Learning from the Wisdom of Crowds by Minimax Entropy

Denny Zhou, John Platt, Sumit Basu and Yi Mao
Microsoft Research, Redmond, WA

Outline

1. Introduction
2. Minimax entropy principle
3. Future work and conclusion

1. Introduction

Machine Learning Meets Crowdsourcing

- To Improve a machine learning model:
  - Add more training examples
  - Create more meaningful features
  - Invent more powerful learning algorithms

More and more efforts, less and less gain

Machine Learning Meets Crowdsourcing

- To Improve a machine learning model:
  - Adding more training examples
  - Creating more meaningful features
  - Inventing more powerful learning algorithms

More and more efforts, less and less gain

Crowdsourcing for Labeling

microWorkers
work & earn or offer a micro job

CloudCrowd
We're working on it. Lots of us.

clickworker

MFreelance ini

Workers
Jobs
Get jobs done by Mini Freelancers
Earn Money working on Mini Jobs

amazon mechanical turk
beta
Artificial Artificial Intelligence

Elance

CrowdFlower

agentanything

oDesk
Love the way you work.

microtask Human powered document processing

SimpleWorkers
What would you do for cash?

Low Cost, but also Low Quality

Norfolk Terrier **Norwich Terrier**

Irish Wolfhound **Scottish Deerhound**

Image Labeling
Average worker accuracy: 68%

amazonmechanical turk
Artificial Artificial Intelligence

(Stanford dogs dataset)

Problem Setting and Notations

Workers: $i = 1, 2, \cdots, m$
Items: $j = 1, 2, \cdots, n$
Categories: $k = 1, 2, \cdots, c$

Response matrix $Z_{m \times n \times c}$

- $z_{ijk} = 1$, if worker i labels item j as category k
- $z_{ijk} = 0$, if worker i labels item j as other (not k)
- $z_{ijk} = unknown$, if worker i does not label item j

Goal: Estimate the ground truth $\{y_{jk}\}$

Toy Example: Binary Labeling

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Worker 1	1	2	1	1	1	2
Worker 2	2	2	1	2	1	1
Worker 3	1	1	2	1	1	2
Worker 4	1	1	1	1	1	2
Worker 5	1	1	1	2	2	2

Problem: What are the true labels of the items?

A Simple Method: Majority Voting

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Worker 1	1	2	1	1	1	2
Worker 2	2	2	1	2	1	1
Worker 3	1	1	2	1	1	2
Worker 4	1	1	1	1	1	2
Worker 5	1	1	1	2	2	2

By majority voting, the true label of item 4 should be class 1:

{workers labeling it as class 1} = 3

{workers labeling it as class 2} = 2

Improve: More skillful workers should have more weight

Dawid & Skene's Method

- Assume that each worker is associated with a $c \times c$ confusion matrix
$$\{p_{kl}^{(i)} = \mathrm{Prob}[z_{ij} = l | y_j = k, i]\}$$
- For any labeling task, the label by a worker is generated according to her confusion matrix

- Maximum Likelihood Estimation (MLE): jointly estimate confusion matrices and ground truth
- Implementation: EM algorithm

Probabilistic Confusion Matrices

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Worker 1	1	2	1	1	1	2
Worker 2	2	2	1	2	1	1
Worker 3	1	1	2	1	1	2
Worker 4	1	1	1	1	1	2
Worker 5	1	1	1	2	2	2

Assume that the true labels are:
Class 1 = {item 1, item 2, item 3}
Class 2 = {item 4, item 5, item 6}

	Class 1	Class 2
Class 1	1	0
Class 2	2/3	1/3

EM in Dawid & Skene's Method

- Initialize the ground truth by majority vote
- Iterate the following procedure till converge:
  - Estimate the worker confusion by using the estimated ground truth
  - Estimate the ground truth by using the estimated worker confusion

Simplified Dawid & Skene's Method

Each worker i is associated with a single number $p_i \in [0,1]$ such that

$$\text{Prob}\left[z_{ij} = y_j | i\right] = p_i$$

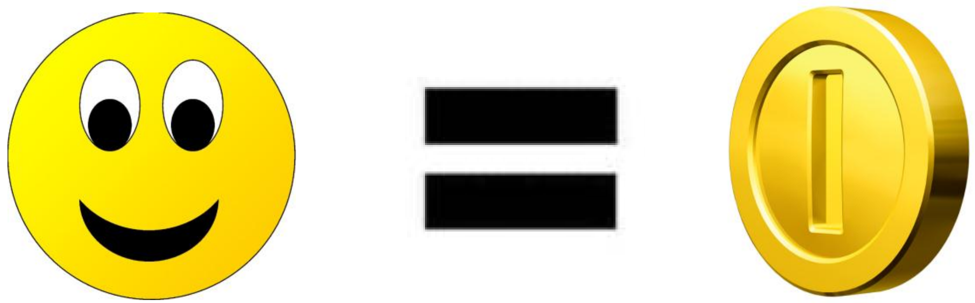
$$\text{Prob}\left[z_{ij} \neq y_j | i\right] = 1 - p_i$$

Simplified Dawid & Skene's Method

Each worker i is associated with a single number $p_i \in [0,1]$ such that

$$\text{Prob}\left[z_{ij} = y_j | i\right] = p_i$$

$$\text{Prob}\left[z_{ij} \neq y_j | i\right] = 1 - p_i$$



worker coin

2. Minimax Entropy Principle

Our Basic Assumption

	item 1	item 2	...	item n
worker 1	z_{11}	z_{12}	...	z_{1n}
worker 2	z_{21}	z_{22}	...	z_{2n}
...	...	...	...	...
worker m	z_{m1}	z_{m2}	...	z_{mn}

Observed labels

	item 1	item 2	...	item n
worker 1	π_{11}	π_{12}	...	π_{1n}
worker 2	π_{21}	π_{22}	...	π_{2n}
...	...	...	...	...
worker m	π_{m1}	π_{m2}	...	π_{mn}

unobserved distributions

Our Basic Assumption

	item 1	item 2	...	item n
worker 1	z_{11}	z_{12}	...	z_{1n}
worker 2	z_{21}	z_{22}	...	z_{2n}
...	...	...	...	...
worker m	z_{m1}	z_{m2}	...	z_{mn}

	item 1	item 2	...	item n
worker 1	π_{11}	π_{12}	...	π_{1n}
worker 2	π_{21}	π_{22}	...	π_{2n}
...	...	...	...	...
worker m	π_{m1}	π_{m2}	...	π_{mn}

Separated distribution per work-item!

Our Basic Assumption

	item 1	item 2	...	item n
worker 1	z_{11}	z_{12}	...	z_{1n}
worker 2	z_{21}	z_{22}	...	z_{2n}
...	...	...	...	...
worker m	z_{m1}	z_{m2}	...	z_{mn}

$$\neq$$

	item 1	item 2	...	item n
worker 1	π_{11}	π_{12}	...	π_{1n}
worker 2	π_{21}	π_{22}	...	π_{2n}
...	...	...	...	...
worker m	π_{m1}	π_{m2}	...	π_{mn}

Separated distribution per work-item!

Maximum Entropy

- To estimate a distribution, it is typical to use the maximum entropy principle

$$\max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

E. T. Jaynes

Column and Row Matching Constraints

	item 1	item 2	...	item n
worker 1	z_{11}	z_{12}	...	z_{1n}
worker 2	z_{21}	z_{22}	...	z_{2n}
...	...	...	...	...
worker m	z_{m1}	z_{m2}	...	z_{mn}

	item 1	item 2	...	item n
worker 1	π_{11}	π_{12}	...	π_{1n}
worker 2	π_{21}	π_{22}	...	π_{2n}
...	...	...	...	...
worker m	π_{m1}	π_{m2}	...	π_{mn}

Column Constraints

$$\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk}$$

column matching

For each item:
Count # workers labeling it as class 1
Count # workers labeling it as class 2

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Worker 1	1	2	1	1	1	2
Worker 2	2	2	1	2	1	1
Worker 3	1	1	2	1	1	2
Worker 4	1	1	1	1	1	2
Worker 5	1	1	1	2	2	2

Row Constraints

$$\sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl}z_{ijk}$$

row matching

For each worker:
Count # misclassifications from class 1 to 2
Count # misclassifications from class 2 to 1

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Worker 1	1	2	1	1	1	2
Worker 1	2	2	1	2	1	1
Worker 3	1	1	2	1	1	2
Worker 4	1	1	1	1	1	2
Worker 5	1	1	1	2	2	2

Maximum Entropy

$$\max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

Subject to

$$\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk}$$ (column constraint)

$$\sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk}$$ (row constraint)

To Estimate True Labels, Can We …

$$\max_{y} \max_{\pi} \quad - \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

Subject to

$$\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk} \qquad \text{(column constraint)}$$

$$\sum_{j=1}^{n} y_{jl} \pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk} \qquad \text{(row constraint)}$$

To Estimate True Labels, Can We …

$$\max_{y} \max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

Subject to

$$\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk}$$ (column constraint)

$$\sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk}$$ (row constraint)

Leading to a uniform distribution for $\{y_{jl}\}$

To Estimate True Labels, Can We …

$$\max_{y}\max_{\pi} -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c}\pi_{ijk}\ln\pi_{ijk}$$

Subject to

$$\sum_{i=1}^{m}\pi_{ijk} = \sum_{i=1}^{m} \text{[illegible]}$$ (column constraint)

$$\sum_{j=1}^{n}y_{jl}\pi_{ijk} = \sum_{j=1}^{n}y_{jl}z_{ijk}$$ (row constraint)

Leading to a uniform distribution for $\{y_{jl}\}$

Minimax Entropy Principle

$$\min_{y} \max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

Subject to

$$\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk}$$ (column constraint)

$$\sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk}$$ (row constraint)

making π_{ij} "peaky" means that z_{ij} is the least random given y_{jl}.

Justification of Minimum Entropy

- Assume true measurement are available:

$$\max_{\pi} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk}$$

Subject to

$$\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} \pi^*_{ijk}$$

true measurements

$$\sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl}\pi^*_{ijk}$$

Justification of Minimum Entropy

- *Theorem*. Minimizing the KL divergence

$$\ell(\pi^*, \pi) = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{\mathrm{KL}}(\pi^*_{ij} \parallel \pi_{ij})$$

is equivalent to minimize entropy.

Lagrangian Dual

- The Lagrangian dual can be written as

$$L = -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c}\pi_{ijk}\ln\pi_{ijk} + \sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_{ij}\left(\sum_{k=1}^{c}\pi_{ijk} - 1\right)$$

$$+ \sum_{j=1}^{n}\sum_{k=1}^{c}\tau_{jk}\sum_{i=1}^{m}(\pi_{ijk} - z_{ijk}) + \sum_{i=1}^{m}\sum_{k=1}^{c}\sum_{l=1}^{c}\sigma_{ikl}\sum_{j=1}^{n}y_{jl}(\pi_{ijk} - z_{ijk})$$

Lagrangian multipliers

Lagrangian Dual

- KKT conditions lead to a closed-form:

$$\pi_{ijk} = \frac{1}{Z} \exp\left\{\tau_{jk} + \sum_l y_{jl}\sigma_{ikl}\right\}$$

Z is the normalization factor given by

$$Z = \sum_k \exp\left\{\tau_{jk} + \sum_l y_{jl}\sigma_{ikl}\right\}$$

Worker Expertise & Task Confusability

- Explanation of dual variables:

$$\pi_{ijk} = \frac{1}{Z} \exp\left\{ \tau_{jk} + \sum_{l} y_{jl} \sigma_{ikl} \right\}$$

item confusability (τ_{jk})

worker expertise (σ_{ikl})

Measurement Objectivity: Item

- *Objective item confusability*. The difference of difficulty between labeling two items should be independent of the chosen workers
- *Mathematical formulation*. Let

$$c(i,j,k) = \frac{\mathbb{P}(Z_{ij} = k | Y_j = l)}{\mathbb{P}(Z_{ij} = l | Y_j = l)}$$

Then the ratio $c(i,j,k)/c(i',j,k)$ should be Independent of the choices of i, i'

Measurement Objectivity: Worker

- *Objective worker expertise*. The difference of expertise between two workers should be independent of the item being labeled
- *Mathematic Formulation*. Let

$$c(i,j,k) = \frac{\mathbb{P}(Z_{ij} = k | Y_j = l)}{\mathbb{P}(Z_{ij} = l | Y_j = l)}$$

Then the ratio $c(i,j,k)/c(i,j',k)$ should be Independent of the choices of j, j'

The Labeling Model Is Objective

Theorem. For deterministic labels, the labeling model given by

$$\pi_{ijk} = \frac{1}{Z} \exp\left\{\tau_{jk} + \sum_{l} y_{jl}\sigma_{ikl}\right\}$$

uniquely satisfies the measurement objectivity principle

Constraint Relaxation

$$\sum_{i=1}^{m} \pi_{ijk} \approx \sum_{i=1}^{m} z_{ijk}$$

column matching

For each item:
Count # workers labeling it as class 1
Count # workers labeling it as class 2

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Worker 1	1	2	1	1	1	2
Worker 1	2	2	1	2	1	1
Worker 3	1	1	2	1	1	2
Worker 4	1	1	1	1	1	2
Worker 5	1	1	1	2	2	2

Constraint Relaxation

$$\sum_{j=1}^{n} y_{jl}\pi_{ijk} \approx \sum_{j=1}^{n} y_{jl}z_{ijk}$$

row matching

For each worker:
Count # misclassifications from class 1 to 2
Count # misclassifications from class 2 to 1

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Worker 1	1	2	1	1	1	2
Worker 1	2	2	1	2	1	1
Worker 3	1	1	2	1	1	2
Worker 4	1	1	1	1	1	2
Worker 5	1	1	1	2	2	2

Constraint Relaxation

$$\min_{y} \max_{\pi,\xi,\zeta} \quad -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} \pi_{ijk} \ln \pi_{ijk} - \sum_{j=1}^{n}\sum_{k=1}^{c} \frac{\xi_{jk}^2}{2\alpha} - \sum_{i=1}^{m}\sum_{k=1}^{c}\sum_{l=1}^{c} \frac{\zeta_{ikl}^2}{2\beta}$$

Subject to

$$\sum_{i=1}^{m} \pi_{ijk} = \sum_{i=1}^{m} z_{ijk} + \xi_{jk}$$

$$\sum_{j=1}^{n} y_{jl}\pi_{ijk} = \sum_{j=1}^{n} y_{jl} z_{ijk} + \zeta_{ikl}$$

Relaxing moment constraints to prevent overfitting

Implementation

- Convert the primal problem to its dual form
- Coordinate descent
  - Split the variables into two blocks: $\{y\}, \{\tau, \sigma\}$
  - Each subproblem is convex and smooth
  - Initialize ground truth by majority vote

Model Selection

- k-fold cross validation to choose (α, β)
  - Split the data matrix into k folds
  - Each fold used as a validation set once
  - Compute average likelihood over validations

We don't need ground truth for model selection!

Experiments: Image Labeling

- 108 bird images, 2 breeds, 39 workers
- Each image was labeled by all workers

From: P. Welinder, S. Branson, S. Belongie and P. Perona. The Multidimensional Wisdom of Crowds. NIPS 2010.

Experiments: Image Labeling

- Experimental results (accuracy, %)

Worker Number	10	20	30
Minimax Entropy	85.18	92.59	93.52
Dawid & Skene	79.63	87.04	87.96
Dawid & Skene (S)*	45.37	57.41	75.93
Majority Voting	67.59	83.33	76.85
Average Worker	62.78		

* Dawid & Skene (S): simplified Dawid and Skene's method

Experiments: Image Labeling

- Experimental results (accuracy, %)

Worker Number	10	20	30
Minimax Entropy	85.18	92.59	93.52
Dawid & Skene	79.63	87.04	87.96
Dawid & Skene (S)	45.37	57.41	75.00
Majority Voting	67.59	83.33	76.85
Average Worker	62.78		

It is risky to model worker expertise by a single number

Experiments: Web Search

- 177 workers and 2665 <query, URL> pairs
- 5 classes: perfect, excellent, good, fair and bad
- Each pair was labeled by 6 workers

Minimax Entropy	88.84
Dawid & Skene	84.09
Majority Voting	77.65
Average worker	37.05

Comparing with More Methods

- Other methods: Raykar et al (JMLR 2010, adding beta/Dirichlet prior), Welinder et al (NIPS 2010, matrix factorization), Karger et al (NIPS, 2011, BP-like iterative algorithm)
- From the evaluation in (Liu et al. NIPS 2012)
  - None of them can outperform Dawid and Skene's
  - Karger et al (NIPS, 2011) is even much worse than majority voting

3. Future Work and Conclusion

Budget-Optimal Crowdsourcing

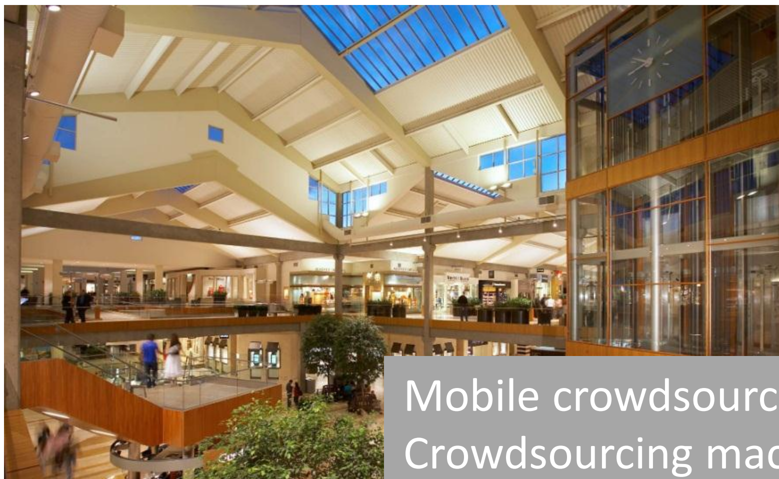
- Assume that we have a budget to get 6 labels. Which one deserves another label, item 2 or 3?
- How about having a budget of 7 labels or even more?

	1st round	2nd round
Item 1	1	1
Item 2	1	-1
Item 3	1	

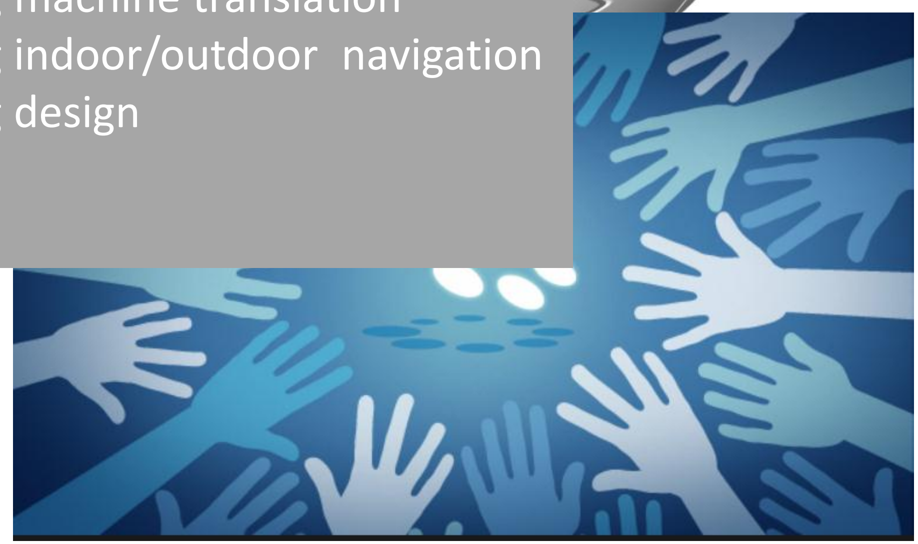
Contextual Minimax Entropy

- Contextual information of items and workers
- (An example) Label a web page as *spam* or *nonspam* by a group of workers
  - For each web page: its URL ends with .edu or not, popularity of its domain, creating time
  - For each worker: education degree, reputation history, working experience

Beyond Labeling



Mobile crowdsourcing platform
Crowdsourcing machine translation
Crowdsourcing indoor/outdoor navigation
Crowdsourcing design
Wikipedia
…



ICML'13 Workshop
Machine Learning Meets Crowdsourcing

ICML | Atlanta

International Conference on Machine Learning

16-21 JUNE 2013 ATLANTA

http://www.ics.uci.edu/~qliu1/MLcrowd_ICML_workshop/

Summary

- Proposed minimax entropy principle for estimating ground truth from noisy labels
- Both task confusability and worker expertise are taken into account in our method
- Measurement objectivity is implied

Acknowledgments

- Daniel Hsu (MSR New England)
- Xi Chen (Carnegie Mellon University)
- Gabriella Kazai (MSR Cambridge)
- Chris Burges (MSR Redmond)
- Chris Meek (MSR Redmond)