A Minimax Optimal Algorithm for Crowdsourcing

Thomas Bonald, Richard Combes
Telecom ParisTech (France), Centrale-Supelec (France)
NIPS 2017

Abstract

We propose a novel lower bound on the minimax estimation error in crowdsourcing, and we propose Triangular Estimation (TE), a low complexity, streaming algorithm to estimate the reliability of workers. We prove that TE is minimax optimal and matches our lower bound. We conclude by assessing the performance of TE and other state-of-the-art algorithms on both synthetic and real-world data sets.

Crowdsourcing

- Crowdsourcing has become a common way to label data
- Simple, repetitive tasks against low payment e.g. Amazon MT
- **Objectives:** find the true labels and detect the spammers.

$$
\begin{array}{c}
G \\
\begin{bmatrix} -1 & -1 & 1 & -1 & 1 \end{bmatrix}
\end{array}
$$

$$
\theta \begin{bmatrix} 0.9 \\ 0.1 \\ -0.7 \end{bmatrix}
\underbrace{\begin{bmatrix} -1 & -1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 1 \\ 1 & 1 & 0 & 1 & -1 \end{bmatrix}}_{\text{tasks}} \text{workers}
$$

Model

- Binary classification tasks: $+1$ or -1
- Ground truth $G(1), \dots, G(t) \in \{+1, -1\}$ i.i.d. uniform
- Answer to task t by worker $i \in \{1, \dots, n\}$:

$$
X_i(t) = \begin{cases} G(t) & \text{w.p.} \quad \alpha \frac{1+\theta_i}{2} \\ -G(t) & \text{w.p.} \quad \alpha \frac{1-\theta_i}{2} \\ 0 & \text{w.p.} \quad 1-\alpha \end{cases}
$$

where $\theta_i \in [-1, 1]$ is the reliability of worker i

- **Objective:** Estimate both the ground truth G and the reliability vector θ by observing only the answers matrix X.

Identifiability and complexity measure

- Observe that the labels are not sufficient to distinguish:
  - $\theta = [\theta_1, \theta_2, 0, \dots, 0]^T$ and $\theta' = [\theta_2, \theta_1, 0, \dots, 0]^T$
  - θ and $-\theta$
- **Proposition:** Any parameter $\theta \in \Theta$ is identifiable, with

$$
\Theta = \left\{ \theta \in [-1, 1]^n : \sum_{i=1}^n 1\{\theta_i \neq 0\} \geq 3, \ \sum_{i=1}^n \theta_i > 0 \right\}
$$

- To study the sample complexity, define

$$
\Theta_{a,b} = \left\{ \theta \in [-1, 1]^n : \min_k \max_{i,j \neq k} \sqrt{|\theta_i \theta_j|} \geq a, \ \sum_{i=1}^n \theta_i \geq b \right\}
$$

Lower bound on the estimation error

- Let $\hat\theta$ be any estimator of $\theta \in \Theta_{a,b}$
- **Theorem 1:** For any small $\epsilon, \delta > 0$, we have

$$
\min_{\theta \in \Theta_{a,b}} \mathbb{P}\left(||\hat\theta - \theta||_\infty \geq \epsilon \right) \geq \delta
$$

whenever $t \leq \max(T_1, T_2)$, where

$$
T_1 = \underbrace{c_1 \frac{1-a}{\alpha^2 a^4 \epsilon^2} \ln\left(\frac{1}{4\delta}\right)}_{\text{absolute value estimation}} \qquad T_2 = \underbrace{c_2 \frac{(1-a)^4 (n-4)}{\alpha^2 a^2 b^2} \ln\left(\frac{1}{4\delta}\right)}_{\text{sign estimation}}
$$

Covariance matrix of answers

- For any $i \neq j$, $C_{ij} = \mathbb{E}(X_i X_j | X_i X_j \neq 0) = \theta_i \theta_j$
- For any $i \neq j \neq k$, $C_{ik} C_{jk} = \theta_i \theta_j \theta_k^2 = C_{ij} \theta_k^2$ so that

$$
\theta_k^2 = \frac{C_{ik} C_{jk}}{C_{ij}} \text{ provided } C_{ij} \neq 0.
$$

- Moreover, $\theta_k \sum_i \theta_i = \theta_k^2 + \sum_{i \neq k} C_{ik}$ so that

$$
\text{sign}(\theta_k) = \text{sign}\left(\theta_k^2 + \sum_{i \neq k} C_{ik} \right)
$$

The TE algorithm

- Compute for all $i \neq j$

$$
\hat C_{ij} = \frac{\sum_t X_i(t) X_j(t)}{\max\left(\sum_t |X_i(t) X_j(t)|, 1 \right)}
$$

- Estimate the absolute value of θ by

$$
|\hat\theta_k| = \sqrt{\left| \frac{\hat C_{i_k k} \hat C_{j_k k}}{\hat C_{i_k j_k}} \right|} \quad \text{with } (i_k, j_k) \in \arg\max_{i \neq j \neq k} |\hat C_{ij}|
$$

- Estimate the sign of θ by

$$
\text{sign}(\hat\theta_k) = \begin{cases} \text{sign}(\hat\theta_{k^\star}^2 + \sum_{i \neq k^\star} \hat C_{ik^\star}) & \text{if } k = k^\star \\ \text{sign}(\hat\theta_{k^\star} \hat C_{kk^\star}) & \text{otherwise} \end{cases}
$$

with $k^\star = \arg\max_k |\hat\theta_k^2 + \sum_{i \neq k} \hat C_{ik}|$

- TE is a streaming algorithm and is not iterative
- Complexity: $\mathcal{O}(n^2)$ time per update and $\mathcal{O}(n^2)$ space.

Minimax optimality of TE

- **Theorem 2:** For any small $\epsilon, \delta > 0$, we have

$$
\max_{\theta \in \Theta_{a,b}} \mathbb{P}\left(||\hat\theta - \theta||_\infty \geq \epsilon \right) \leq \delta
$$

whenever $t \geq \max(T_1', T_2')$, where

$$
T_1' = c_1' \frac{1}{\alpha^2 a^4 \epsilon^2} \ln\left(\frac{6n^2}{\delta}\right) \qquad T_2' = c_2' \frac{n}{\alpha^2 a^2 b^2} \ln\left(\frac{6n^2}{\delta}\right),
$$

Performance on real data

Data sets description

Dataset	# Tasks	# Workers	# Labels	# Labels / W
Bird	108	39	4,212	108
Dog	807	109	8,070	74
Duchenne	159	64	1,221	19
RTE	800	164	8,000	49
Temp	462	76	4,620	61
Web	2,653	177	15,539	88

Prediction error

Dataset	Majority Vote	Expectation Maximization	Belief Propagation	TE	TE+EM
Bird	0.24	0.28	0.28	**0.18**	0.28
Dog	0.18	**0.17**	0.19	0.20	**0.17**
Duchenne	0.28	**0.23**	0.30	0.26	0.28
RTE	0.10	**0.08**	0.50	0.38	0.10
Temp	**0.06**	**0.06**	0.43	0.08	**0.06**
Web	0.14	0.06	**0.02**	0.03	0.06

Conclusion

- TE is a low complexity, streaming algorithm which requires no iterative procedure (such as BP, EM or Power Iteration)
- Surprisingly EM is not necessary at all for minimax optimality