Appendix to the paper "Aggregating Labels from Humans and AIs with Asymmetric Performance"

TAKUMI TAMURA, University of Tsukuba, Japan
HIROYOSHI ITO, University of Tsukuba, Japan
SATOSHI OYAMA, Nagoya City University, Japan
ATSUYUKI MORISHIMA, University of Tsukuba, Japan

A Confusion matrix generation process for synthetic AI workers

First, we define what we term a *symmetric confusion matrix*. For an n-class classification task (where $n \geq 3$), this is an $n \times n$ matrix A:

$$A = \begin{bmatrix} h & l & \dots & l \\ l & h & \dots & l \\ \vdots & \vdots & \ddots & \vdots \\ l & l & \dots & h \end{bmatrix} \tag{1}$$

Here, each row sums to 1, and $h > l$. The off-diagonal elements are therefore defined as:

$$l = \frac{1-h}{n-1} \tag{2}$$

In other words, this is a confusion matrix where all diagonal elements are h and all other elements are l.

Next, we model the AI's *asymmetric performance* with a matrix such as the following:

$$B^{(1\to 2)} = \begin{bmatrix} l & h & \dots & l \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix} \tag{3}$$

This matrix is identical to matrix A from the second row onwards. In the first row, however, the diagonal element is $B_{11}^{(1\to 2)} = l$, and the element $B_{12}^{(1\to 2)}$ is set to h. This represents a bias where the AI systematically misclassifies tasks from class 1 as class 2. Similarly, we also use matrices like $B^{(1\to 3)}$ and $B^{(1\to 4)}$ in our experiments, which represent biases where class 1 is misclassified as class 3 or 4, respectively.

We base our problem setting on a baseline scenario where **the overall performance of a biased AI is equivalent to that of an average human worker**. If an AI worker is represented by a biased confusion matrix $B^{(1\to m)}$ (where m is an integer such that $1 < m \leq n$), its expected overall accuracy, a_{AI}, is given by the following equation, assuming a uniform class distribution[1]:

$$a_{AI} = \frac{l + (n-1)h}{n} \tag{4}$$

[1]Although our `Adult` dataset has a non-uniform class distribution, we used the same equation as assuming a uniform class distribution for `Adult` . This is a practical and common assumption, as the true distribution is typically unknown in a general aggregation setting.

Authors' Contact Information: Takumi Tamura, tamura.takumi.ap@alumni.tsukuba.ac.jp, University of Tsukuba, Tsukuba, Japan; Hiroyoshi Ito, ito@slis.tsukuba.ac.jp, University of Tsukuba, Tsukuba, Japan; Satoshi Oyama, oyama@ds.nagoya-cu.ac.jp, Nagoya City University, Nagoya, Japan; Atsuyuki Morishima, mori@slis.tsukuba.ac.jp, University of Tsukuba, Tsukuba, Japan.

Solving this equation for h yields:

$$h = \frac{n(n-1)a_{AI} - 1}{n(n-2)} \tag{5}$$

This equation for h is a linear function of a_{AI}. Given the constraint $\frac{1}{n} < h \leq 1$, a_{AI} must satisfy the condition:

$$\frac{1}{n} < a_{AI} \leq \frac{n-1}{n} \tag{6}$$

Let μ and σ be the mean and standard deviation of the accuracy of human workers in a given dataset. These two values are calculated directly from the human n dataset (after adjusting for redundancy r). We determine the value of h for the AI's confusion matrix according to the following conditions:

(a) Low Performance $a_{AI} = \mu - \sigma$
(b) Equal Performance $a_{AI} = \mu$
(c) High Performance $a_{AI} = \mu + \sigma$
(c') Max Performance $a_{AI} = \frac{n-1}{n}$

Condition (c') is used as a substitute for condition (c), as (c) frequently fails to satisfy the constraint in Equation (6). Condition (c') represents the theoretical upper bound for a_{AI}. When a confusion matrix is generated based on this condition, it results in $h = 1$ (and $l = 0$).

B Prior Distributions for Bayesian models

Aggregation algorithms based on approximate Bayesian methods require specifying hyperparameters for their prior distributions. This file provides supplementary information on the prior distributions used in this paper and the effect of changes in these priors on the experimental results. In particular, this section explains the prior distributions for BDS and HS-DS. Note that for CBCC, we used the prior distributions as specified in the original implementation from the prior study.

B.1 Prior Distributions for BDS and HS-DS

In this section, we adopt the notation from Liu & Wang [1], and let n to the number of the classes. In BDS (HybridConfusion), we have to set two hyperparametor (prior distiributions) $\boldsymbol{\alpha}$ and Λ.

B.1.1 Class prior. $\boldsymbol{\alpha}$ is the prior distribution for the class distribution $\boldsymbol{\rho}$, and is a vector of length n. Liu & Wang set this parameter as follows:

$$\boldsymbol{\alpha} = [1, 1, ..., 1] \tag{7}$$

This implies the assumption of a uniform class distribution. Since the true class distribution is generally unknown in label aggregation problems, this setting is considered a universal and standard approach. We adopt this value for both BDS and HS-DS in this paper.

B.1.2 Confusion matrix prior. On the other hand, Λ represents the prior distribution for the confusion matrices. Lin Wang point out that the optimal setting is dataset-dependent and use several parameters in their experiments. We introduce a variable τ and set Λ as follows:

$$\Lambda = \begin{bmatrix} \frac{\tau(n-1)}{1-\tau} & 1 & \dots & 1 \\ 1 & \frac{\tau(n-1)}{1-\tau} & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & \frac{\tau(n-1)}{1-\tau} \end{bmatrix}_{n \times n} \tag{8}$$

This implies the assumption of a worker with an average accuracy of τ. Liu & Wang define Λ using a parameter λ; in our formulation, this is equivalent to $\lambda = \frac{n\tau-1}{1-\tau}$.

In our paper, we used $\tau = 0.75$. This corresponds to $\lambda = 3n - 4$ (note that inference can be run with different values of τ by changing the argument `init_worker_accuracy` in the implementation available in the repository).

B.1.3 The effect of changes in τ. While it is expected that changes in the value of τ will have some influence on the experimental results, it was not feasible to run numerous patterns, considering the time required for the experiments (our complete set of experiments took over two weeks to finish). Note that previuos benchmark studies, such as Zheng et al. [3] and Paun et al. [2], also did not investigate multiple patterns for the prior distributions.

Therefore, we provide a case study of the effect of varying τ in a specific case as supplementary information.

Figure 1 shows the results for BDS and HS-DS on the Dog dataset, with $r = 5$ and $a_{AI} = \mu$ as our experiment 1, for different values of τ.

The results show that the experimental outcomes follow a similar trend regardless of the value of τ (although an extremely high value, such as $\tau = 0.95$, causes a decrease in performance in the homo scenario).

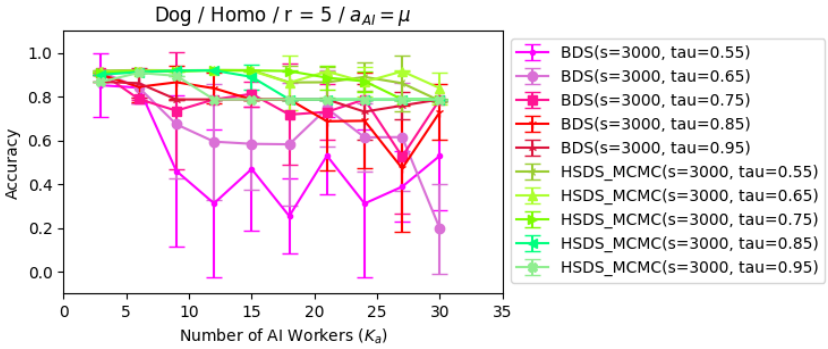
This result can be reproduced by changing `init_worker_accuracy` and running the part of the main experiment.

References

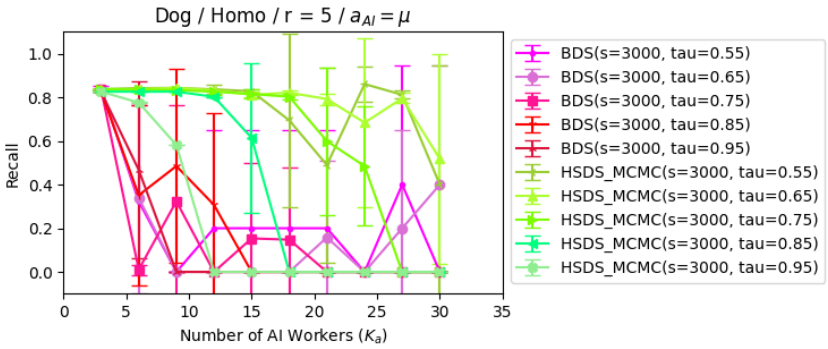
[1] Chao Liu and Yi-Min Wang. 2012. TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*. 17–24.

[2] Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics* 6 (2018), 571–585. doi:10.1162/tacl_a_00040

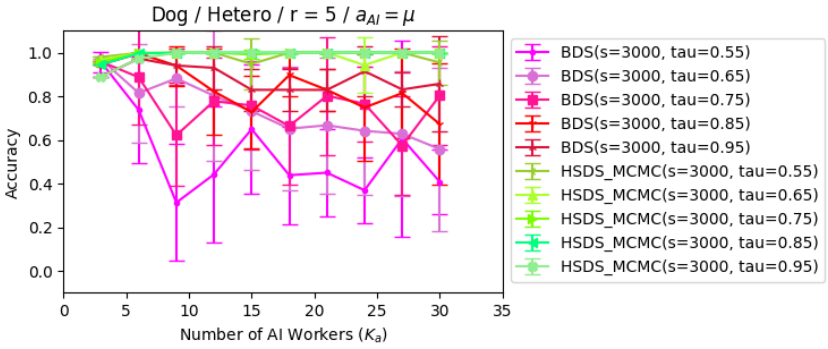
[3] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552. doi:10.14778/3055540.3055547



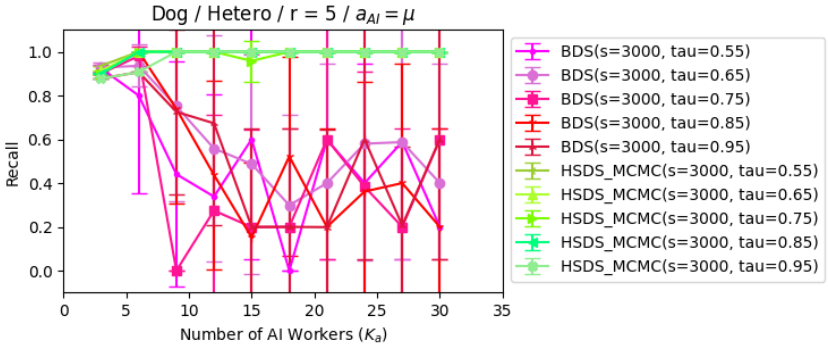
(a) Accuracy changes in the homo scenario



(b) Recall changes in the homo scenario



(c) Accuracy changes in the hetero scenario



(d) Recall changes in the hetero scenario

Fig. 1. The effect of changing τ in the setting of our experiment 1