Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning

Honglei Zhuang (庄弘磊)

LinkedIn Intern & UIUC PhD Student

Joel Young

Engineering Manager

Motivation

- Crowdsourcing is often adopted as a cheap way to collect labeled data for training a classifier
- To save cost, data items to label can be grouped together

Motivation

- Crowdsourcing is often adopted as a cheap way to collect labeled data for training a classifier
- To save cost, data items to label can be grouped together

Tech's Gender Gap Wasn't Always So Bad. Here's How It Got Worse.

The results is a new documentary film called CODE: Debugging the Gender Gap, which explores the glaring lack of American female and minority …

Original Post

- ❑ I am a retired female IT person …
- ❑ May I assume that she's going to do follow-up documentaries…
- ❑ Silly. Word. Games.
- ❑ …

A batch of comments to label

Motivation

- Crowdsourcing is often adopted as a cheap way to collect labeled data for training a classifier
- To save cost, data items to label can be grouped together

Tech's Gender Gap Wasn't Always So Bad. Here's How It Got Worse.

The results is a new documentary film called CODE: Debugging the Gender Gap, which explores the glaring lack of American female and minority …

Original Post

- ❑ I am a retired female IT person …
- ❑ May I assume that she's going to do follow-up documentaries…
- ❑ Silly. Word. Games.
- ❑ …

A batch of comments to label

- However, annotations on items in the same batch may interfere with each other, resulting in *in-batch annotation bias*.

Real-World Example of In-Batch Annotation Bias

Crowdsourcing annotation results on inappropriate content

“Yes” for inappropriate, “No” for acceptable

Batch 1		**Batch 2**	
Yes	***[URL]***	No	Even after doing all this, sometimes it still doesn’t work I might add. It is part of the job.
No	something related: ***[URL]***	No	Chernobyl nuclear power plant was probably engineered using “quality (sic !), speed and cost”…
No	***[URL]*** Its tough to be an engineer!	No	I think a culture where folks can air grievances can be very productive … *[Omitted]*
Yes	***[URL]***	No	Brian, my favorite saying as an engineer is: “I told you so 100 times.” … *[Omitted]*
No	Now from the perspective of the engineer: ***[URL]***	**Yes**	Now from the perspective of the engineer: ***[URL]***

Research Challenges

In a crowdsourcing task where

- Data items are presented to crowds as batches
- Labels are collected for training classifiers

Questions

- Is there annotation interference between items in the same batch?
- How to quantitatively measure the annotation bias?
- Can we leverage the bias for actively assembling data batches to improve classifier performance?

Talk Outline

- Verifying the presence of in-batch annotation bias
- Quantifying and measuring the bias
- Exploiting the bias for batch active learning

Verifying In-Batch Annotation Bias

Question:

Do other data items in the same batch *affect* the annotation of a particular data item?

Methodology:

- Construct two batches, with one comment fixed, but other comments varying
- Compare the annotations for the fixed data item

Our Data Set

LinkedIn comments

On influencer and company posts

Task

Label if a comment is inappropriate (e.g. promotional, profane, blatant soliciting, random greeting)

Ground-truth labels

Labeled by 9 trained LinkedIn employees

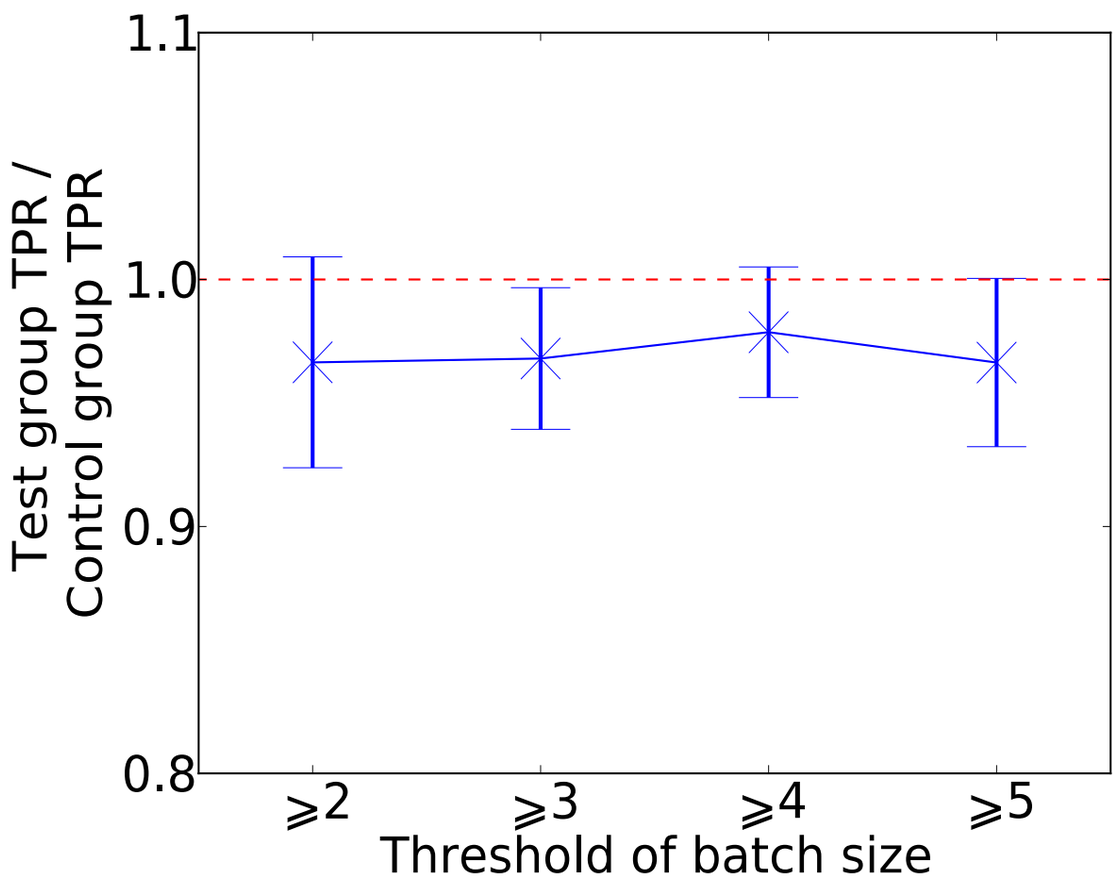
Mean Cohen's *Kappa = 0.79*

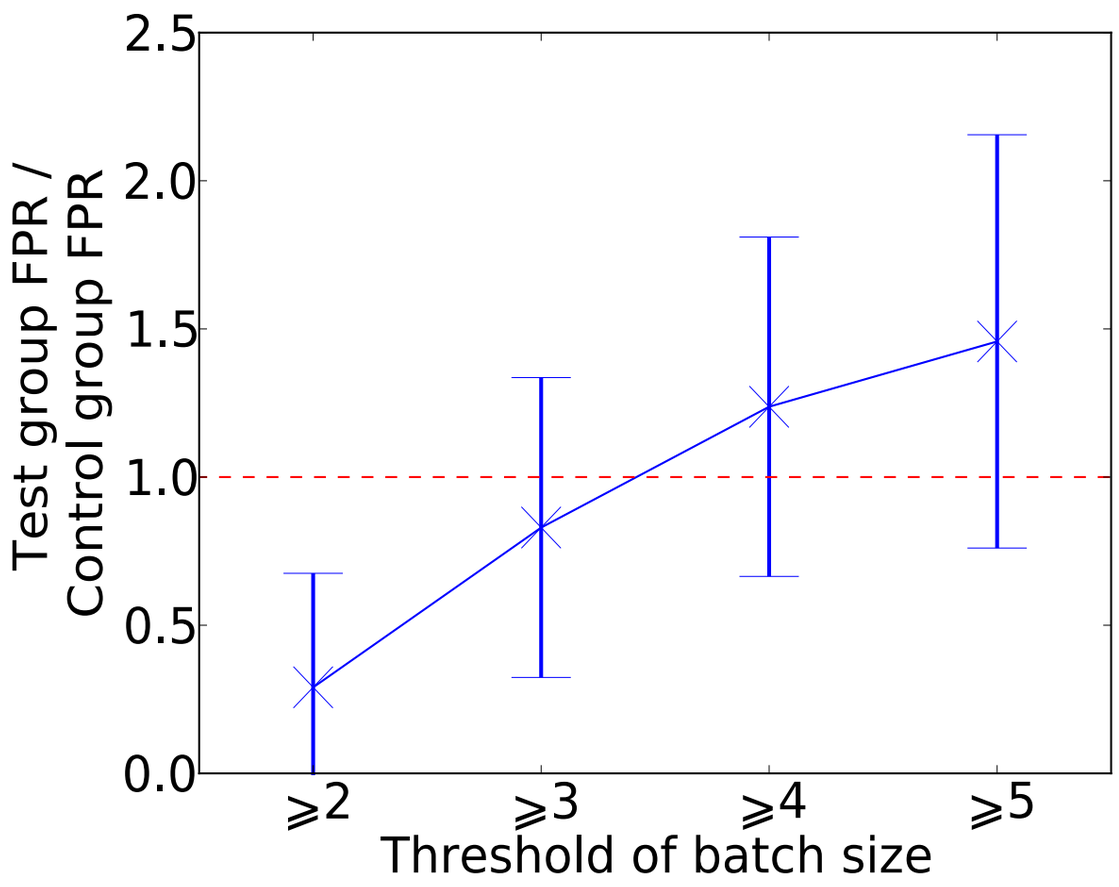
TPR/FPR vs. Batch Size *k*

Comparing crowds' True Positive Rate (TPR) or False Positive Rate (FPR) of the same comment, in batches with different sizes

Test group: batches with size ≥ threshold *t*

Control group: batches with size < threshold *t*



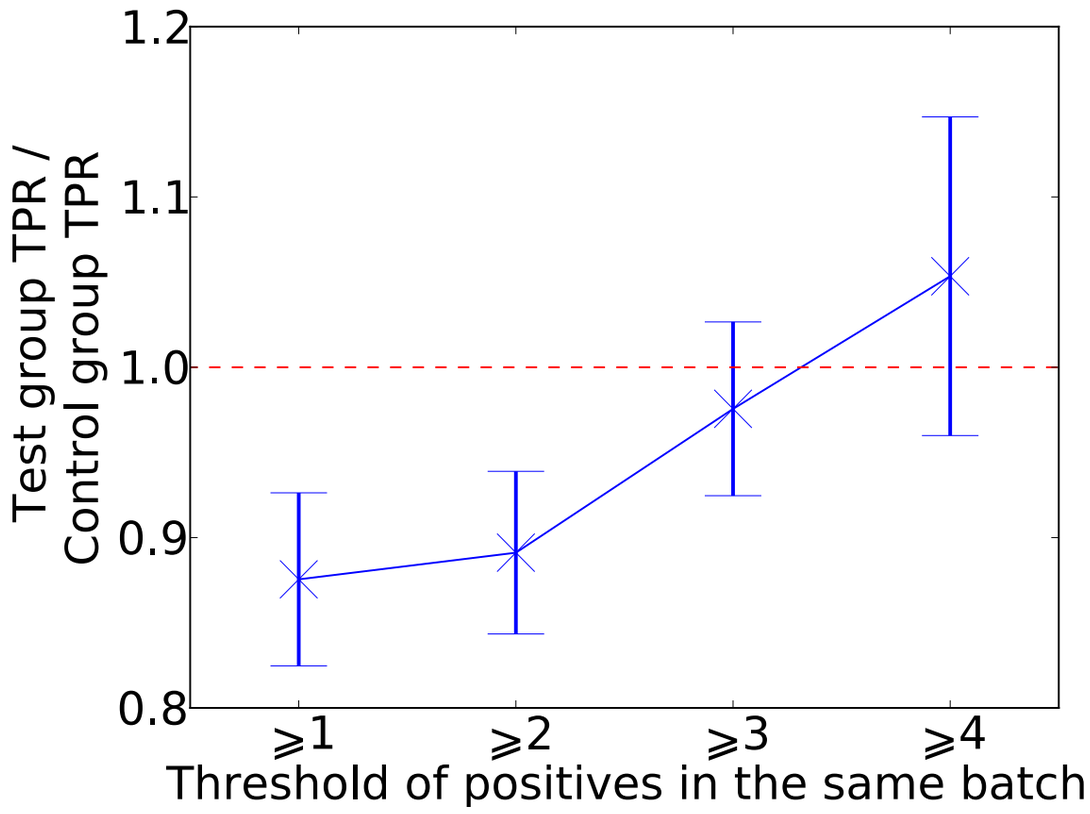


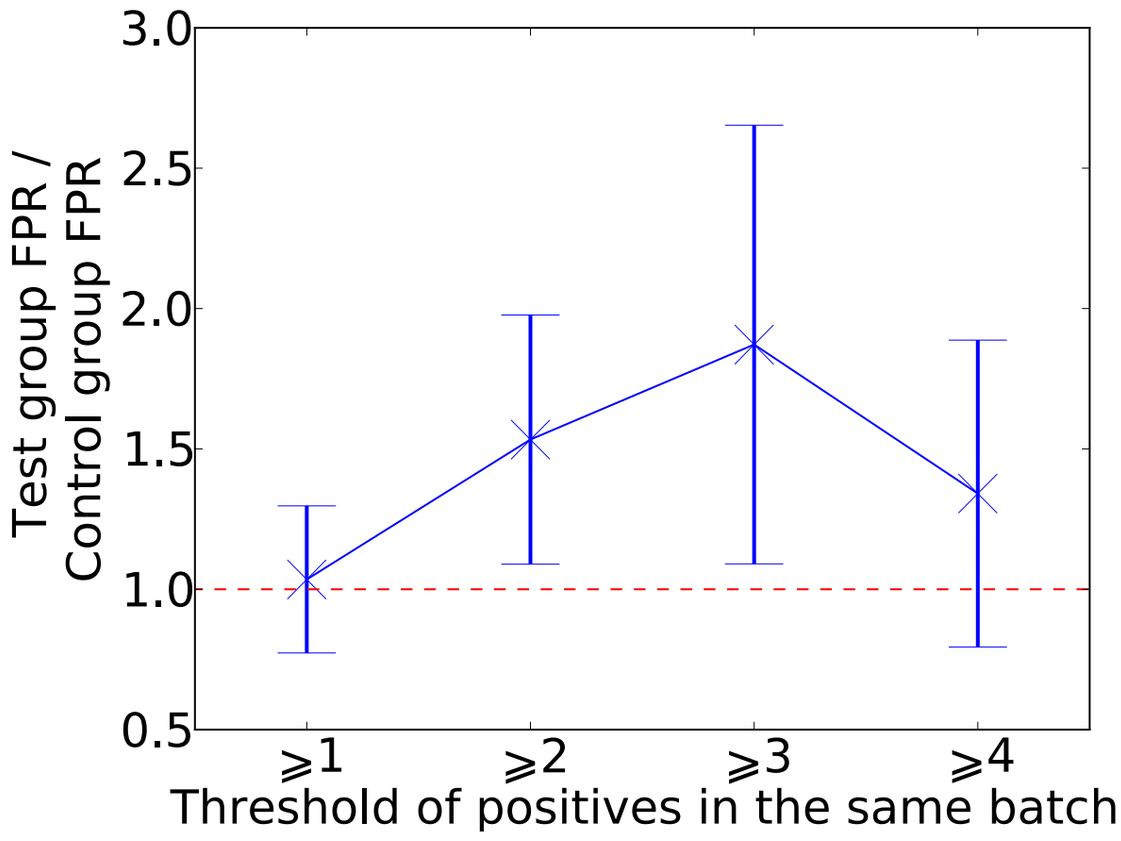
TPR/FPR vs. Number of Positives

Comparing crowds' True Positive Rate (TPR) or False Positive Rate (FPR) of the same comment, in batches with numbers of positive (inappropriate) comments in the same batch

Test group: batches with ≥ *t* other positives

Control group: batches with < *t* other positives

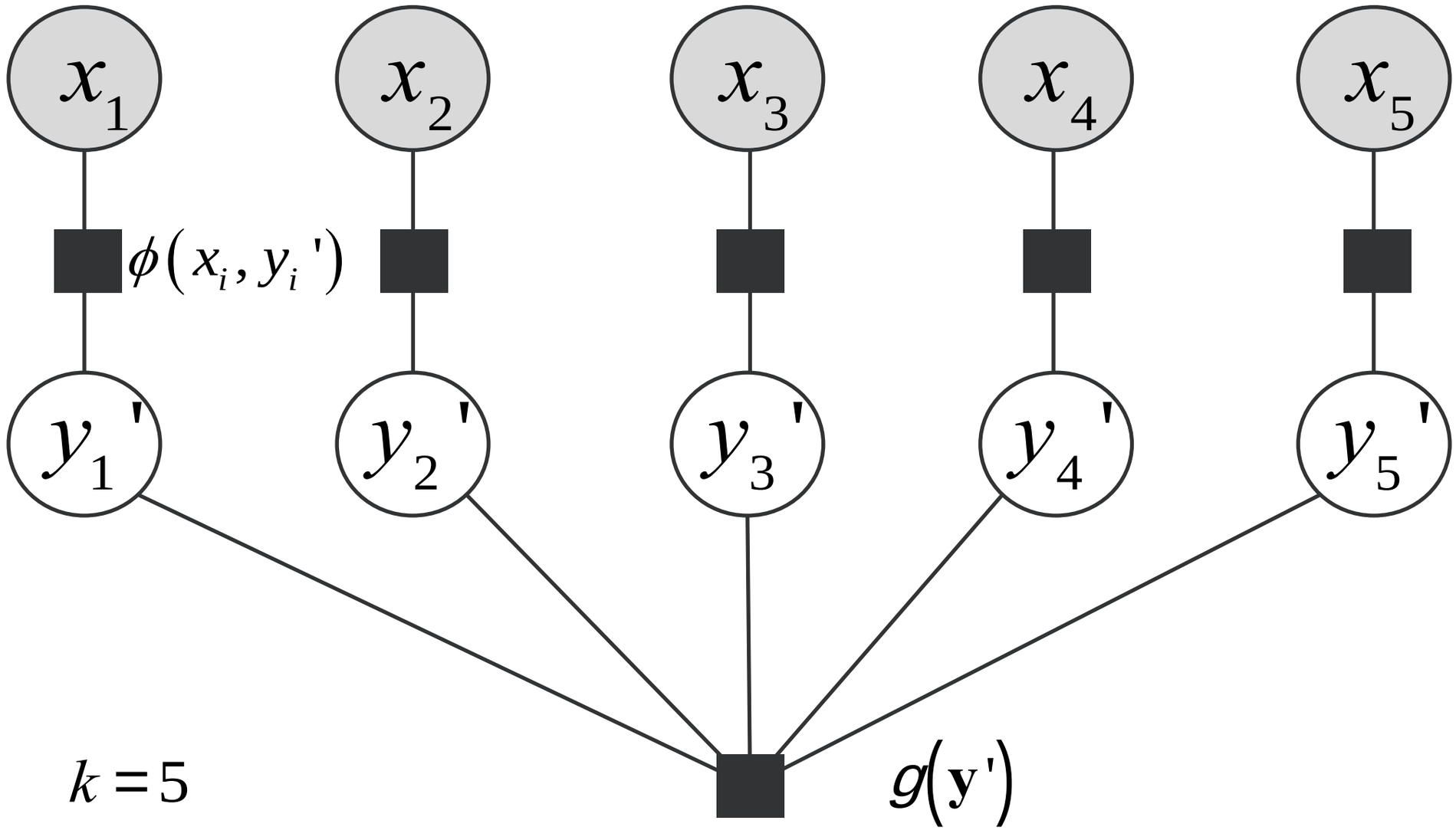




Quantifying and measuring the bias

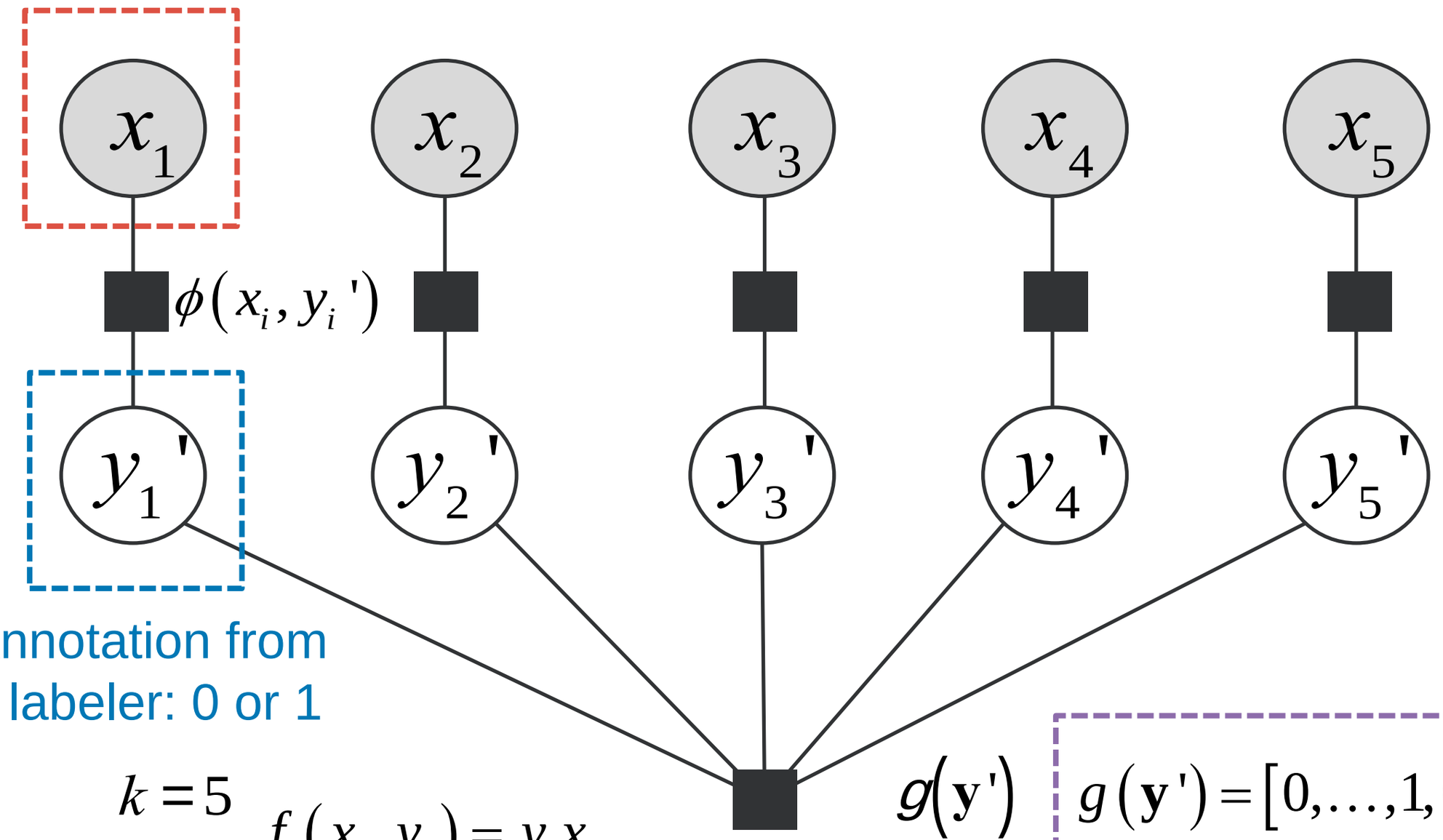
How do we *quantify* this bias?

Annotation Model



Annotation Model

Feature vector



Annotation from a labeler: 0 or 1

$k = 5$

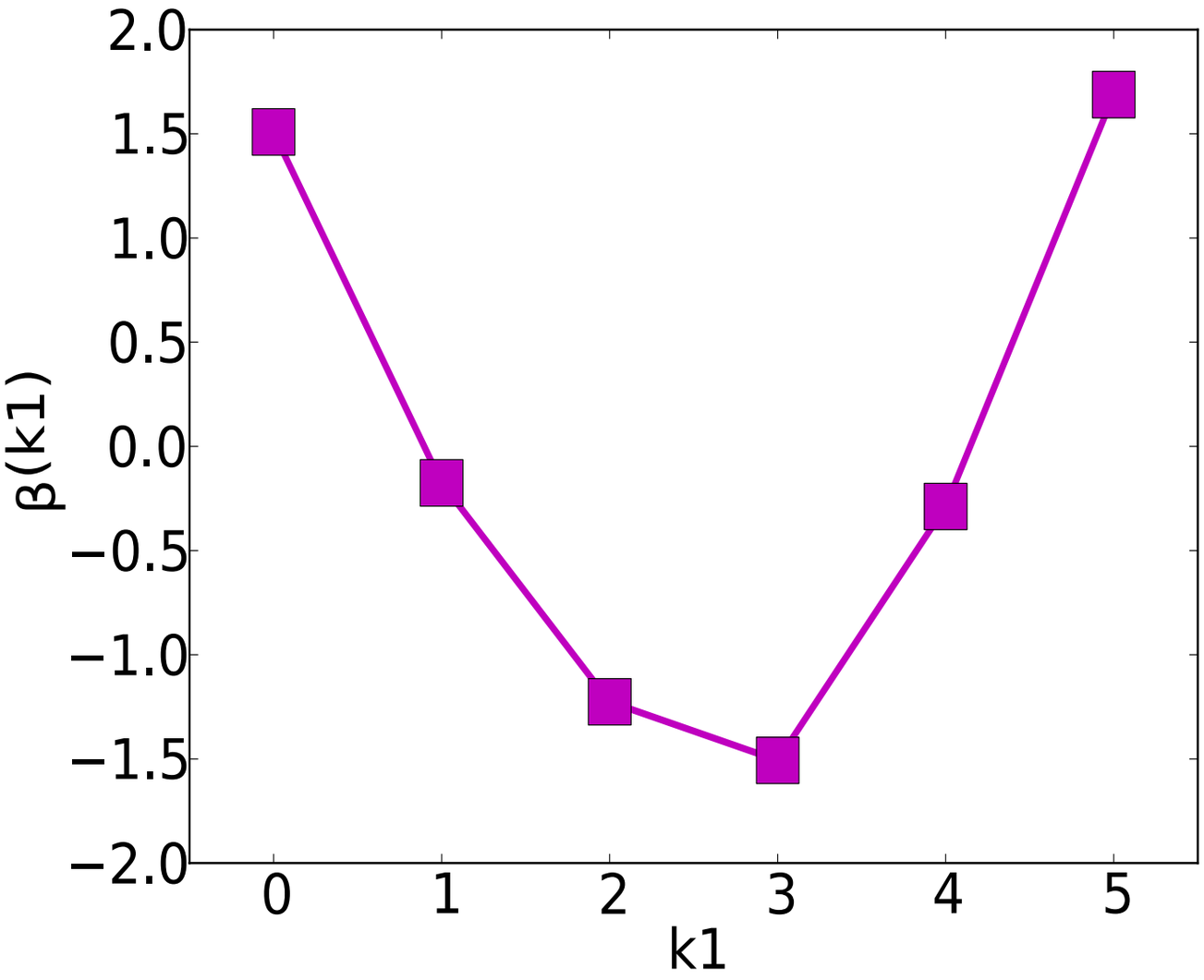
$f(x_i, y_i) = y_i x_i$

$g(\mathbf{y}') = [0, \ldots, 1, 0, \ldots 0]^T$

$$L = \frac{1}{Z} \prod_{i=1}^{k} \underbrace{\exp\left[\alpha^T f(x_i, y_i)\right]}_{\phi(x_i, y_i')} \cdot \underbrace{\exp\left[\beta^T g(\mathbf{y}')\right]}_{\gamma(\mathbf{y}')}$$

(k+1) dimensional vector indicating how many 1s are in *y'*

Trained Annotation Model



Model Learning

- Learned by gradient descent
- Training based on 6,121 annotations made on 1,000 size *k*=5 batches

k_1: Number of positive annotations in the same batch

Exploiting the bias for batch active learning

How do we *leverage* the annotation bias for better classifier performance?

Batch Active Learning

Batch Active Learning:

- Define the labeled data set as *L*, unlabeled as *U*
- In each iteration, select a batch of data items, *A,* from *U* to be labeled by an *oracle*
- Such that classifier performance can be best improved by training on the updated labeled set

Batch Active Learning with Biased Oracles:

Denote the distribution from which the annotation of oracles is drawn as

$$q\left(y_A \mid x_A; \alpha, \beta\right)$$

Active Learning Strategy

Take Discriminative Active Learning - Proposed in [1]

$$A^* = \arg\max_{A \subset U} \max_{y_A} (1-\mu) \underbrace{\sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right)}_{\text{Likelihood of labeled data}} - \mu \underbrace{H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)}_{\text{Entropy of unlabeled data}}$$

Define

$$F\left(A, y_A\right) = \left(1-\mu\right) \sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right) - \mu H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)$$

Yielding Weighted Discriminative Active Learning

$$A^* = \arg\max_{A \subset U} \max_{y_A} \left[q\left(y_A \mid x_A; \alpha, \beta\right)\left(F\left(A, y_A\right) - F\left(\varnothing, .\right)\right)\right]$$

[1] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In NIPS, pages 593–600, 2008

Active Learning Strategy

Take Discriminative Active Learning - Proposed in [1]

$$A^* = \arg\max_{A \subset U} \max_{y_A} (1-\mu) \underbrace{\sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right)}_{\text{Likelihood of labeled data}} - \mu \underbrace{H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)}_{\text{Entropy of unlabeled data}}$$

Define

$$F\left(A, y_A\right) = \left(1-\mu\right) \sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right) - \mu H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)$$

Yielding Weighted Discriminative Active Learning

$$A^* = \arg\max_{A \subset U} \max_{y_A} \left[q\left(y_A \mid x_A; \alpha, \beta\right)\left(F\left(A, y_A\right) - F\left(\varnothing, .\right)\right)\right]$$

[1] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In NIPS, pages 593–600, 2008

Active Learning Strategy

Take Discriminative Active Learning - Proposed in [1]

$$A^* = \arg\max_{A \subset U} \max_{y_A} (1-\mu) \underbrace{\sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right)}_{\text{Likelihood of labeled data}} - \mu \underbrace{H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)}_{\text{Entropy of unlabeled data}}$$

Define

$$F(A, y_A) = (1-\mu) \sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right) - \mu H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)$$

Yielding Weighted Discriminative Active Learning

$$A^* = \arg\max_{A \subset U} \max_{y_A} \left[q(y_A \mid x_A; \alpha, \beta)\left(F(A, y_A) - F(\varnothing, .)\right)\right]$$

[1] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In NIPS, pages 593–600, 2008

Active Learning Strategy

Take Discriminative Active Learning - Proposed in [1]

$$A^* = \arg\max_{A \subset U} \max_{y_A} \underbrace{(1-\mu) \sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right)}_{\text{Likelihood of labeled data}} - \underbrace{\mu H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)}_{\text{Entropy of unlabeled data}}$$

Define

$$F\left(A, y_A\right) = (1-\mu) \sum_{i \in L \cup A} \log p\left(y^{(i)}_{L \cup A} \mid x_i; w^{(t+1)}\right) - \mu H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right)$$

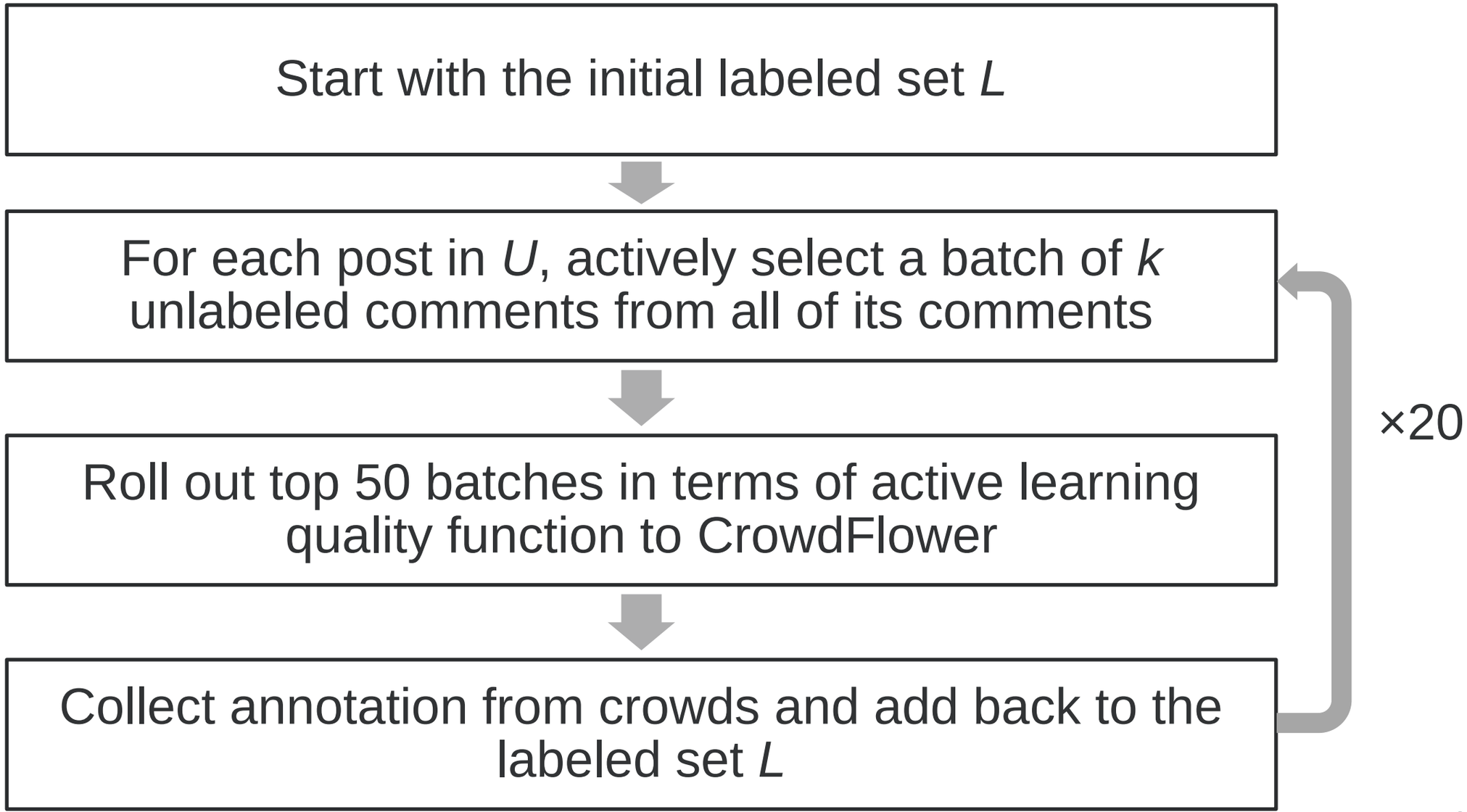
Yielding Weighted Discriminative Active Learning

$$A^* = \arg\max_{A \subset U} \max_{y_A} \left[q\left(y_A \mid x_A; \alpha, \beta\right)\left(F\left(A, y_A\right) - F\left(\varnothing, .\right)\right)\right]$$

[1] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In NIPS, pages 593–600, 2008

Training Setup and Methodology

Data set	Labeled L	Unlabeled U	Test	Validation
Size	30	6,982	1,372	343



Experimental Results

Method	AUC	Rcl@Prc=0.95	Rcl@Prc=0.90
RND	97.73	11.45	43.49
MU	97.86	29.74	48.16
DA	97.82	37.50	52.17
WDA	**98.17**	**51.97**	**56.05**

- **Random (RND)**

  At each iteration, randomly select *b* data items

- **Maximum Uncertainty (MU)**

  Select top-*k* uncertain data items where

$$u(x_i) = \sum_{y_i} -P(y_i \mid x_i; w) \log P(y_i \mid x_i; w)$$

- **Discriminative Active Learning (DA)**

$$A^* = \arg\max_{A \subset U} \max_{y_A} \left[(1-\mu) \sum_{i \in L \cup A} \log p\left(y_{L \cup A}^{(i)} \mid x_i; w^{(t+1)}\right) - \mu H_p\left(y_{U \setminus A} \mid x_{U \setminus A}; w^{(t+1)}\right) \right]$$

Summary

Qualitatively verified the existence of in-batch annotation bias

Designed a factor-graph based annotation model to quantitatively measure that bias

Leveraged that model to enhance our batch active learning algorithm

Verified performance on a real crowdsourcing platform with real operational data

Our mission is to connect the world's professionals to make them more productive and successful.

Jeff Weiner

Our vision is to create economic opportunity for every member of the global workforce.

Jeff Weiner