Semi-Supervised Learning from Crowds Using Deep Generative Model

Kyohei Atarashi
Hokkaido University
atarashi_k@complex.ist.hokudai.ac.jp

Satoshi Oyama
Hokkaido University/RIKEN AIP
oyama@ist.hokudai.ac.jp

Masahito Kurihara
Hokkaido University
kurihara@ist.hokudai.ac.jp

AAAI-18

Introduction

- Previous Methods: only crowd data

- A more pratical problem setting: crowd data + unlabeled data
  - Semi-supervised learning from crowds

Previous Methods

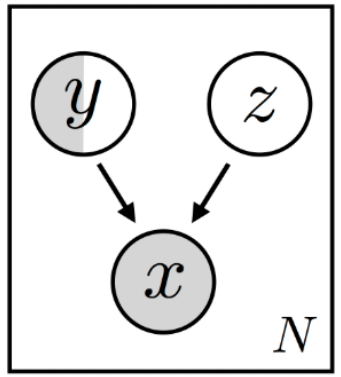
- Infer true labels
  - Crowd label -> “True” label -> Supervised classification model
  - MV, DS, ……
  - Additional information are lost.

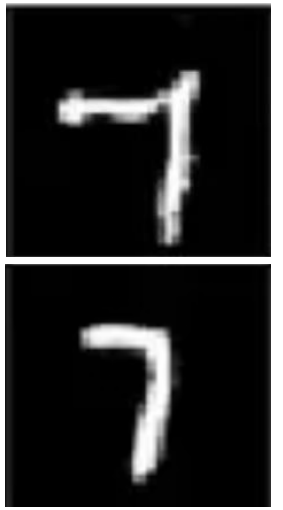
- Learning from crowds
  - Crowd label -> Classifier
  - True labels as latent variables
  - EM

- Semi-Supervised Learning
  - SSVAE-M2

SSVAE-M2

- Model
  - $p(x, y, z) = p(x|y, z)p(y)p(z)$
  - $p(y) = Cat(y|\pi)$
  - $p(z) = N(z|0, I)$
  - $p_\theta(x|z, y) = N(x|\mu(z, y), \sigma^2(z, y))$ decoder



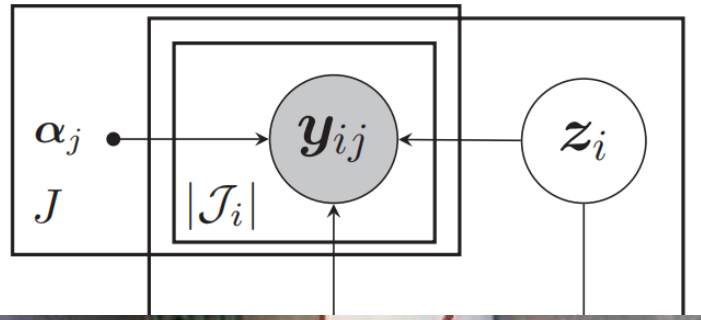


- Inference
  - $q_\phi(\mathbf{y} \mid \mathbf{x}) = \mathrm{Cat}\big(\mathbf{y} \mid \alpha_\phi(\mathbf{x})\big)$ classifier
  - $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) = \mathcal{N}\big(\mathbf{z} \mid \mu_\phi(\mathbf{x}, \mathbf{y}), \sigma^2_\phi(\mathbf{x}, \mathbf{y})\big)$ encoder

- Loss
  - $\mathcal{J}^\alpha = \mathcal{J} + \alpha \mathop{\mathbb{E}}_{\tilde{p}_l(\mathbf{x},\mathbf{y})}\big[-\log\big(q_\phi(\mathbf{y} \mid \mathbf{x})\big)\big] = \mathcal{L}(\mathbf{x}, \mathbf{y}) + \mathcal{U}(\mathbf{x}) + \alpha \mathop{\mathbb{E}}_{\tilde{p}_l(\mathbf{x},\mathbf{y})}\big[-\log\big(q_\phi(\mathbf{y} \mid \mathbf{x})\big)\big]$

Model

$$
\begin{aligned}
&p(\mathcal{X}_c, \mathcal{Y}_c, \mathcal{T}_c, \mathcal{Z}_c, \mathcal{X}_u, \mathcal{T}_u, \mathcal{Z}_u) \\
&= p(\mathcal{X}_c, \mathcal{Y}_c, \mathcal{T}_c, \mathcal{Z}_c) p(\mathcal{X}_u, \mathcal{T}_u, \mathcal{Z}_u) \\
&= p(\mathcal{X}_c \mid \mathcal{T}_c, \mathcal{Z}_c) p(\mathcal{Y}_c \mid \mathcal{T}_c, \mathcal{Z}_c) p(\mathcal{T}_c) p(\mathcal{Z}_c) \\
&\times p(\mathcal{X}_u \mid \mathcal{T}_u, \mathcal{Z}_u) p(\mathcal{T}_u) p(\mathcal{Z}_u) \\
&= \prod_{i \in \Lambda_c} \prod_{j \in \mathcal{J}_i} p(\boldsymbol{x}_i \mid \boldsymbol{t}_i, \boldsymbol{z}_i) p(\boldsymbol{y}_{ij} \mid \boldsymbol{t}_i, \boldsymbol{z}_i) p(\boldsymbol{t}_i) p(\boldsymbol{z}_i) \\
&\times \prod_{i \in \Lambda_u} p(\boldsymbol{x}_i \mid \boldsymbol{t}_i, \boldsymbol{z}_i) p(\boldsymbol{t}_i) p(\boldsymbol{z}_i),
\end{aligned}
$$



f label-

$$
\begin{aligned}
&p(\boldsymbol{t}_i) = \mathrm{Cat}(\boldsymbol{t}_i \mid \boldsymbol{\pi}) \\
&p(\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{z}_i \mid \mathbf{0}, \mathbf{I}), \\
&p(\boldsymbol{x}_i \mid \boldsymbol{t}_i, \boldsymbol{z}_i) = \mathcal{N}\big(\boldsymbol{x}_i \mid \mu_{\boldsymbol{\theta}}(\boldsymbol{t}_i, \boldsymbol{z}_i), \mathrm{diag}(\sigma^2_{\boldsymbol{\theta}}(\boldsymbol{t}_i, \boldsymbol{z}_i))\big) \\
&p(\boldsymbol{y}_{ij} \mid \boldsymbol{t}_i, \boldsymbol{z}_i) = \mathrm{Cat}(\boldsymbol{y}_{ij} \mid f_{\boldsymbol{\alpha}_j}(\boldsymbol{t}_i, \boldsymbol{z}_i))
\end{aligned}
$$

decoder

Optimization and Inference

- Computing the posteriors of ***z*** and ***t*** are intractable, cannot use EM
- Cannot get a classifier directly

$$q(\boldsymbol{t}, \boldsymbol{z} \mid \boldsymbol{x}) = q(\boldsymbol{t} \mid \boldsymbol{x}) q(\boldsymbol{z} \mid \boldsymbol{t}, \boldsymbol{x})$$

classifier encoder

- The marginal log-likelihood

$$\log p(\mathcal{X}_c, \mathcal{Y}_c, \mathcal{X}_u) = \log p(\mathcal{X}_c, \mathcal{Y}_c) + \log p(\mathcal{X}_u)$$

Unlabeled data

$$\log p(\mathcal{X}_u) = \log \int \int p(\mathcal{X}_u, \mathcal{T}_u, \mathcal{Z}_u) d\boldsymbol{t} d\boldsymbol{z}$$

$$\geq \sum_{i \in \Lambda_u} \mathbb{E}_{q(\boldsymbol{t}_i, \boldsymbol{z}_i \mid \boldsymbol{x}_i)} \left[\log p(\boldsymbol{x}_i \mid \boldsymbol{t}_i, \boldsymbol{z}_i) \right]$$

$$- \sum_{i \in \Lambda_u} \mathrm{KL}[q(\boldsymbol{t}_i, \boldsymbol{z}_i \mid \boldsymbol{x}_i) || p(\boldsymbol{t}_i, \boldsymbol{z}_i)]$$

$$:= -\mathcal{U}(\mathcal{X}_u).$$

Labeled data

$$\log p(\mathcal{X}_c, \mathcal{Y}_c) = \log \int \int p(\mathcal{X}_c, \mathcal{Y}_c, \mathcal{T}_c, \mathcal{Z}_c) d\boldsymbol{t} d\boldsymbol{z}$$

$$\geq \mathbb{E}_{q(\mathcal{T}_c, \mathcal{Z}_c \mid \mathcal{X}_c)} \left[\log \frac{p(\mathcal{X}_c, \mathcal{Y}_c, \mathcal{T}_c, \mathcal{Z}_c)}{q(\mathcal{T}_c, \mathcal{Z}_c \mid \mathcal{X}_c)} \right]$$

$$= \sum_{i \in \Lambda_c} \sum_{j \in \mathcal{J}_i} \mathbb{E}_{q(\boldsymbol{t}_i, \boldsymbol{z}_i \mid \boldsymbol{x}_i)} \left[\log p(\boldsymbol{x}_i, \boldsymbol{y}_{ij} \mid \boldsymbol{t}_i, \boldsymbol{z}_i) \right]$$

$$- \sum_{i \in \Lambda_c} \mathrm{KL}[q(\boldsymbol{t}_i, \boldsymbol{z}_i \mid \boldsymbol{x}_i) || p(\boldsymbol{t}_i, \boldsymbol{z}_i)]$$

$$:= -\mathcal{C}(\mathcal{X}_c, \mathcal{Y}_c),$$

Optimization and Inference

- Pseudo classification loss

$$-\sum_{i\in\Lambda_c}\log(q(\bar{\boldsymbol{y}}_i \mid \boldsymbol{x}_i))$$
$$\bar{y}_i = \frac{1}{|\mathcal{J}_i|}\sum_{j\in\mathcal{J}_i}\boldsymbol{y}_{ij}$$

- The prior of $\boldsymbol{t}$ for crowdsouced data

$$p(\boldsymbol{t}_i) = \mathrm{Cat}(\boldsymbol{t}_i \mid \bar{y}_i)$$

- Final objective function

$$\mathcal{J} = \mathcal{C}(\mathcal{X}_c, \mathcal{Y}_c) + \mathcal{U}(\mathcal{X}_u) - \alpha \cdot \sum_{i\in\Lambda_c}\log(q(\bar{\boldsymbol{y}}_i \mid \boldsymbol{x}_i))$$

Experiments

Table 1: Experimental results for the MNIST dataset.

	PC labeling	DS labeling
Proposed	**0.9646**	**0.8647**
MV-LR	0.7162	0.6705
MV-MLP	0.7128	0.6547
PC	0.7756	0.6063
PC-GP	0.7754	0.6055
M2	0.9421	0.8445

Table 2: Experimental results for the Rotten Tomatoes dataset. The accuracies of the MVhard (this model is same as the MV-LR), the MVsoft, the Raykar, the Rayker w/prior and the MA-LR are taken from Rodrigues, Pereira, and Ribeiro (2013).

Model	Accuracy	Model	Accuracy
Proposed	**0.7290**	MVhard	0.7027
MV-MLP	0.6979	MVsoft	0.7165
PC	0.7261	Raykar	0.4867
PC-GP	0.7263	Raykar w/prior	0.7078
M2	0.7096	MA-LR	0.7240

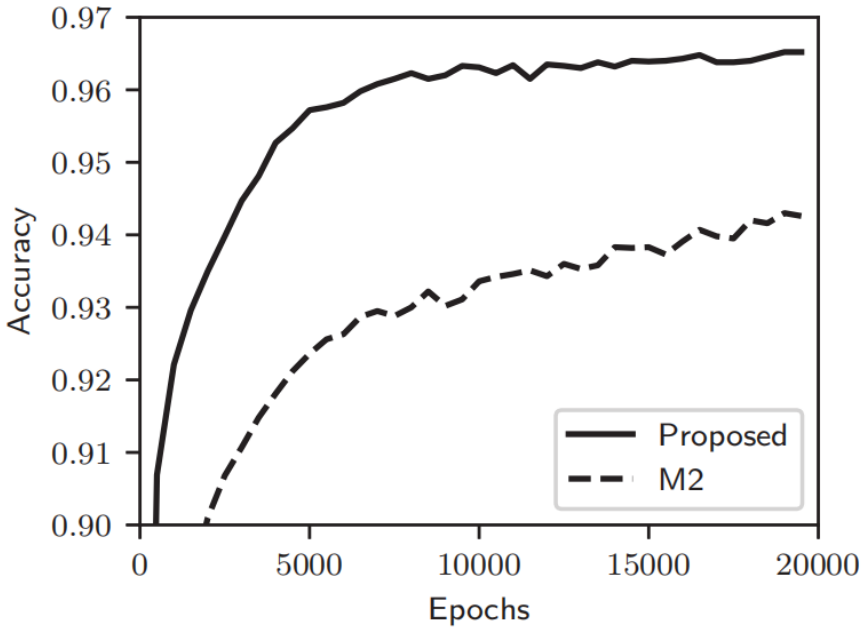


Figure 3: Accuracy for the PC labeling evaluated every 100 epochs between the proposed model and the M2 model.

Experiments

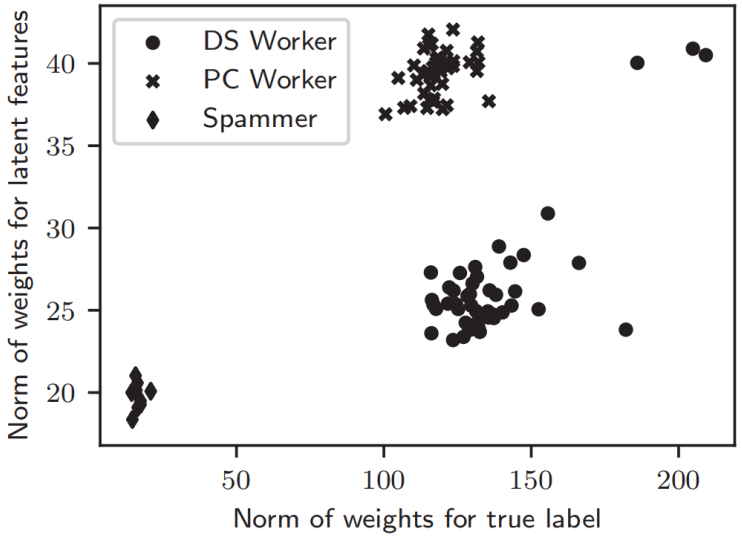


Figure 4: Estimated $\|\boldsymbol{\alpha}^{(t)}\|_1$ and $\|\boldsymbol{\alpha}^{(z)}\|_1$ of 45 DS workers, 45 PC workers and 10 spammers.

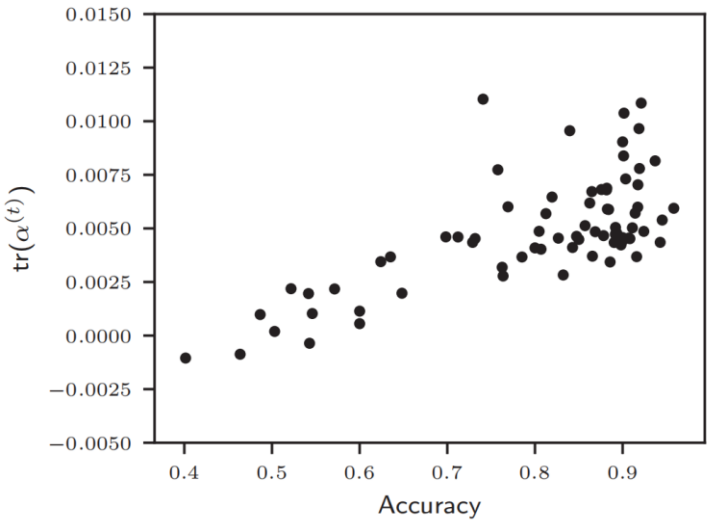


Figure 5: Estimated $\mathrm{diag}(\boldsymbol{\alpha}^{(t)})$ and the accuracy of each worker who labels more than 100 data.