Mixture of Experts Based Multi-Task Supervise Learning from Crowds

Tao Han, Huaixuan Shi, Xinyi Ding, Xi-Ao Ma, Huamao Gu, Yili Fang*

School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China
{hantao, xding, mxa, ghmsjq, fangyl}@zjgsu.edu.cn, 22020100076@pop.zjgsu.edu.cn

Abstract

Existing learning-from-crowds methods aim to design proper aggregation strategies to infer the unknown true labels from noisy labels provided by crowdsourcing. They treat the ground truth as hidden variables and use statistical or deep learning based worker behavior models to infer the ground truth. However, worker behavior models that rely on ground truth hidden variables overlook workers' behavior at the item feature level, leading to imprecise characterizations and negatively impacting the quality of learning-from-crowds. This paper proposes a new paradigm of multi-task supervised learning-from-crowds, which eliminates the need for modeling of items's ground truth in worker behavior models. Within this paradigm, we propose a worker behavior model at the item feature level called Mixture of Experts based Multi-task Supervised Learning-from-Crowds (MMLC), then, two aggregation strategies are proposed within MMLC. The first strategy, named MMLC-owf, utilizes clustering methods in the worker spectral space to identify the projection vector of the oracle worker. Subsequently, the labels generated based on this vector are regarded as the items's ground truth The second strategy, called MMLC-df, employs the MMLC model to fill the crowdsourced data, which can enhance the effectiveness of existing aggregation strategies . Experimental results demonstrate that MMLC-owf outperforms state-of-the-art methods and MMLC-df enhances the quality of existing learning-from-crowds methods.

Introduction

Existing learning-from-crowdsmethods can be broadly classified into two categories: weakly supervised and supervised approaches. In the weakly supervised approach, unknown ground truth are treated as hidden variables. This involves utilizing statistics from workers' noisy answers to calculate results directly. Alternatively, it entails creating worker behavior models and employing unsupervised learning methods such as the EM algorithm to estimate unknown parameters and infer the ground truth. The weakly supervised approach can further be classified into statistical learning and deep learning methods based on whether considering the features of items. Statistical learning methods, such as MV (Imamura, Sato, and Sugiyama 2018), DS (Dawid and Skene 1979), and HDS (Karger, Oh, and Shah 2011; Li and Yu 2014), do not incorporate item features. In contrast, deep learning methods like Training Deep Neural Nets (Gaunt, Borsa, and Bachrach 2016) take item features into account. In the supervised approach, a classifier model is first constructed with item features as the input and ground truth as the output. Then, a worker behavior model is created based on a confusion matrix that establishes the relationship between one item's ground truth and the worker labels. On this basis, supervised learning is implemented using the classifier model and the worker behavior model by treating the worker labels as supervisory information. Finally, the output of the classifier model is used as the inferred ground truth. In recent years, various learning-from-crowds methods based on supervised learning have been proposed, such as Crowdlayer(Rodrigues and Pereira 2018), CoNAL(Chu, Ma, and Wang 2021), and UnionNet(Wei et al. 2022). However, the worker behavior model based on the confusion matrix faces challenges in effectively capturing variations in feature characteristics across different items. Neglecting these variations in worker behavior under different conditions can result in inaccurate representations of worker behavior, consequently impacting the quality of truth inference. For example, in handwritten digit recognition, workers generally have high accuracy. Suppose there are two items: one closely resembles the digit "1," but its ground truth is actually "7," and the other is a normal "7." The former receives many labels as "1," while the latter rarely gets labeled as "1." Under the worker behavior model based on the confusion matrix, it is difficult to model the labeling behavior of such difficult items accurately. Therefore, there is a high probability that the model will interpret the former with "1" as the ground truth, leading to incorrect judgments. The quality of aggregation strategies is influenced by uncertainty from hidden variables, the method's data adaptability, and the accuracy in characterizing worker behavior. The purpose of this paper is to develop a supervised model that can achieve high-quality truth inference with a worker behavior model at the item feature level.

In this paper, we propose Multi-task Supervised Learning-from-crowds (MLC), a novel paradigm for crowdsourcing learning. Unlike the traditional paradigm, MLC does not rely on the ground truth of the items but instead focuses on understanding the unique behavior of individual work-

*Corresponding author

ers across different items. When multiple workers handle the same item, they share the item's features, leading to a multi-task learning paradigm. Within this paradigm, we propose a method called Mixture of Experts-based Multi-task Supervised learning-from-crowds (MMLC). MMLC does not utilize a single classifier but instead creates multiple expert modules. The outputs from these expert modules serve as the *bases* of the worker spectral space. Each worker is represented by his or her projection vector in the spectral space that characterizes their behavior. The worker behavior model provides a more precise depiction of their behavior across different items by accurately modeling the workers' behavior on item features. However, it is important to note that the model itself cannot determine the ground truth. To address this limitation, we introduce two aggregation strategies based on MMLC. The first strategy involves analyzing the distribution of workers' projections in the worker spectral space. We identify the projection of the oracle worker by applying clustering methods, and consider its labels as the ground truth. This approach is referred to as Oracle Worker Finding of MMLC (MMLC-owf). The second strategy leverages the sparsity of crowdsourced data to fill the original dataset with MMLC outputs, generating a new crowdsourcing dataset. Existing learning-from-crowds methods can then be applied within this framework, which is called Data Filling of MMLC (MMLC-df). The main contributions are as follows:

- We introduce a novel paradigm of multi-task supervised learning-from-crowds and propose a novel worker behavior model called MMLC based on feature-level behavior modeling.
- We leverage MMLC to identify the oracle worker for labeling items as the ground truth, referred to as MMLC-owf. Experimental results demonstrate that the labels obtained using this method exhibit higher quality compared to state-of-the-art methods.
- We introduce an aggregation framework called MMLC-df, which leverages the MMLC model to fill sparse crowdsourced data. This framework then applies aggregation methods to determine the ground truth. Experimental results demonstrate that MMLC-df significantly enhances learning-from-crowds methods, leading to higher quality results.

Related Work

Weakly supervised approaches: These approaches model the relationship between workers' noisy responses and the true labels, treating the ground truth as a latent variable and employing weak supervision techniques to deduce it. MV (Sheng et al. 2017) is a commonly used method that assumes the most frequent response as the ground truth, but it fails to account for worker variability. To address this limitation, (Tao, Jiang, and Li 2020) proposed a model that separately considers the majority and minority responses, factoring in the labeling quality of workers. The DS (Dawid and Skene 1979) uses a confusion matrix to characterize worker behavior and estimates parameters with the EM algorithm to infer the ground truth. HDS (Raykar et al. 2010) posits equal chances of erroneous worker choices, refining the confusion matrix with this assumption. GLAD (Whitehill et al. 2009), conversely, factors in both the proficiency of workers and the inherent challenge of tasks, utilizing a sigmoid function to depict worker behavior and the EM algorithm to ascertain the ground truth. Various weakly supervised learning-from-crowds methods integrate deep learning to deduce the ground truth, initiating with strategies that aggregate the noisy labels into an initial answer table. Post this, a neural network is trained for classification, leveraging the curated label set. While these methods (Gaunt, Borsa, and Bachrach 2016; Ghiassi et al. 2022; Zhu, Xue, and Jiang 2023) propose label reliability metrics that significantly influence the outcomes of crowd-based learning. Additionally, (Xu et al. 2024) presents a two-stage method that uses a multi-centroid grouping penalty to incorporate subgroup structures for tasks and workers in inferring the ground truth. While weakly supervised methods have seen successes, they are often hampered by sparse data and the treatment of ground truth as a hidden variable, which limits their accuracy.

Supervised approaches: These methods focus on creating a classifier to link item features with the ground truth and a worker behavior model using a confusion matrix to reflect the relationship between true labels and worker responses. These responses act as supervision, allowing for joint training of both models in a supervised manner, with the classifier's output serving as inferred ground truth(Chu and Wang 2021; Ibrahim, Nguyen, and Fu 2023). Techniques like the Expectation-Maximization (EM) algorithm are used for label aggregation and classifier training. Crowd-layer(Rodrigues and Pereira 2018) replaces the traditional confusion matrix with a crowd layer, integrating label reasoning and classifier training for more accurate results. Tan and Chen(Tanno et al. 2019) enhance model accuracy by applying confusion and labeled transfer matrices alongside classifier predictions. Other approaches(Gao et al. 2022; Cao et al. 2023) introduce worker weight vectors and focus on modeling label reliability to estimate worker abilities. UnionNet(Wei et al. 2022) aggregates worker annotations into a parameter transfer matrix to facilitate training. The CoNAL method(Chu, Ma, and Wang 2021) categorizes noise into common and individual types, effectively managing diverse labeling noise. Despite its strengths, supervised learning faces challenges in precisely characterizing worker behavior, making it difficult to achieve optimal results when candidate answers are poorly differentiated.

Departing from the traditional latent variable approach to ground truth, our method concentrates on the interplay between worker behavior and item features. This shift enables us to develop a supervised learning framework from the crowd, yielding a nuanced worker behavior model. With this model, we pinpoint the oracle workers capable of precise labeling, thus ascertaining the ground truth.

Problem Formulation

Our main goal is to obtain a worker behavior model and achieve joint learning of worker abilities by utilizing multi-task learning from crowds to aggregate the ground truth. Let

$\mathcal{W} = \{w_j\}$ denote the worker set, where w_j represents an individual worker, and $\mathcal{X} = \{x_i\}$ denote the set of items, where x_i represents a single item to be labeled. The labels for each item belong to the category set $\mathcal{K} = \{k\}$. We use y_{ij} to denote the category label assigned by worker w_j to item x_i. We have an indicator function I_{ij}, where $I_{ij} = 1$ if y_{ij} exists and $I_{ij} = 0$ otherwise. Consequently, we obtain the crowdsourced triples dataset $\mathcal{D} = \{\langle x_i, w_j, y_{ij}\rangle | I_{ij} = 1\}$. With regards to learning from crowd in crowdsourcing, we provide the following definition:

Definition 1 (Problem of Learning-from-Crowds (LC Problem) *By modeling and learning from the crowdsourced label dataset $\mathcal{D}$, the problem of learning-from-crowds aims to find a function $g^* : \mathcal{X} \to \mathcal{K}$ such that*

$$g^* = \arg\min_{g \in \mathcal{H}} \sum_{i=1}^{|\mathcal{X}|} \mathcal{L}\left(\hat{z}_i, g(x_i)\right) + \lambda \|g\|_{\mathcal{H}}. \tag{1}$$

Here, $\mathcal{H}$ denotes the hypothesis space of functions, $\|\cdot\|_{\mathcal{H}}$ denotes the norm of hypothesis space, λ is the regularization coefficient, $\mathcal{L}$ is the loss function, and $\hat{z}_i = t_i(\mathcal{D})$ is the estimation of label z_i for item x_i from learning on the dataset. Since the crowdsourced LC problem is an unsupervised learning-from-crowds problem without supervised information, the estimation of ground truth is utilized instead of the goal of learning.

Definition 2 (Problem of Multi-task supervise Learning-from-Crowds (MLC Problem)) *Let $\mathcal{S}_j = \{(x_i, y_{ij})\}_{x_i \in \mathcal{X}_j}$ denote the crowdsourced training dataset for worker w_j, where $\mathcal{X}_j = \{x_i | I_{ij} = 1\}$. The labels provided by worker j can be regarded as the j-th task for the corresponding item. Consequently, we obtain the dataset as $\mathcal{S} = \bigcup_j \mathcal{S}_j$. The problem of multi-task supervised learning-from-crowds is to find a worker behavior function $f^* \in \mathcal{H}$ such that*

$$f^* = \arg\min_{f \in \mathcal{H}} \sum_{w_j \in \mathcal{W}} \frac{1}{|\mathcal{X}_j|} \sum_{x_i \in \mathcal{X}_j} \mathcal{L}\left(y_{ij}, f_{w_j}(x_i)\right) + \lambda \|f\|_{\mathcal{H}}, \tag{2}$$

where $\mathcal{H}$ is a vector-valued reproducing kernel Hilbert space with functions $f : \mathcal{X} \to \mathcal{K}^{|\mathcal{W}|}$ having components $f_j : \mathcal{X} \to \mathcal{K}$.

We can observe that the solution to the MLC problem does not directly address the LC problem. Therefore, we provide two approaches to tackle this issue. The first approach is to identify an oracle worker w_{oracle} based on the distribution of workers. We then consider the labels provided by this oracle worker as the ground truth, that is,

$$g^*(x_i) = f^*_{w_{oracle}}(x_i). \tag{3}$$

The second approach considers the sparsity characteristic of crowdsourced data, where workers do not annotate every item. Consequently, we can utilize the results of MLC to generate a new dataset for inference. The new crowdsourced data can be defined as follows:

$$\mathcal{D}' = \mathcal{D} \cup \left\{ \langle x_i, w_j, \hat{y}_{ij} \rangle \middle| \hat{y}_{ij} = f^*_{w_j}(x_i), I_{ij} = 0 \right\}. \tag{4}$$

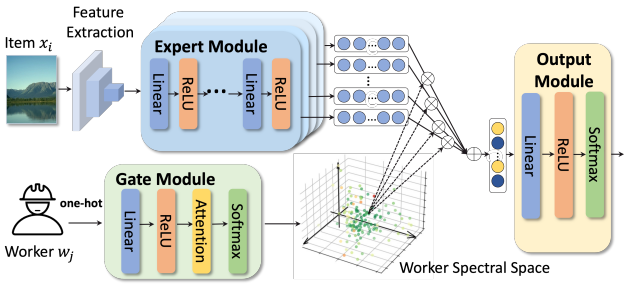

Figure 1: Model Structure of MMLC.

Proposed Methodology

To address the MLC problem, we propose a Mixture of Experts based Multi-task Supervised Learning-from-Crowds (MMLC) model. This model utilizes mixture of experts to characterize the varying attention of workers towards different item features, aiming to capture the feature-level behavior differences of workers when dealing with various items. The framework of the model is shown in Fig. 1. It consists of three main modules: expert module, gate module, and output module.

In the expert module, each item is processed by a feature extractor to obtain an item feature vector x_i. Then, the item feature vector is fed into m expert modules, where each module captures the unique characteristics of worker behavior associated with different feature information. Each expert module performs transformations and compressions on the feature vector, resulting in the output matrix of the expert module: $\boldsymbol{U}(x_i) = (\boldsymbol{u}_1(x_i), \boldsymbol{u}_2(x_i), ..., \boldsymbol{u}_E(x_i))$. Each expert sub-module follows the same structure, consisting of multiple layers of fully connected neural networks with ReLU activation functions in each layer. For each expert sub-module $\boldsymbol{u}_e$, the high-dimensional feature vector x_i is transformed into a low-dimensional vector $\boldsymbol{u}_e(x_i)$.

In the gate module, a gate network is constructed to control the selection of expert modules. This gate network takes worker data as input and generates a projection vector of the worker in the worker spectral space, with a length of E. The *bases* of the worker spectral space are the outputs of the expert sub-modules. Specifically, the module takes the one-hot encoded vector representing each worker w_j as input. After passing through a fully connected ReLU layer, the data proceeds through an attention layer and a softmax layer. Finally, it produces a worker projection vector $\boldsymbol{v}(w_j) = (v_1(w_j), v_2(w_j), ..., v_E(w_j))^T$ with a length of E. The projection of worker w_j in the worker spectral space with the expert sub-modules as the *bases* is:

$$proj_{\boldsymbol{U}(x_i)}(w_j) = \sum_{e=1}^{E} v_e(w_j)\boldsymbol{u}_e(x_i). \tag{5}$$

Here, the attention layer helps reduce the model's dependence on unimportant or redundant features, improving the model's efficiency and accuracy by focusing on the most useful information.

In the output module, the worker's labels for the item are generated. The output module generates labels for each

worker based on their chosen expert modules. It involves mapping worker behavior through the gate network, which includes weighting and summing the outputs of the expert modules. Subsequently, through a fully connected ReLU layer and a softmax layer, the model produces the label output of worker w_j for item x_i as follows:

$$f_{w_j}(x_i|\mathbf{\Theta}) = \boldsymbol{o}\left(proj_{\boldsymbol{U}(x_i)}(w_j)\right), \tag{6}$$

where $\boldsymbol{o}(\cdot)$ denotes the output function, and $\mathbf{\Theta}$ is the parameter set of the functions $\boldsymbol{U}$, $\boldsymbol{v}$, and $\boldsymbol{o}$ within the MMLC model. The MMLC model deals with a classification problem with $|\mathcal{K}|$ categories. The network's output is a $|\mathcal{K}|$-dimensional vector, where each element represents the predicted probability of a category.

The model's loss function combines a cross-entropy loss term with a regularization term. The loss function is formulated as follows:

$$\mathcal{L}_{\mathbf{\Theta}} = -\frac{1}{|\mathcal{D}|}\sum_{w_j\in\mathcal{W}}\sum_{x_i\in\mathcal{X}_j}\sum_{k\in\mathcal{K}} y_{ij}^k \log\left(f_{w_j}(x_i|\mathbf{\Theta})\right) + \lambda\|\mathbf{\Theta}\|_F, \tag{7}$$

The first term denotes the multi-class cross-entropy loss, while the second term represents the regularization of the model's parameter set $\mathbf{\Theta}$ to prevent overfitting. In the equation, λ is the regularization coefficient, and $\|\cdot\|_F$ denotes the Frobenius norm. By minimizing the loss function, we can obtain the final model $\mathcal{M}^* : f(\cdot|\mathbf{\Theta}^*)$. This model uses the function $f_{w_j}(x_i|\mathbf{\Theta}^*)$ to predict the labels of worker w_j for item x_i, where $\mathbf{\Theta}^*$ represents the optimized parameters.

MMLC with Oracle Worker Finding (MMLC-owf): The MMLC model does not directly generate the ground truth for inference. To address this issue, this section proposes a method for inferring the ground truth by identifying the oracle worker's projection vector in the worker spectral space. Specifically, each worker is theoretically associated with a projection in the worker spectral space, representing their unique characteristics. Workers are distributed in the spectral space. We assume the existence of an omniscient oracle worker who possesses a projection vector in the spectral space and is capable of providing the ground truth in the MMLC model. Therefore, by identifying the projection vector of the oracle worker in the worker spectral space, we can consider its output as the inferred truth. If we treat any worker as a random expression of the oracle worker's error, then the center of the worker's distribution projected onto the spectral space can be regarded as the projection vector of the oracle worker, that is,

$$\boldsymbol{v}_{oracle} = \boldsymbol{\tau}\left(\boldsymbol{v}(\mathcal{W})\right), \tag{8}$$

where the function $\boldsymbol{\tau}(\cdot)$ is used to determine the distribution center, which can be found using methods such as kernel density estimation, mean, median, etc. According to the MMLC model, the outcome of the Oracle Worker Finding method (MMLC-owf) for inferring the ground truth of item x_i can be expressed as follows:

$$f_{w_{oracle}}(x_i|\mathbf{\Theta}^*) = \boldsymbol{o}\left(\boldsymbol{U}(x_i)\boldsymbol{v}_{oracle}\right). \tag{9}$$

MMLC with Data Filling (MMLC-df): In addition to the MMLC-owf method, we propose a method using data filling under the MMLC model called MMLC-df, which utilizes the sparsity of crowdsourced data. A new crowdsourced dataset $\mathcal{D}'$ is constructed through data filling as follows:

$$\mathcal{D}' = \mathcal{D} \cup \left\{\langle x_i, w_j, \hat{y}_{ij}\rangle \middle| \hat{y}_{ij} = f_{w_j}(x_i|\mathbf{\Theta}^*), I_{ij} = 0\right\}. \tag{10}$$

Subsequently, any learning-from-crowds method applied to this new crowdsourced dataset can infer higher-quality ground truth compared to that obtained from the original data.

Experiments

We verify the effectiveness of our method through experiments[1]. We compare our learning-from-crowds method MMLC-owf with the following baselines: MV(Sheng, Provost, and Ipeirotis 2008) directly uses majority voting to determine the ground truth; DS(Dawid and Skene 1979) employs a confusion matrix to characterize the labeling behavior of workers and uses the EM algorithm to infer the ground truth; HDS(Karger, Oh, and Shah 2011) simplifies the DS method by assuming that each worker has the same probability of being correct under different truth values and equal probabilities for incorrect options; FDS(Sinha, Rao, and Balasubramanian 2018) is a simple and efficient algorithm based on DS, designed to achieve faster convergence while maintaining the accuracy of truth inference; MaxMIG(Cao et al. 2019) utilizes the EM algorithm to integrate label aggregation and classifier training; CoNAL(Chu, Ma, and Wang 2021) distinguishes between common noise and individual noise by predicting a joint worker confusion matrix using classifiers; CrowdAR(Cao et al. 2023) estimates worker capability features through classifier prediction and models the reliability of joint worker labels.

We compare MMLC-df with the following baselines: G_MV (Sheng 2011) utilizes truth inference results from the MV algorithm to evaluate worker ability and assign new labels accordingly; G_IRT (Baker, Kim et al. 2017) utilizes joint maximum likelihood estimation to estimate parameters of the IRT model, such as worker abilities and item difficulties, and generates new labels based on these parameters; TDG4Crowd (Fang et al. 2023) learns the feature distributions of workers and items separately using worker models and item models. An inference component is used to learn the label distribution and generate new labels.

We identified three representative datasets that exemplify different types of crowdsourcing scenarios and data characteristics. *LableMe* (Rodrigues and Pereira 2018; Russell et al. 2008): This dataset consists of 1000 images categorized into 8 classes, with a total of 2547 labels provided by 59 workers. Each image is represented by 8192-dimensional features extracted using a pre-trained VGG-16 model. *Text* (Dumitrache, Aroyo, and Welty 2018): This dataset comprises 1594 sentences extracted from the CrowdTruth corpus, categorized into 13 groups. The dataset includes 14,228 labels provided by 154 workers. Each sentence is represented by 768-dimensional features extracted using a pre-trained BERT model. *Music* (Rodrigues, Pereira,

[1]Our code is available at https://github.com/Crowds24/MMLC.

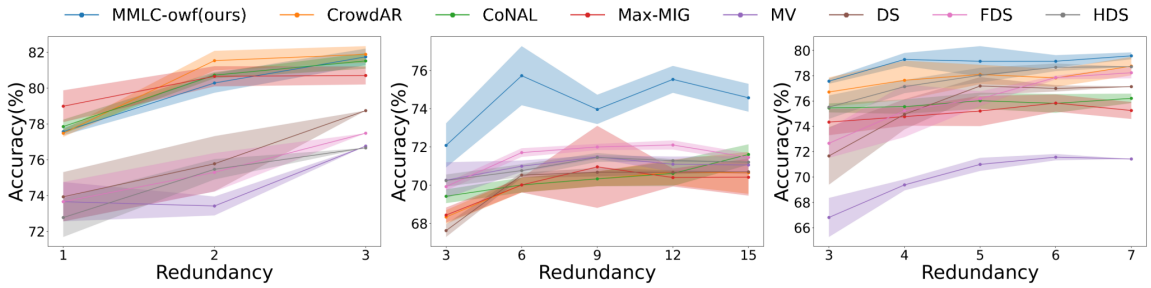

Figure 2: Accuracy Under Various Redundancies. (Left:*LableMe*, Center:*Text*, Right: *Music*)

Method	*LableMe*	*Text*	*Music*
MV	76.76	71.33	71.42
DS	79.73	70.70	77.14
FDS	77.78	71.45	77.57
HDS	76.66	71.20	78.01
Max-MIG	80.02±0.68	70.31±0.23	74.22±0.57
CoNAL	81.46±0.49	72.75±0.49	76.02±0.36
CrowdAR	**82.14±0.36**	70.48±0.42	78.54±0.59
MMLC-owf	81.74±0.47	**74.31±0.39**	**79.14±0.21**

Table 1: Accuracy of Learning-from-Crowds Methods on Three Crowdsourced Datasets.

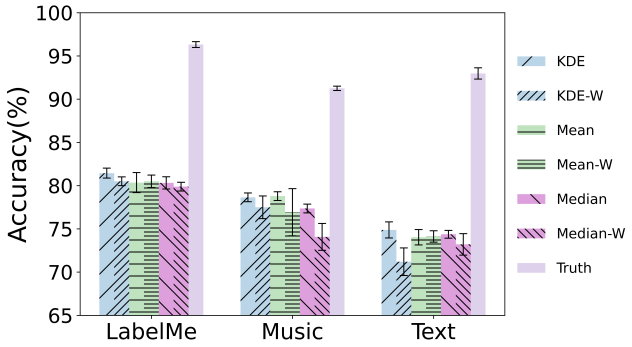

Figure 3: Accuracy of MMLC-owf with Various Clustering Methods on Three Crowdsourced Datasets.

and Ribeiro 2014): This dataset consists of 700 music compositions, each with a duration of 30 seconds, and categorized into 10 groups. It includes 2,945 labels provided by 44 workers. Each music composition is represented by 124-dimensional features extracted using the Marsyas (Rodrigues, Pereira, and Ribeiro 2013) music retrieval tool.

To accommodate the feature scales of the three experimental datasets, our model's architecture varies accordingly. For the *LableMe* dataset, our model employs 16 expert modules, each comprising 3 fully connected ReLU layers, with a final layer output dimension of 32. For the *Text* and *Music* datasets, we utilize 10 expert modules. Each module consists of 3 and 2 fully connected ReLU layers, with output dimensions of 32 and 16, respectively. We adopt the settings from the learning-from-crowds methods Max-MIG, CoNAL, and CrowdAR, we adopt the settings from their source code for the *LableMe* and *Music* datasets. Since there is no source code available for the *Text* dataset, we adopt the settings used in the *LableMe* dataset. The hyperparameters mainly follow the expert configuration from Google's MMoE model(Ma et al. 2018) and the classifier setup of CrowdAR. Regarding the TDG4Crowd data filling algorithm, we utilize the settings from its source code. The remaining methods do not use deep network structures and rely on default settings.

Evaluation of Oracle Worker Finding(MMLC-owf)

Main Result: Our method, MMLC-owf, was evaluated alongside seven other methods through five rounds experiments. The average accuracy results are shown in Tab. 1. In our method, we utilized kernel density estimation (KDE) to compute the projection vector of the oracle worker in the worker spectral space. Our method, MMLC-owf, achieved the highest accuracy in the *Text* and *Music* datasets. In the *LableMe* dataset, it ranked second, with only a 0.4% difference from the top-performing CrowdAR method. Deep learning-based methods typically produce better results when analyzing datasets with high-dimensional item features like *LableMe*. In datasets with fewer features, the advantage of deep learning methods was not significant.

Impact of Redundancy: We examine how varying levels of redundancy affect the accuracy of our method. Due to the varying redundancy of data items, a maximum redundancy parameter R is set. We randomly keep R labels for items with more than R labels and discard the rest. This process generates a dataset with a maximum redundancy R. By conducting five repeated experiments and averaging the accuracy and standard deviation, the results are shown in Fig. 2. As the average number of worker responses increases, all methods show an upward trend in their results. The analysis of various redundancy levels across the datasets indicates that higher redundancy levels are more advantageous for our method. Our method can effectively utilize worker behavior descriptions on datasets with higher redundancy but may face underfitting on datasets with lower redundancy.

Clustering Methods in Oracle Worker Finding: Our method, MMLC-owf, utilizes a clustering method to determine the center of the distribution of the projection vector of workers in the worker spectral space as the projection vector

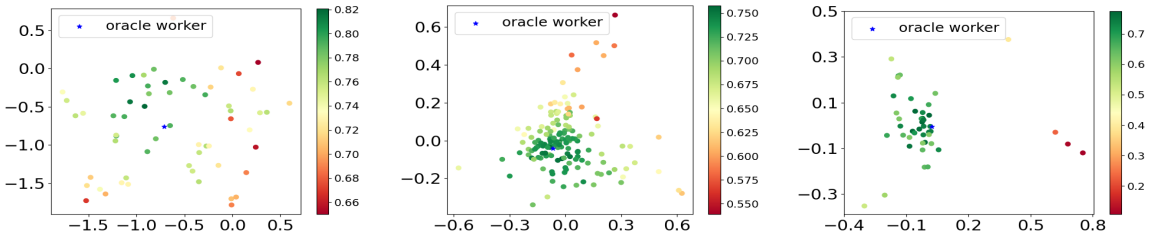

Figure 4: Worker Scatter Plot in Worker Spectral Space.(Left:*LableMe*, Center:*Text*, Right: *Music*)

of the oracle worker. Here, we examine how different clustering methods affect learning-from-crowds outcomes. We compare three clustering methods: kernel density estimation (KDE), Mean, and Median, as well as their worker-weighted variants: KDE-W, Mean-W, and Median-W. Worker weights are calculated based on the proportion of items answered by each worker relative to the total number of items, considering data imbalance. In addition, The parameters of the expert modules and output modules are fixed and we optimize the projection vector in the worker spectral space using the ground truth of the items as the best performance of our model for clustering. This result is referred to as "Truth." By conducting five repeated experiments and averaging the results, as shown in Fig. 3. For example, in the *LableMe* dataset, the oracle worker's projection vector is derived using KDE, KDE-W, Median, Median-W, Mean, and Mean-W. The MMLC-owf uses the oracle worker to generate ground truth. The quality of the ground truth obtained by the following five methods in each dataset is very similar, but the oracle worker generated using KDE produces the highest quality ground truth in each dataset. The "Truth" method, which represents the theoretical upper limit with clustering methods, achieved accuracy rates of 96.32%, 91.25%, and 92.97% on the three datasets respectively.The quality of the ground truth generated by oracle workers using six clustering methods still slightly deviates from theoretical upper limits. This implies that the MMLC-owf method can provide high-quality ground truth by optimally projecting the worker spectral space, approaching the theoretical upper limit. The model has strong expressive ability, with a small gap between the projected spectral space and the ground truth. There is potential to enhance MMLC-owf by choosing a more effective clustering method.

Worker Distribution in Worker Spectral Space: We assume that each worker is an oracle worker with random errors in their expression. The center of the distribution of workers projected onto the spectral space corresponds to the projection vector of the oracle worker. To validate this assumption, we employed the IOSMAP dimensionality reduction method to reduce the worker projection vectors obtained from the MMLC model to 2D, resulting in the scatter plot shown in Fig. 4. We calculated the accuracy of each worker on the dataset, where the closer the point's color on the graph is to green, the worker's accuracy is higher. The closer the point's color is to red, the lower the worker's accuracy. The plot also shows the projection obtained by the KDE method for the oracle worker, represented by blue asterisks. From the distribution of worker projections, although the shapes of the distributions differ across datasets, there is a noticeable trend where workers with higher accuracy tend to cluster closer to the projection of the oracle worker. This observation demonstrates a clear tendency towards aggregation and provides some degree of confirmation for the validity of our assumption.

Evaluation of Data Filling (MMLC-df)

Main Results: We compared MMLC-df with three filling methods: G_MV, G_IRT, and TDG4Crowd. We used the filled data with eight learning-from-crowds methods from the previous section to infer the ground truth. We used three real datasets and applied eight truth value inference methods to infer the ground truth, resulting in a total of 24 scenarios. We conducted five rounds of experiments, and the mean and variance of all ground truth accuracy are presented in Tab. 2. It can be observed that our MMLC-df framework achieves enhanced performance compared to the original data in 100% of the scenarios, with 79.2% of the scenarios achieving best enhancement. On the other hand, G_MV, G_IRT, and TDG4Crowd achieve enhanced results in 50%, 62.5%, and 75% of the scenarios, respectively. In terms of the enhancement magnitude, our method performs the best on the *Text* dataset. While other methods may achieve better results in certain scenarios for other datasets, their performance is unstable, and there are cases where the results deteriorate. This indicates that our MMLC-df framework demonstrates good stability and consistent performance.

Impact of Data Filling's Density: Our MMLC-df framework leverages the sparsity of crowdsourced data for data filling. To clarify, we define the data density of non-empty crowdsourced data as $d_{\mathcal{D}} \in (0, 1]$ and $d_{\mathcal{D}} = \frac{|\mathcal{D}|}{|\mathcal{W}|\times|\mathcal{X}|}$. The data densities of the three original datasets *LabelMe*, *Text*, and *Music* are 0.0431, 0.0579, and 0.0956, respectively. The original data seems sparse. We gradually fill the data until reaching a data density to 1 for the analysis of its impact of data density on the results. We set a threshold for the number of items to be filled, denoted as $n_{interval}$. For workers with items exceeding this threshold, we replace their labeled items with predicted values. By adjusting the threshold from large to small, we gradually fill the data until all workers have completed their items, achieving a data density of 1. Due to the large amount of filled data, deep learning meth-

Data	Method	Original	G_MV	G_IRT	TDG4Crowd	MMLC-df
LableMe	MV	76.76	-0.03±0.41	-2.87±0.58	+1.88±0.41	**+3.89±0.17**
	DS	79.73	-2.87±0.77	+0.06±0.26	-1.05±0.37	**+1.72±0.20**
	FDS	77.78	+0.07±0.92	+0.20±0.62	+0.93±0.23	**+3.38±0.21**
	HDS	76.66	-0.23 ±0.44	+0.18±0.33	+2.02±0.32	**+2.48±0.36**
	Max-MIG	80.02±0.68	+1.90±0.79	+1.76±0.14	+2.23±0.21	**+4.81±0.21**
	CoNAL	81.46±0.49	-1.92±0.41	-0.78±0.68	+0.02±0.72	**+1.73±0.57**
	CrowdAR	82.14±0.36	**+3.21±0.28**	+2.52±0.28	-0.69±0.56	+2.49±0.41
	MMLC-owf	81.74±0.47	-2.91±0.46	+0.05±0.61	-1.41±0.54	**+1.65±0.66**
Text	MV	71.33	+0.53±0.31	-0.26±0.44	-0.01±0.08	**+3.35±0.31**
	DS	70.72	+1.80±0.53	+0.01±0.51	+0.38±0.21	**+4.02±0.26**
	FDS	71.45	+0.88±0.62	-0.63±0.48	-0.16±0.04	**+3.26±0.35**
	HDS	71.21	+0.24±0.17	-0.39±0.34	-1.30±0.48	**+2.18±0.19**
	Max-MIG	70.31±0.23	-1.24±0.70	-1.87±0.17	+0.33±0.64	**+3.88±0.51**
	CoNAL	72.75±0.49	-1.33±0.22	-2.27±0.43	+0.62±0.32	**+2.16±0.61**
	CrowdAR	70.48±0.42	+0.37±0.62	-0.33±0.21	+1.96±0.12	**+3.63±0.39**
	MMLC-owf	74.31±0.39	-2.04±0.41	-0.15±0.37	+0.39±0.26	**+1.46±0.52**
Music	MV	71.42	-0.95±0.35	+6.72±0.23	+5.73±0.28	**+7.58±0.45**
	DS	77.14	-5.95±0.31	+0.27±0.43	+1.57±0.13	**+2.41±0.41**
	FDS	77.57	-6.47±0.64	+0.41±0.57	+0.81±0.08	**+1.57±0.22**
	HDS	78.01	-7.25±0.87	+0.52±0.43	**+0.70±0.25**	+0.12±0.54
	Max-MIG	74.22±0.57	+1.42±0.50	+0.32±0.98	**+5.54±0.72**	+4.77±0.41
	CoNAL	76.02±0.36	**+7.97±0.54**	+4.65±0.54	+5.55±0.51	+6.37±0.66
	CrowdAR	78.54±0.59	+1.46±0.42	+2.31±0.20	+1.97±0.11	**+2.47±0.28**
	MMLC-owf	79.14±0.21	+0.24±0.37	+0.75± 0.32	**+1.62±0.31**	+1.43± 0.42

Table 2: The Change of Accuracy After Data Filling.

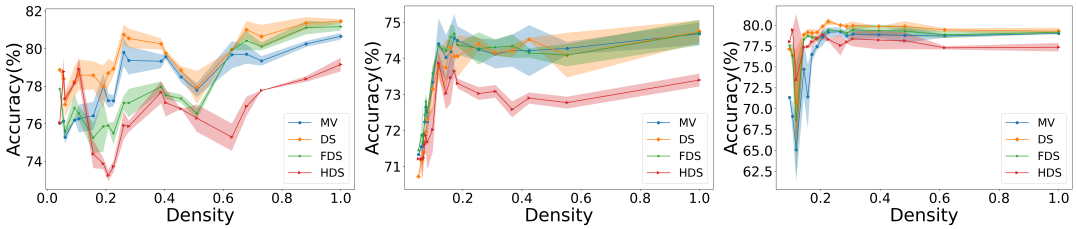

Figure 5: Accuracy with Various Data Filling's Density. (Left:*LableMe*, Center:*Text*, Right: *Music*)

ods can be time consuming. The accuracies obtained by various methods show a similar trend to the density transformation. Therefore, we conducted this experiment using only statistical machine learning methods. The experiment is repeated in five rounds, and the average accuracy and standard deviation are shown in Fig. 5. The trends are generally consistent across all methods, but the variations differ significantly among different datasets. In *Text* dataset, as density increases, the algorithm's accuracy stabilizes rapidly and then reaches a plateau. In the *LableMe* dataset, accuracy fluctuates significantly as density increases. Higher density often improves accuracy. In *Music* dataset, as density increases, accuracy initially fluctuates rapidly before stabilizing. The *Text* data filling performs the best, likely due to the larger scale of this dataset compared to the other two, resulting in a more significant impact.

Conclusion

This paper introduces a novel crowd-learning paradigm called MLC. Within this paradigm, we propose a feature-level worker behavior model called MMLC. Based on this model, we develop two learning-from-crowds methods: MMLC-owf, which uses oracle worker finding, and a framework MMLC-df based on data filling. Experimental results demonstrate the superior performance of MMLC-owf compared to other methods. Furthermore, we assess the theoretical upper performance limit of the MMLC-owf method, demonstrating its potential to enhance clustering method selection and validate its strong performance. The experiments also validate the effectiveness and stability of the MMLC-df framework in enhancing learning-from-crowds methods through data filling. Furthermore, we observed that our model exhibited better performance on datasets with a higher number of annotations per worker.

Acknowledgments

This research has been supported by the Natural Science Foundation of Zhejiang Province, China (Grant No. LQ22F020002, LZ22F020008, and LQ24F020003), the National Natural Science Foundation of China under grant 61976187. Besides, the authors want to thank the anonymous reviewers for the helpful comments and suggestions to improve this paper.

References

Baker, F. B.; Kim, S.-H.; et al. 2017. *The basics of item response theory using R*, volume 969. Springer.

Cao, P.; Xu, Y.; Kong, Y.; and Wang, Y. 2019. Max-mig: an information theoretic approach for joint learning from crowds. *arXiv preprint arXiv:1905.13436*.

Cao, Z.; Chen, E.; Huang, Y.; Shen, S.; and Huang, Z. 2023. Learning from Crowds with Annotation Reliability. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2103–2107.

Chu, Z.; Ma, J.; and Wang, H. 2021. Learning from crowds by modeling common confusions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5832–5840.

Chu, Z.; and Wang, H. 2021. Improve learning from crowds via generative augmentation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 167–175.

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.

Dumitrache, A.; Aroyo, L.; and Welty, C. 2018. Crowdsourcing semantic label propagation in relation classification. *arXiv preprint arXiv:1809.00537*.

Fang, Y.; Shen, C.; Gu, H.; Han, T.; and Ding, X. 2023. TDG4Crowd: test data generation for evaluation of aggregation algorithms in crowdsourcing. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2984–2992.

Gao, Z.; Sun, F.-K.; Yang, M.; Ren, S.; Xiong, Z.; Engeler, M.; Burazer, A.; Wildling, L.; Daniel, L.; and Boning, D. S. 2022. Learning from multiple annotator noisy labels via sample-wise label fusion. In *European Conference on Computer Vision*, 407–422. Springer.

Gaunt, A.; Borsa, D.; and Bachrach, Y. 2016. Training deep neural nets to aggregate crowdsourced responses. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence. AUAI Press*, volume 242251.

Ghiassi, A.; Birke, R.; Chen, L. Y.; et al. 2022. LABNET: A Collaborative Method for DNN Training and Label Aggregation. In *ICAART (2)*, 56–66.

Ibrahim, S.; Nguyen, T.; and Fu, X. 2023. Deep learning from crowdsourced labels: Coupled cross-entropy minimization, identifiability, and regularization. *arXiv preprint arXiv:2306.03288*.

Imamura, H.; Sato, I.; and Sugiyama, M. 2018. Analysis of Minimax Error Rate for Crowdsourcing and Its Application to Worker Clustering Model. *Cornell University - arXiv,Cornell University - arXiv*.

Karger, D.; Oh, S.; and Shah, D. 2011. Iterative Learning for Reliable Crowdsourcing Systems. *Neural Information Processing Systems,Neural Information Processing Systems*.

Li, H.; and Yu, B. 2014. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939.

Raykar, V.; Yu, S.; Zhao, L.; Valadez, G.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning From Crowds. *Journal of Machine Learning Research,Journal of Machine Learning Research*.

Rodrigues, F.; and Pereira, F. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, 1611–1618.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2013. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12): 1428–1436.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2014. Gaussian process classification and active learning with multiple annotators. In *International conference on machine learning*, 433–441. PMLR.

Russell, B. C.; Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2008. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77: 157–173.

Sheng, V. S. 2011. Simple multiple noisy label utilization strategies. In *2011 IEEE 11th International Conference on Data Mining*, 635–644. IEEE.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 614–622.

Sheng, V. S.; Zhang, J.; Gu, B.; and Wu, X. 2017. Majority voting and pairing with multiple noisy labeling. *IEEE Transactions on Knowledge and Data Engineering*, 31(7): 1355–1368.

Sinha, V. B.; Rao, S.; and Balasubramanian, V. N. 2018. Fast dawid-skene: A fast vote aggregation scheme for sentiment classification. *arXiv preprint arXiv:1803.02781*.

Tanno, R.; Saeedi, A.; Sankaranarayanan, S.; Alexander, D. C.; and Silberman, N. 2019. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11244–11253.

Tao, F.; Jiang, L.; and Li, C. 2020. Label similarity-based weighted soft majority voting and pairing for crowdsourcing. *Knowledge and Information Systems*, 62: 2521–2538.

Wei, H.; Xie, R.; Feng, L.; Han, B.; and An, B. 2022. Deep learning from multiple noisy annotators as a union. *IEEE transactions on neural networks and learning systems*.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J.; and Ruvolo, P. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22.

Xu, Q.; Yuan, Y.; Wang, J.; and Qu, A. 2024. Crowdsourcing Utilizing Subgroup Structure of Latent Factor Modeling. *Journal of the American Statistical Association*, 119(546): 1192–1204.

Zhu, K.; Xue, S.; and Jiang, L. 2023. Improving label quality in crowdsourcing using deep co-teaching-based noise correction. *International Journal of Machine Learning and Cybernetics*, 1–14.