HAEM: Obtaining Higher-Quality Classification Task Results with AI Co-Workers

Yu Yamashita,[1] **Hiroyoshi Ito,** [2] **Kei Wakabayashi,** [2] **Masaki Kobayashi,** [3] **Atsuyuki Morishima** [4]

University of Tsukuba, Japan

[1] yu.yamashita.2021b@mlab.info, [2] {ito, kwakaba}@slis.tsukuba.ac.jp, [3] makky@klis.tsukuba.ac.jp,

[4] morishima-office@ml.cc.tsukuba.ac.jp

Abstract

Obtaining high-quality results for a fixed set of classification tasks with a limited budget is a critical issue in crowdsourcing. The introduction of AI models to complement the process should be explored. However, there are few existing approaches to directly address the problem; existing approaches have been proposed in the context of how to train AI models using noisy crowdsourced data. This paper presents a more direct approach for solving the problem of introducing AI to improve the results of human workers for a fixed number of tasks with a limited budget; we deal with an AI model as a co-worker and aggregates the results of both human and AI workers. The proposed "Human-AI EM" (HAEM) algorithm, which extends the Dawid-Skene model, deals with AI models as co-workers, and explicitly computes their confusion matrices to derive higher-quality aggregation results. We conducted an extensive set of experiments and compared HAEM with two methods (MBEM and Dawid-Skene model). We found that AI-powered HAEM shows better performance than the Dawid-Skene Model in most cases and that it shows better performance than MBEM when the AI model does not show very good performance.

Introduction

Obtaining high-quality results for a fixed set of classification tasks with a limited budget is a critical issue in crowdsourcing because they can be malicious or low-skilled, and required in a wide range of applications. A mainstream approach for obtaining high-quality classification labels from non-expert workers is to apply aggregation techniques (Dawid and Skene 1979). However, such techniques are often difficult to apply with a low budget, since they need to assign duplicate tasks to many human workers.

Although an approach worth exploring is the introduction of AI models to complement the process, few existing approaches directly address the problem; existing approaches have been proposed in the context of how to train AI models using noisy crowdsourced data. For example, attempts have been made to obtain better training data by using AI predictions for adaptively aggregating crowd worker results (Khetan, Lipton, and Anandkumar 2017).

This paper presents a more direct approach for solving the problem of introducing AI to improve the results of human workers for a fixed number of tasks with a limited budget (Figure 1). We introduce an AI model that serves as a co-worker, assign adequate tasks to human workers as long as the budget allows, and aggregate the results of both human and AI workers. We propose "Human-AI EM" (HAEM) algorithm, which extends the Dawid-Skene model (Dawid and Skene 1979), deals with AI models as co-workers and explicitly computes their skills to derive higher-quality aggregation results.

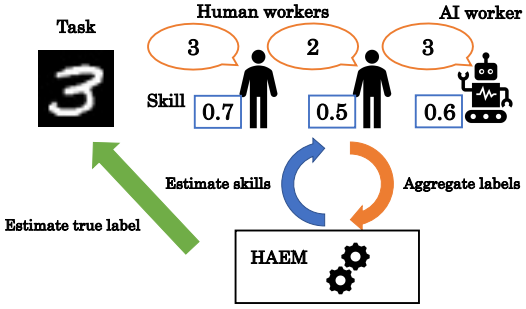

Figure 1: Our proposed method: HAEM

The contributions of this study are as follows. First, this paper applies a novel AI-powered approach to the practical problem. Second, we present HAEM, which is an EM algorithm based on a natural extension of the Dawid-Skene model that deals with relevant variables; it incorporates AI's task result prediction and task features as observed variables and the AI's skill and the AI parameter as latent ones. Third, we conducted an extensive set of experiments with two datasets, two data sizes, two worker models, and two AI models, and compared HAEM with the Dawid-Skene model with only human workers and the state-of-the-art model (Khetan, Lipton, and Anandkumar 2017) for training AI with noisy human labels (MBEM). We found that AI-powered HAEM shows better performance than the Dawid-Skene Model in most cases and that it shows better performance than MBEM when the AI model does not show very good performance. This justifies our approach that deals with human and AI workers.

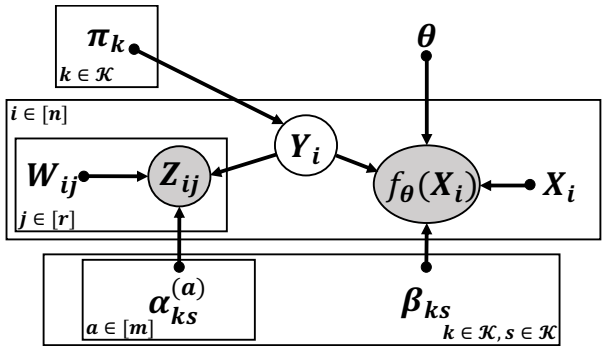

Figure 2: Graphical model of HAEM. Shaded circles indicate the observed random variables, white circles indicate the latent random variables, and small black circles indicate the observed variables. $\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta, \boldsymbol{\pi}$ are the parameters of HAEM.

Proposed Method: HAEM

Input and Output. We are given m human workers, a set of n tasks, a set $\mathcal{K}$ of classification labels, and the task results given by the workers, denoted by Z_{ij}. Here, $Z_{ij} \in \mathcal{K}$ is the label that a worker provided for task i in j-th redundancy. We require that each task $X_{i\in[n]}$ be completed by at least two distinct workers. Then, we want to accurately estimate the true label $\{Y_i\}_{i\in[n]}$ for each task. We assume that we have an ML-based AI worker behind this problem that works as another human worker.

Our Model. The graphical model of HAEM is shown in Figure 2. We estimate the hidden true labels of each task and the parameters of the model based on the EM algorithm. In this algorithm, we maximize the probability of observing the labels of human workers $\mathbf{Z} = \{Z_{ij}\}_{i\in[n],j\in[r]}$, and predictions of AI workers $f_{\boldsymbol{\theta}}(\mathbf{X}) = \{f_{\boldsymbol{\theta}}(X_i)\}_{i\in[n]}$, while estimating the distribution of hidden true labels $\mathbf{Y} = \{Y_i\}_{i\in[n]}$, and the model parameters $\boldsymbol{\Theta} := \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta, \boldsymbol{\pi}\}$, where $\boldsymbol{\alpha} = \{\alpha_{ks}^{(a)}\}_{k\in\mathcal{K},s\in\mathcal{K},a\in[m]}$ is the confusion matrices of human workers, which is the probability that the a-th human worker provides label $s \in \mathcal{K}$ for the task ground truth label $k \in \mathcal{K}$, $\boldsymbol{\beta} = \{\beta_{ks}\}_{k\in\mathcal{K},s\in\mathcal{K}}$ is the confusion matrix of AI worker, $\boldsymbol{\theta}$ is parameters of the AI worker, and $\boldsymbol{\pi}$ is the marginal distribution of the labels.

In the proposed algorithm, we iteratively estimate the probability distribution of the hidden true labels $q(\mathbf{Y})$ and the model parameters $\boldsymbol{\Theta}$, which maximizes the log-likelihood of the observed random variables $\mathbf{Z}$ and $f_{\boldsymbol{\theta}}(\mathbf{X})$. Estimating the optimal proposed distribution $q(\mathbf{Y})$ is called E-step, and estimating the optimal parameters $\boldsymbol{\Theta}$ is called M-step.

Procedure. First, we initialize the posterior distribution of the true labels using weighted majority vote. We set posterior distribution of the hidden true labels $q(\mathbf{Y})$ using weighted majority vote to train the AI worker and estimate $\widehat{\boldsymbol{\theta}}$, let the AI worker predict on all samples, and initialize the posterior distribution of the true labels using weighted majority vote from human workers' label and prediction of the AI worker.

In M-step, we train the AI worker and estimate $\widehat{\boldsymbol{\theta}}$, let the AI worker predict on all samples, estimate the AI worker confusion matrix $\widehat{\boldsymbol{\beta}}$, estimate human workers confusion matrices $\widehat{\boldsymbol{\alpha}}$, and estimate marginal distribution $\widehat{\boldsymbol{\pi}}$. In E-step, we estimate the distribution of the hidden true labels $q(\mathbf{Y})$. We repeat two steps (M-step and E-step) for T times.

Table 1: The accuracies of EM, MBEM(ResNet), HAEM(ResNet), MBEM(AlexNet), and HAEM(AlexNet).

Task	CIFAR-10		Tiny ImageNet
Size	50K	5K	5K
Worker	Synthesized		Real
EM	0.543	0.625	0.732
MBEM(ResNet)	0.614	0.636	0.7088
HAEM(ResNet)	**0.628**	**0.653**	**0.7376**
MBEM(AlexNet)	**0.843**	0.598	0.7176
HAEM(AlexNet)	0.815	**0.801**	**0.7492**

Experiment

Tasks, Workers, and AI Workers. We used CIFAR-10 (Krizhevsky, Hinton et al. 2009) and Tiny ImageNet (Le and Yang 2015) as tasks. It has large data (a total of $n = 50K$ images) and small data ($n = 5K$) in CIFAR-10, and has only small data ($n = 5K$) in Tiny ImageNet. All images belong to 10 classes. We used two worker models: (1) Synthesized worker model and (2) Real worker model taken from Amazon Mechanical Turk (AMT) as workers. In synthesized worker model, each worker is either a hammer (always correct) with probability 0.6 or a spammer (chooses labels uniformly at random). We assign each task to $r = 2$ workers and set $m = 200$. We emplpy ResNet-18 (He et al. 2016) and AlexNet (Krizhevsky, Sutskever, and Hinton 2012) as AI workers. We set $T = 2$.

Result. We compare our method with the Dawid-Skene model with only human workers and the state-of-the-art model (Khetan, Lipton, and Anandkumar 2017) for training AI with noisy human labels (MBEM) in several settings. In Table 1, we plotted accuracies. All accuracies show the average for three times. We found that AI-powered HAEM shows better performance than the Dawid-Skene Model in most cases and that it shows better performance than MBEM when the AI model does not show very good performance.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR16E3 including AIP challenge, Japan.

References

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28(1): 20–28.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, 770–778.

Khetan, A.; Lipton, Z. C.; and Anandkumar, A. 2017. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577* .

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images .

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25: 1097–1105.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N* 7: 7.