Label Aggregation of Composite Crowd Tasks by Worker Ability Constraint Satisfaction

Jiyi Li

University of Yamanashi, Kofu, Japan
garfieldpigljy@gmail.com

Abstract

Quality control is a crucial issue of label data collection by crowdsourcing. Typically, aggregation methods to redundant crowd labels are proposed for estimating high-quality labels from noisy crowd labels. Most of the existing works concentrate on the label aggregation for Single Crowd Tasks (SCTs) which have a single object set with homogeneous question types. However, it is useful for a requester to combine multiple relevant but different crowd tasks into a Composite Crowd Task (CCT) which have heterogeneous question types and (or) multiple object sets for diverse purposes. Instead of the label aggregation on each crowd task respectively, label aggregation methods by bridging multiple SCTs in CCTs can potentially improve the label quality of all tasks. In this paper, we propose a general label aggregation approach for such CCTs by worker ability constraint satisfaction and relaxed optimization. We collected real crowd datasets of CCTs with diverse task settings based on heterogeneous question types, including categorization, pairwise preference comparisons, and pairwise similarity comparisons. The results demonstrate that our approach can effectively bridge the worker information of CCTs to improve the quality of aggregated labels and outperforms the baselines proposed for SCTs.

Introduction

Crowdsourcing has been widely used for collecting diverse label data. The quality of crowd labels is critical due to the ability or diligence of the crowd workers. One method for improving label quality is label aggregation, which is proposed for estimating high-quality labels from noisy crowd labels. Most of the existing works on this topic concentrate on the homogeneous types of questions for a *single* set of objects. For example, for some classification tasks, workers are asked to provide categorical labels to the objects (Snow et al. 2008; Whitehill et al. 2009; Venanzi et al. 2014; Li and de Rijke 2023; Li, Jiang, and Xue 2023; Yang et al. 2024), e.g., which category should be assigned to an object. For some ranking tasks, workers are asked to provide pairwise preference labels, e.g., whether one object is preferred to another (Bradley and Terry 1952; Chen et al. 2013; Jin et al. 2020; Liu, Fang, and Lu 2023; Ferrara et al. 2024). For some clustering and representation learning tasks, workers are asked to provide pairwise similarity labels, i.e., whether two objects are similar or not (Gomes et al. 2011; Yi et al. 2012; Nguyen, Ibrahim, and Fu 2023; Ariu et al. 2024). There are also methods for triplet similarity labels (van der Maaten and Weinberger 2012), numerical data (Li et al. 2014), complex data such as bounding boxes and taxonomy paths (Meir et al. 2024), and so on. When deploying crowd tasks on the web and gathering labels, typically, a single type of question is employed, and only one type of label is collected. In this study, we call such crowdsourcing tasks as Single Crowd Tasks (SCTs).

Object Set	Question Type: Homogeneous $\{\mathcal{Q}_1\}$	Question Type: Heterogeneous $\{\mathcal{Q}_1, \mathcal{Q}_2\}$
Single $\{\mathcal{O}_1\}$	$\{\mathcal{O}_1, \mathcal{Q}_1\}$ (SCT)	① $\{\{\mathcal{O}_1, \mathcal{Q}_1\}, \{\mathcal{O}_1, \mathcal{Q}_2\}\}$ (Our, CCT)
Multiple $\{\mathcal{O}_1, \mathcal{O}_2\}$	② $\{\{\mathcal{O}_1, \mathcal{Q}_1\}, \{\mathcal{O}_2, \mathcal{Q}_1\}\}$ (Our, CCT)	③ $\{\{\mathcal{O}_1, \mathcal{Q}_1\}, \{\mathcal{O}_2, \mathcal{Q}_2\}\}$ (Our, CCT)

Table 1: Settings of our CCTs. Existing works of SCTs focus on homogeneous questions to single object sets. We concentrate on the cases with multiple object sets or (and) heterogeneous question types.

However, in real applications, it is useful for a requester to combine multiple relevant but different crowd tasks into a Composite Crowd Task (CCT) which have heterogeneous question types and (or) multiple object sets for diverse purposes. We use the categorization task, ranking tasks with pairwise preference comparisons, and clustering tasks with pairwise similarity comparisons to raise three examples. First, Case ① in Table 1 consists of two tasks with two different question types, $\mathcal{Q}_1$ and $\mathcal{Q}_2$, for the same object set $\mathcal{O}_1$. Table 2 provides an example with the objects, two heterogeneous types of pairwise comparison questions and labels, and aggregated results. Given a pair of opinions from a set of crowd opinions $\mathcal{O}_1$ collected to solve an issue, a worker can answer two heterogeneous types of questions: "Which of the two opinions do you prefer?" ($\mathcal{Q}_1$: pairwise preference) and "Are two opinions similar or not?" ($\mathcal{Q}_2$: pairwise similarity), rather than only answering single question type. During label aggregation, one question type and its labels can provide auxiliary information to another question type and its labels.

Second, Case ② in Table 1 consists of two tasks with the same question type, $\mathcal{Q}_1$, for two different object sets, $\mathcal{O}_1$

Objects $\mathcal{O}_1$:: Crowd Opinions for Solving an Issue. Issue: How to support foreign tourists with language barrier? o_1: Write signs with their explanations in English; o_2: Make signs into easily understood logos for foreigners; o_3: Develop OCR and translation smartphone app; o_4: Provide information in as many languages as possible; o_5: Provide interpreters who are always with them.	
Crowd Task 1: Pairwise Preference Comparison $\mathcal{Q}_1$	$o_1 \succ_{a^1} o_2$: worker a^1 prefers o_1 to o_2; $o_2 \succ_{a^2} o_3$; $o_3 \succ_{a^2} o_4$; $o_1 \succ_{a^1} o_4$; ...
Estimated Ranking	Pairwise preference aggregation result: $o_1 \succ o_2 \succ o_3 \succ o_4 \succ o_5$
Crowd Task 2: Pairwise Similarity Comparison $\mathcal{Q}_2$	$o_1 \sim_{a^1} o_2$: worker a^1 judges o_1 and o_2 *similar*; $o_3 \neq_{a^1} o_4$; $o_1 \sim_{a^2} o_3$; $o_2 \neq_{a^2} o_5$; ...
Estimated Clusters	Pairwise similarity label aggregation result: $o_1 \sim o_2 \sim o_3$, $o_4 \sim o_5$

Table 2: First example of CCTs: pairwise preference and pairwise similarity comparison tasks for the same object set, Case ①.

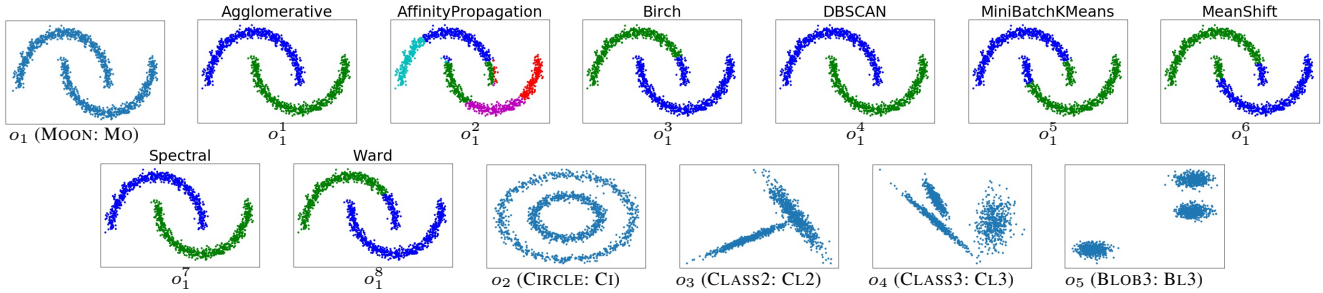

Figure 1: Third example of CCTs: categorization and pairwise preference tasks for different object sets, Case ③, Part 1.

Crowd Task 1: Categorization $\mathcal{Q}_1$	How many clusters in a point set by checking the original visualization?
Object set $\mathcal{O}_1$: $\{o_1, o_2, o_3, o_4, o_5\}$	$y_1^3 = 2$: worker a^3 judges o_1 having two clusters; $y_2^3 = 2$; $y_3^1 = 1$; $y_3^2 = 2$; ...
Estimated Category	Categorical label aggregation: $y_1 = 2$; $y_2 = 2$; $y_3 = 2$; $y_4 = 3$; $y_5 = 3$
Crowd Task 2: Pairwise Preference Comparisons $\mathcal{Q}_2$	Which cluster visualization for point set o_1 is better?
Object set $\mathcal{O}_2$: $\{o_1^1, o_1^2, o_1^3, o_1^4, o_1^5, o_1^6, o_1^7, o_1^8\}$	$o_1^7 \succ_{a^1} o_1^6$: worker a^1 prefers o_1^7 to o_1^6; $o_1^1 \succ_{a^1} o_1^3$; $o_1^5 \succ_{a^2} o_1^2$; ...
Estimated Ranking	Pairwise preference label aggregation: $o_1^1 \approx o_1^4 \approx o_1^7 \succ o_1^3 \approx o_1^8 \succ o_1^5 \approx o_1^6 \succ o_1^2$

Table 3: Third example of CCTs: categorization and pairwise preference tasks for different object sets, Case ③, Part 2.

and $\mathcal{O}_2$. For a set of opinions $\mathcal{O}_1$ for solving the "cheat" issue of "how to effectively prevent students from cheating in exams" and a set of opinions $\mathcal{O}_2$ for solving the "meeting" issue of "how to reduce the number of latecomers for team meetings," a requester wants to obtain two separate ranking lists on the same type of pairwise preference comparison question $\mathcal{Q}_1$ for the "cheat" opinion and "meeting" opinion sets, respectively. One crowdsourcing HIT with a batch of opinion pairs can contain two homogeneous types of questions: "Which of the two opinions do you prefer?" to both the "cheat" and "meeting" opinion sets, rather than only containing the questions to a single set of opinions. The labels of one set of objects can provide auxiliary information on the worker abilities to labels of other set of objects.

Third, for Case ③ in Table 1, Figure 1 and Table 3 provide an example with a categorization task ("How many clusters in a point set?") and pairwise preference comparison tasks ("Which cluster visualization is better?").

Instead of the label aggregation on each crowd task respectively, label aggregation methods by bridging the multiple SCTs in CCTs can potentially improve the label quality of all tasks. CCTs raise a research issue, i.e., how to effectively leverage additional information among SCTs in CCTs to improve the performance of label aggregation for each SCT. In this paper, we propose a general label aggregation approach for CCTs by worker ability constraint satisfaction and relaxed optimization. We assume that the multiple SCTs in CCTs are closely related and have similar difficulties, and thus the workers have similar abilities on them (this is ensured by the CCT design policies introduced in the supplementary material). Our approach has a general formulation and can utilize diverse backbone label aggregation models proposed for SCTs. This study is based on heterogeneous question types, including categorization, pairwise similarity and preference comparisons. In addition, datasets containing the composite crowd labels in CCTs are crucial for investigating this topic. Due to the lack of datasets available, we created novel real crowdsourcing datasets with diverse task settings, including Cases ①, ②, and ③ in Table 1.

The contributions of this paper are as follows: (1) We propose a novel label aggregation method that can effectively leverage the shared information of worker ability in the composite crowd labels of CCTs to improve the quality of estimated labels of SCTs. (2) Each model for SCT from existing work has its own formulation and may not be compatible with other models. We reformulate the existing models for SCTs into backbone models with a unified format so that they can be merged into a unified CCT model with a multi-task objective function in our approach. (3) We create real crowd datasets on CCTs that can be used in the research on this topic, and utilize them in the experiments to verify our approach. Our approach can improve performance in scenarios where either when multiple SCTs in a CCT have comparable levels of difficulty, or worker abilities of one SCT in a

CCT can be easily estimated by label aggregation models.

Related Work

We review the existing work on each type of SCT. The approaches for the categorization task include the expectation-maximization (EM) algorithm (Dawid and Skene 1979), bipartite models (Karger, Oh, and Shah 2011), the maximum entropy principle (Zhou et al. 2012), and Bayesian inference (Liu, Peng, and Ihler 2012; Venanzi et al. 2014) to simultaneously evaluate worker ability and true answers. More sophisticated models incorporate task difficulty (Whitehill et al. 2009), worker-task affinity (Welinder et al. 2010), and its Bayesian treatments (Wauthier and Jordan 2011; Bachrach et al. 2012). There are also other recent works, and we just list some of them (Li and Kashima 2017; Li, Baba, and Kashima 2018a; Kawase, Kuroki, and Miyauchi 2019; Li 2019; Li et al. 2020; Li and de Rijke 2023; Li, Jiang, and Xue 2023; Yang et al. 2024; Li 2024a; Zhang, Jiang, and Li 2024b,a). Besides answer aggregation approaches, there is another type of approach that directly trains the classification models with the noisy crowd labels (Rodrigues and Pereira 2018; Li, Sun, and Li 2022).

For the pairwise preference comparison task (Cattelan 2012), a typical solution is the Bradley-Terry model (Bradley and Terry 1952) and its various extensions or generalizations to diverse settings (Causeur and Husson 2005; Chen and Joachims 2016a,b; Chen et al. 2013; Raman and Joachims 2014; Li, Baba, and Kashima 2018b; Li 2022; Jin et al. 2020; Zuo et al. 2020; Zhang, Li, and Kashima 2022; Liu, Fang, and Lu 2023; Ferrara et al. 2024). There are also other methods, such as matrix completion (Yi et al. 2013; Oh, Thekumparampil, and Xu 2015), and so on.

For the pairwise similarity comparison task, the object embeddings are learned from high-dimensional feature representations or pairwise similarity comparison labels by preserving the neighborhoods or pairwise similarities (Hinton and Roweis 2002; van der Maaten and Hinton 2008; Gomes et al. 2011; Yi et al. 2012; Nguyen, Ibrahim, and Fu 2023; Ariu et al. 2024).

There are also methods for triplet similarity labels (van der Maaten and Weinberger 2012; Li, Endo, and Kashima 2021; Lu et al. 2023), numerical data (Li et al. 2014), complex data such as bounding boxes and taxonomy paths (Meir et al. 2024), text data (Li and Fukumoto 2019; Li 2020, 2024b), and so on. We share our crowdsourcing datasets at https://github.com/garfieldpigljy/ljycrowd. Numerous works exist for various SCTs, and while we do not provide an exhaustive list here, they serve as potential backbone models in our approach. To integrate them seamlessly into our methodology, they need to be adapted into the unified format with worker ability we introduce, allowing for their combination into a multi-task objective function.

There are a few existing works titled multiple tasks for crowds. Zhou, Ying, and He (2019) proposed an optimization framework for dual learning from task and worker. However, it only merged multiple datasets on the homogeneous tasks (e.g., text classification on a Computer News dataset and a Science News dataset are two tasks), and the worker sets are separate for each task. It does not consider our CCTs in which the workers provide labels across tasks. It also does not consider our Cases ① and ③ in Table 1 with heterogeneous tasks.

Label Aggregation for Multi-Crowd-Tasks

In this study, we illustrate CCTs by focusing on the cases involving two object sets or two question types. Note that constructing CCTs is not limited to these examples; it is feasible to create CCTs by combining more than two SCTs.

Definitions and Notations

We have three types of SCTs, i.e., categorization, pairwise preference (ranking), and similarity (clustering) comparison tasks. We denote the set of objects $\mathcal{O} = \{o_i\}_{i=1}^n$ and the set of crowd workers $\mathcal{A} = \{a^l\}_{l=1}^m$. We assume that no feature representations of the objects are available, which is the common setting for crowdsourced label aggregation.

The goal of the categorization task is to assign categories to an object set. The crowd labels $\mathcal{Y} = \{y_i^l\}_{i,l}$ for categories are collected. The questions for the workers can be: "Which category should be assigned to an object?" $y_i^l = c$, $c \in \mathcal{C}$, and $|\mathcal{C}| = \mathcal{K}$, where $\mathcal{C}$ contains the candidate categories. The outputs are the estimated aggregation labels $\hat{\mathcal{Y}} = \{\hat{y}_i\}_{i=1}^n$ from the categorical labels collected.

The goal of the preference task is to obtain a ranking list reflecting the relative preference among the objects. The crowd labels $\mathcal{P} = \{p_{ij}^l\}_{i,j,l}$ of pairwise preference comparisons are collected. The binary questions for the workers can be: "Which of the two objects is preferred to the other one?" $p_{ij}^l = 1$ if o_i is preferred to o_j by worker a^l; otherwise, $p_{ij}^l = 0$. The collected preference labels are aggregated into a single ranking list through the estimation of statistical models. The outputs are the preference scores $\Omega = \{\omega_i\}_{i=1}^n$.

The goal of the similarity task is to obtain proximity relations among the objects. The crowd labels $\mathcal{S} = \{s_{ij}^l\}_{i,j,l}$ of pairwise similarity comparisons are collected. The binary questions for the workers can be "Are these two objects similar?". $s_{ij}^l = 1$ if worker a^l judges that o_i is similar to o_j; otherwise, $s_{ij}^l = 0$. The collected similarity labels are first used to seek for an d-dimensional embedding $\mathbf{x}_i \in \Re^d$ of each object o_i. The outputs are the embeddings $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$. The embeddings can be utilized for clustering the objects using any existing clustering algorithm.

Because the number of all candidate object pairs is huge for both pairwise comparisons, only a small subset of labels is required; each object pair is annotated by multiple workers; each worker only annotates a small subset. The target is to obtain accurate ranking and (or) clustering results using a limited number of pairwise comparison labels.

In addition, for all tasks, worker abilities $\Theta = \{\theta^l\}_{l=1}^m$ are estimated. In our approach, because each worker provides labels to multiple SCTs, Θ is used to bridge the information among multiple SCTs. In summary, for examples, in Case ①, the input can be $\{\mathcal{O}_1, \mathcal{P}_1, \mathcal{S}_1\}$ for one preference task and one similarity task on the same object set $\mathcal{O}_1$, the output can be $\{\Omega_1, \Theta_{\Omega_1}, \mathbf{X}_1, \Theta_{\mathbf{X}_1}\}$. In Case ②, the input can be $\{\mathcal{O}_1, \mathcal{P}_1, \mathcal{O}_2, \mathcal{P}_2\}$ for two preference

tasks or $\{\mathcal{O}_1, \mathcal{S}_1, \mathcal{O}_2, \mathcal{S}_2\}$ for two similarity tasks on two different object sets $\mathcal{O}_1$ and $\mathcal{O}_2$, and the outputs can be $\{\Omega_1, \Theta_{\Omega_1}, \Omega_2, \Theta_{\Omega_2}\}$ or $\{\mathbf{X}_1, \Theta_{\mathbf{X}_1}, \mathbf{X}_2, \Theta_{\mathbf{X}_2}\}$, respectively. In Case ③, the input and output for a categorization task and a preference task on two different object sets can be $\{\mathcal{O}_1, \mathcal{Y}_1, \mathcal{O}_2, \mathcal{P}_2\}$ and $\{\hat{\mathcal{Y}}_1, \Theta_{\hat{\mathcal{Y}}_1}, \Omega_2, \Theta_{\Omega_2}\}$, respectively.

Backbone Models and Reformulations

First, we introduce the backbone models that adapt the existing label aggregation methods of SCTs for diverse types of labels. Each of them has its own formulation and may not be compatible with other methods. Thus, we need to reformulate these methods so that the objective functions of multiple SCTs have unified formats, and can be merged into a unified CCT model with a multi-task objective function in our approach. Because our approach bridges the backbone models of multiple SCTs by the workers who provide labels to multiple SCTs, worker ability modeling is mandatory in the selected backbone models.

For **the categorization task based on categorical labels**, we utilize one of the most typical label aggregation methods, GLAD (Whitehill et al. 2009), because it has worker ability parameters and can be reformulated into our unified formats. The original GLAD method utilizes an Expectation Maximization approach (EM) to obtain maximum likelihood estimates of the potential parameters, i.e., estimated labels $\hat{\mathcal{Y}}$, worker abilities Θ, and task easiness $\Gamma = \{\gamma_i\}_{i=1}^n$. We reformulate it into one objective function. The probability q_i^l that a worker a^l assigns the correct label to an object o_i can be formulated as $q_i^l = 1/\left(1 + (\mathcal{K}-1)e^{-\theta^l \gamma_i}\right)$. We assume that a worker assigns an incorrect answer uniformly at random. We reformulate the crowd labels y_i^l into a $\mathcal{K}$-dimensional one-hot binary vector, only one of its elements is 1, $\mathbf{y}_i^l = (0, \ldots, 0, 1, 0, \ldots, 0)^\top \in \{0,1\}^{\mathcal{K}}$, and reformulate the estimated labels $\hat{y}_i$ into a $\mathcal{K}$-dimensional one-hot binary vector $\hat{\mathbf{y}}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top \in \{0,1\}^{\mathcal{K}}$. $\mathbf{y}_i^{l\top}\hat{\mathbf{y}}_i$ can compute the correctness of label $\mathbf{y}_i^l$ by worker a^l to object o_i. The objective function based on log-likelihood of the crowd labels can be reformulated as follows.

$$\mathcal{L}_{glad} = \sum_{i,l}\left(\mathbf{y}_i^{l\top}\hat{\mathbf{y}}_i \log q_i^l + \left(1 - \mathbf{y}_i^{l\top}\hat{\mathbf{y}}_i\right)\log\left(\frac{1-q_i^l}{\mathcal{K}-1}\right)\right). \tag{1}$$

For the **ranking task based on pairwise preference comparisons**, because we need to bridge multiple SCTs by modeling worker abilities, one typical model is CROWDBT (Chen et al. 2013), which extends Bradley-Terry (BT) model (Bradley and Terry 1952) in the crowdsourced setting. The probability q_{ij} that o_i is preferred over o_j can be defined based on the latent variables of the preference scores $q_{ij} = \Pr(o_i \succ o_j) = 1/\left(1 + \exp(-(\omega_i - \omega_j))\right)$, and the objective function of CrowdBT in our formulation is as follows.

$$\begin{aligned}\mathcal{L}_{cbt} = -\frac{1}{n_p}\sum_{p_{ij}^l \in \mathcal{Y}} &\left\{p_{ij}^l \log\left(\theta^l q_{ij} + (1-\theta^l)(1-q_{ij})\right)\right. \\ &\left.+(1-p_{ij}^l)\log\left(\theta^l(1-q_{ij}) + (1-\theta^l)q_{ij}\right)\right\} \\ &- \frac{\lambda_0}{n}\sum_{o_i \in \mathcal{O}}(\log q_{0i} + \log q_{i0}).\end{aligned} \tag{2}$$

We modify the original CROWDBT model by incorporating the term for negative labels into the objective function. The third term with an auxiliary score ω_0 ($\omega_0 \notin \Omega$) is used to address the scale-invariant problem in the objective function and make the scores in the objective identifiable. We set λ_0 as a general value, i.e., $\lambda_0 = 1$.

There is also a recent model HBTL (Jin et al. 2020), which extends the BT model by modeling the scale factor of judgment noise for each evaluator that can be regarded as the individual worker abilities Θ in our model. The probability model is $q_{ij}^l = 1/\left(1 + \exp(-\theta^l(\omega_i - \omega_j))\right)$, and objective function of HBTL is as follows.

$$\mathcal{L}_{hbt} = -\frac{1}{n_p}\sum_{p_{ij}^l \in \mathcal{P}}\left\{p_{ij}^l \log q_{ij}^l + (1-p_{ij}^l)\log(1-q_{ij}^l)\right\}. \tag{3}$$

The model also addresses the identifiable objective problem of the scores by setting a constraint $\sum_{o_i \in \mathcal{O}} \omega_i = 0$.

For the **clustering task based on pairwise similarity comparisons**, we utilize and verify two backbone models for estimating the object similarities ϕ_{ij}^l based on embedding $\mathbf{X}$, by using Gaussian kernel and heavy-tailed Student-t kernel with α degrees of freedom, respectively. We add worker ability parameters into these SCT models so that they can be merged into our CCT model.

$$\phi_{ij}^l = \exp\left(\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\theta^l}\right), \text{ or } \phi_{ij}^l = \left(1 + \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\theta^l \alpha}\right)^{-\frac{\alpha+1}{2}} \tag{4}$$

$$\begin{aligned}\mathcal{L}_{sim} = -\frac{1}{n_s}\sum_{s_{ij}^l \in \mathcal{S}} &\{s_{ij}^l \log \phi_{ij}^l \\ &+(1-s_{ij}^l)\log(1-\phi_{ij}^l)\} + \frac{\lambda_s}{nd}||\mathbf{X}||_2^2,\end{aligned} \tag{5}$$

Proposed Label Aggregation Method

The proposed general label aggregation model of a CCT is based on the backbone models of the SCTs contained in the CCT. We denote the loss function of a backbone model as $\mathcal{F}(\mathbf{W}, \Theta)$, where Θ contains the parameters related to worker abilities; $\mathbf{W}$ denotes the other parameters in the specific task, e.g., the object easiness Γ for a categorization task, the preference scores Ω for a ranking task, or the embeddings $\mathbf{X}$ for a clustering task.

To improve label aggregation performance of each SCT by effectively leveraging the additional information among multiple SCTs, we consider the overlaps and relations among the workers for these SCTs. Based on the design policies of the CCTs, we assume that workers in different SCTs have similar abilities. When estimating the worker abilities for an SCT, we can utilize the estimated abilities of the same workers of another SCT as auxiliary information.

Our idea is represented as a worker ability constraint satisfaction problem to find worker abilities Θ^1 and Θ^2 that satisfy the following constraint. Here, $\mathcal{A}_1$ and $\mathcal{A}_2$ are the two worker sets for two SCTs; we utilize superscript to the symbols of parameters for representing the index of a task.

$$\theta^{1l} = \theta^{2l}, \ \forall a^l \in \mathcal{A}_1 \cap \mathcal{A}_2. \tag{6}$$

In practice, there might be neither an exact solution nor a unique solution satisfying Eq. (6). Worker abilities in different SCTs are also not exactly the same. Thus, we convert the constraint to a relaxed optimization problem,

$$\underset{\Theta^1,\Theta^2,a^l\in\mathcal{A}_1\cap\mathcal{A}_2}{\arg\min} \mathcal{G}(\Theta^1,\Theta^2). \tag{7}$$

where $\mathcal{G}$ computes the distances among the abilities of the same workers in different SCTs. We provide three options. One is based on the square of Euclidean norm, one uses the KL-divergence, and one utilizes the L1 norm.

$$\mathcal{G}_{euc}(\Theta^1,\Theta^2) = \frac{1}{|\mathcal{A}_1\cap\mathcal{A}_2|}||\mathcal{H}(\Theta^1)-\mathcal{H}(\Theta^2)||_2^2. \tag{8}$$

$$\begin{aligned}\mathcal{G}_{kl}(\Theta^1,\Theta^2) = -\frac{1}{|\mathcal{A}_1\cap\mathcal{A}_2|}(&\mathcal{H}(\Theta^1)\log\mathcal{H}(\Theta^2)\\ &+\mathcal{H}(\Theta^2)\log\mathcal{H}(\Theta^1)).\end{aligned} \tag{9}$$

$$\mathcal{G}_{l1}(\Theta^1,\Theta^2) = \frac{1}{|\mathcal{A}_1\cap\mathcal{A}_2|}||\mathcal{H}(\Theta^1)-\mathcal{H}(\Theta^2)||_1. \tag{10}$$

$\mathcal{H}$ selects the parameters of the workers who provide labels in both SCTs. There is a risk that the worker ability constraint becomes harmful when the worker abilities in different SCTs are not consistent. Another option is that, instead of using the constraint based on the distance, we relax the constraint by the relative order of worker abilities. If a worker a^{l_1} has higher ability than another worker a^{l_2} in one SCT, we impose the constraint that a^{l_1} has higher ability than a^{l_2} in another SCT. We implement this idea as another relaxed constraint.

$$\begin{aligned}&\mathcal{G}_{rank}(\Theta^1,\Theta^2) =\\ &\sum_{\Theta^1,\Theta^2,a^l\in\mathcal{A}_1\cap\mathcal{A}_2} (\mathcal{I}(\theta^{1l_1}>\theta^{1l_2})\log(1+e^{-\tau(\theta^{2l_1}-\theta^{2l_2})})\\ &\qquad +\mathcal{I}(\theta^{2l_1}>\theta^{2l_2})\log(1+e^{-\tau(\theta^{1l_1}-\theta^{1l_2})}))\end{aligned} \tag{11}$$

where $\mathcal{I}$ is the indicator function. The constant $\tau>0$ controls how strict the constraint is. The computation cost of $\mathcal{G}_{rank}$ is higher than the distance based constraints $\mathcal{G}_{euc}$, $\mathcal{G}_{kl}$ and $\mathcal{G}_{l1}$, but it does not give a significant impact on the scalability of the entire optimization.

We combine this constraint with the models of SCTs, as a regularization term. When there are two SCTs, the general formulation of the loss function $\mathcal{L}$ of our model is

$$\mathcal{L} = \lambda^1\mathcal{F}(\mathbf{W}^1,\Theta^1)+\lambda^2\mathcal{F}(\mathbf{W}^2,\Theta^2)+\lambda\mathcal{G}(\Theta^1,\Theta^2). \tag{12}$$

In Case ①, $\mathbf{W}^1=\Omega_1$, $\Theta^1=\Theta_{\Omega_1}$, $\mathbf{W}^2=\mathbf{X}_1$, and $\Theta^2=\Theta_{\mathbf{X}_1}$ for one preference task and one similarity task on the same object set. In Case ②, $\mathbf{W}^1=\Omega_1$, $\Theta^1=\Theta_{\Omega_1}$, $\mathbf{W}^2=\Omega_2$, and $\Theta^2=\Theta_{\Omega_2}$ for two preference tasks; $\mathbf{W}^1=\mathbf{X}_1$, $\Theta^1=\Theta_{\mathbf{X}_1}$, $\mathbf{W}^2=\mathbf{X}_2$, and $\Theta^2=\Theta_{\mathbf{X}_2}$ for two similarity tasks on two different object sets. In Case ③, $\mathbf{W}^1=\{\hat{\mathcal{Y}}_1,\Gamma_1\}$, $\Theta^1=\{\Theta_{\hat{\mathcal{Y}}_1},\Theta_{\Gamma_1}\}$, $\mathbf{W}^2=\Omega_2$, and $\Theta^2=\Theta_{\Omega_2}$ for one categorization task and one preference task on two different object sets. All the above objective functions are convex. The optimal solutions are found by iterative updating $\mathbf{W}$ and Θ based on gradients. Furthermore, the formulations can also be generalized for more than two tasks, by denoting $\mathcal{U}$ as the set of SCTs, i.e.,

$$\mathcal{L} = \sum_{u=1}^{|\mathcal{U}|}\lambda^u\mathcal{L}^u+\lambda\mathcal{G}(\mathcal{H}(\Theta^1),...,\mathcal{H}(\Theta^{|\mathcal{U}|})). \tag{13}$$

Experiments

Dataset Creation

To investigate the performance of the label aggregation approaches in the scenario of CCTs, we require real datasets that contain crowd labels. Thus, we create two groups of real crowd datasets. The first group has two SCTs with comparable difficulty; the second group has an SCT in which the worker abilities can be easily estimated by label aggregation models. We use these two groups to illustrate the proper scenarios of our approach.

The first group contains the CCTs with pairwise preference and pairwise similarity tasks for Case ① and Case ②. We select two datasets from (Baba, Li, and Kashima 2020), "cheat" and "meeting" datasets, that contain crowd opinions for solving real-life issues in society. For example, the issue of the "cheat" dataset is "How to effectively prevent students from cheating in exams?" A crowd opinion in the text can be "use different orders for the problems of different students." The issue of the "meeting" dataset is "How can we reduce the number of latecomers for team meetings?". A crowd opinion in the text can be "increase the number of advance reminders". We utilize the subsets of crowd opinions in these two datasets as two object sets.

For fair evaluations, we do not use the crowd pairwise preference labels in (Baba, Li, and Kashima 2020) and recollect new pairwise crowd labels for the CCTs using the selected opinion subsets, so that the same workers annotate in both two SCTs. We collect four crowd datasets. For Case ① (one preference task and one similarity task on the same object set), we create dataset ⓐ based on "cheat" opinions and dataset ⓑ based on "meeting" opinions. For Case ② (two preference tasks or two similarity tasks on two different object sets), based on two object sets of "cheat" opinions and "meeting" opinions, we create dataset ⓒ with two preference tasks, and dataset ⓓ with two similarity tasks. The examples of labels and descriptions of the preference and similarity tasks can be found in Table 2 and Section "Definitions and Notations". Table 4 shows the datasets statistics.

The second group contains a CCT with one categorization task and one pairwise preference task on two different object sets for Case ③. As shown in Figure 1 and Table 3, in Task 1, the original visualization of a point set o_1' (MOON:MO)) is shown to a worker and asks the number of the clusters in o_1'; in Task 2, the cluster visualizations generated by different clustering methods are shown in a pairwise manner; the question of pairwise preference comparison is "Which cluster visualization for point set o_1' is better?" The categorization labels of five point sets consist of the data for the categorization task; the preference labels for one point set consist of the data for one ranking task; the categorization data (i.e., categorical labels for all MO, CI, CL2, CL3, and BL3) and one ranking data (e.g., preference labels for MO) consist of one dataset (e.g., named MO). There are five datasets in total in this group. Examples of labels and descriptions of the categorization and preference tasks can be found in Figure 1, Table 3 and Section "Definitions and Notations." Table 7 lists datasets statistics. We share the data and code at https://github.com/garfieldpigljy/CompositeCrowdTasks.

Datasets	$\lvert\mathcal{A}_1\rvert = \lvert\mathcal{A}_2\rvert$	$\mathcal{O}_1$ ($\lvert\mathcal{O}_1\rvert$)	$\lvert\mathcal{P}_1\rvert$ or $\lvert\mathcal{S}_1\rvert$	$\mathcal{O}_2$ ($\lvert\mathcal{O}_2\rvert$)	$\lvert\mathcal{P}_2\rvert$ or $\lvert\mathcal{S}_2\rvert$
ⓐ Case ①: One Pairwise Preference Task and One Pairwise Similarity Task	189	cheat (40)	7438	cheat (40)	7438
ⓑ Case ①: One Pairwise Preference Task and One Pairwise Similarity Task	189	meeting (40)	7424	meeting (40)	7465
ⓒ Case ②: Two Pairwise Preferences Tasks	99	cheat (26)	3222	meeting (26)	3225
ⓓ Case ②: Two Pairwise Similarities Tasks	94	cheat (26)	3217	meeting (26)	3219

Table 4: Statistics of the two pairwise comparison dataset group (Case ① and ②). For Case ①, Task 1 is the preference task.

		Baseline	Our CCTs			
	Task 1	Single	EUC	KL	L1	Rank
ⓐ Case ①	HBTL	0.6906 ± 0.0180	**0.6953 ± 0.0178***	0.6932 ± 0.0170*	0.6942 ± 0.0183*	0.6941 ± 0.0178*
cheat	HBTL	0.6906 ± 0.0180	**0.6945 ± 0.0181***	0.6932 ± 0.0172*	0.6939 ± 0.0184*	**0.6945 ± 0.0179***
One Preference	CROWDBT	0.6885 ± 0.0185	**0.6897 ± 0.0183***	0.6893 ± 0.0180*	0.6895 ± 0.0184*	0.6895 ± 0.0180*
One Similarity	CROWDBT	0.6885 ± 0.0185	0.6892 ± 0.0181*	**0.6894 ± 0.0181***	**0.6894 ± 0.0181***	0.6893 ± 0.0183*
ⓑ Case ①	HBTL	0.6737 ± 0.221	0.6772 ± 0.0219*	0.6768 ± 0.0212*	**0.6787 ± 0.0223***	0.6772 ± 0.0207*
meeting	HBTL	0.6737 ± 0.0221	**0.6781 ± 0.0215***	0.6771 ± 0.0205*	0.6778 ± 0.0213*	0.6766 ± 0.0206*
One Preference	CROWDBT	0.6693 ± 0.0230	0.6714 ± 0.0226*	0.6710 ± 0.0230*	**0.6719 ± 0.0230***	0.6709 ± 0.0229*
One Similarity	CROWDBT	0.6693 ± 0.0230	**0.6711 ± 0.0228***	**0.6711 ± 0.0230***	**0.6711 ± 0.0230***	**0.6711 ± 0.0230***
ⓒ Case ②	HBTL	0.6953 ± 0.0334	0.7027 ± 0.0384*	0.7025 ± 0.0372*	0.7022 ± 0.0360*	**0.7028 ± 0.0364***
cheat	HBTL	0.6953 ± 0.0334	0.7016 ± 0.0358*	0.7027 ± 0.0371*	0.7024 ± 0.0347*	**0.7036 ± 0.0372***
meeting	CROWDBT	0.6959 ± 0.0353	0.6989 ± 0.0356*	0.6988 ± 0.0363*	0.6991 ± 0.0352*	**0.6998 ± 0.0366***
Two Preference	CROWDBT	0.6959 ± 0.0353	0.6985 ± 0.0362*	0.6988 ± 0.0363*	0.6985 ± 0.0360*	**0.6989 ± 0.0363***
ⓓ Case ②	Student-t	0.5259 ± 0.0292	0.5333 ± 0.0293*	0.5319 ± 0.0292*	**0.5336 ± 0.0281***	0.5266 ± 0.0292*
cheat	Student-t	0.5259 ± 0.0292	**0.5343 ± 0.0297***	0.5320 ± 0.0293*	0.5340 ± 0.0293*	0.5330 ± 0.0289*
meeting	Gaussian	0.5266 ± 0.0292	**0.5340 ± 0.0284***	0.5322 ± 0.0287*	0.5338 ± 0.0296*	0.5328 ± 0.0293*
Two Similarity	Gaussian	0.5266 ± 0.0292	**0.5340 ± 0.0287***	0.5322 ± 0.0287*	0.5334 ± 0.0298*	0.5329 ± 0.0291*

Table 5: Main results for the two pairwise comparison dataset group. "*" mark indicates the cases that the our CCT models are statistically significantly better than the corresponding baseline SCT models in a Wilcoxon signed-rank test ($p < 0.05$).

		Baseline	Our CCTs			
	Task 2	Single	EUC	KL	L1	Rank
ⓐ Case ①	Student-t	0.5410 ± 0.0217	**0.5455 ± 0.0202***	0.5449 ± 0.0204*	0.5453 ± 0.0199*	0.5438 ± 0.0201*
cheat	Gaussian	0.5406 ± 0.0214	0.5425 ± 0.0197	**0.5427 ± 0.0191**	0.5424 ± 0.0196*	0.5426 ± 0.0195*
One Preference	Student-t	0.5410 ± 0.0217	0.5451 ± 0.0183*	**0.5455 ± 0.0186***	0.5446 ± 0.0187*	0.5449 ± 0.0185*
One Similarity	Gaussian	0.5406 ± 0.0214	0.5427 ± 0.0194	0.5426 ± 0.0186	**0.5429 ± 0.0195***	**0.5429 ± 0.0184***
ⓑ Case ①	Student-t	**0.5307 ± 0.0154**	0.5297 ± 0.0165	0.5294 ± 0.0161	0.5294 ± 0.0166	0.5297 ± 0.0157
meeting	Gaussian	**0.5302 ± 0.0151**	0.5291 ± 0.0166	0.5288 ± 0.0172	0.5293 ± 0.0167	0.5292 ± 0.0169
One Preference	Student-t	0.5307 ± 0.0154	0.5319 ± 0.0151	0.5301 ± 0.0149	**0.5320 ± 0.0151**	0.5304 ± 0.0146
One Similarity	Gaussian	**0.5302 ± 0.0151**	0.5288 ± 0.0149	0.5290 ± 0.0148	0.5286 ± 0.0147	0.5290 ± 0.0148
ⓒ Case ②	HBTL	0.6655 ± 0.0320	0.6691 ± 0.0334*	0.6694 ± 0.0319*	**0.6700 ± 0.0342***	0.6686 ± 0.0329*
cheat	CROWDBT	0.6678 ± 0.0310	**0.6718 ± 0.0305***	0.6712 ± 0.0296*	0.6712 ± 0.0309*	0.6712 ± 0.0299*
meeting	HBTL	0.6655 ± 0.0320	0.6699 ± 0.0327*	0.6693 ± 0.0324*	**0.6713 ± 0.0324***	0.6709 ± 0.0324*
Two Preference	CROWDBT	0.6678 ± 0.0310	0.6711 ± 0.0295*	**0.6712 ± 0.0296***	0.6711 ± 0.0298*	**0.6712 ± 0.0296***
ⓓ Case ②	Student-t	0.5458 ± 0.0372	0.5507 ± 0.0337*	0.5479 ± 0.0359*	**0.5520 ± 0.0355***	0.5498 ± 0.0358
cheat	Gaussian	0.5456 ± 0.0369	0.5507 ± 0.0359*	0.5486 ± 0.0363*	**0.5513 ± 0.0349***	0.5494 ± 0.0339
meeting	Student-t	0.5458 ± 0.0372	**0.5524 ± 0.0348***	0.5489 ± 0.0352*	0.5499 ± 0.0356*	0.5489 ± 0.0364
Two Similarity	Gaussian	0.5456 ± 0.0369	**0.5518 ± 0.0368***	0.5492 ± 0.0352	0.5502 ± 0.0346*	0.5493 ± 0.0349

Table 6: Main results for the two pairwise comparison dataset group. The format of the results is the same with Table 5.

Experimental Settings

We set common values to the hyperparameters to verify the capability of our approach without heavy hyperparameter tuning. Crowdsourcing requesters can thus use our approach conveniently. The weight hyperparameters in the loss function Eq.(12) of our approach are set to $\lambda^1 = \lambda^2 = \lambda = 1$.

We utilize the backbone models as the baseline models of the SCTs, i.e., GLAD for categorization task, CROWDBT and HBTL for pairwise preference task, and Gaussian kernel and Student-t kernel based models for pairwise similar-

	MOON	CIRCLE	CLASS2	CLASS3	BLOB3
$\lvert\mathcal{P}_2\rvert$	381	390	436	327	371

Table 7: Statistics of the categorization and preference dataset group (Case ③). For other statistics, $|\mathcal{O}_1|$ = 5; $|\mathcal{A}_1|$ = 113; $|\mathcal{Y}_1|$ = 113; $|\mathcal{O}_2|$ = 8. $|\mathcal{A}_2|$ = 113.

	Baseline	Our CCTs			
	SCTs	EUC	KL	L1	RANK
MO	0.435	**0.644** ± **0.127***	0.637 ± 0.130*	0.643 ± 0.127*	0.640 ± 0.128*
CI	0.364	0.574 ± 0.148*	0.580 ± 0.144*	**0.586** ± **0.142***	0.579 ± 0.149*
CL2	0.261	0.531 ± 0.140*	0.529 ± 0.130*	0.530 ± 0.136*	**0.536** ± **0.131***
CL3	0.462	**0.537** ± **0.131***	**0.537** ± **0.138***	0.534 ± 0.136*	0.535 ± 0.137*
BL	0.455	0.500 ± 0.138*	0.496 ± 0.141*	**0.502** ± **0.133***	0.494 ± 0.142*

Table 8: Main results for task 2 of the categorization and preference dataset group. The category accuracy of Task 1 that is not list in the Table is always perfect (i.e., equal to 1). "*" mark indicates the cases that our CCT models are statistically significantly better than the corresponding baseline SCT models in a Wilcoxon signed-rank test ($p < 0.05$).

ity task. For the pairwise similarity task, the dimension of embeddings is assigned to $d = 20$, the hyperparameters in Eqs.(4) and (5) are set to $\lambda_s = 0.01$, $\alpha = d - 1$. We evaluate the approaches on their capability to estimate the pairwise comparison labels of all object pairs by training the models with only a small number of labeled object pairs. In detail, in one experimental trial, we randomly select a subset of all object pairs for each SCT respectively, with a sampling rate $r = 0.1$. We evaluate the average performance of 100 trials.

We use three performance evaluation metrics: *category accuracy*, *pairwise preference accuracy*, and *pairwise similarity accuracy*. The category accuracy is the accuracy of the estimated labels in the categorization task. The pairwise preference accuracy is the accuracy of the estimated preference. For an object pair o_i and o_j, if o_i is preferred to o_j in the ground truth labels, the estimated preference score $\hat{\omega_i}$ should be higher than $\hat{\omega_j}$. The pairwise similarity accuracy is the accuracy of the relationships between the estimated similarity of two object pairs. Without a similar-dissimilar threshold, we cannot assign a similarity label to the estimated similarity $\hat{s}_{ij}$. Instead, for a similar object pair (o_{i_1}, o_{j_1}) and a dissimilar object pair (o_{i_2}, o_{j_2}) in the ground truth, if the estimated similarity $\hat{s}_{i_1 j_1}$ is higher than $\hat{s}_{i_2 j_2}$, we determine that the relationships between the estimated similarities of these two object pairs are correct. The ground truths of pairwise preference (similarity) labels are obtained by applying majority voting to all crowd preference (similarity) labels of an object pair in the entire dataset.

Results: Preference and Similarity Datasets

Tables 5 and 6 list the main results for the group of two pairwise comparison datasets. Note that because we collect the datasets for Case ① and ② separately, the preference (similarity) results on the same object set in Table 5 (or 6).ⓐ, ⓑ, ⓒ, and ⓓ are not comparable with each other. Our approach statistically significantly outperforms the baselines in most of the results. It shows that our label aggregation approach can effectively leverage the additional information in the labels from CCTs and outperforms the baselines for the corresponding SCTs. Note that this dataset group is a combination of two tasks with comparable difficulty. Our approach is thus proper for this scenario. Furthermore, the four types of constraints $\mathcal{G}$ have comparable results. We do not exactly have special recommendations from them when using our approach, but considering the number of times it can achieve the best results, $\mathcal{G}_{euc}$ can be selected; considering the number of times it can reach results better than the baselines significantly, $\mathcal{G}_{l1}$ can be selected; if the worker abilities on different SCTs are not comparable or a requester is not so sure about this issue, $\mathcal{G}_{rank}$ can be selected.

Results: Categorization and Preference Datasets

For the evaluations in Case ③, Table 8 lists the main results for Task 2 of the categorization and preference dataset group. The category accuracy of Task 1 is always perfect because the correct labels as well as worker abilities of this categorization task are easy to be estimated by the label aggregation model GLAD. In Table 8, our approach prominently improves the pairwise preference accuracy of Task 2. It shows that our approach can effectively enhance the label aggregation performance by bridging the worker abilities. All four constraints $\mathcal{G}$ have comparable results. Our approach is thus very appropriate for the scenario when worker abilities of one SCT in a CCT are easy to be estimated by label aggregation models.

Conclusion

In this paper, in contrast to the traditional SCTs, we propose a novel paradigm for crowdsourced label collection, i.e., CCTs, which is flexible for the requesters to design crowd tasks. We propose a novel label aggregation method that can effectively leverage the shared information of worker ability on the composite crowd labels to improve the quality of estimated labels. We reformulate existing models into a unified format so that they can be merged into a unified CCT model with a multi-task objective function in our approach. The experimental results demonstrate that our approach can effectively bridge the worker information of CCTs to improve the quality of aggregated labels and outperform the baselines proposed for SCTs.

In the experiments, we used the datasets with various types of data and tasks, including categorical labels, pairwise preference comparison labels, and pairwise similarity comparison labels. We consider the diverse CCT scenarios including heterogeneous question types and single object sets (Case ①), homogeneous question types and multiple object sets (Case ②), and heterogeneous question types and multiple object sets (Case ③), which shows the wide coverage of the CCT scenarios of our work. Our datasets and experiments provide board testing and verify the generalizability. In future work, we will consider other domains and more complex scenarios. The design of CCTs can be free and more complex.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP23K28092.

References

Ariu, K.; Ok, J.; Proutiere, A.; and Yun, S. 2024. Optimal clustering from noisy binary feedback. *Machine Learning*, 113(5): 2733–2764.

Baba, Y.; Li, J.; and Kashima, H. 2020. CrowDEA: Multi-View Idea Prioritization with Crowds. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP'20)*, volume 8, 23–32.

Bachrach, Y.; Minka, T.; Guiver, J.; and Graepel, T. 2012. How to grade a test without knowing the answers: a Bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML'12)*, 819–826.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Cattelan, M. 2012. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 412–433.

Causeur, D.; and Husson, F. 2005. A 2-dimensional extension of the Bradley–Terry model for paired comparisons. *Journal of Statistical Planning and Inference*, 135(2): 245–259.

Chen, S.; and Joachims, T. 2016a. Modeling Intransitivity in Matchup and Comparison Data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16)*, 227–236.

Chen, S.; and Joachims, T. 2016b. Predicting Matchups and Preferences in Context. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, 775–784.

Chen, X.; Bennett, P. N.; Collins-Thompson, K.; and Horvitz, E. 2013. Pairwise Ranking Aggregation in a Crowdsourced Setting. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13)*, 193–202.

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).

Ferrara, A.; Bonchi, F.; Fabbri, F.; Karimi, F.; and Wagner, C. 2024. Bias-aware ranking from pairwise comparisons. *Data Mining and Knowledge Discovery*, 1–25.

Gomes, R. G.; Welinder, P.; Krause, A.; and Perona, P. 2011. Crowdclustering. In *Advances in Neural Information Processing Systems 24 (NIPS'11)*, 558–566.

Hinton, G. E.; and Roweis, S. 2002. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems (NIPS'02)*, volume 15.

Jin, T.; Xu, P.; Gu, Q.; and Farnoud, F. 2020. Rank Aggregation via Heterogeneous Thurstone Preference Models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'20)*, volume 34, 4353–4360.

Karger, D. R.; Oh, S.; and Shah, D. 2011. Iterative Learning for Reliable Crowdsourcing Systems. In *Advances in Neural Information Processing Systems (NIPS'11)*, 1953–1961.

Kawase, Y.; Kuroki, Y.; and Miyauchi, A. 2019. Graph Mining Meets Crowdsourcing: Extracting Experts for Answer Aggregation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'19)*, 1272–1279.

Li, D.; and de Rijke, M. 2023. Extending Label Aggregation Models with a Gaussian Process to Denoise Crowdsourcing Labels. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'23)*, 729–738.

Li, H.; Jiang, L.; and Xue, S. 2023. Neighborhood Weighted Voting-Based Noise Correction for Crowdsourcing. *ACM Trans. Knowl. Discov. Data*, 17(7).

Li, J. 2019. Budget Cost Reduction for Label Collection with Confusability Based Exploration. In *Proceedings of the 26th International Conference on Neural Information Processing (ICONIP'19)*, 231–241.

Li, J. 2020. Crowdsourced Text Sequence Aggregation Based on Hybrid Reliability and Representation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20)*, 1761–1764.

Li, J. 2022. Context-Based Collective Preference Aggregation for Prioritizing Crowd Opinions in Social Decision-Making. In *Proceedings of the ACM Web Conference 2022 (WWW'22)*, 2657–2667.

Li, J. 2024a. A Comparative Study on Annotation Quality of Crowdsourcing and LLm Via Label Aggregation. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, 6525–6529.

Li, J. 2024b. Human-LLM Hybrid Text Answer Aggregation for Crowd Annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, 15609–15622.

Li, J.; Baba, Y.; and Kashima, H. 2018a. Incorporating Worker Similarity for Label Aggregation in Crowdsourcing. In *Proceedings of the 27th International Conference on Artificial Neural Networks (ICANN'18)*, 596–606.

Li, J.; Baba, Y.; and Kashima, H. 2018b. Simultaneous Clustering and Ranking from Pairwise Comparisons. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'18)*, 1554–1560.

Li, J.; Endo, L. R.; and Kashima, H. 2021. Label Aggregation for Crowdsourced Triplet Similarity Comparisons. In *Proceedings of the 28th International Conference on Neural Information Processing (ICONIP'21)*, 176–185.

Li, J.; and Fukumoto, F. 2019. A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for

Ground Truth Creation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP (AnnoNLP'19)*, 24–28.

Li, J.; and Kashima, H. 2017. Iterative Reduction Worker Filtering for Crowdsourced Label Aggregation. In *Proceedings of the 18th International Conference on Web Information Systems Engineering (WISE'17)*, 46–54.

Li, J.; Kawase, Y.; Baba, Y.; and Kashima, H. 2020. Performance as a Constraint: An Improved Wisdom of Crowds Using Performance Regularization. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'20)*, 1534–1541.

Li, J.; Sun, H.; and Li, J. 2022. Beyond Confusion Matrix: Learning from Multiple Annotators with Awareness of Instance Features. *Mach. Learn.*, 112(3): 1053–1075.

Li, Q.; Li, Y.; Gao, J.; Su, L.; Zhao, B.; Demirbas, M.; Fan, W.; and Han, J. 2014. A Confidence-Aware Approach for Truth Discovery on Long-Tail Data. *Proc. VLDB Endow.*, 8(4): 425–436.

Liu, Q.; Peng, J.; and Ihler, A. 2012. Variational Inference for Crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS'12)*, 692–700.

Liu, Y.; Fang, E. X.; and Lu, J. 2023. Lagrangian inference for ranking problems. *Operations research*, 71(1): 202–223.

Lu, X.; Li, J.; Takeuchi, K.; and Kashima, H. 2023. Multiview Representation Learning from Crowdsourced Triplet Comparisons. In *Proceedings of the ACM Web Conference 2023 (WWW)*, 3827–3836.

Meir, R.; Nguyen, V.-A.; Chen, X.; Ramakrishnan, J.; and Weinsberg, U. 2024. Efficient Online Crowdsourcing with Complex Annotations. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'24)*, 38(9): 10119–10127.

Nguyen, T.; Ibrahim, S.; and Fu, X. 2023. Deep Clustering with Incomplete Noisy Pairwise Annotations: A Geometric Regularization Approach. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, volume 202, 25980–26007.

Oh, S.; Thekumparampil, K. K.; and Xu, J. 2015. Collaboratively Learning Preferences from Ordinal Data. In *Advances in Neural Information Processing Systems 28 (NIPS'15)*, 1909–1917.

Raman, K.; and Joachims, T. 2014. Methods for Ordinal Peer Grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*, 1037–1046.

Rodrigues, F.; and Pereira, F. C. 2018. Deep learning from crowds. volume 32 of *Proceedings of the AAAI conference on artificial intelligence (AAAI'18)*, 1611–1618.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, 254–263.

van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov): 2579–2605.

van der Maaten, L.; and Weinberger, K. 2012. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 1–6.

Venanzi, M.; Guiver, J.; Kazai, G.; Kohli, P.; and Shokouhi, M. 2014. Community-Based Bayesian Aggregation Models for Crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, 155–164.

Wauthier, F. L.; and Jordan, M. I. 2011. Bayesian Bias Mitigation for Crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS'11)*, 1800–1808.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. J. 2010. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems (NIPS'10)*, 2424–2432.

Whitehill, J.; Wu, T.; Bergsma, J.; Movellan, J. R.; and Ruvolo, P. L. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems 22 (NIPS'09)*, 2035–2043.

Yang, Y.; Zhao, Z.-Q.; Wu, G.; Zhuo, X.; Liu, Q.; Bai, Q.; and Li, W. 2024. A Lightweight, Effective, and Efficient Model for Label Aggregation in Crowdsourcing. *ACM Trans. Knowl. Discov. Data*, 18(4).

Yi, J.; Jin, R.; Jain, S.; and Jain, A. 2013. Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach. In *Proceeding of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*.

Yi, J.; Jin, R.; Jain, S.; Yang, T.; and Jain, A. K. 2012. Semi-Crowdsourced Clustering: Generalizing Crowd Labeling by Robust Distance Metric Learning. In *Advances in Neural Information Processing Systems 25 (NIPS'12)*, 1772–1780.

Zhang, G.; Li, J.; and Kashima, H. 2022. Improving Pairwise Rank Aggregation via Querying for Rank Difference. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA'22)*, 1–9.

Zhang, W.; Jiang, L.; and Li, C. 2024a. IWBVT: Instance Weighting-based Bias-Variance Trade-off for Crowdsourcing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS'24)*.

Zhang, W.; Jiang, L.; and Li, C. 2024b. KFNN: K-Free Nearest Neighbor For Crowdsourcing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS'24)*.

Zhou, D.; Basu, S.; Mao, Y.; and Platt, J. 2012. Learning from the Wisdom of Crowds by Minimax Entropy. In *Advances in Neural Information Processing Systems (NIPS'12)*, volume 25.

Zhou, Y.; Ying, L.; and He, J. 2019. Multi-Task Crowdsourcing via an Optimization Framework. *ACM Trans. Knowl. Discov. Data*, 13(3).

Zuo, X.; Li, J.; Zhou, Q.; Li, J.; and Mao, X. 2020. AffectI: A Game for Diverse, Reliable, and Efficient Affective Image Annotation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*, 529–537.