Semi-crowdsourced Clustering with Deep Generative Models

Yuen Luo, Tian Tian, Jiaxin Shi, Jun Zhu, Bo Zhang
Department of Computer Science and Technology, Tsinghua University, Beijing, China
luoye15@mails.tsinghua.edu.cn. Code available at https://github.com/xinmei9322/semicrowd

1. Learning from crowds
   - Distribute micro-tasks to web workers in parallel, fast with relatively low cost
   - Comparing pairs is easier for non-experts → pairwise constraints

2. (Semi-) crowd clustering
   - Bayesian clustering [Gomes et al., NIPS 2011]
     → Cost grows quadratically as N grows. Not scalable!
   - SemiCrowd [Yi et al., NIPS 2012]
     → Linear similarity function, ignores the noise and inter-worker variations
   - Multiple Clustering Views from Multiple Uncertain Experts [Chang et al., ICML 2017]
     → Discriminative clustering, does not use the information in unlabeled samples

3. Guide the learning of DGMs with statistical relational models
   - DGM: Raw data observations $x_n$ corresponding latent variable $k_n$, cluster index $z_n$
     $\mu(Z | \pi) = \prod_{n=1}^N \mathcal{N}(x_n | \mu_k, \Sigma_k)$, $p(X|Z, \mu, \Sigma) = \prod_{n=1}^N \mathcal{N}(x_n | \mu_k, \Sigma_k)$
   - Relational model: M workers, accuracy parameters: sensitivity $\alpha$ and specificity $\beta$
     $\rho(x_n | k_n) = \text{Ber}_n^{(\alpha)}(x_n, k_n^{(\alpha)}(1 - \rho^{-1}(x_n)))$
   - Variational message passing for conjugate structures and amortized learning of deep components

4. Simple version: Amortized inference
   - Let $\Theta = \{\pi, \mu, \Sigma, \alpha, \beta\}$, maximize the variational lower bound $\mathcal{L}$
     $\log p(D, \Theta) \geq \mathbb{E}_{q(Z)}[\log p(Z, X, D, \Theta, \gamma)] - \log q(Z, X|D) = \mathcal{L}(D, \Theta, \gamma)$
   - Inference networks
     $q(z_n | x_n, \phi) = \text{Cat}(z_n | \pi(x_n, \phi))$
     $q(x_n | z_n, \phi) = \mathcal{N}(x_n | \mu_k, \Sigma_k)$
   - Analytically sum over the discrete $z_n$, use the parameterization trick for $u_n$

5. Natural gradient stochastic variational inference
   - Local partial optimizers for $\Theta$: $q(\Theta) = \prod_{k=1}^K q(\Theta_k)$
     $\nabla_{\Theta_k} \mathcal{L}(D, \Theta, \gamma) = \mathbb{E}_{q(Z)}[\nabla_{\Theta_k} \mathbb{E}[\log p(D | X, Z, \Theta)]]$
     $\mathbb{E}_{q(Z)}[\nabla_{\Theta_k} \log q(Z | D, X, \Theta)]$
     where $\nabla_{\Theta_k} \mathbb{E}_{q(Z)}$ = $\sum_{n=1}^N \nabla_{\Theta_k} \sum_{z_n} q(z_n | x_n) q(x_n | z_n) q(z_n | x_n, \phi)$
   - Final objective: $J(D, \Theta, \gamma)$ = $\mathcal{L}(D, \Theta, \gamma) - \mathbb{E}_{q(Z)}(\log q(Z | D, X, \Theta))$
   - Update the global variational parameters $\Theta_k$ by natural gradients
   - For other parameters $\phi, \gamma$, compute the gradients $\nabla_{\phi, \gamma} J(D, \Theta, \gamma)$ and $\nabla_{\phi, \gamma} \mathcal{L}(D, \Theta, \gamma)$

6. Outperforms competing methods
   - Face dataset, 640 images from 20 people with different poses (straight, left, right, up).

7. Crowdsourced real annotations from Amazon Mechanical Turks
   - Method       Accuracy    NMI    Time
     SCOC       0.97 ± 0.048  0.17 ± 0.025  20.8s
     BaselineDC  0.97 ± 0.048  0.17 ± 0.025  20.8s
     SemiCrowd  0.97 ± 0.048  0.17 ± 0.025  20.8s
     SCOC       0.97 ± 0.048  0.17 ± 0.025  20.8s
     BaselineDC  0.97 ± 0.048  0.17 ± 0.025  20.8s
     SemiCrowd  0.97 ± 0.048  0.17 ± 0.025  20.8s

Figure 1: Schematic of Bayesian crowdclustering (from Gomes et al., NIPS 2011)

Figure 2: (Semi-) crowd clustering

Figure 3: MNIST visualization of generated samples of 99 clusters during training SemiCrowd. Each column represents a cluster, whose infected proportion ($x_n$) is reflected by brightness.

Figure 4: Clustering results as CIFAR-10: (left) unsupervised; (right) with noisy annotations.