Planning to Control Crowd-Sourced Workflows

Daniel S. Weld
University of Washington

30,000’ View

- Crowdsourcing is huge & growing rapidly
  - Virtual organizations
  - Flash teams with mixed human & machine members

- Automatic organization of work
  - Reduce labor required by 30-85%
Crowdsourcing

- Performing work by **soliciting effort** from many people
- **Combining the efforts** of volunteers/part-time workers (each contributing a small portion) to produce a large or significant result

Crowdsourcing Successes

- 190 M reviews of 4.4 M businesses
- Answers to 7.1 M prog. questions
- Universal reference for anything
Citizen Science

800,000 volunteers – Hubble images
Discovered “Hanny’s Voorwerp” black-hole
“Pea galaxies”

Crowdsourced bird count & identification
Migration shift -> effect of climate change

Game to find 3D structure of proteins.
Solved 15 year outstanding AIDS puzzle

Labor Marketplaces
Will Grow to $5B by 2018 [Staffing Industry Analysts]

- 2.7 million workers
- 540,000 requestors
- 35M hours worked in 2012

60% Growth Hours / week

Charts from Paroo’s (Paroo.com) blog, phone from gizmodo
Example Job on Mechanical Turk

Write a descriptive caption for this picture, then submit.

A partial view of a pocket calculator together with some coins and a pen.

Submit $0.05

Big Work from Micro-Contributions

• Challenges
  – Small work units
  – Reliability & skill of individual workers vary

• Therefore
  – Use a workflow to aggregate results & ensure quality
  – Manage workers with (unreliable) workers
Ex: Iterative Improvement

Initial caption

Ex: Iterative Improvement

Initial caption

[Liole et al., 2010]
Ex: Iterative Improvement

initial caption

improve

improved

original

vote

output

Ex: Iterative Improvement

initial caption

improve

improved

original

vote

output
Iterative Improvement

[Little et al, 2010]

First version
A partial view of a pocket calculator together with some coins and a pen.

After 8 iterations
A CASIO multi-function, solar powered scientific calculator.
A blue ball point pen with a blue rubber grip and the tip extended.
Six British coins; two of £1 value, three of 20p value and one of 1p value.
Seems to be a theme illustration for a brochure or document cover treating finance - probably personal finance.

Figure from [Little et al, 2010]
“You (misspelled) (several) (words). Please spellcheck your work next time. I also notice a few grammatical mistakes. Overall your writing style is a bit too *phoney*. You do make some good (points), but they got lost amidst the (writing). (signature)"

According to our ground truth, the highlighted words should be “flowery”, “get”, “verbiage” and “B-” respectively.

**Workflow Control Problem**

- How many voters?
- How many times?
- Adaptive, Decision-Theoretic Control
Outline

✓ Introduction
  • Case Study: Controlling Iterative Improvement
  • Case Study: Controlling Taxonomy Generation
  • In Progress: Controlling ML Annotation

TurKontrol
POMDP Control of Iterative Improvement

Peng Dai  Chris Lin
Both co-advised with Mausam
Artificial Intelligence 101

Agent

Sensors

POMDP

Actuators

Environment

Percepts

Actions

Markov Decision Process

Input:

World State

$s = <x, y>$

Actions

$P(s' | s, a)$

Cost $c$

Observe: Next State $s' = <x', y'>$

Reward $= f(s, a, s')$

Output:

Construct policy $\pi : S \rightarrow A$, that chooses best action for each state 
I.e., actions that maximize expected reward – costs over time

While learning action & reward probabilities
(Reinforcement learning)
Partially- Observable Markov Decision Process

Input:

- **Belief State** $P(s)$
- **Actions** $P(s' | s, a)$
- **Observe: Noisy Sensor** $= f(s')$

Output:

- Construct **policy**, $\pi : S \rightarrow A$, that chooses best action for each state
  
  i.e., actions that maximize expected reward – costs over time

  While **learning** action & reward probabilities
  
  (Reinforcement learning)

Solving the POMDP

**Constructing the policy**, $\pi$, to choose the best action

- **Many algorithms**
  
  - Point-based methods
  
  - UCT on discretized space
  
  - Lookahead search with beta distribution belief states

  $$Q^*(s, a) = \sum_{s'} P(s' | s, a) \left[ R(s, a, s') + \gamma \max_a Q^*(s, a) \right]$$

- **Exploration / exploitation problem**
  
  - $\varepsilon$-greedy
  
  - UCB / Multi-armed bandit
From (Hidden)

World State \( <x, y> \) coords

Actions Move
Grasp

Costs Power used

Reward

Quality \( Q_1, Q_2 \in (0, 1) \)

Improve caption task
Vote best caption

\( \$$ \) paid to workers

F(quality returned)

Belief State

Quality \( \alpha_1 \)

Quality \( \alpha_2 \)
Transition Model of Voting Action

Learned using Expectation Maximization

Worker votes that artifact 1 is better

POMDP for Iterative Improvement

Submit

Initial artifact (α)

Need improving?

Y

N

α

α' generate improve job estimate quality of α'

more voting?

Y

N

α α'

make ballot job update quality estimates

better of α and α'
POMDP for Iterative Improvement

submit

initial artifact ()

need improving?

Y

α generate improve job

α’ estimate quality of α’

more voting?

Y

α α’ make ballot job

update quality estimates

N

better of α and α’
POMDP for Iterative Improvement

submit

initial artifact ()

need improving ?

α

generate improve job

α’

estimate quality of α’

more voting ?

Y

N

α, α’

update quality estimates

better of α and α’

POMDP for Iterative Improvement

submit

initial artifact ()

need improving ?

α

generate improve job

α’

estimate quality of α’

more voting ?

Y

N

α, α’

update quality estimates

better of α and α’
POMDP for Iterative Improvement

submit

initial artifact ()

need improving?

α

generate improve job

α’

estimate quality of α’

more voting?

N

Y

α, α’

make ballot job

update quality estimates

better of α and α’
POMDP for Iterative Improvement

- Submit
- Initial artifact ( )
- Need improving?
  - Yes: Generate improve job
  - No: Estimate quality of \( \alpha' \)
  - More voting?
    - Yes: Make ballot job
    - No: Update quality estimates

Better of \( \alpha \) and \( \alpha' \)
POMDP for Iterative Improvement

submit

initial artifact ()

need improving?

N

Y

α

generate improve job

α’
estimate quality of α’

more voting?

Y

N

α, α’

make ballot job

update quality estimates

better of α and α’
Comparison

40 images, same average cost

Controlling quality: POMDP 30% less labor

Allocation of Human Labor
Human Labor Redirected

![Graph showing comparison between POMDP and Hand Coded methods]

Lessons So Far

- Reduced labor costs
- Improved quality

- POMDP planning
  - Update belief states about uncertain world
  - Model sensing actions

- Expectation maximization & prob inference
Outline

✓ Introduction
✓ Case Study: Controlling Iterative Improvement
  • Case Study: Controlling Taxonomy Generation
  • In Progress: Controlling ML Annotation

Cascade
Crowdsourcing Taxonomy Creation

Lydia Chilton
Co-advised with James Landay
Image Data Sets

- images (100)
  - sild (10)
  - birds (6)
  - leopards (3)
  - tiger (3)
  - weas (2)
- people (25)
  - sport (8)
  - bus (6)
  - world travel (2)
  - surfing (2)
- tree (18)
  - rock formations (12)
  - architecture (10)
  - mammoths (9)
  - historical landmarks (7)
  - castle (2)
  - hanging (5)
  - winter landscapes (5)
  - sculptures (6)
  - polaroid culture (4)
  - scuba diving (4)
  - snorkeling (3)
  - airplane (2)
  - boats (2)
  - island (2)

Q&A Site Responses

- traveling (100)
  - air travel tips (49)
    - preparation for flying (38)
    - comfortable flying (13)
    - airport tips (26)
    - airport shortcuts (17)
    - flight (25)
      - flight comfort (11)
      - seating in airlines (5)
      - flight layovers (2)
    - in flight meals (6)
    - airport food (4)
    - best picks for airport and airline food (4)
    - where to sit on long flight (2)
    - membership discounts (2)
    - international phone usage (3)
    - international data plans (2)
    - insider tips (49)
      - making friends with locals (6)
      - airport amenities (4)
Crowdsourcing Taxonomy Generation

*Is Hard!*

- Good taxonomy requires a global perspective
- But workers see only a tiny fraction of data...?

Iterative Improvement?

**Task:** Add Tips to the Hierarchy of Travel Advice

**Problems**
1. The growing hierarchy becomes overwhelming
2. Workers confused

**Lesson:** Decompose the task into smaller steps
Initial Approach 2: Category Comparison

Lesson: Don’t compare abstractions

Cascade Overview

[Chilton et al., CHI-13]

Use the crowd to:
1. Generate category names
2. Select the best categories
3. Place the data into the best categories

Use machines to:
4. Infer global structure of categories

Problem
Without context it’s hard to judge relationships:
• TSA liquids vs. removing liquids
• Packing vs. what to bring
Example Input: 100 Random Colors

Step 0. Sample Data

Step 1. Generate Categories

For each color

Task

What category do you suggest for this color?

Crowd responses

This generates an initial set of category names.
Step 2. Select Best Categories

For each color	

Task

Crowd responses

What is the best category for this color?

<table>
<thead>
<tr>
<th>Category</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqua</td>
<td>1/5</td>
</tr>
<tr>
<td>Greenish</td>
<td>4/5</td>
</tr>
<tr>
<td>Lime</td>
<td>0/5</td>
</tr>
<tr>
<td>Pastel</td>
<td>0/5</td>
</tr>
</tbody>
</table>

An early filter for spam and vague categories

---

Step 3. Label Data

For each color and category

Task

Crowd responses

What categories does this belong to?

<table>
<thead>
<tr>
<th>Category</th>
<th>Fits</th>
<th>Doesn’t Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>Greenish</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>Yellow</td>
<td>✘</td>
<td>✔</td>
</tr>
<tr>
<td>Pink</td>
<td>✘</td>
<td>✘</td>
</tr>
</tbody>
</table>

This determines category membership.
Step 4. Global Structure Inference

Determine parent/child relations; eliminate duplicates.

Finally, ... Recurse

May lead to new tags & recomputing taxonomy
Evaluation

**Quality**

- Inter-annotator agreement
- \( \text{avg} \)

**Cost**

- \( \text{avg} \)
- Cost: Cascade vs. 4 Experts
- Cost: Experts vs. 3 other experts

**Deluge**

(Decision-Theoretic Control of Cascade)

Jonathan Bragg

Co-advised with Mausam
**Why is Cascade Expensive?**

<table>
<thead>
<tr>
<th>Generate</th>
<th>SelectBest</th>
<th>Categorize</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 workers</td>
<td>5 workers</td>
<td></td>
</tr>
<tr>
<td># initial items</td>
<td># initial items</td>
<td># categories</td>
</tr>
</tbody>
</table>

But do we really need all these questions? What’s the best order to ask them?

---

**POMDP Model Agent Belief State**

*World state* = taxonomy & labels applying to item

![Diagram showing POMDP model with world states and beliefs](image_url)
POMDP Model Agent Belief State

**Belief state** includes ... distribution over taxonomies
label probabilities for item

![POMDP Model Diagram](image)

Learn & refine taxonomy during execution
Too complex for off-the-shelf POMDP solver

---

Decision Cycle for New Item

![Decision Cycle Diagram](image)

Probabilistic inference to update
- Posterior probabilities
- Co-occurrence model for labels
- Worker accuracy

Ask about label with max VOI

[Bragg, Mausam & W HCOMP-13]
Performance of Decision-Theoretic Model

Now crowd is cheaper than experts!

Reaches same performance as Cascade with only 13% as many voting jobs
Lessons So Far

- Decision-theoretic planning
  - Probabilistic inference
  - Expectation maximization

- Reduced labor & improved quality
  - Iterative Improvement
  - Taxonomy Generation
  - ???

Outline

- Introduction
- Case Study: Controlling Iterative Improvement
- Case Study: Controlling Taxonomy Generation
- In Progress: Controlling ML Annotation
Information Omnivore Project

• Large Scale Information Extraction

• Train via 2 kinds of Weak Supervision
  – Align Corpus to Background Knowledge Base
    [Wu & W CIKM-07; ... Koch et al. EMNLP-14]
  – Identify & Extract Events from Newswire
    [Zhang & W EMNLP-13; Zhang, Soderland & W TACL-15]

• Augment with Crowdsourced Annotations
  – Eg: “Calling himself Guccifer, Marcel-Lehel Lazar rampaged through the email accounts of rich and powerful Americans…”

• Train via Semi-Distant Supervision
  – Align Corpus to Background Knowledge Base
    [Wu & W CIKM-07; ... Koch et al. EMNLP-14]
  – Identify & Extract Events from Newswire
    [Zhang & W EMNLP-13; Zhang, Soderland & W TACL-14]
Observation

• Vast proportion of micro-task crowdsourcing... is used to create training data for ML classifiers
  – Chris Caliston-Burch (UPenn) $250,000 on MTurk
  – LDC: 44 FT employees just creating NLP training data
  – Google, MSFT – internal CS: each larger than MTurk

• Common approach
  – Get two humans to annotate
  – If they agree, ... done
  – Else recruit a third to arbitrate

\[ U(p(x|y, \theta)) = \left\| E_{p(x|y)}(\beta_x) - E_{p(x|y, q(y))}(\beta_x) \right\|_2 \]
\[ \approx \left\| E \left( \frac{1}{S-1} \sum_{s=1}^{S} z_s^{(s)} \right)^T \left[ (\gamma_{1,1}^{(s-1)}, Z^{(s-1)}) - (\gamma_{1,1}^{(s)}, Z^{(s)}) \right] \right\|_2 \]

\[ Q(\alpha, \beta) = E[\log p(\lambda | \alpha, \beta)] \]
\[ = E \left[ \log \prod_{i} \left( \prod_{j} p(q_{ij} | r_i, \beta_j) \right) \right] \]
\[ = \sum_{i} E[\log \gamma_{1,1}] + \sum_{i,j} E[\log \gamma_{ij} | \alpha_i, \beta_j] \]

\[ p(z|L, \theta) = \prod_{j \in [M]} p(q_{ij}|0) y_{ij}^{(1)} (1 - q_{ij})^{1-x_{ij}} d_{ij} \]
\[ = \prod_{j \in [M]} \psi_j(z_{ij}) \]
How should one best spend a fixed annotation budget... *when training an ML classifier?*

Unilabel?

9 examples with labels that are 75% accurate?
2/3 Relabel?

3 examples each with 3 labels – consensus 84% accurate?

Or Even?

1 example with 9 labels – consensus 98% accurate?
**Existing Data Sets?**

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Features</th>
<th># Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Breast Cancer</td>
<td>9</td>
<td>699</td>
</tr>
<tr>
<td>(b) Bank Note Authentication</td>
<td>4</td>
<td>1372</td>
</tr>
<tr>
<td>(c) Seismic Bumps</td>
<td>18</td>
<td>2584</td>
</tr>
<tr>
<td>(d) EEG Eye State</td>
<td>14</td>
<td>14980</td>
</tr>
<tr>
<td>(e) Sonar</td>
<td>60</td>
<td>208</td>
</tr>
<tr>
<td>(f) Breast Cancer Diagnostic</td>
<td>30</td>
<td>569</td>
</tr>
<tr>
<td>(g) Hill-Valley</td>
<td>100</td>
<td>606</td>
</tr>
<tr>
<td>(h) Hill-Valley with Noise</td>
<td>100</td>
<td>606</td>
</tr>
<tr>
<td>(i) Internet Ads</td>
<td>1558</td>
<td>2359</td>
</tr>
<tr>
<td>(j) Gisette</td>
<td>5000</td>
<td>6000</td>
</tr>
<tr>
<td>(k) Farm Ads</td>
<td>54877</td>
<td>4143</td>
</tr>
<tr>
<td>(l) Spambase</td>
<td>57</td>
<td>4601</td>
</tr>
</tbody>
</table>

- Relabeling is better
- Unilabeling is better

Question:

What is the dataset for unilabeling and relabeling?
Factors that Affect Relabeling Efficacy

Inductive Bias of Classifier

“Strong” $\rightarrow$ limited expressiveness

“Weak” $\rightarrow$ can learn many different concepts

Worker Accuracy

Budget

If Data was Clean

True Concept: 65 and older $\rightarrow$ “Senior Citizen”
With Noisy Annotation

True Concept: 65 and older -> “Senior Citizen”

(low expressiveness)
Strong Inductive Bias Classifier

True Concept: 65 and older -> “Senior Citizen”
Overfitting to Noise

(high expressiveness)
Weak Inductive Bias Classifier

True Concept: 65 and older -> “Senior Citizen”

Conjecture

• Relabeling more important for classifiers with weak inductive bias
  (e.g., in domains with myriad features)
Experiments on Synthetic Data

- Weaker Inductive Bias
- Relabeling Better

Number of Features (≈VC Dimension)

Relabeling Accuracy / Unilabeling Accuracy

- 2/3 Relabeling
- 3/5 Relabeling
- 4/7 Relabeling

Revisiting the Real Data

- < 100 features
- > 100 features

Dataset

- 2/3 Relabeling
- 3/5 Relabeling
- 4/7 Relabeling
Factors that Affect Relabeling Efficacy

Inductive Bias of Classifier
Worker Accuracy

Accuracy of Training Data

Figure 7: The increase in the aggregate accuracy of the training set data when using relabeling instead of unilabeling for various worker accuracies.

Figure 8: When the workers are moderately accurate ($p = 0.7$), the difference in the upper bound on error due to typical relabeling strategies is greatest. However, this intuition is faulty as we now explain. Indeed, (Ipeirotis et al. 2013) have shown that typical relabeling strategies have the maximum effect on the accuracy of a training set (not on the resulting classifier), when workers are of intermediate abilities. Consider Figure 7, which plots the increase in aggregate accuracy of the training data when relabeling instead of unilabeling as a function of worker accuracy. The three peaks happen between $0.73 - 0.79$. We also observe that 2/3-relabeling only improves accuracy by about 0.1 in the best case, whereas 4/7-relabeling can get to an almost 0.2 increase. Furthermore, as the amount of relabeling is increased, the peak in accuracy gain moves to the left, suggesting that strategies with increasing amounts of relabeling have their maximum effect as the workers become less accurate.

But these past results only apply to the quality of a training set, not the accuracy of the resulting classifier. By considering the accuracy of the classifier, we must address the confounding factor that eschewing relabeling frees budget to be spent labeling new examples. To study this scenario further, we continue the analysis technique from the previous section to produce Figure 8, which compares various upper bound curves for different settings of worker accuracy in the setting of VC=1 and budget = 1000. Consider the difference in error bound as $m$ ranges between 333 (when every example is labeled 3 times) to 1000 (when thrice as many examples are labeled once). This delta is much greater when the workers are moderately accurate ($p = 0.75$) than for other settings of worker skill. We see similar patterns for other settings of labeling redundancy and VC dimension. These differences in error bound support the belief that typical relabeling strategies are most likely to reduce classifier error when $p$ is not an extreme value.

Simulated Datasets

To confirm these insights, we again present experimental analysis using our artificial Gaussian datasets, and use varying settings of worker accuracy. As in the previous section, we fix the number of features to be $l = 50$, and the budget to be $b = 500$. We train using decision trees and set the maximum depth to be 10. For this experiment, instead of averaging over 1000 runs, we average over 2000 runs in order to create tight confidence intervals across varying worker accuracies.

Figure 9 shows our results. The more highly redundant approaches, 4/7- and 3/5- relabeling, clearly have their maximum benefit when workers are 65% accurate. On the other hand, 2/3-relabeling has its maximum benefit somewhere between $p = 0.65$ and $p = 0.75$. These results (and similar ones that we find using different classifiers like logistic regression) mirror our intuition and our theoretical analysis. Thus, choosing the correct amount of relabeling redundancy is a complex decision which ideally should be informed by knowledge of worker accuracy.

The Effect of Budget

We now investigate the effect of budget on relabeling power. Intuitively, one might think that as the budget increases, relabeling will become the more effective strategy, because in the extreme case of when the budget is infinitely large, we should clearly label each example infinitely many times. Such a strategy allows us to train the classifier using the en-
Factors that Affect Relabeling Efficacy

Inductive Bias of Classifier
Worker Accuracy

Future Work
Relax Assumptions
Complete Decision-Theoretic Control
Outline

✓ Introduction
✓ Case Study: Controlling Iterative Improvement
✓ Case Study: Controlling Taxonomy Generation
  • Future Challenges

Other Challenges

• Usually assume *workers choose* job to perform

  ![Amazon Mechanical Turk](https://i.imgur.com/3Q5Q5Q.png)

  ![Artificial Intelligence](https://i.imgur.com/3Q5Q5Q.png)

• What if employer can *assign* jobs to best workers?
  – Google internal crowdsourcing
    • Street-view/maps, knowledge graph, search relevance
    • Task routing (expert / novice) in citizen science
Matching Jobs to Workers

- Set of jobs, each with difficulty
- Set of workers, each with
  - Skill
  - Capacity (bound on # jobs)
  - Independent errors (conditioned on difficulty)
- Minimize overall error wrt fixed budget

Knapsack?
  - “Pack” jobs with workers

Unknown Difficulty v Skill

- If skill levels are known...
  - Assigning unknown problem is like MAB “arm”
  - Once find hard problem (workers disagree), add expert
- If difficulty is known...
  - Assigning unknown worker is like MAB “arm”
- Exploration / Exploitation Tradeoff
  - [Bragg, Kolobov, Mausam & W HCOMP-14]
Additional Challenges

• Balancing worker desires w/ central needs
  – Frenzy [Chilton et al. CHI-14]

• Optimizing for time

• Balancing work, teaching & testing
Additional Challenges

- Balancing worker desires w/ central needs
- Optimizing for time
- Interleaving work, education & testing
- Workers improving job instructions
- Aggregation when majority is wrong
  - Bayesian truth serum
  - MicroTalk – focused argumentation

Related Work

- DT Crowdsourcing / Active Learning with Noise
  - GalaxyZoo – [Kamar & Horvitz 2012]
  - BBMC – [Wauthier & Jordan 2011]
  - ITS – Poppovic & Brunskill
  - [Sheng et al. 2008, Donmez et al. 2009;
    – Etc.
- Crowdsourcing Global Structure
  - Mobi – [Zhang et al. 2012]
  - Context Trees - [Verroios & Bernstein 2014]
- Information Omnivore
  - [Angeli & Manning 2014, Pershina et al. 2014]
Conclusion

• **Crowdsourcing is huge & growing rapidly**
  – Specialized communities, citizen science & labor mkts

• **Decision theoretic planning – large potential**
  – Reduce required labor by 30-85%
  – Sequential decision making is crucial
  – Must model uncertainty & noisy sensors

• **Many open questions...**
Extra Slides