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.

Model

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

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

where \( \theta_i \in [-1, 1] \) is the reliability of worker \( i \)
- Objective: Estimate both the ground truth \( G \) and the reliability vector \( \theta \) by observing only the answers matrix \( X \).

Identifiability and complexity measure

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

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

To study the sample complexity, define

\[
\Theta_{a,b} = \left\{ \theta \in [-1, 1]^n : \min\limits_{k, l \neq k} \sum_{i \in 1}^{n} |\theta_i| \geq a , \sum_{i=1}^{n} \theta_i \geq b \right\}
\]

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

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

whenever \( t \leq \max\left( T_1, T_2 \right) \), where

\[
T_1 = \frac{c_1(1-a)4(n-4)}{\alpha^2 a^2 \epsilon^2} \ln \left( \frac{1}{\delta} \right)
\]

sign estimation

\[
T_2 = \frac{c_2(1-a)4(n-4)}{\alpha^2 a^2 b^2} \ln \left( \frac{1}{\delta} \right)
\]

Covariance matrix of answers

- For any \( i \neq j \), \( C_{ij} = \mathbb{E}\left( X_i X_j | X_i X_j \neq 0 \right) = \theta_i \theta_j \)
- For any \( i \neq j \neq k \), \( C_{ik} \neq C_{ij} \neq 0 \) so that

\[
C_{ij} = \frac{C_{ik} C_{jk}}{C_{ij}} \quad \text{provided } C_{ij} \neq 0.
\]

Moreover, \( \theta_i \sum_{j \neq i} \theta_j = \theta_i^2 + \sum_{j \neq k} C_{ik} \) so that

\[
\text{sign}(\theta_i) = \text{sign} \left( \theta_i^2 + \sum_{j \neq k} C_{ik} \right)
\]

The TE algorithm

- Compute for all \( i \neq j \)

\[
\hat{C}_{ij} = \max \left( \sum_{t=1}^{T(t)} X_i(t) X_j(t), 1 \right)
\]

- Estimate the absolute value of \( \theta \) by

\[
|\theta_k| = \sqrt{\frac{\hat{C}_{ik} \hat{C}_{kj}}{\hat{C}_{kk}}} \quad \text{with } (i_k, j_k) \in \arg\max_{l \neq k} |\hat{C}_{l,l}|
\]

- Estimate the sign of \( \theta \) by

\[
\text{sign}(\theta_k) = \begin{cases} 
\text{sign}(\theta_k^2 + \sum_{j \neq k} \hat{C}_{ik}) & \text{if } k = k^* \\
\text{sign}(\theta_k^2 + \sum_{j \neq k} \hat{C}_{ik}) & \text{otherwise}
\end{cases}
\]

with \( k^* = \arg\max_{k} |\theta_k| \)

- 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

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

whenever \( t \geq \max\left( T_1, T_2 \right) \), where

\[
T_1 = \frac{1}{\alpha^2 a^2 \epsilon^2} \ln \left( \frac{6n^2}{\delta} \right)
\]

Performance on real data

Data sets description

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

Prediction error

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Majority</th>
<th>Expectation</th>
<th>Belief</th>
<th>TE</th>
<th>TE+EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vote</td>
<td>0.24</td>
<td>0.28</td>
<td>0.28</td>
<td>0.18</td>
<td>0.28</td>
</tr>
<tr>
<td>Dog</td>
<td>0.18</td>
<td>0.17</td>
<td>0.19</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>Duchenne</td>
<td>0.28</td>
<td>0.23</td>
<td>0.30</td>
<td>0.26</td>
<td>0.28</td>
</tr>
<tr>
<td>RTE</td>
<td>0.10</td>
<td>0.08</td>
<td>0.50</td>
<td>0.38</td>
<td>0.10</td>
</tr>
<tr>
<td>Temp</td>
<td>0.06</td>
<td>0.06</td>
<td>0.43</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>Web</td>
<td>0.14</td>
<td>0.06</td>
<td>0.02</td>
<td>0.03</td>
<td>0.06</td>
</tr>
</tbody>
</table>

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