# Scalable Active Learning by Approximated Error Reduction

**Weijie Fu**, Meng Wang, Shijie Hao, Xindong Wu
Hefei University of Technology, China
fwj.edu@gmail.com

# Outline



01 Problem Background

02 Scalable Active Learning

03 Experiment and Conclusion

# Outline



01 **Problem Background**

- The aim and limitation of existing active learning.

02 Scalable Active Learning

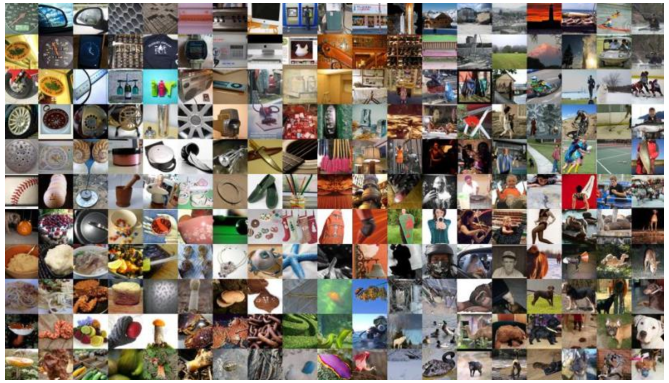
03 Experiment and Conclusion

# Problem background



## Classification

- Identifying the categories of unlabeled instances
  - computer vision, handwriting recognition, speech recognition, document classification



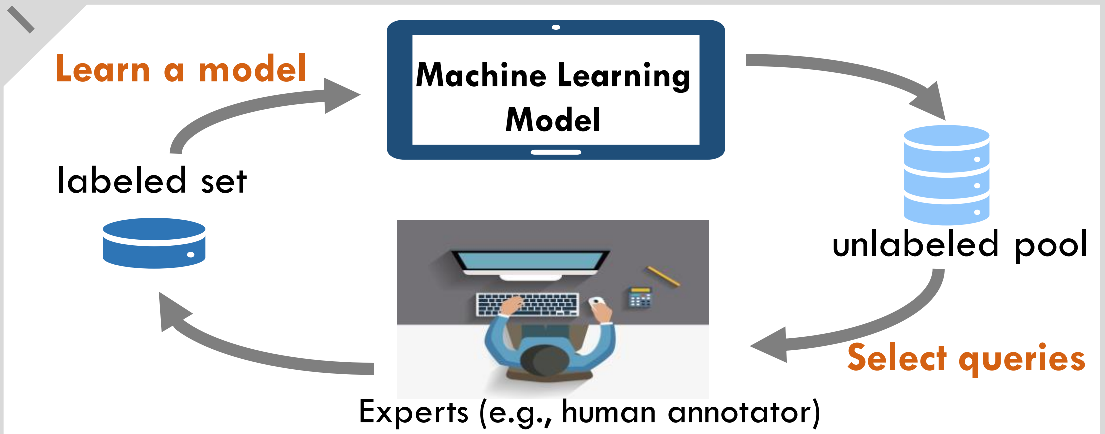


- Difficulties of this problem
  - Quality of labeled instances
  - Expensive costs of collecting labels

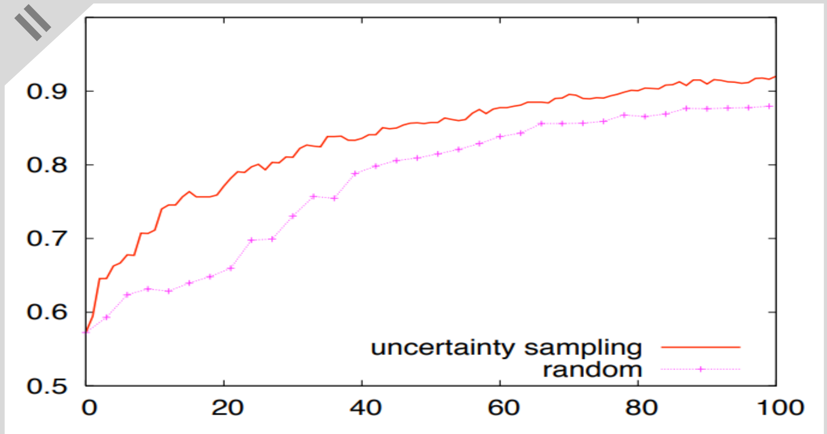
} **Active Learning**

# Problem background

## Active Learning




Procedures of Active Learning



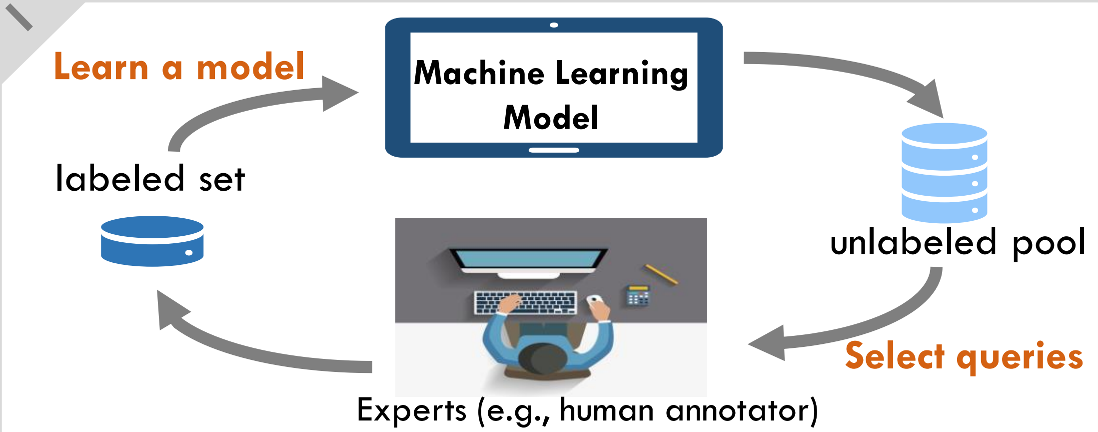

Performance of Active Learning

- Common procedures in the cycle.
  - Label prediction based on current **semi-supervised classifier**.
  - Measure estimation based on **query selection criterion**.
  - Query labeling by experts.

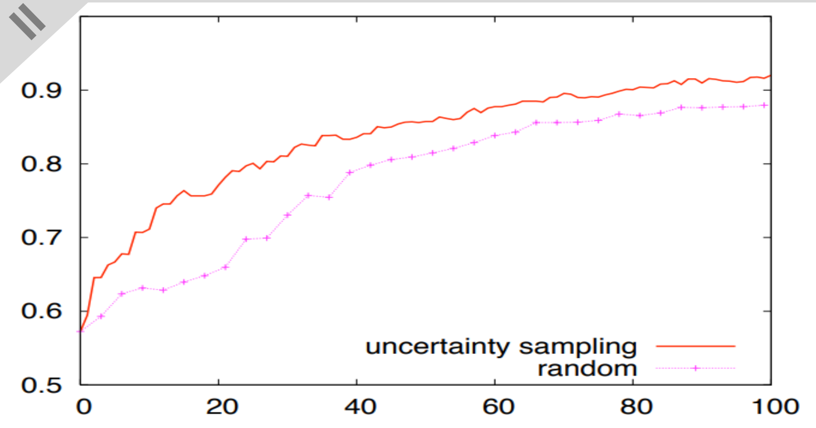
# Problem background



## Active Learning




Procedures of Active Learning




Performance of Active Learning

Common procedures in the cycle.

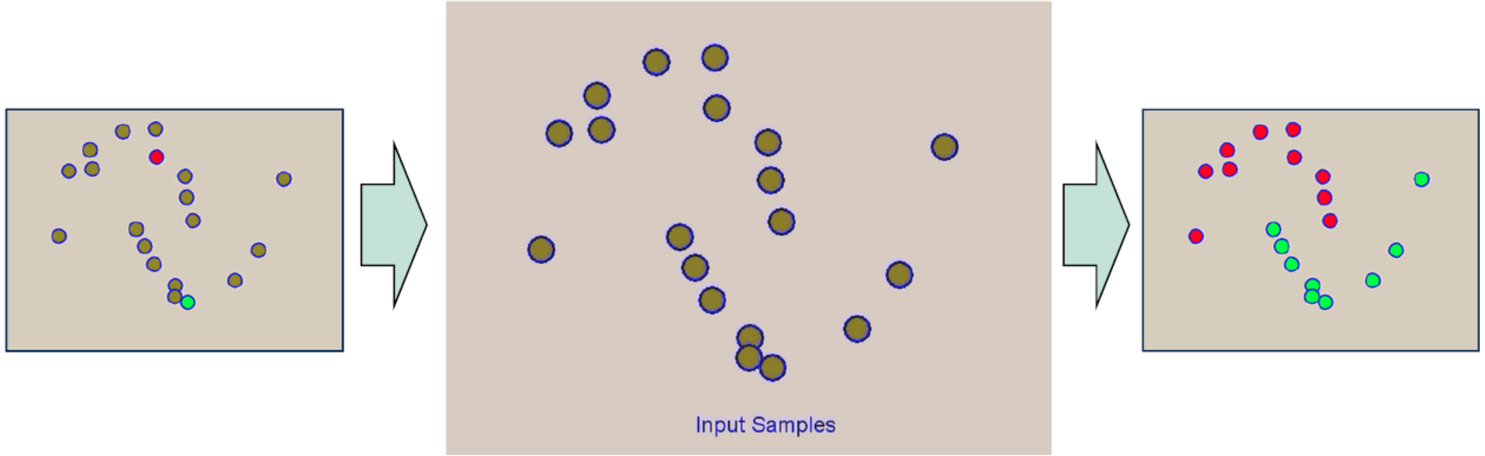
- Label prediction based on current **semi-supervised classifier**.
- Measure estimation based on **query selection criterion**.
- Query labeling by experts.

**Keypoints**

# Problem background



- **Semi-Supervised Classifier** (learn from labeled and unlabeled data)
  - Graph-based Classifier
    - Illustration:




    - Graph Construction + Label Propagation
    - Advantages:
      - Easy for explanation; Analytic solution …

# Problem background



## Semi-Supervised Classifier

### Graph-based Classifier

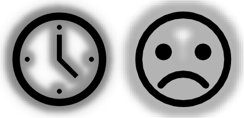
- **Formulation:**

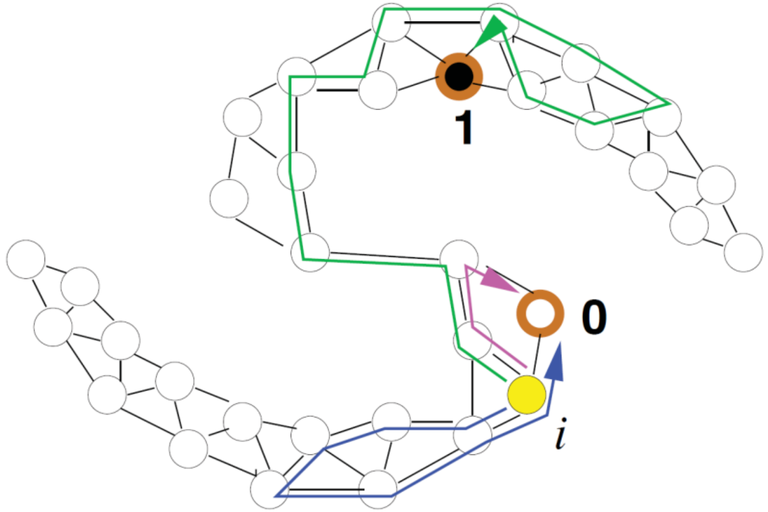
$$\textit{minimize}\ \sum_i^n \|\boldsymbol{f}_i - \boldsymbol{y}_i\|^2 + \frac{\lambda}{2}\sum_{i,j=1}^n W_{ij}\left\|\frac{1}{\sqrt{D_{ii}}}\boldsymbol{f}_i - \frac{1}{\sqrt{D_{jj}}}\boldsymbol{f}_j\right\|^2 \Rightarrow \textit{minimize}\ \|\mathbf{F}-\mathbf{Y}\|_F^2 + \lambda \mathrm{tr}[\mathbf{F}^T(\mathbf{I}-\mathbf{W})\mathbf{F}]$$

- Optimal Solution:

$$\underline{\mathbf{F} = [\mathbf{I} + \lambda(\mathbf{I}-\mathbf{W})]^{-1}\mathbf{Y}}$$

- **Procedure:**
  - I Label Propagation
    - matrix inversion with a cubic cost $O(N^3)$.
  - II Graph Construction
    - adjacency matrix with a quadratic cost $O(N^2)$.





# Problem background



## Query Selection Criterion

### A. Expected Error Reduction (EER)

- **Definition:** choose the instance with the largest error reduction ⇒ tradeoff on error reduction
- **Formulation:**

$$\operatorname{argmax}_k \ \underbrace{\hat{\mathcal{E}}(f)}_{\text{current estimated error}} - \underbrace{\hat{\mathcal{E}}(f^{+y_{k,i}})}_{\text{expected generalization error}}, \quad \text{where } k \in 1{:}N, i \in 1{:}C.$$

# Problem background



## Query Selection Criterion

### A. Expected Error Reduction (EER)

- **Definition**: choose the instance with the largest error reduction ⇒ tradeoff on error reduction.
- **Formulation**:

$$\underbrace{\operatorname{argmax}_k \ \hat{\mathcal{E}}(f)}_{\text{current estimated error}} - \underbrace{\hat{\mathcal{E}}(f^{+y_{k,i}})}_{\text{expected generalization error}}, \quad \text{where } k \in 1{:}N, i \in 1{:}C.$$

- **Procedure**:

At each iteration,
  For each unlabeled instance
    Suppose this instance is labeled, and re-train the classifier.
    Re-infer the soft labels exactly for hard labels.
    Estimate the expected error.
  End
Select the instance whose expected error reduction is largest.

# Problem background

## Query Selection Criterion

### A. Expected Error Reduction (EER)

- **Definition**: choose the instance with the largest error reduction ⇒ tradeoff on error reduction.
- **Formulation**:

$$\underbrace{\text{argmax}_k \ \hat{\mathcal{E}}(f)}_{\text{current estimated error}} - \underbrace{\hat{\mathcal{E}}(f^{+y_{k,i}})}_{\text{expected generalization error}}, \quad \text{where } k \in 1{:}N, i \in 1{:}C.$$

- **Procedure**:

```
At each iteration,
  For each unlabeled instance
      Suppose this instance is labeled, and re-train the classifier.
      Re-infer the soft labels exactly for hard labels.
      Estimate the expected error.
  End
Select the instance whose expected error reduction is largest.
```

- **Cost**:

$O(N + t_{train}) \times N$, where $t_{train}$ is the time cost of training the classifier, e.g., $O(N^3)$ for many SSL classifier.

# Problem background



## Query Selection Criterion

### A. Expected Error Reduction (EER)

### B. Uncertainty Sampling

- **D**efinition: choose the instance with the largest uncertainty.
- **P**rocedure:

At each iteration,
~~For each unlabeled instance~~ avoid model retraining.
Infer the labels of unlabeled instances.
Estimate the uncertainty.
~~End~~
Select the instance with the largest uncertainty.

- Analysis:
  - Cost: reduce the time cost to $O(N)$ without re-training.
  - Effectiveness: ignore the influence of labels; outliers may be selected.

# Problem background



## Limitations

- Semi-supervised Classifier
  - A. Graph-based Classifier
    - Large time cost of **graph construction** and **mode training.**
- Query Selection Criterion
  - A. Expected Error Reduction

    *(Perform well at either tuning decision boundaries or discovering new classes)*
    - Large time cost of **model re-training** and **label re-inference**.
  - B. Uncertainty Sampling
    - Ignore the **influence** of labels on the **classifier** and **other instances**.

**Scalable Active Learning**

# Outline



01 Problem Background

02 **Scalable Active Learning**

- An alternative to select high-quality queries efficiently.

03 Experiment and Conclusion

# Motivations

## Efficient Semi-Supervised Classifier

- Reduce the time cost of graph-based learning.
- Keep a high classification accuracy.

## Scalable Query Selection Criterion

- Cut down the time cost of query selection.
- Keep the high quality of selected instances.

# Motivations

## Efficient Semi-Supervised Classifier

- Cut down the time cost of graph-based learning.
- Keep a high classification accuracy.

## Scalable Query Selection

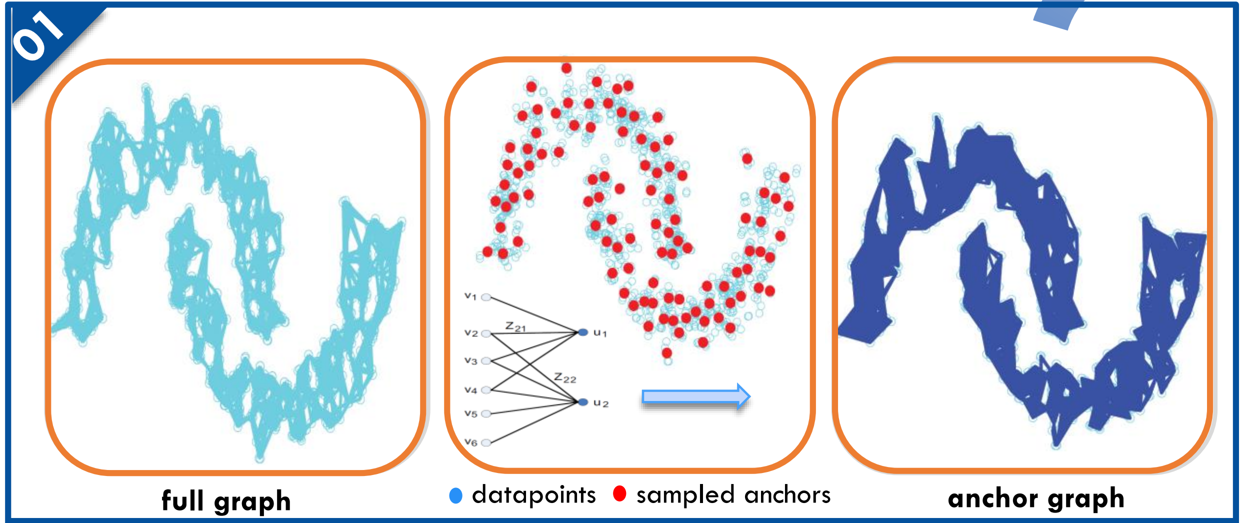
- Cut down the time cost of query selection.
- Keep the high quality of selected instances.

# Scalable Active Learning

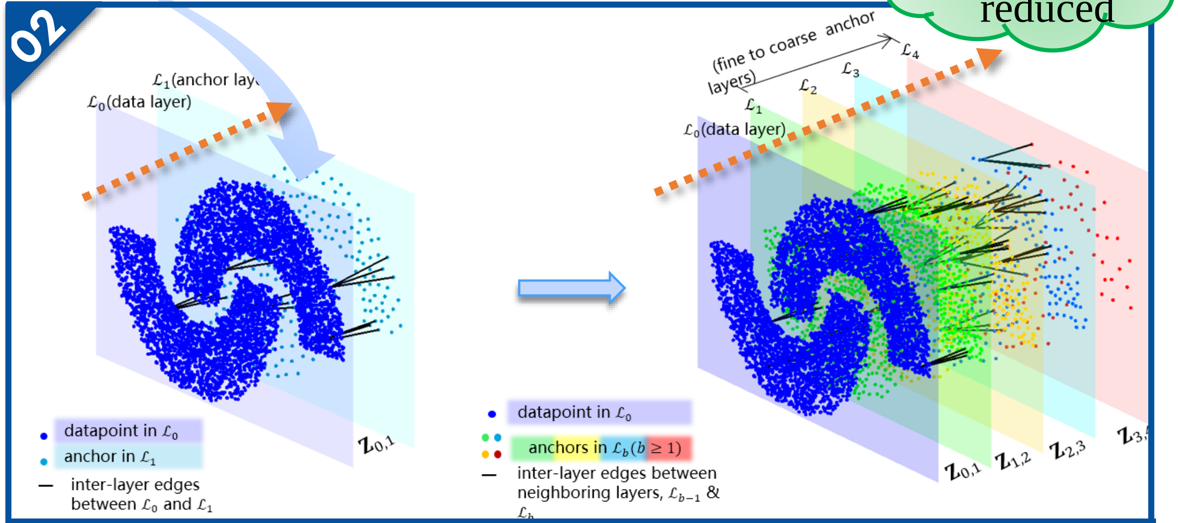
## Efficient Semi-Supervised Classifier

- Hierarchical Anchor Graph
- Learning with Hierarchical Anchor Graph Regularization (HAGR)

Sizes are generally reduced



from Full-Graph to Anchor Graph



from Anchor Graph to Hierarchical Anchor Graph

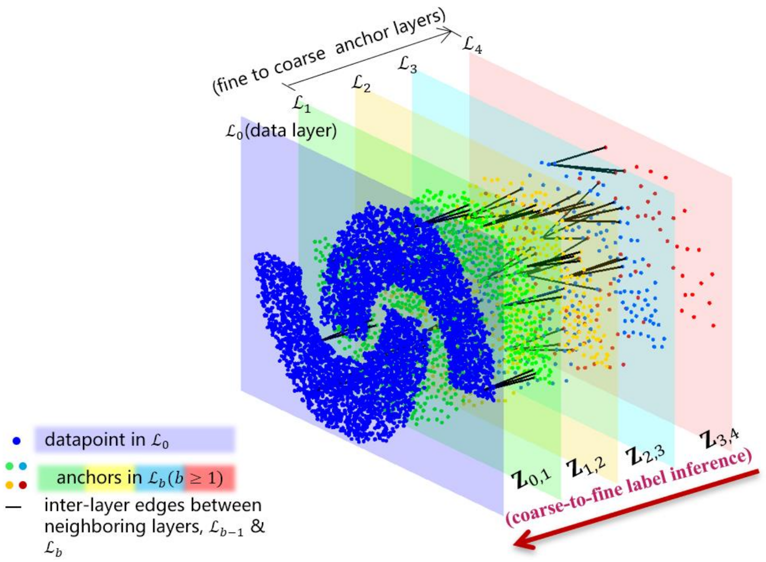
Figure from Wei Liu et al, ICML 2010.

# Scalable Active Learning



## Efficient Semi-Supervised Classifier

- Hierarchical Anchor Graph
- Learning with **H**ierarchical **A**nchor **G**raph **R**egularization **(HAGR)**
  - Formulation: $\sum_{i=1}^{l}\left\|\mathbf{Z}_{i.}^{\mathrm{H}}\mathbf{A}-\mathbf{y}_i\right\|^2+\frac{\lambda}{2}\sum_{i,j}^{n}W_{ij}(\mathbf{Z}_{i.}^{\mathrm{H}}\mathbf{A}-\mathbf{Z}_{j.}^{\mathrm{H}}\mathbf{A})^2.$
    - **Label smoothing** (Laplacian matrix) based on the **finest** anchors with $\mathbf{W}=\mathbf{Z}^{0,1^{\mathrm{T}}}\mathbf{Z}^{0,1}$.
    - **Label inference** (hierarchically) from the **coarsest** anchors with $\mathbf{Z}^{\mathrm{H}}=\mathbf{Z}^{0,1}\mathbf{Z}^{1,2}\dots\mathbf{Z}^{h-1,h}$.
  - Solution: $\mathbf{A}=\left(\mathbf{Z}_{\mathrm{L}}^{\mathrm{H}^{\mathrm{T}}}\mathbf{Z}_{\mathrm{L}}^{\mathrm{H}}+\lambda\tilde{\mathbf{L}}\right)^{-1}\mathbf{Z}_{\mathrm{L}}^{\mathrm{H}^{\mathrm{T}}}\mathbf{Y}_{\mathrm{L}}$
  - Time cost: reduced to $O(NN_h^2+N_h^3)$.



# Motivations

- Efficient Semi-Supervised Classifier
  - Cut down the computational cost of graph construction and model optimization.
  - Keeping a satisfying performance on the classification accuracy.

- Scalable Query Selection Criterion
  - Cut down the computational cost of query selection.
  - Keeping a satisfying performance on the quality of selected instances.

# Scalable Active Learning



## Scalable Query Selection

### Approximated Error Reduction (AER)

Definition:

- an approximated estimation of expected error reduction with **limited computations**.

Formulation:

average estimated error

$$\mathrm{argmax}_{\mathbf{x}_q}\ I_q \times \left(\frac{\mathcal{E}r_{\langle q\rangle}}{I_{\langle q\rangle}}\right)^{1-\varepsilon}$$

where $I_q$ is the **expected impact** over **all instances**, $\varepsilon$ is the hyper-parameter and $\frac{\mathcal{E}r_{\langle q\rangle}}{I_{\langle q\rangle}}$ is the **approximated ratio** between the **error reduction** and the **expected impact** over **nearby instances**.

Interpretation:

- $error\ reduction = expected\ impact \times \frac{expected\ error\ reduction}{expected\ Impact}$.
- $approximated\ error\ reduction = expected\ impact \times \left(\frac{expected\ error\ reduction}{expected\ Impact}\right)^{1-\varepsilon}_{nearby\ datapoints}$.
- Setting $\epsilon$ as **average estimated error** within (0, 1) ⇒ **Adaptive tradeoff between two terms with the error decreasing**.

# Scalable Active Learning



## Scalable Query Selection

### **A**pproximated **E**rror **R**eduction (**AER**)

Formulation:

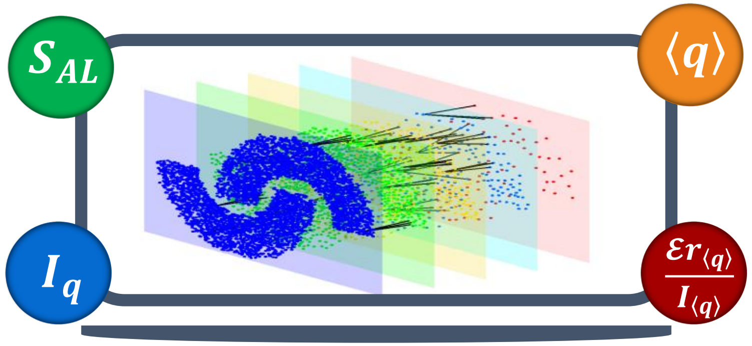
$$\text{argmax}_{\mathbf{x}_q}\ I_q \times \left(\frac{\mathcal{E}r_{\langle q\rangle}}{I_{\langle q\rangle}}\right)^{1-\varepsilon}, q \in S_{AL}$$

where $I_q$ is the expected impact over all instances, $\frac{\mathcal{E}r_{\langle q\rangle}}{I_{\langle q\rangle}}$ is the approximated ratio between the error reduction and the expected impact over nearby instances, and $\varepsilon$ is the tradeoff parameter.

**Keypoints**:

Candidate set

Expected impact over all instances




Nearby datapoints of the candidate $\mathbf{x}_q$

Approximated ratio over nearby datapoints

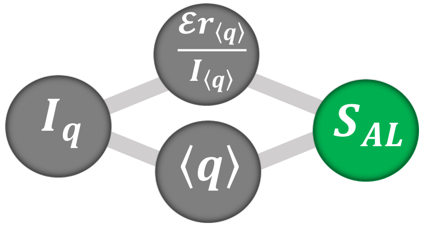
# Scalable Active Learning



## Scalable Query Selection

**A**pproximated **E**rror **R**eduction (**AER**)

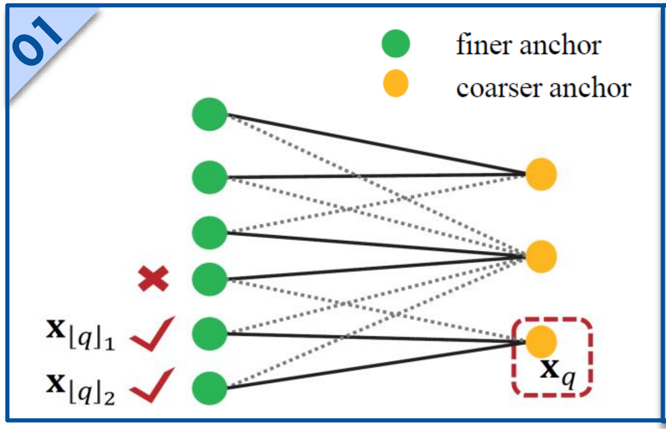
**Keypoints:**




Candidate set

**Details:**

**Hierarchical expansion of candidates**




### Hierarchical Expansion

1. Initialize candidates with all the coarsest anchors.
2. Once $\mathbf{x}_q$ is labeled, the connected finer anchors whose nearest coarser anchor is $\mathbf{x}_q$, is added into the candidate set.

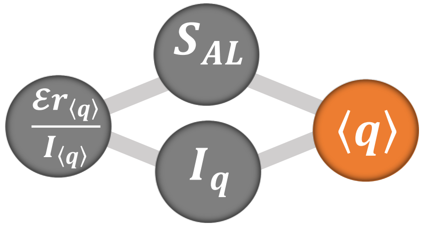
# Scalable Active Learning



## Scalable Query Selection

Approximated Error Reduction (**AER**)

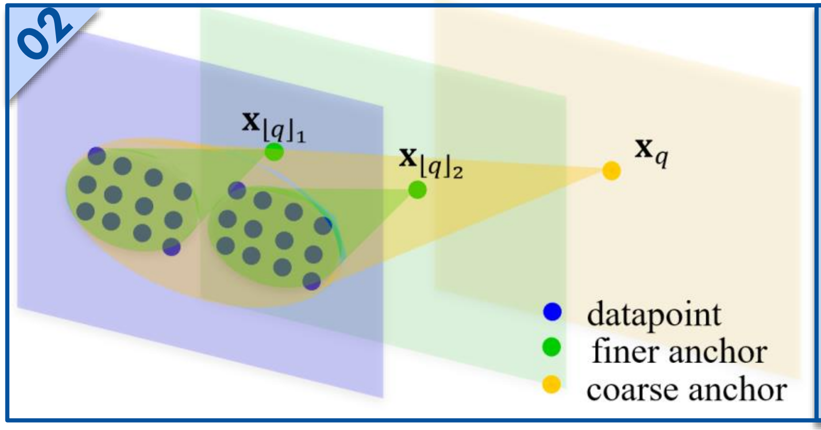
**Keypoints:**




Nearby datapoints of $\mathbf{x}_q$

**Details:**

**Hierarchical assignment of nearby datapoints**




### Hierarchical Assignment

1. Initialize nearby points by assigning datapoints to its nearest anchors.
2. Once $\mathbf{x}_q$ is labeled, its nearby datapoints $\langle q \rangle$ are re-assigned to the nearest finer anchors $x_{\lfloor q \rfloor_1}$ and $x_{\lfloor q \rfloor_2}$ .

the coarser candidates requires more nearby datapoints to estimate their approximated ratio effectively.

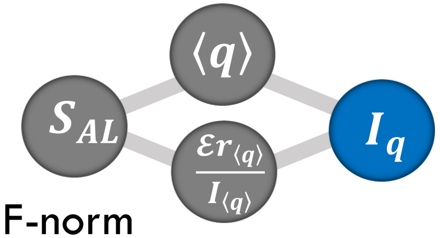
# Scalable Active Learning



## □ Scalable Query Selection

**A**pproximated **E**rror **R**eduction (**AER**)

**Keypoints:**




Expected impact over all instances

**Details:**

Only soft labels are required.

**Fast computation of the expected impact**

Linear classifier with F-norm

03

**HAGR**: $I_q = \sum_{r=1}^{C} f_{qr} \left\| \mathbf{Z}^{\mathrm{H}} \left( \mathbf{A}^{+\hat{y}_{\mathrm{qr}}} - \mathbf{A} \right) \right\|_{\mathrm{F}}^{2}$

$$\left\| \mathbf{Z}^{\mathrm{H}} \left( \mathbf{A}^{+\hat{y}_{\mathrm{qr}}} - \mathbf{A} \right) \right\|_{\mathrm{F}}^{2}$$

$$= \mathrm{trace}\left[ \left( \mathbf{A}^{+\hat{y}_{\mathrm{qr}}} - \mathbf{A} \right) \boldsymbol{\Delta} \left( \mathbf{A}^{+\hat{y}_{\mathrm{qr}}} - \mathbf{A} \right) \right]$$

$$\mathbf{A} = \left( {\mathbf{Z}_q^{\mathrm{H}}}^{\mathrm{T}} \mathbf{Z}_q^{\mathrm{H}} + \mathbf{M} \right)^{-1} \left( {\mathbf{Z}_q^{\mathrm{H}}}^{\mathrm{T}} \mathbf{Y}_q + {\mathbf{Z}_{\mathrm{L}}^{\mathrm{H}}}^{\mathrm{T}} \mathbf{Y}_{\mathrm{L}} \right)$$

$$\boxed{\boldsymbol{\Delta} = {\mathbf{Z}_{\mathrm{L}}^{\mathrm{H}}}^{\mathrm{T}} \mathbf{Z}_{\mathrm{L}}^{\mathrm{H}}} \qquad \mathbf{M} = {\mathbf{Z}_{\mathrm{L}}^{\mathrm{H}}}^{\mathrm{T}} \mathbf{Z}_{\mathrm{L}}^{\mathrm{H}} + \lambda \tilde{\mathbf{L}}$$

remaining time cost to data size is avoided

### Fast Computation

Let $\widetilde{\mathbf{M}} =$ … $, \alpha_q = {\mathbf{Z}_q^{\mathrm{H}}}^{\mathrm{T}} \mathbf{M} \mathbf{Z}_q^{\mathrm{H}}, \beta_q = \frac{1}{1+\alpha_q},$

then $\left( \mathbf{Z}_q^{\mathrm{H}}\right.$ … $\left.\right)^{-1}$ … ${\mathbf{Z}_q^{\mathrm{H}}}^{\mathrm{T}} \mathbf{Z}_q^{\mathrm{H}} \mathbf{M})$

matrix inversion lemma

**Time Cost:**

For $N_q$ candidates, the time cost of expected impact estimation is $O\left( N_h^2 N_q + N_h^3 + N_h N_q C + N_h^2 C + N_q C^2 \right) \approx O(N_h^2 N_q)$.

direct matrix operations !

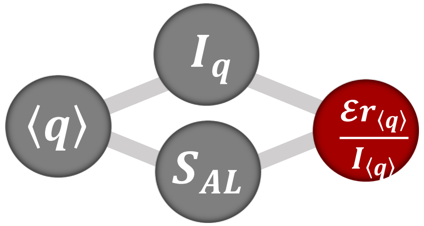
# Scalable Active Learning



## Scalable Query Selection

Approximated Error Reduction (**AER**)

**Keypoints**:




**Approximated ratio**

**Details:**

**Fast estimation of the approximated ratio**

04

expected impact over $\langle q\rangle$

$$I_{\langle q\rangle} \approx \frac{I_q}{1+\mu}$$

$\mu$:degree of the impact overflow to nearby datapoints.

expected error reduction over $\langle q\rangle$

$$\mathcal{E}r_{\langle q\rangle} = \sum_{i=1}^{N_{\langle q\rangle}} \eta_i \ell(f_i, \hat{f}_i) \approx \eta \sum \ell(f_i, \hat{f}_i)$$

$\eta$:degree of the expected error will be reduced.

### Fast Estimation

$$\frac{\mathcal{E}r_{\langle q\rangle}}{I_{\langle q\rangle}} = \eta \cdot \frac{\mathcal{E}_{\langle q\rangle}}{\frac{I_q}{1+\mu}} = \eta(1+\mu) \times \frac{\mathcal{E}_{\langle q\rangle}}{I_q}$$

**Time Cost:**

For $N_q$ candidates, the time cost of approximated ratio estimation is $O(NC)$.

# Scalable Active Learning



## Scalable Query Selection

### Approximated Error Reduction (AER)

Objective formulation: $\operatorname{argmax}_{\mathbf{x}_q} I_q \times \left(\frac{\mathcal{E}r_{\langle q \rangle}}{I_{\langle q \rangle}}\right)^{1-\varepsilon}, q \in S_{AL}$

Final formulation: $\boldsymbol{\operatorname{argmax}_{\mathbf{x}_q} I_q^{\varepsilon} \times \mathcal{E}_{\langle q \rangle}^{1-\varepsilon}, q \in S_{AL}}$

average estimated error

$S_{AL}$ $\langle q \rangle$ $I_q$ $\frac{\mathcal{E}r_{\langle q \rangle}}{I_{\langle q \rangle}}$

uncertainty over nearby instances

expected impact over all instances

Pros:

- AER enables an efficient estimation of error reduction without re-inferring labels of instances.
- The expected impact can be calculated for all candidates via direct matrix operations rather than multiple iterations.
- Apart from the **similar time cost to** that of the **uncertainty sampling**, the remaining time cost of our AER-based approach is independent of data sizes during the query selection.
- AER focuses on global impact first and pays attention to local uncertainty later, which provides an opportunity to achieve comparable or even higher accuracies than the EER-based approach.

# Outline

# Experiment and Conclusion



## Experiment

- Efficient Semi-Supervised Learning on Hierarchical Anchor Graph
- Scalable Active Learning with Approximated Error Reduction

Classification error rates (%) on **USPS-Train** (7,291 samples) with $l = 100$ labeled samples. $m = 1000$ for four versions of AGR. The running time of `k-means` clustering is 7.65 seconds.

| Method | Error Rate (%) | Running Time (seconds) |
|---|---|---|
| 1NN | 20.15±1.80 | 0.12 |
| LGC with 6NN graph | 8.79±2.27 | 403.02 |
| GFHF with 6NN graph | 5.19±0.43 | 413.28 |
| random `AnchorGraphReg`$^0$ | 11.15±0.77 | 2.55 |
| random `AnchorGraphReg` | 10.30±0.75 | 8.85 |
| `AnchorGraphReg`$^0$ | 7.40±0.59 | 10.20 |
| `AnchorGraphReg` | **6.56±0.55** | 16.57 |

# Experiment and Conclusion



## □ Experiment

- Efficient Semi-Supervised Learning on Hierarchical Anchor Graph
- Scalable Active Learning with Approximated Error Reduction

Classification accuracies (%) with different number of labeled samples on the MNIST8M dataset.

| # of labeled samples | 1NN | LSVM | AGR -30,000 | EAGR -30,000 | HAGR -30,000-5,000 | HAGR -300,000-30,000-5000 |
|---|---|---|---|---|---|---|
| 100 | $60.16 \pm 1.96$ | $59.67 \pm 2.19$ | $89.87 \pm 1.78$ | $90.27 \pm 0.18$ | $89.46 \pm 1.24$ | $\mathbf{91.36 \pm 0.70}$ |
| 200 | $68.66 \pm 1.29$ | $64.46 \pm 2.37$ | $91.15 \pm 0.59$ | $91.76 \pm 0.57$ | $90.85 \pm 0.50$ | $\mathbf{92.46 \pm 0.42}$ |
| 300 | $72.78 \pm 0.81$ | $66.79 \pm 2.25$ | $92.21 \pm 0.51$ | $92.37 \pm 0.51$ | $91.66 \pm 0.42$ | $\mathbf{93.05 \pm 0.37}$ |
| 400 | $75.33 \pm 0.60$ | $68.33 \pm 1.97$ | $92.47 \pm 0.44$ | $92.73 \pm 0.38$ | $92.16 \pm 0.36$ | $\mathbf{93.43 \pm 0.37}$ |
| 500 | $77.24 \pm 0.55$ | $70.65 \pm 1.49$ | $92.70 \pm 0.41$ | $93.05 \pm 0.29$ | $92.50 \pm 0.29$ | $\mathbf{93.78 \pm 0.24}$ |
| 600 | $78.58 \pm 0.54$ | $72.64 \pm 1.36$ | $92.80 \pm 0.34$ | $93.17 \pm 0.27$ | $92.64 \pm 0.26$ | $\mathbf{93.90 \pm 0.27}$ |
| 700 | $79.87 \pm 0.70$ | $73.80 \pm 1.27$ | $93.12 \pm 0.31$ | $93.41 \pm 0.30$ | $92.92 \pm 0.28$ | $\mathbf{94.10 \pm 0.25}$ |
| 800 | $81.02 \pm 0.50$ | $73.87 \pm 1.18$ | $93.19 \pm 0.23$ | $93.51 \pm 0.15$ | $93.06 \pm 0.16$ | $\mathbf{94.21 \pm 0.15}$ |
| 900 | $81.76 \pm 0.49$ | $73.97 \pm 0.96$ | $93.29 \pm 0.36$ | $93.63 \pm 0.21$ | $93.18 \pm 0.26$ | $\mathbf{94.28 \pm 0.16}$ |
| 1000 | $82.51 \pm 0.42$ | $76.95 \pm 1.13$ | $93.49 \pm 0.22$ | $93.79 \pm 0.15$ | $93.37 \pm 0.16$ | $\mathbf{94.39 \pm 0.12}$ |

The comparison of time costs (in seconds) of AGR, EAGR, and HAGR methods on the MNIST8M dataset.

| Dataset | AGR-30,000 | EAGR-30,000 | HAGR-30,000-5000 | HAGR-300,000-30,000-5,000 |
|---|---|---|---|---|
| MNIST8M | 665.07 | 662.60 | 104.97 | 137.54 |

# Experiment and Conclusion

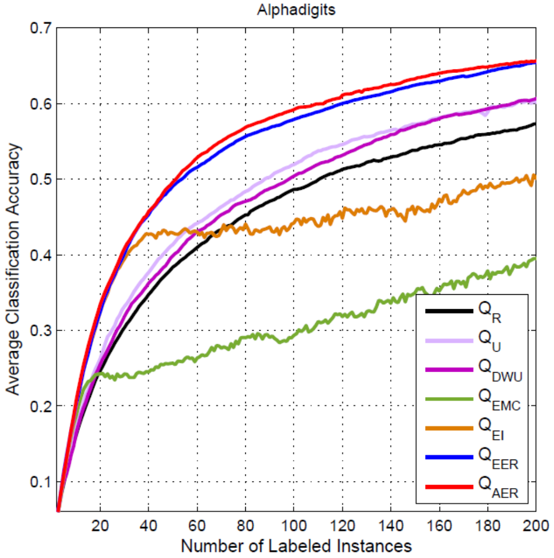


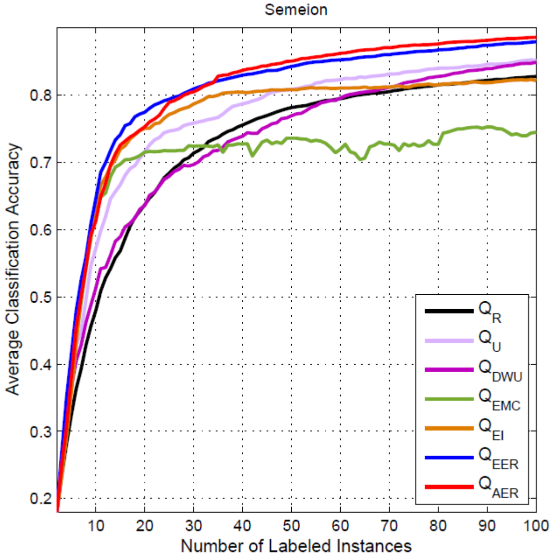
## ☐ Experiment

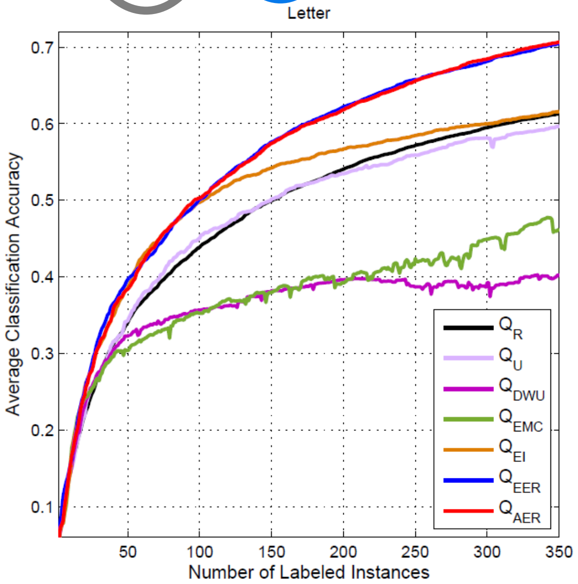
- Efficient Semi-Supervised Learning on Hierarchical Anchor Graph
- Scalable Active Learning with Approximated Error Reduction

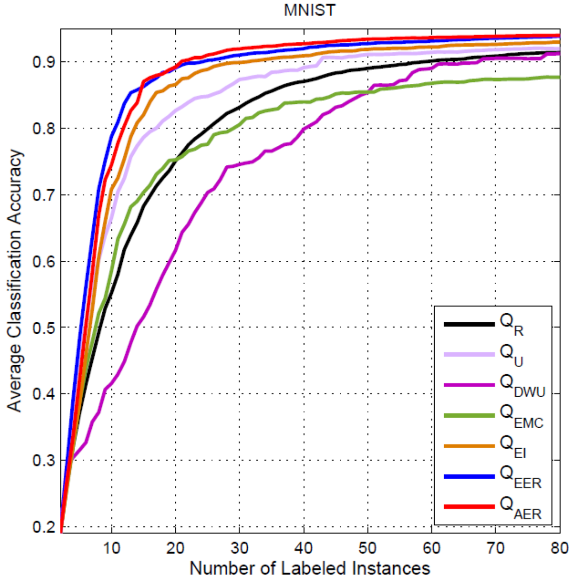
| Dataset | $Q_U$ | $Q_{DWU}$ | $Q_{EMC}$ | $Q_{EI}$ | $Q_{EER}$ | $Q_{AER}$ |
|---|---|---|---|---|---|---|
| Alphadigits | 0.01 | 0.03 | 0.04 | 0.04 | 1.80 | 0.04 |
| Semeion | 0.01 | 0.01 | 0.02 | 0.02 | 0.67 | 0.02 |
| Letter | 0.03 | 0.10 | 0.11 | 0.17 | 13.81 | 0.20 |
| MNIST | 0.04 | 0.18 | 0.15 | 0.17 | 30.16 | 0.21 |

Uncertainty
Model Change
Expected Error Reduction
Approximated Error Reduction



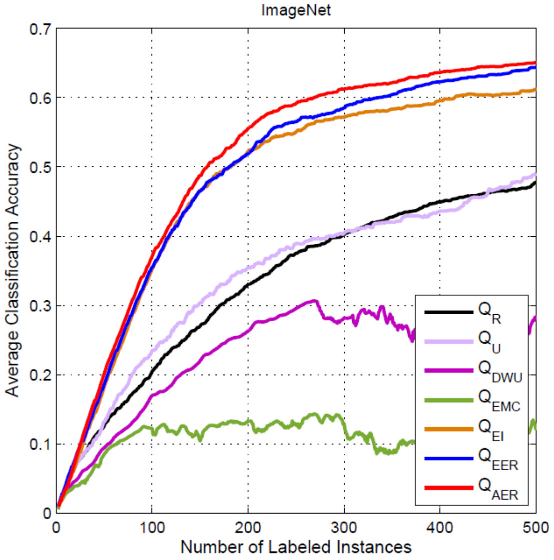
# Experiment and Conclusion

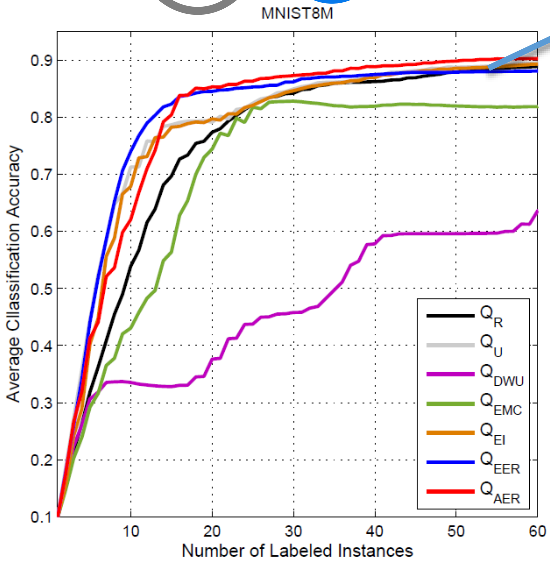


## Experiment

- Efficient Semi-Supervised Learning on Hierarchical Anchor Graph
- Scalable Active Learning with Approximated Error Reduction

| Dataset | $Q_U$ | $Q_{DWU}$ | $Q_{EMC}$ | $Q_{EI}$ | $Q_{EER}$ | $Q_{AER}$ |
|---|---|---|---|---|---|---|
| ImageNet | 2.15 | 10.05 | 4.19 | 7.61 | 90.73 | 9.35 |
| MNIST8M | 4.22 | 14.47 | 10.97 | 12.19 | 198.73 | 16.15 |




Since AER pays attention to local region based on local uncertainty, it can leads to better performance than EER.

Uncertainty
Model Change
Expected Error Reduction
Approximated Error Reduction

## □ Conclusion

Our Work

- Semi-supervised classifier on hierarchical anchor graph.
- Query selection criterion with approximated error reduction.
- Scalable active learning strategy for efficient classification.

# Reference




- Weijie Fu, Meng Wang, Shijie Hao, and Xindong Wu. 2018. Scalable Active Learning by Approximated Error Reduction. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Meng Wang, Weijie Fu, Shijie Hao, Hengchang Liu, and Xindong Wu. 2017. Learning on big graph: Label inference and regularization with anchor hierarchy. IEEE Transactions on Knowledge and Data Engineering 29, 5 (2017), 1101–1114.
- Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. 2003. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In Proceedings of the International Conference on Machine Learning, Vol. 3.
- Wei Liu, Junfeng He, and Shih Fu Chang. 2010. Large graph construction for scalable semi-supervised learning. In Proceedings of the International Conference on Machine Learning. 679–686.
- Oisin Mac Aodha, Neill D. F. Campbell, Jan Kautz, and Gabriel J. Brostow. 2014. Hierarchical subquery evaluation for active learning on a graph. In Proceedings of the Conference on Computer Vision and Pattern Recognition. 564–571.
- Wenbin Cai, Ya Zhang, and Jun Zhou. 2013. Maximizing expected model change for active learning in regression. In Proceedings of the IEEE International Conference on Data Mining. 51–60.
- Gautam Dasarathy, Robert Nowak, and Xiaojin Zhu. 2015. S2: An efficient graph based active learning algorithm with application to nonparametric classification. In Proceedings of the Annual Conference on Learning Theory. 503–522.
- Ajay J Joshi, Fatih Porikli, and Nikolaos P Papanikolopoulos. 2012. Scalable active learning for multiclass image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 11 (2012), 2259–2273.
- Xin Li and Yuhong Guo. 2013. Adaptive active learning for image classification. In Proceedings of the Conference on Computer Vision and Pattern Recognition. 859–866.
- Marius Muja and David G Lowe. 2014. Scalable nearest neighbor algorithms for high dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 11 (2014), 2227–2240.
- Meng Wang, Weijie Fu, Shijie Hao, Dacheng Tao, and Xindong Wu. 2016. Scalable semi-supervised learning by efficient anchor graph regularization. IEEE Transactions on Knowledge and Data Engineering 28, 7 (2016), 1864–1877.
- Burr Settles. 2010. Active learning literature survey. University of Wisconsin, Madison 52, 55-66 (2010), 11.

# Thank you for attention!

**Weijie Fu**, Meng Wang, Shijie Hao, Xindong Wu
Hefei University of Technology, China

https://github.com/fuweijie/AER
fwj.edu@gmail.com