Crowdsourcing via Tensor Augmentation and Completion (TAC)

Presenter: Yao Zhou           joint work with: Dr. Jingrui He

Arizona State University      Arizona State University
Roadmap

- Background
- Related work
- Crowdsourcing based on TAC
- Experimental results
- Conclusion
Crowdsourcing in machine learning

- Training a supervised machine learning model needs training labels
- Many crowdsourcing platforms provide services to collect labels information.
An example of crowdsourcing

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W</td>
<td>D</td>
<td>?</td>
<td>W</td>
<td>?</td>
<td>W</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>?</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>W</td>
<td>D</td>
<td>?</td>
<td>W</td>
<td>?</td>
</tr>
</tbody>
</table>

Give labels of wild animals (W) and domestic animals (D) questions mark (?) means no answer

Lynx (wildcat)

Tabby (domestic cat)
Key problem of crowdsourcing

- How to infer the true labels from a large number of labels collected from crowd?

- **Pros:**
  - Low cost: Collecting large amounts of labels is economic.

- **Cons:**
  - Low quality: Collected labels from the crowd (non-expert) are noisy.
  - Missing labels: Some workers are not willing to label all of the items.

Give labels of wild animals (W) and domestic animals (D) questions mark (?) means no answer
Some related work

- **MV**
  - Majority Voting, a simple baseline.

- **DS-EM** [Dawid and Skene, 1979]
  - Infer worker’s ability matrix and true labels.
  - Two-coin model for a binary labelling task.

- **GLAD** [Whitehill et al., 2009].
  - Infer the worker’s ability, item difficulty and item true labels simultaneously.

- **DS-MF** [Liu et al., 2012].
  - Employ variational Bayesian inference using meanfield algorithm.

- **MMCE** [Zhou et al., 2012].
  - Employ the minimax entropy principle to infer worker ability, item difficulty and true labels at the same time.

Structural information of labels is not utilized!!
Roadmap

- Background
- Related work
- Crowdsourcing based on TAC
- Experimental results
- Conclusion
Tensor augmentation

- Notation:
  - Re-organize labels of crowds as a three-way tensor: \( T = \{0, 1\}^{N_w \times N_i \times N_c} \)
  - Based on worker’s labelling decision, generate an index set: \( \Omega = \{0, 1\}^{N_w \times N_i \times N_c} \)
    - Workers: \( i = 1, 2, \ldots, N_w \)
    - Items: \( j = 1, 2, \ldots, N_i \)
    - Classes: \( k = 1, \ldots, N_c \)

- The ground truth layer:
  - Extra tensor slice of size \( N_i \times N_c \).
  - Augmented on tensor along the worker dimension.
Tensor augmentation and completion (TAC)

Goal of TAC:
- Complete the augmented tensor $\mathcal{X}$

Main principle of TAC:
- **Rank minimization**
  \[
  \min_{x} \text{rank}(\mathcal{X}) \\
  \text{s.t.} : \mathcal{X}_{\Omega} = \mathcal{T}_{\Omega}
  \]
  (NP-hard)

- **Trace norm minimization**
  \[
  \min_{x} \| \mathcal{X} \|_* \\
  \text{s.t.} : \mathcal{X}_{\Omega} = \mathcal{T}_{\Omega}
  \]
Tensor augmentation and completion (TAC)

Definition of trace norm for an n-way tensor [Liu et.al 2012]:

\[ \|X\|_* = \sum_{l=1}^{n} \alpha_l \|X_{(l)}\|_* \]

s.t. \[ \sum_{l=1}^{n} \alpha_l = 1, \alpha_l \geq 0, l = 1, \ldots, n \]

Here, \( X_{(l)} \) represents for the unfold of a tensor \( X \). The reverse operation is fold.

\[
X_1 = \begin{bmatrix}
1 & 4 \\
2 & 5 \\
3 & 6
\end{bmatrix}, \quad X_2 = \begin{bmatrix}
13 & 16 \\
14 & 17 \\
15 & 18
\end{bmatrix}
\]

Tensor: \( X \in R^{3 \times 2 \times 2} \)

\[
X_{(1)} = \begin{bmatrix}
1 & 4 & 13 & 16 \\
2 & 5 & 14 & 17 \\
3 & 6 & 15 & 18
\end{bmatrix}, \quad X_{(2)} = \begin{bmatrix}
1 & 2 & 3 & 13 & 14 & 15 \\
4 & 5 & 6 & 16 & 17 & 18
\end{bmatrix}
\]

Unfolded matrices:
\( X_{(1)} \in R^{3 \times 4} \)
\( X_{(2)} \in R^{2 \times 6} \)
\( X_{(3)} \in R^{2 \times 6} \)

Reference: Ji Liu et.al. Tensor completion for estimating missing values in visual data. TPAMI 2012
Tensor augmentation and completion (TAC)

- Relaxed objective of TAC with regularization:

\[
\min_{\mathcal{X}, M_l} \sum_{l=1}^{n} \alpha_l \|M_l\|_* + \frac{\beta_l}{2} \|\mathcal{X}(l) - M_l\|_F^2 + R(\mathcal{X}(\Omega))
\]

s.t. \( \mathcal{X}_\Omega = T_\Omega \)

- Index of the ground truth layer

- Intermediate relaxed matrices

Solution:
- Block Coordinate Descend (BCD)
- Four blocks of variables: \( \mathcal{X}, M_1, M_2, M_3 \)
Updating $M_l$

Sub-problem:

$$\min_{M_l} : \frac{\alpha_l}{\beta_l} \|M_l\|_* + \frac{1}{2} \|\mathcal{X}(l) - M_l\|^2_F$$

Closed form solution, proved by [Cai et.al. 2009]:

$$D_\tau(\mathcal{X}(l)) = U\Sigma_\tau V^T$$

Here, $\Sigma_\tau = diag(\{\sigma_i - \tau\}_+)$ and $\tau = \frac{\alpha_l}{\beta_l}$

Updating $X$

- Two formulations:
  - Prior guided ground truth inference (PG-TAC)

$$\min_{\hat{\mathcal{X}}, \mathcal{M}_l} \sum_{l=1}^{n} \alpha_l \| M_l \|_\star + \frac{\beta_l}{2} \| \mathcal{X}(l) - M_l \|_F^2 + \frac{\gamma}{2} \| \mathcal{X}_{i_g} - S \|_F^2,$$
  
  s.t. : $\mathcal{X}_\Omega = \mathcal{T}_\Omega$

  - Slack variable

  - Relaxed simplex ground truth inference (RS-TAC)

$$\min_{\mathcal{X}, \mathcal{M}_l} \sum_{l=1}^{n} \alpha_l \| M_l \|_\star + \frac{\beta_l}{2} \| \mathcal{X}(l) - M_l \|_F^2 + \frac{\gamma}{2} \sum_{j=1}^{N_i} \xi_j^2$$

  s.t. : $\sum_{k=1}^{N_e} \mathcal{X}_{ijk} - 1 = \xi_j, \mathcal{X}_{ijk} \geq 0, i = i_g, \forall j = 1, \ldots, N_i$

  s.t. : $\mathcal{X}_\Omega = \mathcal{T}_\Omega$

Prior Statistics
Slack variable
Updating $X$

Prior guided ground truth inference (PG-TAC)

$\min_{\mathcal{X}} \sum_{l=1}^{n} \frac{\beta_l}{2} \|\mathcal{X}(l) - M_l\|_F^2 + \frac{\gamma}{2} \|\mathcal{X}_{ig}^:: - S\|_F^2$

s.t. $\mathcal{X}_\Omega = T_\Omega$

Solution:
Elements of tensor $X$ can be divided into three sets $\{C_1, C_2, C_3\}$

Elements of set $C_1$:
$\mathcal{X}_{ijk} = T_{ijk}$

$(i, j, k) \in \Omega$

Elements of set $C_2$:
$\mathcal{X}_{ijk} = \left(\frac{\sum_{l=1}^{n} \beta_l f_{old_l}(M_l)}{\sum_{l=1}^{n} \beta_l}\right)_{ijk}$

$(i, j, k) \notin \Omega$ and $i \neq i_g$

Elements of set $C_3$:
$\mathcal{X}_{ijk} = \left(\frac{\sum_{l=1}^{n} \beta_l f_{old_l}(M_l)}{\sum_{l=1}^{n} \beta_l + \gamma}\right)_{ijk} + \left(\frac{\gamma S}{\sum_{l=1}^{n} \beta_l + \gamma}\right)_{jk}$

$(i, j, k) \notin \Omega$ and $i = i_g$
Updating $X$

- **Prior guided ground truth inference (RS-TAC)**

$$\min_{\mathcal{X}, M_i} \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathcal{X}(l) - M_l||^2_F + \frac{\gamma}{2} \sum_{j=1}^{N_i} \xi_j^2$$

s.t. $\sum_{k=1}^{N_c} \mathcal{X}_{ijk} - 1 = \xi_j, \mathcal{X}_{ijk} \geq 0, i = i_g, \forall j = 1, \ldots, N_i$

$$\mathcal{X}_\Omega = \mathcal{T}_\Omega$$

- **Solution:**

Elements of tensor $X$ can be divided into three sets $\{C_1, C_2, C_3\}$

- **Elements of set $C_1$:**

  $$\mathcal{X}_{ijk} = \mathcal{T}_{ijk}$$

  $(i, j, k) \in \Omega$

- **Elements of set $C_2$:**

  $$\mathcal{X}_{ijk} = \left( \frac{\sum_{l=1}^{n} \beta_l fold_l(M_l)}{\sum_{l=1}^{n} \beta_l} \right)_{ijk}$$

  $(i, j, k) \notin \Omega$ and $i \neq i_g$

- **Elements of set $C_3$:**

  $$\mathcal{X}_{ijk} = \left( \frac{\sum_{l=1}^{n} \beta_l fold_l(M_l)}{\sum_{l=1}^{n} \beta_l} \right)_{ijk} + \frac{\gamma}{\gamma N_c + \sum_{l=1}^{n} \beta_l} \left( \frac{1 - \sum_{k=1}^{N_c} (fold_l(M_l))_{ijk}}{\sum_{l=1}^{n} \beta_l} \right)_{ijk}$$

  $(i, j, k) \notin \Omega$ and $i = i_g$
Roadmap

- Background
- Related work
- Crowdsourcing based on TAC
- Experimental results
- Conclusion
Experimental Results

Synthetic Data Set:

- **Notations:**
  - # of Workers: \( N_w \)
  - # of Items: \( N_i \)
  - # of Classes: \( N_c \)
  - Probability of not giving labels \( q \)

- **Initial configuration:**
  - \( N_w = 50, N_i = 400 \)
  - \( N_c = 4, q = 0.7 \)

- **Four configurations:**
  - (a). \( N_w \in [20, 90] \)
  - (b). \( N_c \in [2, 8] \)
  - (c). \( N_i \in [50, 1000] \)
  - (d). \( q \in [0, 0.95] \)
Experimental Results

Real-world Data Set:

<table>
<thead>
<tr>
<th>Data Set</th>
<th># classes</th>
<th># items</th>
<th># workers</th>
<th># total labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTE [Snow et al., 2008]</td>
<td>2</td>
<td>800</td>
<td>164</td>
<td>8000</td>
</tr>
<tr>
<td>Temp [Snow et al., 2008]</td>
<td>2</td>
<td>462</td>
<td>76</td>
<td>4620</td>
</tr>
<tr>
<td>Web [Zhou et al., 2012]</td>
<td>5</td>
<td>2653</td>
<td>177</td>
<td>15567</td>
</tr>
<tr>
<td>Dog [Zhou et al., 2012]</td>
<td>4</td>
<td>807</td>
<td>109</td>
<td>7354</td>
</tr>
<tr>
<td>Spam [Zhou et al., 2015]</td>
<td>2</td>
<td>149</td>
<td>18</td>
<td>1901</td>
</tr>
<tr>
<td>Age [Han et al., 2014]</td>
<td>7</td>
<td>1002</td>
<td>165</td>
<td>10020</td>
</tr>
</tbody>
</table>

Table 1: For Dog data set, the unqualified workers, who have only labeled a small amount of images, are remained. For web data set, 12 items have been removed due to lack of true labels. For age data set, data has been discretized into 7 bins: [0, 9], [10, 19], [20, 29], [30, 39], [40, 49], [50, 59], [60, 100].

References:
Hu Han, et.al. Demographic estimation from face images: Human vs. machine performance. TPAMI, 2015.
Experimental Results

Real-world Data Set results:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RTE</td>
<td>10.31</td>
<td>7.25</td>
<td><strong>6.63</strong></td>
<td>7.00</td>
<td>7.50</td>
<td>7.13</td>
<td>7.00</td>
<td>7.25</td>
</tr>
<tr>
<td>Temp</td>
<td>6.39</td>
<td>5.84</td>
<td>5.84</td>
<td>5.63</td>
<td>5.63</td>
<td>5.84</td>
<td>5.41</td>
<td>5.84</td>
</tr>
<tr>
<td>Dog</td>
<td>17.91</td>
<td>15.86</td>
<td><strong>15.74</strong></td>
<td>–</td>
<td>16.23</td>
<td>15.86</td>
<td>15.74</td>
<td>15.74</td>
</tr>
<tr>
<td>Spam</td>
<td>19.80</td>
<td>13.42</td>
<td><strong>12.75</strong></td>
<td>18.12</td>
<td><strong>12.75</strong></td>
<td>14.10</td>
<td>12.75</td>
<td>13.42</td>
</tr>
<tr>
<td>Age</td>
<td>34.88</td>
<td>39.62</td>
<td>36.33</td>
<td>35.73</td>
<td><strong>31.14</strong></td>
<td>31.24</td>
<td>32.44</td>
<td><strong>31.14</strong></td>
</tr>
</tbody>
</table>

Table 2: Comparison results of all methods on six real data sets in error rate (in percentage)

References:
Conclusion

- Two novel methods PG-TAC and RS-TAC:
  - Augment the data tensor with a ground truth layer.
  - Utilize the structural information of crowd labels.
  - Infer the true labels of items in binary and multi-class settings.

- Experimental results:
  - Six real data sets.
  - Outperform state-of-the-art methods.
Thank you!

&

Questions?