Evaluating Crowdsourcing Participants in the Absence of Ground-Truth

Ramanathan Subramanian[1], Romer Rosales[2], Glenn Fung[3], Jennifer Dy[1]
Northeastern Univ. Boston, MA USA
LinkedIn, Mountain View, CA USA
Siemens Healthcare, Malvern, PA, USA

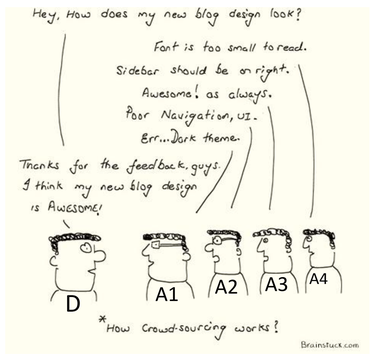
Outline

Data can be acquired, shared, and processed by an increasingly larger number of entities, in particular people. The distributed nature of this phenomenon has contributed to the development of many crowdsourcing projects. This scenario is prevalent in most forms of expert/non-expert group opinion and rating tasks (including many forms of internet or on-line user behavior), where a key element is the aggregation of observations-opinions from multiple sources.

This poster describes an approach for learning to label/annotate by modeling participant expertise. The main focus is on proposing and evaluating a novel approach to identify unhelpful or adversarial participants. The paper presents an approach for participant (also referred to as annotators) evaluation based on a multi-annotator statistical model and provides experimental validation of the approach with respect to number of adversaries, degree of adversarial behavior, and effect on the classification accuracy.

Summary Problem Definition

- Supervised/semi-supervised learning where:
  - Ground-truth exist but is not available or is very expensive
  - Multiple annotation sources (participants) are available and label some data points
- Problem:
  - How to evaluate participants
- Example use:
  - Identify/eliminate unhelpful or adversarial participants



Learning with a Single Participant/Annotator

- Predict a binary event y given (related) observations about examples $\mathbf{x}$

Training data

$$\left\{\begin{array}{lll} \text{Examples:} & X = [\mathbf{x}_1^T, \mathbf{x}_2^T, ..., \mathbf{x}_N^T] & \mathbf{x}_i \in \mathbb{R}^D \\ \text{Labels:} & Y = [y_1, y_2, ..., y_N]^T & y \in \{+,-\} \end{array}\right\}$$

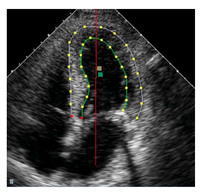
Examples	Labels
$\mathbf{x}_1$	+
$\mathbf{x}_2$	-
$\mathbf{x}_3$	+
⋮	⋮
$\mathbf{x}_k$	-
⋮	⋮
$\mathbf{x}_N$	+

"ground-truth"

Learning → $p(y|\mathbf{x})$

Fundamental Limitations

- Example scenario: Coronary artery disease diagnosis
  - CAD can be diagnosed by measuring and scoring regional heart-wall motion in echocardiography images.



Cardiologists in agreement Heart-level (A/N)	% of cases N=240
5	48.75%
4 or more	62.09%
3 or more	70.42%

  - No objective ground-truth
  - Quality of diagnosis highly dependent on skill and training of reader. High variability in quality of interpretation
- Increasingly common scenario
  - Product quality from user reviews, conference paper rating from multiple reviews, quality of on-line contributions from user history, …
- What is common in these problems?
  - Ground-truth exists but is unverifiable/subjective
    - Or it is verifiable but at a high cost
  - Multiple annotator participants exist
    - Multiple inexpensive labels may be available, but will depend on labeler idiosyncrasies.
- We could address these problems using the standard supervised/unsupervised approach
  - Simply aggregate training data
  - Majority voting

  (But not ideal)

Try to Learn by Discovering the Expertise in Everyone

- Basic idea:
  - Face the unavailability of ground-truth
  - Learn different aspects from different participants/annotators
  - Employ labeler identity→ learn to discriminate
    - Create participant/annotator-specific models

Examples	L1	L2	L3	L4	L5	Ground-truth
$\mathbf{x}_1$	+	+	+	-	+	X
$\mathbf{x}_2$	-	+	-	-	+	X
$\mathbf{x}_3$	-	+	+	-	-	X
$\mathbf{x}_k$	-	+	-	-	-	X
$\mathbf{x}_N$	+	+	+	-	+	X

Challenges when Modeling Multiple Participants

- Some participants/annotators may:
  - be more reliable than others
  - be more reliably for certain types of data
  - be malicious
  - be correlated with others
  - **have varying accuracy depending on the data instance observed**
- Questions (e.g.; medical setting)
  - How should the patients be diagnosed when doctors disagree?
  - How do we evaluate the doctors' diagnoses?
  - How do we evaluate doctors?

Modeling Multiple Participants

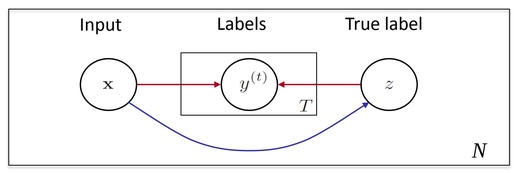


Figure 1: Graphical Model for $\mathbf{x}$, $\mathbf{y}$, and z respectively inputs, annotator-specific labels, and ground truth label (for simplicity $\alpha, \beta, \{\gamma_t\}$, and $\{\mathbf{w}_t\}$, with $t \in \{1, .., T\}$, are excluded)

$$p(Y, Z|X) = \prod_i p(z_i|\mathbf{x}_i) \prod_t p(y_i^{(t)}|\mathbf{x}_i, z_i)$$

Classification Model: Logistic Regression

$$p(z_i = 1|\mathbf{x}_i) = (1 + \exp(-\alpha^T \mathbf{x}_i - \beta))^{-1}$$

Multi-labeler Model: Bernoulli || Gaussian

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = (1 - \eta_t(\mathbf{x}))^{|y_i^{(t)} - z_i|} \eta_t(\mathbf{x})^{1-|y_i^{(t)} - z_i|}$$

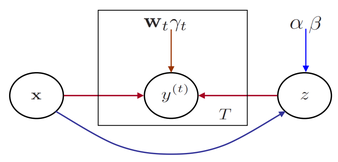
$\eta_t(\mathbf{x})$: probability that labeler t is correct for (given input)

$$p(y_i^{(t)}|\mathbf{x}_i, z_i) = \mathcal{N}(y_i^{(t)}; z_i, \sigma_t(\mathbf{x}_i))$$

$\sigma_t(\mathbf{x})$: how labeler t deviates from the true value (given input)

- Bernoulli/Gaussian model parameters, as a function of the input

$$\eta_t(\mathbf{x}) = (1 + \exp(-\mathbf{w}_t^T \mathbf{x}_i - \gamma_t))^{-1}$$



- Participant/annotator specific distributions

	$p(y_i^{(t)}\|\mathbf{x}_i, \text{A})$	$p(y_i^{(t)}\|\mathbf{x}_i, z_i)$
L1:	{N,A,A,N,A,N,A,A,N,A,A}	{N,N,N,N,N,N,A,N,N,A,N}
L2:	{A,A,A,N,A,A,A,A,N,A,A}	{N,N,A,N,N,N,N,N,N,N,N}
L3:	{A,N,A,N,A,N,A,N,A,N,A}	{A,N,A,N,A,N,A,N,A,N,A}

Formally Evaluating Participants

- Known ground-truth (GT):
  - Relatively simple solution (can compare against GT)
- Unknown ground-truth
  - No ground-truth but we could rely on label redundancy
  - Improve performance by annotator expertise modeling
  - Underlying idea (for a single point): evaluate participant/annotator k by determining how easy/difficult it is to predict its label from the labels given by other annotators

$$p(y^{(k)}|\{y^{(t\backslash k)}\}, \mathbf{x}) \longrightarrow$$ Worse … Better

0 … 1

Evaluating Participants - Formulation Details

- Partially annotated data distribution given by model

$$p(Y_O|X) = \prod_i \sum_{z_i} p(z_i|\mathbf{x}_i) \prod_{t|t \in \mathcal{T}_i} p(y_i^{(t)}|\mathbf{x}_i, z_i)$$

Partial labels

$\mathcal{T}_i$: set of annotators who annotated i-th point

- Using model independence assumptions

$$p(y^{(k)}|\{y^{(t\backslash k)}\}, \mathbf{x}) = \frac{p(\{y^{(t)}\}|\mathbf{x})}{p(\{y^{(t\backslash k)}\}|\mathbf{x})}$$

$$= \frac{\sum_z p(\overbrace{\{y^{(t)}\}}^{\text{all annot.}}|z, \mathbf{x}) \overbrace{p(z|\mathbf{x})}^{\text{GT estimate}}}{\sum_z p(\underbrace{\{y^{(t\backslash k)}\}}_{\text{w/o k-th annot.}}|z, \mathbf{x}) p(z|\mathbf{x})}$$

- In the limit of known ground-truth we have that annotators are mutually independent:

$$p(y^{(k)}|\{y^{(t\backslash k)}\}, \mathbf{x}) = p(y^{(k)}|\mathbf{x})$$

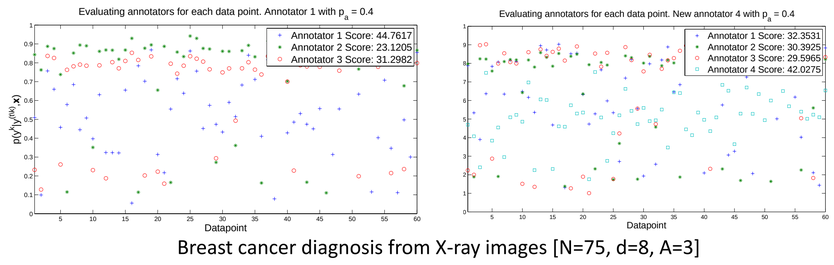
Experimental Setup

- Train multi-annotator model on 70% of the data
  - Employs real annotators and (when applicable) simulated adversaries
- Test on remaining 30% of the data
- Results show averages and deviations on 10 repeated experiments
- Annotator score (sum): log probability of annotator k label assignments given remaining annotator label assignments

$$S_k = -\sum_i \log p(y_i^{(k)}|\{y_i^{(t\backslash k)}\}, \mathbf{x}_i)$$

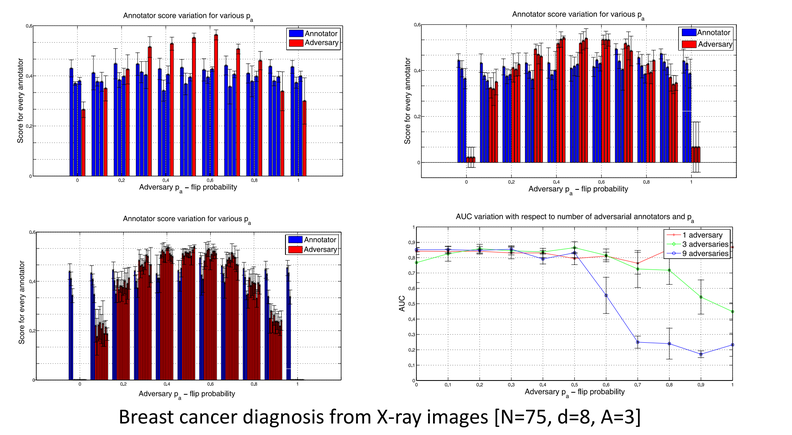
- Helpful/adversarial participants added by generating a random label flip with varying probability p_a

Conditional Label Probability Values - Breast Cancer

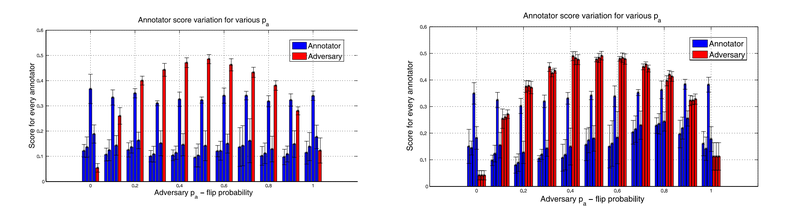


Breast cancer diagnosis from X-ray images [N=75, d=8, A=3]

Experiments [score/p_a/A] - Multiple Participant Data

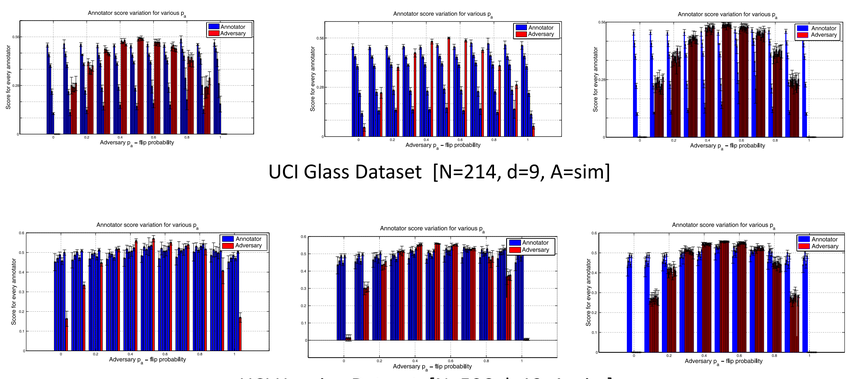


Breast cancer diagnosis from X-ray images [N=75, d=8, A=3]



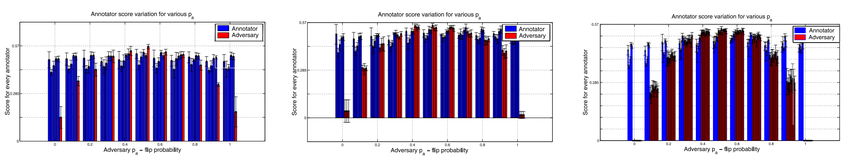
Atrial Fibrillation Dataset (inference from text) [N=1081, d=231, A=4]

Experiments [score/p_a/A]. UCI Data



UCI Glass Dataset [N=214, d=9, A=sim]

UCI Housing Dataset [N=506,d=13, A=sim]



UCI Ionosphere Dataset [N=351,d=34, A=sim]

Experiments [AUC/p_a/A]. Multiple Datasets

