Graph-based reasoning for crowdsourcing

Alexander Ihler

Dept. of Computer Science
UC Irvine
In the era of big data...

- Predictions and models need more and more data!
  - But, the data need to be *clean*, or even *annotated*
  - Where does it come from?

Experts: Time-consuming & expensive

Average Joes: Cheap but unreliable
In the era of big data...

- Predictions and models need more and more data!
  - But, the data need to be *clean*, or even *annotated*
  - Where does it come from?

Experts:
- Time-consuming & expensive

Crowdsourcing:
- Combine many non-experts
Wisdom of the crowd

“Who wants to be a millionaire?”

Options for getting help:
Example applications

ML: Building data sets

Education: Peer grading

Decision-making
We have proposed a Bayesian generative probabilistic model for the annotation process. Given only binary labels of images from many different annotators, it is possible to infer not only the underlying class (or value) of the image, but also parameters such as image difficulty and annotator competence and bias. Furthermore, the model represents both the images and the annotators as multidimensional entities, with different high level attributes and strengths respectively. Experiments suggest that there are at least three different groups of annotators that did not provide answers consistent with any other annotators. This is a common phenomenon on MTurk, where a small percentage of the annotators will provide bad quality labels or, more generally, it could be used for a softer definition of ground truth. Furthermore, our findings suggest that there are at least three different groups of annotators, those who separate: (1) ducks from everything else, (2) ducks and grebes from everything else, and (3) grebes from everything else. Interestingly, the first group of annotators was better at separating out Canada geese than Red-necked grebes have shorter necks and look more duck-like in most poses. There were also a few outlier annotators that did not provide answers consistent with any other annotators. This is a common phenomenon on MTurk, where a small percentage of the annotators will provide bad quality labels or, more generally, it could be used for a softer definition of ground truth. Furthermore, our findings suggest that there are at least three different groups of annotators, those who separate: (1) ducks from everything else, (2) ducks and grebes from everything else, and (3) grebes from everything else.

We also compared the labels predicted by the different models to select images containing at least one “duck”. The estimated parameters for each image are marked with darker gray means the model estimated the annotator to be more competent. Notice how the annotators’ decision planes fall roughly into three clusters, marked by the blue circles and discussed in Section 5.2.

### Conclusions

- **Assignment graph**
  - (How to assign?)
  - How to exploit?

- **Pre-crowd experts**
  - How to use
  - Allocation tradeoffs

- **Post-crowd experts**
  - Which items to evaluate?
perceive the task. This could be used to select annotators that are experts on certain tasks and to
ings suggest that annotators fall into different groups depending on their expertise and on how they
Besides estimating ground truth classes from binary labels, our model provides information that is
labels by integrating the labels provided by several annotators with different skills, and it does so
modeled in multidimensional space. Ultimately, our model can accurately estimate the ground truth
competence level and widely different biases, and that the annotators' classification criterion is best

Interestingly, the first group of annotators was better at separating out Canada geese than Red-necked

Results from the experiment, shown in Figure

Figure 6: Estimated image and annotator parameters on the Waterbirds dataset. The annotators were asked
to select images containing at least one “duck”. The estimated

Acknowledgements

P.P. and P.W. were supported by ONR MURI Grant #N00014-06-1-0734 and EVOLUT.ONR2. S.B. was sup-
ported by NSF CAREER Grant #0448615, NSF Grant AGS-0941760, ONR MURI Grant #N00014-08-1-0638,
and ONR MURI Grant #N00014-06-1-0734.

6 Conclusions

We have proposed a Bayesian generative probabilistic model for the annotation process. Given
\( \text{\textit{parameters}} \) for each image are marked with

\( \text{\textit{coordinates}} \) of some example

\( \text{\textit{workers}} \)
Crowdsourcing

- Goal: estimate true $z_i$
  - Minimize bit-wise error
  $$\min_{\hat{z}_i} \mathbb{E}\left[\sum_i 1(\hat{z}_i \neq z_i)\right]$$
- Challenge: workers are
  - diverse
  - adversarial
- Worker’s reliability $q_j$
  $$q_j = \text{prob}[L_{ij} = z_i] = 1 - \text{prob}[L_{ij} = -z_i]$$
  $q_j < 1/2$  $q_j \approx 1/2$  $q_j \approx 1$

Tasks: $z_i \in \{\pm 1\}$

Workers: $q_j$

Adversaries
Spammers
Experts
Model assumptions

- Worker model
  \[ p(L_{ij} | z_i, q_j) = \begin{cases} 
  q_j, & \text{if } L_{ij} = z_i \text{ (correct)} \\
  1 - q_j, & \text{if } L_{ij} \neq z_i \text{ (wrong)} 
  \end{cases} \]

- Prior on worker reliability
  \[ q_j \sim p(q_j) \]

- Joint posterior distribution
  \[ p(z, q | L) \propto \prod_j p(q_j)^{c_j} (1 - q_j)^{d_j - c_j} \]

- Estimator
  \[ \hat{z}_i = \arg \max_{z_i} \left\{ p(z_i | L) \equiv \sum_{z_{[N]} \backslash i} \int_q p(z, q | L) dq \right\} \]

  - Minimizes bit-wise error rate

\(d_j: \) total \# of images by \(j\)

\(c_j: \) \# of correct images by \(j\):

\[ c_j = \sum_i 1[L_{ij} = z_i] \]
Graphical model for crowdsourcing

- Integrating out $q_j$ analytically couples the labels:

$$p(z|L) = \int_q p(z, q|L) dq = \prod_j \int_0^1 p(q_j) q_j^{c_j} (1 - q_j)^{d_j - c_j} dq_j$$

$$\overset{\text{def}}{=} \prod_j \psi_j(c_j) \quad \text{(Discrete factor graph)}$$

where $\psi_j(c_j) = \int_0^1 p(q_j) q_j^{c_j} (1 - q_j)^{d_j - c_j} dq_j$

Tasks

[Images of tasks]

Variables

[Images of variables]

Workers

[Images of workers]

Factors

[Images of factors]
Discrete graphical models

- Lots of work on approximate inference
  - Variational methods: scalable and easily distributed

- Belief propagation variants
  - Classical algorithm (Pearl ‘86) with many variations
  - “Message passing” or “local decomposition” algorithms
  - Provide good estimates and/or confidence intervals

Variables → Factors:

\[ m_{i \rightarrow j}(z_i) \propto \prod_{j' \neq j} m_{j' \rightarrow i}(z_i) \]

Factors → Variables:

\[ p(z_i) \approx \prod_{j \in \partial_i} m_{j \rightarrow i}(z_i) \]

\[ m_{j \rightarrow i}(z_i) \propto \sum_{z_{\partial j \setminus \{i\}}} \psi_j \prod_{i' \neq i} m_{i' \rightarrow j}(z_{i'}) \]
Effects of the prior

- Worker model

\[ p(L_{ij} | z_i, q_j) = \begin{cases} 
q_j, & \text{if } L_{ij} = z_i \text{ (correct)} \\
1 - q_j, & \text{if } L_{ij} \neq z_i \text{ (wrong)} 
\end{cases} \]

- Prior on worker reliability \( q_j \sim p(q_j) \)

Some different priors:

- Beta(1,1)
- Beta(3,1)
- Spammer–hammer
Special priors

- Deterministic prior
  - No diversity; sensitive to adversaries
  - BP => majority voting

- Haldane prior
  - High diversity (variance); emphasizes adversaries
  - BP => [Karger, Oh, Shah 2011]

Optimistic

Pessimistic
Experiments

- Random \((l,r)\) regular bipartite graph
  - 1000 tasks
  - True prior: spammer-hammer
  - Vary # of workers per image \((l)\), fixing # images per worker \((r)\)
Experiments

- Random \((l,r)\) regular bipartite graph
  - 1000 tasks
  - True prior: spammer-hammer
  - Vary # of workers per image \((l)\), fixing # images per worker \((r)\)

\[
\ell (\text{fixed } r = 5)
\]

\[
\text{Algorithm Prior}
\]
Experiments

- Random \((l,r)\) regular bipartite graph
  - 1000 tasks
  - True prior: spammer-hammer
  - Vary # of workers per image \((l)\), fixing # images per worker \((r)\)
Experiments

- Random \((l, r)\) regular bipartite graph
  - 1000 tasks
  - True prior: spammer-hammer
  - Vary # of workers per image \((l)\), fixing # images per worker \((r)\)
Extensions

- Sensitivity & specificity models of workers:

  Reliability Model:
  \[ q_j = \text{prob}[L_{ij} = z_i] \]
  \[ = 1 - \text{prob}[L_{ij} = -z_i] \]

  Sensitivity & Specificity Model (Dawid & Skene 79):
  \[ s_j = \text{prob}[L_{ij} = +1|z_i = +1], \quad \text{(sensitivity)} \]
  \[ t_j = \text{prob}[L_{ij} = -1|z_i = -1]. \quad \text{(specificity)} \]

- Model selection by marginal likelihood:

  \[ K = \frac{p(L|M_1)}{p(L|M_2)} = \frac{\sum_z \int_q p(z, q, L|M_1)dq}{\sum_z \int_q p(z, q, L|M_2)dq} \]

- Incorporating item features, label priors, ...
Bluebird dataset [Welinder et al., 2010]

- Indigo Bunting vs. Blue Grosbeak
  - 108 tasks (images)
  - 39 workers

[Image: A graph showing the error rate over the number of workers per task for different methods: 'KOS', 'EM -Beta(2,1)', 'BP-Beta(2,1)', 'MV'. The x-axis represents the number of workers per task, ranging from 5 to 20, and the y-axis represents the error rate, ranging from 0.1 to 0.5.]

Thanks to P. Welinder and S. Belongie for data and code.
Bluebird dataset [Welinder et al., 2010]

- Indigo Bunting vs. Blue Grosbeak
  - 108 tasks (images)
  - 39 workers

Welinder, Branson, Belongie, Perona, NIPS 2010

Thanks to P. Welinder and S. Belongie for data and code.
Crowdsourcing and graphs

(1) Assignment graph
- (How to assign?)
- How to exploit?

Workers:

Tasks:

(2) Pre-crowd experts
- How to use
- Allocation tradeoffs

(3) Post-crowd experts
- Which items to evaluate?
Effects of bias

- Ex: price estimation data set
  - 80 items in 10 categories
  - 155 + 287 participants

- Predictions are heavily biased!
  - Need some control items to estimate bias...
Expert advice (pre-crowd)

• Control questions = expert-labeled data points
  – Can tell us about worker accuracy, bias, etc
  – At the same time, they consume worker resources

Control questions:
• Known answers
• Evaluate the workers

Target questions:
• Unknown answers
• Want to answer
Expert advice (pre-crowd)

• Control questions = expert-labeled data points
  – Can tell us about worker accuracy, bias, etc
  – At the same time, they consume worker resources

Control questions:
  • Known answers
  • Evaluate the workers

Target questions:
  • Unknown answers
  • Want to answer
Expert advice (pre-crowd)

• Control questions = expert-labeled data points
  – Can tell us about worker accuracy, bias, etc
  – At the same time, they consume worker resources

Control questions:
• Known answers
• Evaluate the workers

Target questions:
• Unknown answers
• Want to answer
Goal: optimal $k$ to minimize error on the target questions?

$l$: # of all questions asked workers (budget)
$k$: # of control questions asked workers
$r$: # of workers per target items, $r = \frac{m}{n_t}(\ell - r)$. 

Problem setup

[Liu, Steyvers, Ihler; NIPS 2013]
**Some simple models**

[Page 106]

Bias-only Model

$$x_{ij} = \mu_i^* + b_j^* + \xi_{ij}, \quad \xi_{ij} \sim \mathcal{N}(0, \sigma^*^2),$$

Bias of worker j  
Label of item i

Bias-variance Model (heteroskedastic)

$$x_{ij} = \mu_i^* + b_j^* + \xi_{ij}, \quad \xi_{ij} \sim \mathcal{N}(0, \sigma_j^{*2}),$$

Bias of worker j  
Label of item i  
Variance of worker j

Variance-only Model (zero bias)

$$x_{ij} = \mu_i^* + \xi_{ij}, \quad \xi_{ij} \sim \mathcal{N}(0, \sigma_j^{*2})$$
Estimation methods

Two-stage estimation:

Scoring workers:
\[ \hat{\nu}_j = \arg \max_{\nu_j} \sum_{i \in \partial_j} \log p(x_{ij} | \mu^*_i, \nu_j) \]

Predicting target items:
\[ \hat{\mu}_i = \arg \max_{\mu_i} \sum_{j \in \partial_i} \log p(x_{ij} | \mu_i, \hat{\nu}_j) \]

Joint estimation:

\[ [\mu, \nu] = \arg \max_{\{\mu, \nu : \mu_C = \mu^*_C\}} \sum_{i,j} \log p(x_{ij} | \mu_i, \nu_j) \]

\( C \) : control questions
Estimation methods

Two-stage estimation:

Control questions \( \mu_i^* \) Target questions \( \hat{\mu}_i \)

\( \hat{\nu}_j \)

Workers

\( k^* \approx [\sqrt{a\ell + a^2} - a] \approx \sqrt{a\ell} \)

Bias-only: \( a=1 \)

Joint estimation:

Control questions \( \mu_i^* \) Target questions \( \hat{\mu}_i \)

\( \hat{\nu}_j \)

Workers

(Bias only, regular, random bipartite graph...)

\( k^* \approx \sqrt{b\bar{\ell}} = \ell / \sqrt{n_t} \)

Fewer control questions than 2-stage:

If \( \ell \leq n_t \), then \( \ell / \sqrt{n_t} \leq \sqrt{\ell} \)
Point Spreads of NFL Games

- 245 NFL games, 386 workers (Massey et al. 11)
- Random subsample, varying $\ell$

Bias-only model:

Two-stage estimator: $k^* = \sqrt{\ell}$

Joint estimator: $k^* = \ell / \sqrt{n_t}$
Guessing product prices

80 household items
155 UC Irvine students

(c) Optimal $k$ vs. $\ell$
Crowdsourcing and graphs

(1) Assignment graph
- (How to assign?)
- How to exploit?

Workers:

Tasks:

(2) Pre-crowd experts
- How to use
- Allocation tradeoffs

(3) Post-crowd experts
- Which items to evaluate?
Expert advice (post-crowd)

• Identify items that need help
• Target using expert budget

• Can target different types of items:

The *most uncertain* item

The *most “influential”* item

*Quantify effects and optimize...*
Expert label set selection

• How should we select the set of items to label?

• Minimize conditional variance?
  \[
  \min_{C : |C| \leq K} \mathbb{E}(\text{var}(\mu_{\neg C} | X, \mu_C) | X)
  \]

• Or, minimize entropy?
  \[
  \min_{C : |C| \leq K} \mathbb{H}(\mu_{\neg C} | X, \mu_C)
  \]

• Both “hard”: use a Laplace approximation to the posterior

• Set selection: use a greedy selection
  – Objective is submodular

\[
i^* = \arg \max_i \{\sigma_{ii} + \sum_{k \neq i} \rho_{ik}^2 \sigma_{kk}\}
\]

\[
i^* = \arg \max_i \{\sigma_{ii}\} \quad \text{Local only!}
\]
Expert advice (post-crowd)

- Compare improvement rates:
  - “Local” choice (most uncertain)
  - “Local + global” choice (estimate combined effect)

---

Simulated data:

Real data (age prediction)

Number of expert labels

Avg MSE on Unchecked Items

[Liu, Ihler, Fisher; Fusion 2015]
Conclusions & future work

• Crowdsourcing + graphical models

• Exploiting graph structure
  – Efficient, message passing algorithms
  – Build on successful methods in graphical models

• Leveraging early expert information
  – Before the crowd: control questions
  – Balance info about crowd with info about tasks

• Leveraging expert information after the crowd
  – Assessment: which tasks still need measurements?
  – Phase transition: finding out about workers versus tasks
Acknowledgements

<table>
<thead>
<tr>
<th>Collaborators</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qiang Liu</td>
<td>(UCI, MIT, Dartmouth)</td>
</tr>
<tr>
<td>Jian Peng</td>
<td>(MIT, UIUC)</td>
</tr>
<tr>
<td>Mark Steyvers</td>
<td>(UCI CogSci)</td>
</tr>
<tr>
<td>John Fisher</td>
<td>(MIT)</td>
</tr>
<tr>
<td>Kanan Pavandale</td>
<td>(UCI BioSci)</td>
</tr>
<tr>
<td>Brian Sato</td>
<td>(UCI BioSci)</td>
</tr>
<tr>
<td>INF 231 students</td>
<td></td>
</tr>
</tbody>
</table>

Thanks!