Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Presentation:
Kamran Ghasedi
Spring 2015
Crowdsourcing
Problem definition
K-coin tossing problem
Markov Decision Process (MDP) and Optimal Policy
Bayesian Setup
Accuracy Maximization
Optimization problem
Stage-wise expected reward
Optimistic Knowledge Gradient
Experiment on simulated data
Experiment on real dataset
Conclusion
Paper

Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Dengyong Zhou  
Microsoft Research, Redmond, WA 98052

Qiang Liu  
University of California, Irvine, CA 92697

John C. Platt  
Microsoft Research, Redmond, WA 98052

Christopher Meek  
Microsoft Research, Redmond, WA 98052

Abstract

We propose a method to aggregate noisy ordinal labels collected from a crowd of workers or annotators. Eliciting ordinal labels is important in tasks such as judging web search quality and rating products. Our method is motivated by the observation that workers usually have difficulty distinguishing between two adjacent ordinal classes whereas distinguishing between two classes which are far away from each other is much easier. We formulate our method as min-

An advanced approach for label aggregation is suggested by Dawid & Skene (1979). They assume that each worker has a latent confusion matrix for labeling. The off-diagonal elements represent the probabilities that a worker mislabels an arbitrary item from one class to another while the diagonal elements correspond to her accuracy in each class. Worker confusion matrices and true labels are jointly estimated by maximizing the likelihood of observed labels. One may further assume a prior distribution over worker confusion matrices and perform Bayesian inference (Raykar et al., 2010; Liu et al., 2012; Chen et al., 2013).
Demand for humans’ labels
Crowdsourcing services

- Microsoft
- Facebook
- Google
- Baidu
- Amazon
Amazon Mechanical Turk

Find an interesting task → Work → Earn money

Fund your account → Load your tasks → Get results

Example of Tagging of an Image
Provide 3 tags for this image.

Instructions:
- You must provide 3 tags for this image.
- Each tag must be a single word.
- No tag can be longer than 25 characters.
- The tags must describe the image, the contents of the image, or some relevant content.

Image:

Tag 1:
Tag 2:
Tag 3:
Crowds vs. experts labeling

More data beats clever algorithms!
Crowdsourced labels may be highly noisy.
**Toy example**

<table>
<thead>
<tr>
<th></th>
<th>Orange (O)</th>
<th>Mandarin (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
<td>O</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

Orange (O) vs. Mandarin (M)
### Toy example

<table>
<thead>
<tr>
<th></th>
<th>Orange (O)</th>
<th>Mandarin (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
<td>M</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
<td>O</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

**True labels?**

Orange (O) vs. Mandarin (M)
### Notations

<table>
<thead>
<tr>
<th>Workers</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>$j$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>$j$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$i$</td>
<td>...</td>
</tr>
<tr>
<td>$i$</td>
<td>1</td>
</tr>
<tr>
<td>$i$</td>
<td>2</td>
</tr>
<tr>
<td>$i$</td>
<td>...</td>
</tr>
<tr>
<td>$i$</td>
<td>$j$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Observed worker labels**

**Unobserved true labels:** $y_j$
Assumptions

Worker ability  Item difficulty

Combination
1. Develop a method to aggregate general multiclass labels
2. Adapt general method to ordinal labels
Examples on multiclass labeling

- Image categorization
- Speech recognition
Assume that each worker has a latent confusion matrix for labeling.

- The off-diagonal elements represent the probabilities that a worker mislabels an arbitrary item from one class to another.
- Diagonal elements correspond to its accuracy in each class.

Worker confusion matrices and true labels are jointly estimated by maximizing the likelihood of observed labels.

One may further assume a prior distribution over worker confusion matrices and perform Bayesian inference.
Two fundamental tensors

represent an observed confusion from class \( c \) to class \( k \) by worker \( i \) on item \( j \)

Empirical confusion tensor

\[
\hat{\phi}_{ij}(c, k) = Q(Y_j = c) \mathbb{I}(x_{ij} = k)
\]

Expected confusion tensor

\[
\phi_{ij}(c, k) = Q(Y_j = c) P(X_{ij} = k | Y_j = c)
\]

represent an expected confusion from class \( c \) to class \( k \) by worker \( i \) on item \( j \)

\( P: \) worker label distribution \hspace{1cm} \( Q: \) true label distribution
Entropy of the observed labels conditioned on the true labels

Reminder:

\[ H(X|Y) = - \sum_{j,c} Q(Y_j = c) \sum_{i,k} P(X_{ij} = k|Y_j = c) \times \log P(X_{ij} = k|Y_j = c). \]
Multiclass maximum conditional entropy

Given the true labels $Q$, estimate $P$ by

$$\max_P H(X|Y)$$

enforce the expected confusion for each worker matches to its empirical confusion.

**worker constraints**

$$\sum_j \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = 0, \forall i, k, c$$

**item constraints**

$$\sum_i \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = 0, \forall j, k, c$$

enforce the expected confusion for each item matches to its empirical confusion.
**Intuitively:** it means that we believe that the observed labels are the least random given the true labels.

**Theoretically:** minimum conditional entropy can be understood as maximum likelihood.
Lagrangian of maximization problem

\[ L = H(X \mid Y) + L_\sigma + L_\tau + L_\lambda \]

\[
L_\sigma = \sum_{i,c,k} \sigma_i(c,k) \sum_j \left[ \phi_{ij}(c,k) - \hat{\phi}_{ij}(c,k) \right]
\]

\[
L_\tau = \sum_{j,c,k} \tau_j(c,k) \sum_i \left[ \phi_{ij}(c,k) - \hat{\phi}_{ij}(c,k) \right]
\]

\[
L_\lambda = \sum_{i,j,c} \lambda_{ijc} \left[ \sum_k P(X_{ij} = k \mid Y_j = c) - 1 \right]
\]

constraints
The $(c,k)$-th entry represents how likely worker $i$ labels a randomly chosen item in class $c$ as class $k$.

\[
P(X_{ij} = k | Y_j = c) = \frac{1}{Z_{ij}} \exp[\sigma_i(c, k) + \tau_j(c, k)]
\]

$Z_{ij}$ normalization factor

The $(c,k)$-th entry represents how likely item $j$ in class $c$ is labeled as class $k$ by a randomly chosen worker.
Dual form of minimax problem

Lagrangian:

$$\log \left\{ \prod_j \sum_c Q(Y_j = c) \prod_i P(X_{ij} = x_{ij} \mid Y_j = c) \right\}$$

1. This only generates deterministic labels
2. Equivalent to maximizing complete likelihood
1. Develop a method to aggregate general multiclass labels
2. Adapt general method to ordinal labels
Example of ordinal labeling

**search results**

- Machine learning - Wikipedia, the free encyclopedia
  - Machine learning: a branch of artificial intelligence concerned with the construction and study of systems that can learn from data. For example, a machine
    - Definition, Generalization, Machine learning and Human interaction

- Machine Learning | Coursera
  - Machine Learning. Learn about the most effective machine learning techniques, and gain practice implementing them and getting them to work for yourself.

- Machine Learning | Stanford Online
  - What is the format of the class? The class will consist of lecture videos, which are broken into small chunks, usually between eight and twelve minutes each.

- Machine learning | Define Machine learning at Dictionary.com
  - World English Dictionary machine learning — n a branch of artificial intelligence in which a computer generates rules underlying or based on raw data that has been
Proceed to ordinal labels

• Formulate assumptions which are specific for ordinal labeling
• Coincide with the previous multiclass method in the case of binary labeling
Assumption for ordinal labeling

likely to confuse

unlikely to confuse
Formulating the assumption using pairwise comparison

Reference label

≥, <

Indirect label comparison

True label

≥, <

Worker label
Ordinal minmax conditional entropy

Jointly estimate $P$ and $Q$ by

$$\min_Q \max_P H(X|Y)$$

subject to

worker constraints

$$\sum_{c\Delta s} \sum_{k\Delta s} \sum_i \left[ \phi_{ij}(c,k) - \hat{\phi}_{ij}(c,k) \right] = 0, \forall i, s$$

item constraints

$$\sum_{c\Delta s} \sum_{k\Delta s} \sum_i \left[ \phi_{ij}(c,k) - \hat{\phi}_{ij}(c,k) \right] = 0, \forall j, s$$

$\Delta$: take on values $<$ or $\geq$

$\forall$: take on values $<$ or $\geq$
Ordinal minmax conditional entropy

Jointly estimate $P$ and $Q$ by

$$\min_{Q} \max_{P} H(X|Y)$$

subject to

- **worker constraints**
  $$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{j} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = 0, \forall i, s$$

- **item constraints**
  $$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{i} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = 0, \forall j, s$$

reference label
true label
worker label
Ordinal minmax conditional entropy

Jointly estimate $P$ and $Q$ by

$$\min_{Q} \max_{P} H(X|Y)$$

subject to

$$\sum_{c \Delta s} \sum_{k \Delta s} \sum_{j} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = 0, \forall i, s$$

$$\sum_{c \Delta s} \sum_{k \Delta s} \sum_{i} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = 0, \forall j, s$$

worker constraints

item constraints

difference from multiclass

reference label

true label

worker label
Partition the Cartesian product of the label set

\{(c, k) | c < s, k < s\}, \{(c, k) | c < s, k \geq s\},
\{(c, k) | c \geq s, k < s\}, \{(c, k) | c \geq s, k \geq s\}.

(a) Partitioning with \( s = 1 \)

<table>
<thead>
<tr>
<th></th>
<th>(0, 0)</th>
<th>(0, 1)</th>
<th>(0, 2)</th>
<th>(0, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0)</td>
<td>(0, 1)</td>
<td>(1, 0)</td>
<td>(1, 1)</td>
<td>(1, 2)</td>
</tr>
<tr>
<td>(1, 0)</td>
<td>(1, 1)</td>
<td>(1, 2)</td>
<td>(1, 3)</td>
<td>(2, 0)</td>
</tr>
<tr>
<td>(2, 0)</td>
<td>(2, 1)</td>
<td>(2, 2)</td>
<td>(2, 3)</td>
<td>(3, 0)</td>
</tr>
<tr>
<td>(3, 0)</td>
<td>(3, 1)</td>
<td>(3, 2)</td>
<td>(3, 3)</td>
<td></td>
</tr>
</tbody>
</table>

(b) Partitioning with \( s = 2 \)

<table>
<thead>
<tr>
<th></th>
<th>(0, 0)</th>
<th>(0, 1)</th>
<th>(0, 2)</th>
<th>(0, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0)</td>
<td>(0, 1)</td>
<td>(0, 2)</td>
<td>(0, 3)</td>
<td>(0, 4)</td>
</tr>
<tr>
<td>(1, 0)</td>
<td>(1, 1)</td>
<td>(1, 2)</td>
<td>(1, 3)</td>
<td>(1, 4)</td>
</tr>
<tr>
<td>(2, 0)</td>
<td>(2, 1)</td>
<td>(2, 2)</td>
<td>(2, 3)</td>
<td>(2, 4)</td>
</tr>
<tr>
<td>(3, 0)</td>
<td>(3, 1)</td>
<td>(3, 2)</td>
<td>(3, 3)</td>
<td>(3, 4)</td>
</tr>
</tbody>
</table>

(c) Partitioning with \( s = 3 \)

<table>
<thead>
<tr>
<th></th>
<th>(0, 0)</th>
<th>(0, 1)</th>
<th>(0, 2)</th>
<th>(0, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0)</td>
<td>(0, 1)</td>
<td>(0, 2)</td>
<td>(0, 3)</td>
<td>(0, 4)</td>
</tr>
<tr>
<td>(1, 0)</td>
<td>(1, 1)</td>
<td>(1, 2)</td>
<td>(1, 3)</td>
<td>(1, 4)</td>
</tr>
<tr>
<td>(2, 0)</td>
<td>(2, 1)</td>
<td>(2, 2)</td>
<td>(2, 3)</td>
<td>(2, 4)</td>
</tr>
<tr>
<td>(3, 0)</td>
<td>(3, 1)</td>
<td>(3, 2)</td>
<td>(3, 3)</td>
<td>(3, 4)</td>
</tr>
</tbody>
</table>
Explaining the ordinal constraints

For example, let $\Delta = \prec, \succ = \succeq$:

$$\sum_{c < s} \sum_{k \geq s} \hat{\phi}_{ij}(c, k) = Q(Y_j < s)I(x_{ij} \geq s)$$

counting mistakes in ordinal sense
By the KKT conditions, the dual problem leads to

\[ P(X_{ij} = k | Y_j = c) = \frac{1}{Z_{ij}} \exp[\sigma_i(c, k) + \tau_j(c, k)] \]

worker ability

\[ \sigma_i(c, k) = \sum_{s \geq 1} \sum_{\Delta, \nabla} \sigma_{is}^{\Delta, \nabla} \mathbb{I}(c\Delta s, k\nabla s) \]

item difficulty

\[ \tau_j(c, k) = \sum_{s \geq 1} \sum_{\Delta, \nabla} \tau_{js}^{\Delta, \nabla} \mathbb{I}(c\Delta s, k\nabla s) \]

structured
Regularization

Two goals:
1. Prevent over fitting
2. Fix the deterministic label issue to generate probabilistic labels
Regularized minmax conditional entropy

Jointly estimate $P$ and $Q$ by

$$\min_Q \max_P H(X|Y) + \text{regularization terms}$$

subject to

**worker constraints**

$$\sum_{c, s} \sum_{k, \nabla s} \sum_{i, j} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] \approx 0, \forall i, s$$

**item constraints**

$$\sum_{c, s} \sum_{k, \nabla s} \sum_{i, j} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] \approx 0, \forall j, s$$
Regularized minmax conditional entropy

Jointly estimate $P$ and $Q$ by

$$\min_Q \max_P H(X|Y) - H(Y) - \frac{1}{\alpha} \Omega(\xi) - \frac{1}{\beta} \Psi(\zeta)$$

subject to

**worker constraints**

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{i} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = \xi_{is}$$

**item constraints**

$$\sum_{c \Delta s} \sum_{k \nabla s} \sum_{j} \left[ \phi_{ij}(c, k) - \hat{\phi}_{ij}(c, k) \right] = \zeta_{js}$$

$$H(Y) = -\sum_{j, c} Q(Y_j = c) \log Q(Y_j = c).$$
Dual problem

\[
\max_{\sigma, \tau, Q} \sum_{j,c} Q(Y_j = c) \sum_i \log P(X_{ij} = x_{ij} | Y_j = c) + H(Y) - \alpha \Omega(\sigma) - \beta \Psi(\tau)
\]

1. This generates probabilistic labels
2. Equivalent to maximizing marginal likelihood
Choosing regularization parameters

- Cross-validation: 5 or 10 folds
- Random split
- Compare the likelihood of worker labels

Don’t need ground truth labels for cross-validation!
Experiments: metric

- Evaluation metrics
  - L0 error: $L_0 = \mathbb{I}(y \neq \hat{y})$
  - L1 error: $L_1 = |y - \hat{y}|$
  - L2 error: $L_2 = |y - \hat{y}|^2$
Experiments: baselines

- Compare regularized minimax condition entropy to
  - Majority voting
  - Dawid-Skene method (1979, see also its Bayesian version in Raykar et al. 2010, Liu et al. 2012, Chen et al. 2013)
Web search data

Search results for "machine learning"
Web search data

- Some facts about the data:
  - 2665 query-URL pairs and a relevance rating scale from 1 to 5
  - 177 non-expert workers with average error rate 63%
  - Each query-URL pair is judged by 6 workers
  - True labels are created via consensus from 9 experts
  - Dataset created by Gabriella Kazai of Microsoft
Result for web search data

<table>
<thead>
<tr>
<th>Method</th>
<th>L0 Error</th>
<th>L1 Error</th>
<th>L2 Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority vote</td>
<td>0.269</td>
<td>0.428</td>
<td>0.930</td>
</tr>
<tr>
<td>Dawid &amp; Skene</td>
<td>0.170</td>
<td>0.205</td>
<td>0.539</td>
</tr>
<tr>
<td>Latent trait</td>
<td>0.201</td>
<td>0.211</td>
<td>0.481</td>
</tr>
<tr>
<td>Entropy multiclass</td>
<td>0.111</td>
<td>0.131</td>
<td>0.419</td>
</tr>
<tr>
<td>Entropy ordinal</td>
<td>0.104</td>
<td>0.118</td>
<td>0.384</td>
</tr>
</tbody>
</table>
Probabilistic labels vs. error rate

![Graph showing L0, L1, and L2 error rates for different probability ranges](image-url)
Price prediction data

<table>
<thead>
<tr>
<th>Price Range</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 – $50</td>
<td>1</td>
</tr>
<tr>
<td>$51 – $100</td>
<td>2</td>
</tr>
<tr>
<td>$101 – $250</td>
<td>3</td>
</tr>
<tr>
<td>$251 – $500</td>
<td>4</td>
</tr>
<tr>
<td>$501 – $1000</td>
<td>5</td>
</tr>
<tr>
<td>$1001 – $2000</td>
<td>6</td>
</tr>
<tr>
<td>$2001 – $5000</td>
<td>7</td>
</tr>
</tbody>
</table>
Price prediction data

• Some facts about the data:
  – 80 household items collected from stores like Amazon and Costco
  – Prices predicted by 155 students of UC Irvine
  – Average error rate 69% and systematically biased
  – Dataset created by Mark Steyvers of UC Irvine
Result for price prediction data

<table>
<thead>
<tr>
<th></th>
<th>L0 Error</th>
<th>L1 Error</th>
<th>L2 Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority vote</td>
<td>0.675</td>
<td>1.125</td>
<td>1.605</td>
</tr>
<tr>
<td>Dawid &amp; Skene</td>
<td>0.650</td>
<td>1.050</td>
<td>1.517</td>
</tr>
<tr>
<td>Latent trait</td>
<td>0.688</td>
<td>1.063</td>
<td>1.504</td>
</tr>
<tr>
<td>Entropy multiclass</td>
<td>0.675</td>
<td>1.150</td>
<td>1.643</td>
</tr>
<tr>
<td>Entropy ordinal</td>
<td>0.613</td>
<td>0.975</td>
<td>1.492</td>
</tr>
</tbody>
</table>
Conclusion

- Minmax conditional entropy principle for crowdsourcing
- Adjacency confusability assumption in ordinal labeling
- Ordinal labeling model with structured confusion matrices
Thanks for your attention