 | 0 | 0.723 | **0.731** | 0.727 |
| | 0.5 | 0.378 | 0.54 | 0.498 | 0.555 | 0.773 | 0.780 | **0.781** |
| | 1 | 0.562 | 0.669 | 0.659 | 0.738 | 0.804 | 0.792 | **0.794** |
| | 2 | 0.772 | 0.806 | 0.771 | 0.807 | **0.815** | 0.801 | 0.811 |
| | 5 | 0.773 | 0.822 | 0.803 | **0.836** | **0.836** | 0.818 | 0.828 |
| | 10 | 0.808 | 0.835 | 0.830 | **0.865** | 0.851 | 0.829 | 0.853 |

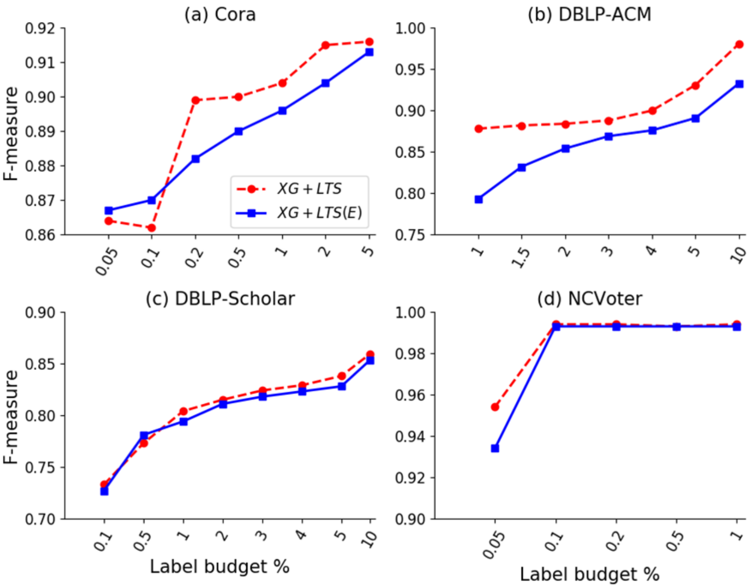
# Results: Different Label Budgets



# Results: Different $\alpha$

| Dataset | Label Budget $\zeta$ (% of $\|X\|$) | XG + US $\alpha = 0$ | XG+LTS $\alpha = 0.5$ | XG+LTS $\alpha = 1$ | XG+LTS $\alpha = 2$ | XG+LTS $\alpha = 5$ | XG + DS $\alpha \to \infty$ |
|---|---|---|---|---|---|---|---|
| Cora | 0.01 | 0 | 0.637 | 0.857 | 0.861 | 0.867 | **0.878** |
| | 0.05 | 0.827 | 0.851 | 0.864 | 0.870 | 0.883 | **0.885** |
| | 0.1 | 0.823 | 0.863 | 0.862 | 0.873 | **0.887** | 0.886 |
| | 0.5 | 0.873 | 0.893 | **0.900** | 0.895 | 0.895 | 0.893 |
| | 1 | 0.870 | 0.896 | 0.902 | **0.904** | 0.898 | 0.894 |
| | 5 | 0.907 | 0.912 | **0.915** | 0.913 | 0.902 | 0.898 |
| NCVoter | 0.01 | 0 | 0.403 | 0.324 | 0.403 | 0.752 | **0.875** |
| | 0.05 | 0 | 0.903 | 0.954 | 0.989 | **0.993** | 0.991 |
| | 0.1 | 0 | 0.989 | **0.994** | 0.993 | 0.993 | 0.993 |
| | 0.5 | 0 | 0.993 | **0.994** | 0.993 | 0.993 | 0.991 |
| | 1 | 0 | 0.993 | 0.993 | 0.993 | 0.992 | **0.994** |
| | 5 | 0.993 | 0.993 | **0.997** | 0.993 | 0.994 | 0.993 |
| DBLP-ACM | 0.1 | 0 | 0 | 0 | 0 | 0 | **0.397** |
| | 0.5 | 0 | 0.382 | 0.702 | **0.720** | 0.651 | 0.632 |
| | 1 | 0 | 0.813 | **0.878** | 0.778 | 0.730 | 0.721 |
| | 2 | 0.403 | 0.851 | **0.884** | 0.867 | 0.789 | 0.783 |
| | 5 | 0.874 | **0.935** | 0.931 | 0.889 | 0.837 | 0.833 |
| | 10 | 0.926 | **0.983** | 0.981 | 0.937 | 0.893 | 0.899 |
| DBLP-Scholar | 0.1 | 0 | 0.586 | 0.723 | 0.733 | **0.741** | 0.731 |
| | 0.5 | 0.555 | 0.764 | 0.773 | **0.794** | 0.790 | 0.780 |
| | 1 | 0.738 | 0.793 | 0.804 | **0.808** | 0.793 | 0.792 |
| | 2 | 0.807 | 0.810 | **0.815** | 0.813 | 0.799 | 0.801 |
| | 5 | 0.836 | **0.838** | 0.836 | 0.831 | 0.821 | 0.818 |
| | 10 | **0.865** | 0.859 | 0.851 | 0.844 | 0.837 | 0.829 |

Australian National University



# Results: Label Cost

- Comparison of label budgets w.r.t. classification results with a desired FM value, where XG+LTS has $\alpha = 1$.

| Dataset | Cora | DBLP-ACM | DBLP-Scholar | NCVoter |
|---|---|---|---|---|
| CART | 5% | 10% | 10% | 3% |
| XG | 4% | 8% | 2% | 2% |
| XG + RS | 5% | 12% | 5% | 2% |
| XG + US | 2% | 7% | 2% | 7% |
| XG + DS | 3% | 10% | 2% | **0.03%** |
| XG + LTS | **0.5%** | **4%** | **0.9%** | **0.03%** |
| FM values | 0.9 | 0.9 | 0.8 | 0.9 |

Australian National University

- We propose a novel active learning framework, namely Learning To Sample (LTS).
- Our sampling model incorporates uncertainty and diversity of samples into a unified process for optimization.
- The experimental results show that our active learning approach significantly outperforms all the baselines when the label budget is limited.

# Thank You!

# Q & A

Email: Jingyu.shao@anu.edu.au