

# Domain-Weighted Majority Voting for Crowdsourcing

Dapeng Tao<sup>✉</sup>, Jun Cheng, Zhengtao Yu<sup>✉</sup>, Kun Yue, and Lizhen Wang, *Member, IEEE*

**Abstract**—Crowdsourcing labeling systems provide an efficient way to generate multiple inaccurate labels for given observations. If the competence level or the “reputation,” which can be explained as the probabilities of annotating the right label, for each crowdsourcing annotators is equal and biased to annotate the right label, majority voting (MV) is the optimal decision rule for merging the multiple labels into a single reliable one. However, in practice, the competence levels of annotators employed by the crowdsourcing labeling systems are often diverse very much. In these cases, weighted MV is more preferred. The weights should be determined by the competence levels. However, since the annotators are anonymous and the ground-truth labels are usually unknown, it is hard to compute the competence levels of the annotators directly. In this paper, we propose to learn the weights for weighted MV by exploiting the expertise of annotators. Specifically, we model the domain knowledge of different annotators with different distributions and treat the crowdsourcing problem as a domain adaptation problem. The annotators provide labels to the source domains and the target domain is assumed to be associated with the ground-truth labels. The weights are obtained by matching the source domains with the target domain. Although the target-domain labels are unknown, we prove that they could be estimated under mild conditions. Both theoretical and empirical analyses verify the effectiveness of the proposed method. Large performance gains are shown for specific data sets.

**Index Terms**—Crowdsourcing, domain knowledge, theoretical guarantee, weighted majority voting (MV).

Manuscript received July 13, 2017; revised December 7, 2017 and April 5, 2018; accepted May 10, 2018. Date of publication June 5, 2018; date of current version December 19, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61772455, Grant 61572486, Grant 61772508, Grant 61472168, Grant 61732005, Grant 61672271, and Grant U1713213, in part by the Yunnan Natural Science Funds under Grant 2016FB105, in part by the Guangdong Technology Project under Grant 2016B010108010, Grant 2016B010125003, and Grant 2017B010110007, in part by the Shenzhen Technology Project under Grant JCYJ20170413152535587, Grant JSGG20160331185256983, and Grant JSGG20160229115709109, in part by the Program for Excellent Young Talents of Yunnan University under Grant WX069051, and in part by the Project of Innovative Research Team of Yunnan Province, CAS Key Technology Talent Program, Shenzhen Engineering Laboratory for 3-D Content Generating Technologies under Grant [2017]476. (*Corresponding authors: Dapeng Tao; Jun Cheng.*)

D. Tao, K. Yue, and L. Wang are with the School of Information Science and Engineering, Yunnan University, Yunnan 650504, China (e-mail: dapeng.tao@gmail.com; lzhwang@ynu.edu.cn).

J. Cheng is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, and also with The Chinese University of Hong Kong, Hong Kong (e-mail: jun.cheng@siat.ac.cn).

Z. Yu is with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China (e-mail: ztyu@hotmail.com).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNNLS.2018.2836969

## I. INTRODUCTION

**L**ABEL, as the supervisory information, is crucial for many modern machine learning methods. Training very complex machine learning models, such as deep networks [1], usually requires large amounts of labeled data [2]. The labeling of very large data sets is becoming a bottleneck for the progress of the machine learning community. Fortunately, crowdsourcing [3] labeling systems provide cheap and fast ways for obtaining large amounts of labeled data, but at the price that the labels obtained via crowdsourcing are often erroneous [4]–[11].

To reduce the label error, multiple annotators are encouraged to provide labels on given observations or objects. Then, more reliable labels can be obtained by employing an aggregation rule [12]–[15]. This strategy is not new and is well known as the wisdom of crowds [16]. For example, when an important and difficult medical problem comes, multiple medical experts are usually independently consulted to reach a reliable decision. Here, we assume that all the experts are biased such that they are more likely to make the right decision. If all the experts have the same competence level or “reputation,” which means that they have the same probability of making the right decision, majority voting (MV) is the optimal rule for aggregating their decisions [17]. If the competence levels are different, the expert with a higher competence level should be counted more. Weighted MV is, therefore, more suitable.

The widely employed crowdsourcing labeling systems are quite different from the above decision-making scenario. Take Amazon Mechanical Turk ([www.mturk.com](http://www.mturk.com)) as an example. It provides a market, where tasks for labeling are distributed and anonymous annotators can choose to complete the task to earn some payment. For example, as shown in Fig. 1, annotators could earn a few cents by clicking the capital city of Australia that they believe to be true. There are key differences lying between the annotators and the medical experts, e.g., the annotators are usually not experts and may be from very different backgrounds. The biases and competence levels of the annotators are thus different and unknown. This brings a challenge of calculating the weights for MV.

Several aggregation methods have been proposed by considering the differences of the competence levels of the nonexpert crowdsourcing annotators, e.g., the popular Dawid–Skene model and its variants [18]–[22], and probabilistic graphical models [19], [23]–[25]. However, they may suffer from the following criticized problems: 1) they are all generative probabilistic models and critically depend on the prior

Select the capital city of Australia

Annotators: 1 1 1 ...




	Sydney	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	Canberra	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Melbourne	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Fig. 1. Illustration of crowdsourcing labeling. Annotators could earn a few cents by clicking the capital city of Australia that they believed to be true.

distribution on the competence level of the annotator; 2) they are observation-independent; 3) most of them employ the EM algorithm [26] and have the local optimality issue; and 4) it is also hard to theoretically guarantee their performances.

In this paper, we propose a simple but efficient method to learn the weights for weighted MV, which does not need any specific prior distribution on the competence level of the annotator, is observation-dependent, and is also theoretically guaranteed to learn the optimal weights under mild conditions. The proposed method is motivated by the work in [10] and [27], where the label noise problem [28]–[33] is studied from a viewpoint of domain adaptation [34]–[38]. The clean sample domain without label noise is treated as the target domain, and the sample domain contaminated with label noise is treated as the source domain. This domain-adaptation framework is powerful in addressing label noise problems once the relationship between clean and noisy domains is clear.

Specifically, in this paper, we model the domain knowledge of different annotators with different distributions and treat the crowdsourcing problem as a multiple-domain adaptation problem. Each annotator is associated with a source domain, whose labels are likely to be noisy. The target domain is assumed to contain the ground-truth labels. Intuitively, the weights for weighted MV can be easily obtained by matching the source domains with the target domain. However, the target-domain labels are usually unknown. If we assume that the label noise is caused by two situations: 1) the observations are difficult to discriminate such that an annotator will have a probability to annotate an incorrect label and 2) that there exist some annotators who (partially) submit random answers to maximize the payment, and also assume that there is no annotator who deliberately submit incorrect labels, we could model the label noise as the observation-dependent random flip label noise and obtain better-than-guessing labels for the target domain. It is easy to prove that the right labels for target domain can be learned with a sufficiently large training sample.

It should be noted that the marginal distribution of the features for both source and target domains is the same. We, therefore, learn the weights for weighted MV by matching the labeling information between source and target domains.

Simple learning algorithms for estimating the weights are proposed, implementing which only requires learning several classifiers, e.g., support vector machines (SVMs). We prove that the weights learned by the proposed methods will converge to the optimal ones as the training sample size goes to infinity and that the estimated labels will converge to the ground-truth labels under mild conditions, which are independent of the number of annotators and provide some insights on choosing annotators. Experiments on both synthetic and real-world data verify the effectiveness of the proposed methods.

The main results and the contributions are summarized in the following.

- 1) We are the first to address the crowdsourcing problem from a viewpoint of multiple-domain adaptation and provide simple and easy-to-understand weighted MV methods for crowdsourcing.
- 2) Efficient learning algorithms for learning the weights for weighted MV are proposed for crowdsourcing. The learned weights will converge to the optimal ones with theoretical guarantees.
- 3) The proposed aggregating rule is observation-dependent. Even if an annotator has not provided any labels for some observations, the proposed method has the ability to account her/his potential altitude for aggregating the final labels because her/his domain knowledge has been modeled.

The rest of this paper is organized as follows. In Section II, we set up the crowdsourcing problem and briefly introduce some background of learning with label noise to better understand the crowdsourcing problem. In Section III, we present algorithm-dependent generalization bounds for MTL. The proofs of our results are presented in Section IV. Experiments and discussions for verifying the efficiency and effectiveness are provided in Section V. Finally, Section VI concludes this paper.

## II. PROBLEM SETUP

In this paper, we study the crowdsourcing problem that there are  $n$  observations and  $m$  annotators and that each annotator will choose one label from  $C$  label classes for each observation. Note that our method also applies to the case that the annotators only partially completed the labeling task. However, to compare with the baseline of MV and for simplicity, we only discuss the fully completed case throughout this paper.

We consider the classical statistical learning problem, where the value of a discrete random variable  $Y$  is to be predicted based on an observation of another real random variable  $X$  and some noisy observations of  $Y$ , where  $Y$  takes values in  $\{1, \dots, C\}$ . Let  $\{X_1, \dots, X_n\}$  be  $n$  i.i.d. random variables having a range contained in  $\mathcal{X}$ , which is the so-called feature space. Let  $\{Y_1, \dots, Y_n\}$  be the ground-truth label variables for  $\{X_1, \dots, X_n\}$ , which is usually unavailable. In addition, let  $\{(Y_{11}, \dots, Y_{1m}), \dots, (Y_{n1}, \dots, Y_{nm})\}$  be the corresponding random variables representing the labels annotated by the annotators, e.g.,  $Y_{ij}$  represents the label to the  $i$ th random variable  $X_i$  annotated by the  $j$ th annotator, which are usually noisy. Let  $\{x_1, \dots, x_n\}$

and  $\{(\tilde{y}_{11}, \dots, \tilde{y}_{1m}), \dots, (\tilde{y}_{n1}, \dots, \tilde{y}_{nm})\}$  be the examples of the random variables  $\{X_1, \dots, X_n\}$  and  $\{(Y_{11}, \dots, Y_{1m}), \dots, (Y_{n1}, \dots, Y_{nm})\}$ . Note that throughout this paper, we will use the notation  $\tilde{y}$  to denote labels that contain noise. One problem of crowdsourcing is to design aggregation rules to predict the right labels for  $\{x_1, \dots, x_n\}$ .

When we have the examples  $\{x_1, \dots, x_n\}$  and  $\{(\tilde{y}_{11}, \dots, \tilde{y}_{1m}), \dots, (\tilde{y}_{n1}, \dots, \tilde{y}_{nm})\}$ , many aggregating rules could be employed to aggregate the multiple noisy labels  $(\tilde{y}_{i1}, \dots, \tilde{y}_{im})$  for the observation of  $x_i$  into a refined or more reliable label. A natural and widely used aggregation rule is MV [39]–[41], which can be formulated as

$$\hat{y}_i = \arg \max_{c \in \{1, \dots, C\}} \sum_{j=1}^m 1_{(\tilde{y}_{ij}=c)} \quad (1)$$

where  $1_{(\cdot)}$  is the indicator function and  $1_{(A)}$  is equal to 1 only when  $A$  holds true otherwise it is equal to 0. Note that throughout this paper, we will use the notation  $\hat{y}$  to denote the estimated value for  $y$ .

Since the annotators might be from very different backgrounds, the competence levels may vary very much. The annotator with a higher competence level should be counted more for the final decision. Weighted MV is, therefore, more suitable for aggregating the noisy labels collected via the crowdsourcing labeling systems. Weighted MV rule [14], [22], [42] can be written as

$$\hat{y}_i = \arg \max_{c \in \{1, \dots, C\}} \sum_{j=1}^m w_j 1_{(\tilde{y}_{ij}=c)} \quad (2)$$

where  $w_i$  is the weight assigned for the  $i$ th annotator.

Note that the quality of the aggregated labels depends not only on the annotated labels but also heavily on the aggregation rule. Thus, the accuracy of the final aggregated labels depends much on the accuracy of the estimated weights. However, since the annotators are anonymous and the ground-truth labels are usually unknown, it is hard to calculate the weights directly. Several methods based on the Dawid–Skene model [18] and probabilistic graphical models [19] have been proposed to infer a marginal posterior distribution over the ground-truth labels by considering the differences of the competence levels of the annotators. However, those aggregation methods are independent of the observations and are also lack of consistent guarantees.<sup>1</sup>

In this paper, we will propose an efficient method to estimate the weights for weighted MV with a consistent guarantee. Specifically, we treat the crowdsourcing problem as a multiple-domain adaptation problem. Each annotator is associated with a source domain. The target domain is assumed to contain the ground-truth labels. Intuitively, the weights for weighted MV can be easily obtained by matching the source domains with the target domain. To show that the proposed method is feasible, we need to introduce some background of learning with label noise [29], [43]–[46] and justify that the right labels could be estimated under mild conditions.

<sup>1</sup>We say that an algorithm has a consistent guarantee if its estimated parameters will converge to the optimal ones, as the training sample size goes to infinity.

*Definition 1 (Observation-Dependent Random Label Noise):* Let  $(X, Y)$  and  $(X, \tilde{Y})$  be the variables representing the clean and noisy data, respectively. The label noise is called observation-dependent random label noise if the right label  $Y = j$  randomly flips to  $\tilde{Y} = i$  depending on  $X$ , which can be modeled as  $P(\tilde{Y} = i | Y = j, X) = \rho_{ij}(X) < 1/C$  for all  $i, j \in \{1, \dots, C\}$ .

The following result [43], [45] plays a central role in supporting our proposed method.

*Theorem 1:* For observation-dependent symmetric random label noise, i.e.,  $\rho_{ij}(X) = \rho_{ji}(X)$  for all  $i, j \in \{1, \dots, C\}$ , any algorithm employing a symmetric loss function that is consistent for classification on the noisy distribution is also consistent on the clean distribution.<sup>2</sup>

Theorem 1 means that a classifier for the right labels can be easily estimated if the label noise introduced by annotators is of observation-dependent symmetric random label noise. We further show that the label noise introduced by annotators can be modeled as observation-dependent random label noise under mild conditions.

We list three conditions that may introduce label noise.

*Assumption 1:* There exist some annotators who (partially) submit random answers to maximize the payment.

*Assumption 2:* The observations are difficult to discriminate such that an annotator will have a probability to annotate an incorrect label.

*Assumption 3:* There exist some annotators who deliberately submit incorrect answers.

If the label noise is caused only by Assumption 1, we could model the label noise as the symmetric random label noise (the flip rates are independent of observations). If the label noise is caused only by Assumptions 1 and 2, we could model the label noise as the observation-dependent random label noise. If all the three assumptions hold, the label noise is hard to model. Our theoretical analysis is focused on Assumptions 1 and 2, which look like strong but would easily happen in practice. Note that existing results on learning with label noise [47]–[51] have good performance on learning the right label for different kinds of label noise. Once we have the target-domain data, the research problem can be focused on estimating the weights for weighted MV.

### III. LEARNING ALGORITHMS

Before presenting our main learning algorithms in this section, we present some preparations to easily understand them. We treat the crowdsourcing problem as a domain-adaptation problem and model the domain knowledge of different annotators with different distributions. Intuitively, the weights for MV can be obtained by matching the source domains with the target domain. It should be noted that the distribution of the features, i.e.,  $X$ , for both source and target domains is the same. We, therefore, learn the weights by matching the labeling information between source and target domains.

<sup>2</sup>A loss function is a symmetric loss function if for any given  $x$ , the sum of the loss on  $(x, 1), \dots, (x, C)$  is a constant. Note that this is hold true for the 0-1 loss. More details can be found in [45].



We characterize the domain knowledge of each annotator by exploiting their labeling function. Lemma 1 could be easily obtained according to Theorem 1.

*Lemma 1:* If the label noise is only caused by Assumption 1, a learning algorithm that is consistent for classification on the noisy distribution is also consistent for classification on the clean distribution.

Lemma 1 shows that under Assumption 1, it is easy to design a learning algorithm to learn a labeling function  $g$  for the target domain, which will converge to the optimal classifier for predicting the ground-truth labels, as the training sample size goes to infinity. Let  $f_j$  be the labeling function for the  $j$ th annotator. The weight for the  $j$ th annotator could be easily learned by matching its labeling function  $f_j$  with the labeling function  $g$ , i.e.,

$$f_j = \arg \min \|g - w_j f_j\|_2^2. \quad (3)$$

We summarize the algorithm for learning the weights as the labeling function-weighted MV (LFWMV) learning algorithm, as shown in Algorithm 1.

---

**Algorithm 1** LFWMV

---

**Input:**  $\{x_1, \dots, x_n\}$  and  $\{(\tilde{y}_{11}, \dots, \tilde{y}_{1m}), \dots, (\tilde{y}_{n1}, \dots, \tilde{y}_{nm})\}$

**Output:**  $\{w_1, \dots, w_m\}$  and  $\{\hat{y}_1, \dots, \hat{y}_n\}$

- 1:  $\hat{g} \leftarrow \text{train } \{(x_1, \tilde{y}_{11}), \dots, (x_1, \tilde{y}_{1m}), \dots, (x_n, \tilde{y}_{n1}), \dots, (x_n, \tilde{y}_{nm})\};$
  - 2: **for**  $j = 1, \dots, m$  **do**
  - 3:    $\hat{f}_j \leftarrow \text{train } \{(x_1, \tilde{y}_{1j}), \dots, (x_n, \tilde{y}_{nj})\};$
  - 4:    $\hat{w}_j = \arg \min_{w_j} \|\hat{g} - w_j \hat{f}_j\|_2^2;$
  - 5: **end for**
  - 6: Normalize  $\{\hat{w}_1, \dots, \hat{w}_n\};$
  - 7: **for**  $i = 1, \dots, m$  **do**
  - 8:   Predict  $\hat{y}_i$  according to  $\sum_{j=1}^m \hat{w}_j \hat{f}_j(x_i);$
  - 9: **end for**
- 

Note that the domain knowledge of different annotators is different, which may partially focus on different parts (or viewpoints) of the observations. It is, therefore, more reasonable to take the feature information into consideration for crowdsourcing; while traditional crowdsourcing methods usually only consider the annotated labels. In LFWMV, the domain knowledge of each annotator is modeled and measured by exploiting the labeling functions, which depend on the observation  $x$  because they are learned from the pairs of  $(x, y)$ . However, more efforts should be placed on learning feature to boost the performance. For example, Ash and Schapire [35] proposed to match domains (via matching label) by learning features and the performance of multiple-domain adaptation has been largely improved.

To further exploit the domain knowledge of annotators, we propose to learn features to better match the source domains of annotators with the target domain and learn the weight  $w_j$  for the  $j$ th annotator by minimizing

$$\begin{aligned} & \| [g(\phi(x_1)), \dots, g(\phi(x_n))] \\ & - w_j [f_j(\phi_j(x_1)), \dots, f_j(\phi_j(x_n))] \|_2^2 \end{aligned} \quad (4)$$

---

**Algorithm 2** DWMV

---

**Input:**  $\{x_1, \dots, x_n\}$  and  $\{(\tilde{y}_{11}, \dots, \tilde{y}_{1m}), \dots, (\tilde{y}_{n1}, \dots, \tilde{y}_{nm})\}$

**Output:**  $\{w_1, \dots, w_m\}$  and  $\{\hat{y}_1, \dots, \hat{y}_n\}$

- 1:  $\{\hat{g}, \phi\} \leftarrow \text{train } \{(\phi(x_1), \tilde{y}_{11}), \dots, (\phi(x_1), \tilde{y}_{1m}), \dots, (\phi(x_n), \tilde{y}_{n1}), \dots, (\phi(x_n), \tilde{y}_{nm})\};$
  - 2: **for**  $j = 1, \dots, m$  **do**
  - 3:    $\{\hat{f}_j, \phi_j\} \leftarrow \text{train } \{(\phi_j(x_1), \tilde{y}_{1j}), \dots, (\phi_j(x_n), \tilde{y}_{nj})\};$
  - 4:    $\hat{w}_j = \arg \min_{w_j} \|[g(\phi(x_1)), \dots, g(\phi(x_n))] - w_j [f_j(\phi_j(x_1)), \dots, f_j(\phi_j(x_n))]\|_2^2;$
  - 5: **end for**
  - 6: Normalize  $\{\hat{w}_1, \dots, \hat{w}_n\};$
  - 7: **for**  $i = 1, \dots, m$  **do**
  - 8:   Predict  $\hat{y}_i$  according to  $\sum_{j=1}^m \hat{w}_j \hat{f}_j(\phi_j(x_i));$
  - 9: **end for**
- 

TABLE I

ILLUSTRATION OF A SUPERIOR OF OBSERVATION-DEPENDENT MV, WHICH MAY CORRECT INCORRECT LABELS BEFORE AGGREGATION

Observations	$o_1$	$o_2$	$o_3$	$o_4$
Ground-truth label	2	1	1	1
Labelled by annotator #1	1	1	1	<b>2</b>
Labelled by annotator #2	2	1	2	1
Labelled by annotator #3	2	1	1	2
Majority voting results	2	2	1	2
Labels predicted by $\hat{f}_1$	1	1	1	<b>1</b>
Labels predicted by $\hat{f}_2$	2	1	2	1
Labels predicted by $\hat{f}_3$	2	1	1	2
Domain based majority voting results	2	1	1	1

as shown in Algorithm 2 and is named as domain-weighted MV (DWMV). Note that, in this paper, we do not focus on how to learn the features. Some more sophisticated feature learning techniques would be investigated in the future.

The proposed aggregating rules are observation-dependent. If an annotator has not provided any labels for some observations, the proposed method has the ability to account the potential altitude of the annotator because its labeling function has been modeled. Moreover, the observation-dependent methods may correct some incorrect labels because of the properties of learning algorithms, e.g., algorithmic robustness [52], which will output similar labels if the input observations are similar. This may help to boost the aggregation performance. In Table I, we employ MV to show an advantage of the observation-dependent method, in which the domain-based MV results outperform that of the traditional MV results because the labeling function  $\hat{f}_1$  learned for annotator #1 has provided a correct label for  $o_4$ . More empirical verifications will be shown in Section V.

We have provided simple and easy-to-understand weighted MV methods for crowdsourcing. Efficient learning algorithms for learning the weights for weighted MV are proposed. We will provide theoretical guarantees for the proposed methods in Section IV. For example, why do not we learn the ground-truth labels by employing the labeling function  $g$  directly? Is the proposed methods useful? Will the learned weights converge to the optimal ones?

#### IV. THEORETICAL ANALYSIS

Before empirically verifying the effectiveness of the proposed methods, in this section, we provide theoretical guarantees to further understand them. Specifically, in Sections IV-A–IV-C, we justify that: 1) the optimal labeling function  $g$  for predicting the ground-truth labels can be learned under mild conditions; 2) the proposed MV methods are superior to directly employing  $\hat{g}$  in predicting the ground-truth labels; and 3) we prove that the weights learned by the proposed method will be identifiable and converge to the optimal ones with a convergence rate of order  $O(\sqrt{1/n})$ , where  $n$  is the training sample size.

##### A. Consistency for Learning $g$

We focus on justifying that the optimal labeling function  $g$  can be learned if the label noise caused by the annotators is of (observation-dependent) symmetric random label noise. This result is not new and has been proven in the community of learning with label noise. For example, it has been proven that any learning algorithm employing a classification-calibrated loss function [53], [54] is robust to symmetric random label noise [29, Th. 9 and Corollary 10], and that any learning algorithm employing a symmetric loss function is also robust to observation-dependent symmetric random label noise [43, Corollary 3]. This algorithmic robustness means that a learning algorithm is consistent for classification, as the noisy (caused by label noise) distribution is also consistent on the clean distribution. We have discussed in Section II that if the label noise in crowdsourcing is caused only by Assumption 1, the label noise is of the symmetric random label noise. If the label noise is caused only by Assumptions 1 and 2, the label noise is of the observation-dependent random label noise. Thus, the Bayesian classifier  $g$  for classifying the right labels in the target domain can be learned with sufficiently large training sample size.

Note that this result is different from the consistent property of MV, where the obtained labels will converge to the ground-truth labels as the number of annotators goes to infinity. Our results essentially state that the obtained labels will converge to the ground-truth labels as the number of labeled examples goes to infinity. However, the convergence rate can be quite slow, which means that with finite training examples, the performance of directly employing  $\hat{g}$  to predict the ground-truth label may be poor.

##### B. Advantages of the Proposed Methods

We further justify that the proposed MV methods are superior to directly employing  $\hat{g}$  to predict the ground-truth labels when the sample size is finite. To do this, we will study the reconstruction error between  $\hat{g}$  and  $\hat{f}_1, \dots, \hat{f}_m$ , i.e.,

$$\min_{\lambda_1, \dots, \lambda_m} \|\hat{g} - \lambda_1 \hat{f}_1 - \dots - \lambda_m \hat{f}_m\| \quad (5)$$

and the dependence property among them. Note that the parameters  $\lambda_1, \dots, \lambda_m$  are only identifiable when  $\{\hat{f}_1, \dots, \hat{f}_m\}$  are linearly independent and several works [55]–[57] have

exploited this reconstruction property to refine domain adaptation. However, we assume that  $\{\hat{f}_1, \dots, \hat{f}_m\}$  are linearly dependent, which means we can use a subset of  $\{\hat{f}_1, \dots, \hat{f}_m\}$  to achieve the same reconstruction for  $\hat{g}$ . This implies that we could aggregate a subset of  $\{\hat{f}_1, \dots, \hat{f}_m\}$  to generate a final rule that has the same performance as  $\hat{g}$ , which is conflict with the wisdom of crowds that more annotators will result in a higher performance. Note that although in practice, it is likely that  $\{\hat{f}_1, \dots, \hat{f}_m\}$  are linearly independent, our aggregation rules for predicting the ground-truth labels in Algorithms 1 and 2 are still superior to employing  $\hat{g}$  directly, because they much more emphasize the “common part” among the annotators.

We also prove that if the target-domain labeling function is linearly combined by the source-domain labeling functions provided by the annotators, the combination of the classification errors of the annotators will be smaller than the classification error of directly employing the labeling function  $\hat{g}$  for the target domain.

Given an observation  $x$ , let  $y$  be the ground-truth label. The prediction error on the observation  $x$  for the  $j$ th annotator can be formulated as

$$\epsilon_j(x) = \mathbb{E}_{X, Y \sim D_j} [1_{(Y \neq y)} | X = x] = P_j(Y \neq y | X = x) \quad (6)$$

where  $D_j$  is the distribution for the data associated with the  $j$ th annotator and  $P_j(Y | X)$  is the probability that  $Y$  is predicted for the observation  $X$ .

Similarly, the prediction error for  $g$ , the labeling function for predicting the ground-truth labels, can be formulated as

$$\epsilon(x) = \mathbb{E}_{X, Y \sim D_g} [1_{(Y \neq y)} | X = x] = P_g(Y \neq y | X = x) \quad (7)$$

where  $D_g$  is the distribution for the data associated with the aggregation rule  $g$ .

The conditional probability  $P(Y | X)$  can be estimated by a simple probabilistic classification method [58], where the corresponding link function maps the outputs of the learned predictor to the interval  $[0, 1]$ , and thus the output can be interpreted as probabilities. For example, the output of logistic regression can be viewed as a conditional probability

$$P(Y | X, f) = L(f, X, Y) = \frac{1}{1 + \exp(-Yf(X))} \quad (8)$$

which is convex with respect to the labeling function  $f$ .

We have Proposition 1.

*Proposition 1:* Assume that the target-domain labeling function is linearly combined by the source-domain labeling functions provided by the annotators, i.e.,

$$g = \sum_{j=1}^m w_j f_j \quad (9)$$

where  $w_1, \dots, w_m$  are the combination parameters. Let the link function be convex with respect to the labeling function. For every observation  $x$ , we have

$$\sum_{j=1}^m w_j \epsilon_j(x) \leq \epsilon(x) \quad (10)$$

where the equality holds when  $\epsilon_1(x) = \dots = \epsilon_m(x)$ .

The proof of Proposition 1 is straightforward. Let  $L(f, x, y)$  be the link function. We have

$$\begin{aligned} P_g(Y|X) &= L(g, X, Y) = L\left(\sum_{j=1}^m w_j f_j, X, Y\right) \\ &\leq \sum_{j=1}^m w_j L(f_j, X, Y) \\ &= \sum_{j=1}^m w_j P_j(Y|X). \end{aligned} \quad (11)$$

Let the ground-truth label for the observation  $x$  be  $y$ , we then have

$$\begin{aligned} \epsilon(x) &= P_g(Y \neq y|X = x) = 1 - P_g(Y = y|X = x) \\ &\geq 1 - \sum_{j=1}^m \beta_j P_j(Y = y|X = x) \\ &= \sum_{j=1}^m \beta_j - \sum_{j=1}^m \beta_j P_j(Y = y|X = x) \\ &= \sum_{j=1}^m \beta_j P_j(Y \neq y|X = x) = \sum_{j=1}^m \beta_j \epsilon_j(x). \end{aligned} \quad (12)$$

The proof ends.

### C. Convergence for Learning the Weights

In this section, we prove that the weights learned by the proposed method will be identifiable and converge to the optimal ones with a fast convergence rate. Specifically, we focus on the algorithm of DWMV.

We present some preliminaries first. We have

$$\begin{aligned} &\arg \min_w \|[g(\phi(X_1)), \dots, g(\phi(X_n))] \\ &\quad - w[f(\phi(X_1)), \dots, f(\phi(X_n))]\|_2^2 \\ &= \arg \min_w \left( \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) \right. \\ &\quad \left. - w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right)^2. \end{aligned} \quad (13)$$

We, therefore, study the convergence property by exploiting

$$\hat{\mathcal{D}}(w) = \left( \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right)^2 \quad (14)$$

and

$$\mathcal{D}(w) = \left( \mathbb{E} \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - \mathbb{E} \frac{w}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right)^2. \quad (15)$$

Let the optimal weight be defined by

$$w^* = \arg \min_w \mathcal{D}(w) \quad (16)$$

which is not available, because the marginal distribution for  $X$  is unknown. Because  $\hat{\mathcal{D}}$  is an unbiased estimator for  $\mathcal{D}$ , we could estimate  $w^*$  by

$$\hat{w} = \arg \min_w \hat{\mathcal{D}}(w). \quad (17)$$

We are interested in analyzing the distance between  $\hat{w}$  and  $w^*$  and will give an upper bound to  $\mathcal{D}(\hat{w}) - \mathcal{D}(w^*)$ . We show that the upper bound will converge to zero fast as the sample size  $n$  goes to infinity.

**Theorem 2:** Let  $w \in [-r, r]$ ,  $r \in \mathbb{R}_+$ . For any  $\delta > 0$ , with probability at least  $1 - \delta$ , we have

$$\mathcal{D}(\hat{w}) - \mathcal{D}(w^*) \leq \frac{8rC^4}{\sqrt{n}} + 4(1+r)C^4 \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (18)$$

Since  $\mathcal{D}(w^*) \leq \mathcal{D}(\hat{w})$ , Theorem 2 shows that the learned weight  $\hat{w}$  will converge to the optimal one  $w^*$  as the sample size  $n$  goes to infinity. Also, the convergence rate is of order  $O((1/n)^{1/2})$ .

We are now going to prove Theorem 2.

Even though we do not know  $w^*$ , we have

$$\begin{aligned} \mathcal{D}(\hat{w}) - \mathcal{D}(w^*) &= \mathcal{D}(\hat{w}) - \hat{\mathcal{D}}(\hat{w}) + \hat{\mathcal{D}}(\hat{w}) - \hat{\mathcal{D}}(w^*) + \hat{\mathcal{D}}(w^*) - \mathcal{D}(w^*) \\ &\leq \mathcal{D}(\hat{w}) - \hat{\mathcal{D}}(\hat{w}) + \hat{\mathcal{D}}(w^*) - \mathcal{D}(w^*) \\ &\leq 2 \sup_w |\mathcal{D}(w) - \hat{\mathcal{D}}(w)| \end{aligned} \quad (19)$$

where the first inequality holds because  $\hat{w}$  is the empirical minimizer of  $\hat{\mathcal{D}}(\hat{w}, \alpha)$  and  $\hat{\mathcal{D}}(\hat{w}) \leq \hat{\mathcal{D}}(w^*)$ .

Furthermore, we have

$$\begin{aligned} &2 \sup_w |\mathcal{D}(w) - \hat{\mathcal{D}}(w)| \\ &= 2 \sup_w \left| \left( \mathbb{E} \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - w \mathbb{E} \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right. \right. \\ &\quad \left. \left. - \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right) \right. \\ &\quad \left. \times \left( \mathbb{E} \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - w \mathbb{E} \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right) \right. \\ &\quad \left. + \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right| \\ &\leq 4C^2 \sup_w \left| \left( \mathbb{E} \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - \mathbb{E} \frac{w}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right. \right. \\ &\quad \left. \left. - \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right) \right| \end{aligned} \quad (20)$$

where the inequality holds because  $g(\phi(x))$ ,  $f(\phi(x)) \in \{1, \dots, C\}$  and  $0 \leq \mathcal{D}(w)$ ,  $\hat{\mathcal{D}}(w) \leq C$ .

In statistical learning theory [59], Rademacher complexity [60] is defined to up bound the term on the right-hand

side of the above inequality. The Rademacher complexity of a function class  $F$  on the feature space  $\mathcal{X}$  is defined as

$$\mathfrak{R}(H) = \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \quad (21)$$

where  $\sigma_1, \dots, \sigma_n$  are i.i.d. Rademacher variables that are uniformly distributed in  $\{-1, +1\}$  and  $X_1, \dots, X_n$  are i.i.d. variables drawn from  $\mathcal{X}$ .

Let

$$\begin{aligned} h(w, X) &= \sup_w \left| \left( \mathbb{E} \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - w \mathbb{E} \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right. \right. \\ &\quad \left. \left. - \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right) \right|. \end{aligned} \quad (22)$$

It can be found that  $h(w, X)$  is a random variable, and we will upper bound it by employing McDiarmid's inequality [61].

*Theorem 3:* Let  $X = [X_1, \dots, X_n]$  be an i.i.d. sample and  $X^i$  a new sample with the  $i$ th example in  $X$  being replaced by an independent example  $X'_i$ . If there exists  $b_1, \dots, b_n > 0$  such that  $h : \mathcal{X}^n \rightarrow \mathbb{R}$  satisfies the following conditions:

$$|h(X) - h(X^i)| \leq b_i \quad \forall i \in \{1, \dots, n\}. \quad (23)$$

Then, for any  $X \in \mathcal{X}^n$  and  $\epsilon > 0$ , the following inequalities hold:

$$P(h(X) - \mathbb{E}h(X) \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n b_i^2}\right). \quad (24)$$

We further assume that  $w$  is upper bounded by  $r$ , which means  $w \in [-r, r]$ . It is easy to check that

$$\begin{aligned} |h(w, X) - h(w, X^i)| &\leq \sup_{w \in [-r, r]} \frac{1}{n} |g(\phi(X_i)) f(\phi(X_i)) - g(\phi(X'_i)) f(\phi(X'_i)) \\ &\quad + w f(\phi(X_i))^2 - w f(\phi(X'_i))^2| \\ &\leq \frac{(1+r)C^2}{n}. \end{aligned} \quad (25)$$

Employing McDiarmid's inequality, we, therefore, have

$$P(h(w, X) - \mathbb{E}h(w, X) \geq \epsilon) \leq \exp\left(\frac{-2n\epsilon^2}{(1+r)^2 C^4}\right). \quad (26)$$

Let

$$\exp\left(\frac{-2n\epsilon^2}{(1+r)^2 C^4}\right) = \delta. \quad (27)$$

We have

$$\epsilon = (1+r)C^2 \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (28)$$

Thus, with probability at least  $1 - \delta$ , we have

$$h(w, X) - \mathbb{E}h(w, X) \leq (1+r)C^2 \sqrt{\frac{\log(1/\delta)}{2n}}. \quad (29)$$

In the next, we are going to upper bound  $\mathbb{E}h(w, X)$ . Let  $X'$  be i.i.d. copy of  $X$ . We have

$$\begin{aligned} \mathbb{E}h(w, X) &= \mathbb{E} \sup_w \left| \left( \mathbb{E} \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) - w \mathbb{E} \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right. \right. \\ &\quad \left. \left. - \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right) \right| \\ &\leq \mathbb{E} \sup_w \left| \left( \frac{1}{n} \sum_{i=1}^n g(\phi(X'_i)) f(\phi(X'_i)) - \frac{w}{n} \sum_{i=1}^n f(\phi(X'_i))^2 \right. \right. \\ &\quad \left. \left. - \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right) \right| \end{aligned}$$

which is symmetric, meaning that its density function is even. Let  $\sigma_i$  be independent Rademacher variables, which are uniformly distributed from  $\{-1, 1\}$ . We have that

$$\sup_w \left| \left( \frac{1}{n} \sum_{i=1}^n g(\phi(X'_i)) f(\phi(X'_i)) - \frac{w}{n} \sum_{i=1}^n f(\phi(X'_i))^2 \right. \right. \\ \left. \left. - \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i))^2 \right) \right|$$

and

$$\sup_w \left| \left( \frac{1}{n} \sum_{i=1}^n \sigma_i g(\phi(X'_i)) f(\phi(X'_i)) - \frac{w}{n} \sum_{i=1}^n \sigma_i f(\phi(X'_i))^2 \right. \right. \\ \left. \left. - \frac{1}{n} \sum_{i=1}^n \sigma_i g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n \sigma_i f(\phi(X_i))^2 \right) \right|$$

have the same distribution. Then, inequality (30) can be written as

$$\begin{aligned} \mathbb{E}h(w, X) &\leq \mathbb{E} \sup_w \left| \left( \frac{1}{n} \sum_{i=1}^n \sigma_i g(\phi(X'_i)) f(\phi(X'_i)) - w \frac{1}{n} \sum_{i=1}^n \sigma_i f(\phi(X'_i))^2 \right. \right. \\ &\quad \left. \left. - \frac{1}{n} \sum_{i=1}^n \sigma_i g(\phi(X_i)) f(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n \sigma_i f(\phi(X_i))^2 \right) \right| \\ &\leq \mathbb{E} \frac{1}{n} \sum_{i=1}^n \sigma_i g(\phi(X'_i)) f(\phi(X'_i)) + \mathbb{E} \sup_w w \frac{1}{n} \sum_{i=1}^n \sigma_i f(\phi(X'_i))^2 \\ &\quad + \mathbb{E} \frac{1}{n} \sum_{i=1}^n \sigma_i g(\phi(X_i)) f(\phi(X_i)) + \mathbb{E} \sup_w w \frac{1}{n} \sum_{i=1}^n \sigma_i f(\phi(X_i))^2 \\ &= \mathbb{E} \sup_w w \frac{1}{n} \sum_{i=1}^n \sigma_i f(\phi(X'_i))^2 + \mathbb{E} \sup_w w \frac{1}{n} \sum_{i=1}^n \sigma_i f(\phi(X_i))^2 \\ &\leq \frac{r}{n} \left( \mathbb{E} \sum_{i=1}^n \sigma_i f(\phi(X'_i))^2 + \mathbb{E} \sum_{i=1}^n \sigma_i f(\phi(X_i))^2 \right) \\ &\leq \frac{r}{n} \left( \sqrt{\mathbb{E} \left( \sum_{i=1}^n \sigma_i f(\phi(X'_i))^2 \right)^2} \right. \\ &\quad \left. + \sqrt{\mathbb{E} \left( \sum_{i=1}^n \sigma_i f(\phi(X_i))^2 \right)^2} \right) \\ &\leq \frac{r}{n} (\sqrt{n}C^2 + \sqrt{n}C^2) \leq \frac{2rC^2}{\sqrt{n}} \end{aligned} \quad (30)$$



where

$$\sqrt{\mathbb{E}\left(\sum_{i=1}^n \sigma_i f(\phi(X'_i))^2\right)} \leq \sqrt{n}C^2 \quad (31)$$

because  $\sigma_1, \dots, \sigma_n$  are i.i.d. and  $\mathbb{E}\sigma_i\sigma_j = 1$  only when  $i = j$ , otherwise  $\mathbb{E}\sigma_i\sigma_j = \mathbb{E}\sigma_i\mathbb{E}\sigma_j = 0$  for  $i \neq j$ .

Combining inequalities (19), (20), (22), (29), and (30), for any  $\delta > 0$ , with probability at least  $1 - \delta$ , we have

$$\begin{aligned} & \mathcal{D}(\hat{w}) - \mathcal{D}(w^*) \\ & \leq 2 \sup_w |\mathcal{D}(w) - \hat{\mathcal{D}}(w)| \\ & \leq 4C^2 \sup_w \left| \left( \mathbb{E} \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) - w \mathbb{E} \frac{1}{n} \sum_{i=1}^n f(\phi(X_i)) \right. \right. \\ & \quad \left. \left. - \frac{1}{n} \sum_{i=1}^n g(\phi(X_i)) + w \frac{1}{n} \sum_{i=1}^n f(\phi(X_i)) \right) \right| \\ & = 4C^2 h(w, X) \\ & \leq 4C^2 \mathbb{E} h(w, X) + 4(1+r)C^4 \sqrt{\frac{\log(1/\delta)}{2n}} \\ & \leq \frac{8rC^4}{\sqrt{n}} + 4(1+r)C^4 \sqrt{\frac{\log(1/\delta)}{2n}}. \end{aligned} \quad (32)$$

The proof of Theorem 2 ends.

## V. EXPERIMENTS

In this section, we employ some simulated and real-world data sets to compare the performances of the proposed methods with those of other baselines and some state-of-the-art methods. The baselines are the MV and the method by directly employing the labeling function  $\hat{g}$ . The state-of-the-art methods are the multiclass labeling algorithm proposed in [42] (referred to as KOS), the SVD-based algorithm proposed in [62] (referred to as GhostSVD), and a recent method combining the spectral method and EM algorithm [21] (referred to as Opt-DS). Those two methods are independent of the instances. GhostSVD contributes to estimate the quality of the labellers based on relating the top eigenvector of the matrix  $\mathbb{E}[UU^\top]$ , where  $U$  represents the matrix of noisy labels. They consider the structure information in labels and, therefore, can outperform the MV method. Our proposed model also aims to estimate the quality of each labeller but additionally considers the information of instances. The following experiments empirically show the superior of our proposed method.

Since we learn the weights for weighted MV, we generate five annotators by flipping the ground-truth labels according to some flip rates to verify the effectiveness of the proposed methods. The classifiers used are support vector machines SVMs. They are learned by employing libsvm [63] and the parameters are set as follows. We employ linear functions and set the parameter  $C$  to be 1000. All the means and standard deviations are calculated on the performances of 100 repetitions.

### A. Synthetic Data Sets

We generate binary classification data as follows. The observations and classifiers for providing ground-truth

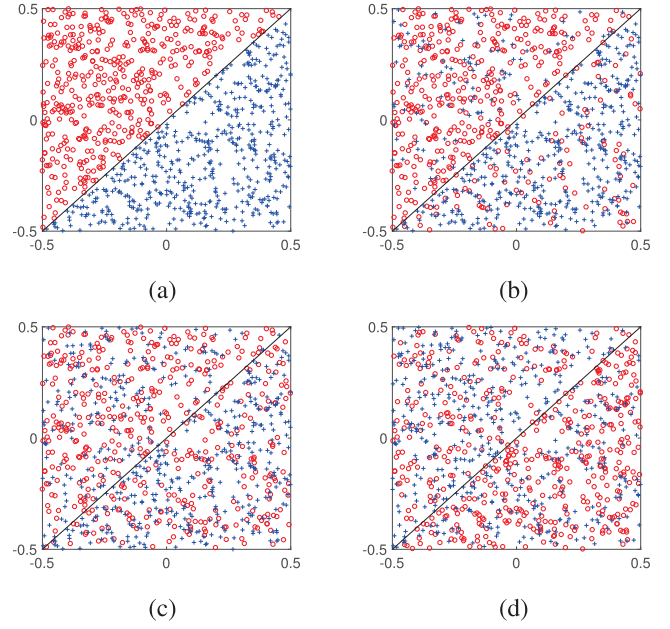


Fig. 2. Illustration of binary synthetic data sets when the dimension is 2. Red circles: sampled points labeled with class 1. Blue crosses: sampled points labeled with class 2. Black lines across the figures: classifiers for the ground-true labels. (a)  $\rho = 0$ . (b)  $\rho = 0.2$ . (c)  $\rho = 0.4$ . (d)  $\rho = 0.6$ .

labels are 10-D vectors and are sampled from uniform distribution, i.e.,

$$x, f \sim \mathcal{U}(-0.5, 0.5)^{10} \quad (33)$$

where  $\mathcal{U}(-0.5, 0.5)$  represents the uniform distribution taking values in the range  $(-0.5, 0.5)$ . The ground-truth labels are generated by  $y = \text{sign}(\langle f, x \rangle)$ . The labels provided by the  $j$ th annotators are generated by randomly flipping the ground-truth labels according to some given flip rate  $\rho_j$ . In Fig. 2, we show some 2-D examples. We can see that as the flip rate becomes larger, it becomes harder to search for the optimal classifier.

The results are shown in Table II. The largest two performances are marked in bold. Overall, the results show that the proposed methods, LFWMV and DWMV, obtain large performance gains. The state-of-the-arts crowdsourcing methods perform poorly on those synthetic data sets because the number of annotators is small. Also they are independent of the feature  $X$ . This implies that the feature information is useful for aggregating the annotated labels.

Specifically, comparing the performances in the cases  $\rho_1 = \rho_2 = \rho_3 = 0.2, \rho_4 = \rho_5 = 0.4$  (in the first row of Table II),  $\rho_1 = \rho_2 = \rho_3 = 0.2, \rho_4 = \rho_5 = 0.6$  (in the second row of Table II), and  $\rho_1 = \rho_2 = 0.2, \rho_3 = \rho_4 = \rho_5 = 0.4$  (in the third row of Table II), we can conclude that the proposed methods are much more robust to label noise than the baselines. The proposed methods outperform the MV method by almost 20%. Surprisingly, comparing the performances in the first and second rows of Table II, we see that the performances of the proposed methods do not decrease when adding more label noise. However, the performances of MV have decreased greatly. This implies that the proposed methods are effective on learning accurate weights. Comparing the



TABLE II  
MEANS AND STANDARD DEVIATIONS (PERCENTAGE) OF THE AGGREGATION ACCURACIES OF THE BASELINES, STATE-OF-THE-ART METHODS, AND THE PROPOSED METHODS ON THE SYNTHETIC DATA SET, WHERE  $n = 1000$  AND  $m = 5$

Noise rate for annotators	GhostSVD	KOS	MV	Opt-DS	$\hat{g}$	LFWMV	DWMV
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.4$	$88.81 \pm 1.05$	$88.74 \pm 1.44$	$86.81 \pm 1.16$	$89.46 \pm 1.05$	<b><math>96.86 \pm 0.74</math></b>	<b><math>96.06 \pm 0.94</math></b>	$96.03 \pm 0.92$
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.6$	$89.00 \pm 1.32$	$88.92 \pm 1.16$	$77.31 \pm 1.21$	$89.56 \pm 0.97$	$95.66 \pm 1.35$	<b><math>95.96 \pm 0.98</math></b>	<b><math>96.00 \pm 0.95</math></b>
$\rho_1 = \rho_2 = 0.2,$ $\rho_3 = \rho_4 = \rho_5 = 0.4$	$84.56 \pm 1.17$	$84.57 \pm 1.14$	$81.53 \pm 1.25$	$84.31 \pm 1.22$	<b><math>96.34 \pm 0.92</math></b>	$94.86 \pm 1.51$	<b><math>94.91 \pm 1.50</math></b>
$\rho_1 = 0.1, \rho_2 = 0.2, \rho_3 = 0.3,$ $\rho_4 = 0.4, \rho_5 = 0.5$	$90.12 \pm 0.92$	$90.11 \pm 0.92$	$84.94 \pm 1.09$	$90.91 \pm 1.02$	<b><math>96.62 \pm 0.91</math></b>	$95.95 \pm 1.12$	<b><math>96.01 \pm 1.10</math></b>
$\rho_1 = 0.2, \rho_2 = 0.3, \rho_3 = 0.4,$ $\rho_4 = 0.5, \rho_5 = 0.6$	$79.40 \pm 1.58$	$79.51 \pm 1.57$	$69.81 \pm 1.39$	$80.27 \pm 1.76$	$93.39 \pm 2.04$	<b><math>94.36 \pm 1.49</math></b>	<b><math>94.39 \pm 1.48</math></b>

TABLE III  
MEANS AND STANDARD DEVIATIONS (PERCENTAGE) OF THE AGGREGATION ACCURACIES OF THE BASELINES, STATE-OF-THE-ART METHODS, AND THE PROPOSED METHODS ON THE UCI HEART DATA SET, WHERE  $n = 270$  AND  $m = 5$

Noise rate for annotators	GhostSVD	KOS	MV	Opt-DS	$\hat{g}$	LFWMV	DWMV
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.4$	<b><math>88.52 \pm 2.06</math></b>	$88.08 \pm 1.97$	$86.84 \pm 1.91$	<b><math>88.81 \pm 2.07</math></b>	$81.04 \pm 3.88$	$82.61 \pm 2.09$	$82.42 \pm 1.81$
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.6$	<b><math>89.11 \pm 2.10</math></b>	$88.46 \pm 2.06$	$77.60 \pm 2.50$	<b><math>89.17 \pm 1.86</math></b>	$78.12 \pm 3.95$	$82.41 \pm 1.85$	$82.44 \pm 1.73$
$\rho_1 = \rho_2 = 0.2,$ $\rho_3 = \rho_4 = \rho_5 = 0.4$	<b><math>84.42 \pm 2.54</math></b>	$83.91 \pm 2.64$	$81.51 \pm 2.52$	<b><math>83.40 \pm 2.48</math></b>	$79.21 \pm 4.29$	$81.43 \pm 2.50$	$81.62 \pm 2.06$
$\rho_1 = 0.1, \rho_2 = 0.2, \rho_3 = 0.3,$ $\rho_4 = 0.4, \rho_5 = 0.5$	<b><math>90.23 \pm 2.02</math></b>	$89.80 \pm 1.93$	$85.09 \pm 2.28$	<b><math>90.21 \pm 2.30</math></b>	$80.73 \pm 4.11$	$82.49 \pm 1.89$	$82.20 \pm 1.73$
$\rho_1 = 0.2, \rho_2 = 0.3, \rho_3 = 0.4,$ $\rho_4 = 0.5, \rho_5 = 0.6$	$78.61 \pm 2.91$	$78.21 \pm 2.81$	$69.12 \pm 2.57$	$78.91 \pm 3.12$	$76.47 \pm 3.87$	<b><math>78.66 \pm 5.34</math></b>	<b><math>80.25 \pm 2.56</math></b>

performances in the last three rows of Table II, we can see that the proposed methods are more robust to large label noise than the baselines because the performances decrease slower than those of baselines.

In the last two rows of Table II, where  $\rho_1 = 0.1, \rho_2 = 0.2, \rho_3 = 0.3, \rho_4 = 0.4$ , and  $\rho_5 = 0.5$  (in the last but one row of Table II) and  $\rho_1 = 0.2, \rho_2 = 0.3, \rho_3 = 0.4, \rho_4 = 0.5$ , and  $\rho_5 = 0.6$  (in the last row of Table II), we further empirically study the cases where all the annotators are distinct. The performances are consistent with the performances in the first three rows of Table II that the proposed methods are more robust to label noise than the baselines.

## B. Real-Word Data Sets

To further verify the effectiveness of the proposed methods, we compare them with the baselines and state-of-the-art methods on three real-world data sets: the UCI Heart data set, MNIST data set, and CIFAR-10 data set.

1) *Experiments on the UCI Heart Data Set:* We conduct empirical experiments on the UCI Heart data set,<sup>3</sup> where the dimensionality of the feature is of 13 and the sample size is of 270 (120 of them are positive examples and 150 of them are negative examples).

The performances are presented in Table III. We can see that the proposed methods LFWMV and DWMV only outperform the baselines in the case where noise level is high (i.e., the last row in Table III), that the performances of the state-of-the-art methods GhostSVD and Opt-DS outperform in all the other cases, and that MV sometimes outperforms the

proposed methods. This may be caused by the reason that there is some structural label information in the UCI Heart data set and that the training sample size is too small to fully exploit the feature information for aggregating labels.

Specifically, we can find that in the first and last but one rows of Table III, where the label flip rates are  $\rho_1 = \rho_2 = \rho_3 = 0.2, \rho_4 = \rho_5 = 0.4$  and  $\rho_1 = 0.1, \rho_2 = 0.2, \rho_3 = 0.3, \rho_4 = 0.4, \rho_5 = 0.5$ , respectively, the baseline MV method outperforms the proposed methods. This is because in those two cases, the label flip rates are relatively small. In the other rows of Table III, where the label flip rates are larger, the proposed methods outperform all the baselines and state-of-the-art methods.

2) *Experiments on the MNIST Data Set:* The MNIST handwritten digit data set contains 60000 training images and 10000 test images [64]. We choose digit 0 as negative examples and digit 1 as positive examples and randomly sample 500 images for each class from training images. To extract CNN features, we use the LeNet-5 [64] implemented in Caffe [65] by modifying the last fully connected layer to obtain 16-D features.

The performances on the MNIST data set are presented in Table IV. We see that the proposed methods greatly outperform the MV method. The accuracies of the proposed methods are all above 99%. This implies that the feature information is essential for some crowdsourcing problem.

3) *Experiments on the CIFAR-10 Data Set:* The CIFAR-10 data set contains 60000 training images and 10000 test images, which are  $32 \times 32$  RGB images in 10 classes [66]. We choose label 6 (i.e., frog) as negative examples and label 8 (i.e., ship) as positive examples and randomly sample 500 images for each class from training images. To extract

<sup>3</sup>We employ the UCI Heart data set, which is preprocessed and made available online by Gunnar Rätsch (<http://theoval.cmp.uea.ac.uk/matlab>).

TABLE IV

MEANS AND STANDARD DEVIATIONS (PERCENTAGE) OF THE AGGREGATION ACCURACIES OF THE BASELINES, STATE-OF-THE-ART METHODS, AND THE PROPOSED METHODS ON THE MNIST DATA SET, WHERE  $n = 1000$  AND  $m = 5$

Noise rate for annotators	GhostSVD	KOS	MV	Opt-DS	$\hat{g}$	LFWMV	DWMV
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.4$	88.99 $\pm$ 1.40	88.94 $\pm$ 1.31	86.88 $\pm$ 1.06	89.58 $\pm$ 0.84	<b>99.98 <math>\pm</math> 0.04</b>	99.90 $\pm$ 0.27	<b>99.90<math>\pm</math> 0.26</b>
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.6$	89.04 $\pm$ 1.33	89.04 $\pm$ 1.29	77.33 $\pm$ 1.46	89.58 $\pm$ 1.06	<b>99.98<math>\pm</math> 0.05</b>	<b>99.95<math>\pm</math> 0.10</b>	99.93 $\pm$ 0.11
$\rho_1 = \rho_2 = 0.2,$ $\rho_3 = \rho_4 = \rho_5 = 0.4$	84.52 $\pm$ 1.13	84.50 $\pm$ 1.13	81.32 $\pm$ 1.25	84.30 $\pm$ 1.26	99.43 $\pm$ 0.28	<b>99.86 <math>\pm</math> 0.74</b>	<b>99.93<math>\pm</math>0.12</b>
$\rho_1 = 0.1, \rho_2 = 0.2, \rho_3 = 0.3,$ $\rho_4 = 0.4, \rho_5 = 0.5$	90.17 $\pm$ 0.90	90.18 $\pm$ 0.90	85.11 $\pm$ 1.18	90.89 $\pm$ 1.00	<b>99.99 <math>\pm</math>0.03</b>	<b>99.88<math>\pm</math> 0.26</b>	99.84 $\pm$ 0.31
$\rho_1 = 0.2, \rho_2 = 0.3, \rho_3 = 0.4,$ $\rho_4 = 0.5, \rho_5 = 0.6$	79.59 $\pm$ 1.65	79.50 $\pm$ 1.60	69.20 $\pm$ 1.50	80.09 $\pm$ 1.79	<b>99.96<math>\pm</math>0.07</b>	<b>99.64<math>\pm</math>7.65</b>	99.08 $\pm$ 3.32

TABLE V

MEANS AND STANDARD DEVIATIONS (PERCENTAGE) OF THE AGGREGATION ACCURACIES OF THE BASELINES, STATE-OF-THE-ART METHODS, AND THE PROPOSED METHODS ON THE CIFAR-10 DATA SET, WHERE  $n = 1000$  AND  $m = 5$

Noise rate for annotators	GhostSVD	KOS	MV	Opt-DS	$\hat{g}$	LFWMV	DWMV
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.4$	89.01 $\pm$ 1.39	89.07 $\pm$ 1.30	86.94 $\pm$ 1.05	89.71 $\pm$ 1.06	94.03 $\pm$ 0.44	<b>97.92<math>\pm</math> 0.58</b>	<b>97.05<math>\pm</math>2.28</b>
$\rho_1 = \rho_2 = \rho_3 = 0.2,$ $\rho_4 = \rho_5 = 0.6$	88.99 $\pm$ 1.34	88.96 $\pm$ 1.32	77.09 $\pm$ 1.30	89.47 $\pm$ 1.08	91.53 $\pm$ 1.13	<b>98.02<math>\pm</math> 0.60</b>	<b>97.20<math>\pm</math>2.10</b>
$\rho_1 = \rho_2 = 0.2,$ $\rho_3 = \rho_4 = \rho_5 = 0.4$	84.62 $\pm$ 1.11	84.63 $\pm$ 1.12	81.62 $\pm$ 1.26	84.32 $\pm$ 1.21	92.99 $\pm$ 0.68	<b>97.37 <math>\pm</math> 1.77</b>	<b>96.12<math>\pm</math> 3.12</b>
$\rho_1 = 0.1, \rho_2 = 0.2, \rho_3 = 0.3,$ $\rho_4 = 0.4, \rho_5 = 0.5$	90.23 $\pm$ 0.96	90.25 $\pm$ 0.99	85.07 $\pm$ 1.12	90.91 $\pm$ 1.10	93.53 $\pm$ 0.56	<b>98.03<math>\pm</math>0.66</b>	<b>95.37<math>\pm</math> 5.69</b>
$\rho_1 = 0.2, \rho_2 = 0.3, \rho_3 = 0.4,$ $\rho_4 = 0.5, \rho_5 = 0.6$	79.23 $\pm$ 1.64	79.22 $\pm$ 1.66	68.84 $\pm$ 1.39	79.86 $\pm$ 1.97	90.17 $\pm$ 2.42	<b>97.25<math>\pm</math>1.67</b>	<b>91.84<math>\pm</math> 8.57</b>

CNN features, we use the Caffe reference network for CIFAR-10 by modifying the last fully connected layer to obtain 16-D features.

The performances on the CIFAR-10 data set are presented in Table V. In Table IV, LFWMV and DWMV outperform MV, but  $\hat{g}$  outperforms LFWMV and DWMV. On the CIFAR-10 data set, the proposed methods clearly outperform all the baselines and state-of-the-art methods. This suggests that the proposed methods may not perform well for all cases and that feature learning may also play an important role in the proposed methods. However, the experiment results show that the proposed methods can greatly improve the aggregation performance.

Comparing all the performances in synthetic and real-world data sets from Tables II–V, we can find that the MV method has similar performances for the same noise levels in different data sets, while the performances of the proposed methods differ. This implies that the proposed methods have the ability to exploiting useful information from the features to significantly boost the performance.

We can find that the performances of baselines are always similar in different situations. This is because the structure information in labels is almost the same or these methods are not very sensitive to the change of label structures.

The performances in the synthetic and real-word data sets also verify our theoretical conclusion that the proposed methods LFWMV and DWMV outperform the baseline  $\hat{g}$  under mild conditions.

## VI. CONCLUSION

In this paper, we studied the aggregation rule of weighted MV for the crowdsourcing problem. Since the annotators may have very different backgrounds and are anonymous and

that the ground-truth labels are usually unknown, it is hard to directly calculate the weights. We are the first to learn the weights from a viewpoint of multiple-domain adaptation. Two simple learning algorithms were presented. The effectiveness of the proposed methods are both theoretically and empirically verified.

Specifically, we treat the domain with ground-truth labels as the target domain and the domains with noisy labels annotated by the annotators as the source domain. The weights are learned by matching the source domains with the target domain. Although the labels for the target domain are unknown, we theoretically justify that they can be consistently learned under mild conditions. However, the convergence rate may be slow. Thus, for the finite sample problem, we did not directly employ the labeling function learned for the target domain to assign labels. Instead, we designed the LFWMV and DWMV methods as the aggregation rules, which integrate the wisdom of crowd by exploiting the feature information as well.

We conclude with a future work that will exploit feature learning for each annotator to more accurately learn the weights for the proposed weighted MV.

## REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, May 2015.
- [2] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” *J. Mach. Learn. Res.*, vol. 15, no. 4, pp. 315–323, 2011.
- [3] V. C. Raykar *et al.*, “Learning from crowds,” *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Apr. 2010.
- [4] J. Whitehill, T.-F. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, “Whose vote should count more: Optimal integration of labels from labelers of unknown expertise,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2035–2043.

- [5] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1953–1961.
- [6] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *J. Mach. Learn. Res.*, vol. 13, pp. 491–518, Feb. 2012.
- [7] N. Shah, D. Zhou, and Y. Peres, "Approval voting and incentives in crowdsourcing," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 10–19.
- [8] N. B. Shah and D. Zhou, "Double or nothing: Multiplicative incentive mechanisms for crowdsourcing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.
- [9] J. Zhang, V. S. Sheng, T. Li, and X. Wu, "Improving crowdsourced label quality using noise correction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1675–1688, May 2018.
- [10] R. Wang, T. Liu, and D. Tao, "Multiclass learning with partially corrupted labels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 9, pp. 2568–2580, Jun. 2018.
- [11] X. Yu, T. Liu, M. Gong, and D. Tao. (2017). "Learning with biased complementary labels." [Online]. Available: <https://arxiv.org/abs/1711.09535?context=cs>
- [12] P. Young, "Optimal voting rules," *J. Econ. Perspectives*, vol. 9, no. 1, pp. 51–64, 1995.
- [13] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2424–2432.
- [14] H. Li and B. Yu. (2014). "Error rate bounds and iterative weighted majority voting for crowdsourcing." [Online]. Available: <https://arxiv.org/abs/1411.4086>
- [15] A. D. Procaccia and N. Shah, "Is approval voting optimal given approval votes?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1801–1809.
- [16] J. Surowiecki, *The Wisdom of Crowds*. New York, NY, USA: Random House LLC, 2005.
- [17] S. Nitzan and J. Paroush, "Optimal decision rules in uncertain dichotomous choice situations," *Int. Econ. Rev.*, vol. 23, no. 2, pp. 289–297, 1982.
- [18] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.
- [19] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 692–700.
- [20] D. Zhou, S. Basu, Y. Mao, and J. C. Platt, "Learning from the wisdom of crowds by minimax entropy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2195–2203.
- [21] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing," *J. Mach. Learn. Res.*, vol. 17, pp. 1–44, Jan. 2016.
- [22] T. Tian and J. Zhu, "Max-margin majority voting for learning from crowds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1621–1629.
- [23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, Nov. 1999.
- [24] R. K. Vinayak and B. Hassibi, "Crowdsourced clustering: Querying edges vs triangles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1316–1324.
- [25] J. Ye, J. Li, M. G. Newman, R. B. Adams, Jr., and J. Z. Wang. (2017). "Probabilistic multigraph modeling for improving the quality of crowdsourced affective data." [Online]. Available: <https://arxiv.org/abs/1701.01096>
- [26] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, vol. 382. Hoboken, NJ, USA: Wiley, 2007.
- [27] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.
- [28] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
- [29] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1196–1204.
- [30] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu. (2016). "Making deep neural networks robust to label noise: A loss correction approach." [Online]. Available: <https://arxiv.org/abs/1609.03683>
- [31] A. Menon, B. V. Rooyen, C. S. Ong, and B. Williamson, "Learning from corrupted binary labels via class-probability estimation," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 125–134.
- [32] G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott, "Classification with asymmetric label noise: Consistency and maximal denoising," *Electron. J. Statist.*, vol. 10, no. 2, pp. 2780–2824, 2016.
- [33] J. Cheng, T. Liu, K. Ramamohanarao, and D. Tao. (2017). "Learning with bounded instance- and label-dependent label noise." [Online]. Available: <https://arxiv.org/abs/1709.03768>
- [34] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1041–1048.
- [35] J. T. Ash, R. E. Schapire, and B. E. Engelhardt. (2016). "Unsupervised domain adaptation using approximate label matching." [Online]. Available: <https://arxiv.org/abs/1602.04889>
- [36] S. Li, S. Song, and G. Huang, "Prediction reweighting for domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1682–1695, Jul. 2017.
- [37] X. Yu, T. Liu, M. Gong, K. Zhang, and D. Tao. (2017). "Transfer learning with label noise." [Online]. Available: <https://arxiv.org/abs/1707.09724>
- [38] T. Liu, Q. Yang, and D. Tao, "Understanding how feature structure transfers in transfer learning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 2365–2371.
- [39] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver, "How to grade a test without knowing the answers—A Bayesian graphical model for adaptive crowdsourcing and aptitude testing," in *Proc. 29th Int. Conf. Mach. Learn.*, Jan. 2012, pp. 1183–1190.
- [40] H. E. Landemore, "Why the many are smarter than the few and why it matters," *J. Public Deliberation*, vol. 8, no. 1, 2012, Art. no. 7.
- [41] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2008, pp. 254–263.
- [42] D. R. Karger, S. Oh, and D. Shah, "Efficient crowdsourcing for multi-class labeling," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 1, pp. 81–92, 2013.
- [43] A. K. Menon, B. van Rooyen, and N. Natarajan. (2016). "Learning from binary labels with instance-dependent corruption." [Online]. Available: <https://arxiv.org/abs/1605.00751>
- [44] S. Sukhbaatar and R. Fergus. (2014). "Learning from noisy labels with deep neural networks." [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.749.1795&rep=rep1&type=pdf>
- [45] A. Ghosh, N. Manwani, and P. S. Sastry, "Making risk minimization tolerant to label noise," *Neurocomputing*, vol. 160, pp. 93–107, Jul. 2015.
- [46] Q. Miao, Y. Cao, G. Xia, M. Gong, J. Liu, and J. Song, "RBoost: Label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2216–2228, Nov. 2016.
- [47] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2691–2699.
- [48] W. Gao, L. Wang, Y.-F. Li, and Z.-H. Zhou, "Risk minimization in the presence of label noise," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1575–1581.
- [49] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li. (2017). "Learning from noisy labels with distillation." [Online]. Available: <https://arxiv.org/abs/1703.02391>
- [50] Y. Duan and O. Wu, "Learning with auxiliary less-noisy labels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1716–1721, Jul. 2017.
- [51] B. Han, I. W. Tsang, L. Chen, C. P. Yu, and S.-F. Fung, "Progressive stochastic learning for noisy labels," *IEEE Trans. Neural Netw. Learn. Syst.*, doi: [10.1109/TNNLS.2018.2792062](https://doi.org/10.1109/TNNLS.2018.2792062).
- [52] H. Xu and S. Mannor, "Robustness and generalization," *Mach. Learn.*, vol. 86, no. 3, pp. 391–423, 2012.
- [53] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.
- [54] C. Scott, "Calibrated asymmetric surrogate losses," *Electron. J. Statist.*, vol. 6, pp. 958–992, May 2012.
- [55] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 819–827.
- [56] A. Iyer, S. Nath, and S. Sarawagi, "Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 530–538.

- [57] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, "Domain adaptation with conditional transferable components," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2839–2848.
- [58] M. Sugiyama, T. Suzuki, and T. Kanamori, "Density ratio estimation: A comprehensive review (statistical experiment and its related topics)," *RIMS Kokyuroku*, vol. 1703, pp. 10–31, Aug. 2010.
- [59] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.
- [60] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Mar. 2003.
- [61] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.
- [62] A. Ghosh, S. Kale, and R. P. McAfee, "Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content," in *Proc. 12th ACM Conf. Electron. Commerce*, 2011, pp. 167–176.
- [63] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
- [64] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [65] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: <https://arxiv.org/abs/1408.5093>
- [66] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.



**Dapeng Tao** received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, and the Ph.D. degree from the South China University of Technology, Guangzhou, China.

He is currently a Professor with the School of Information Science and Engineering, Yunnan University, Kunming, China. He has authored or co-authored over 60 scientific articles. He has served over 10 international journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS

ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, and *Information Sciences*. His current research interests include machine learning, computer vision, and robotics.



**Jun Cheng** received the B.Eng. and M.Eng. degrees from the University of Science and Technology of China, Hefei, China, in 1999 and 2002, respectively, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2006.

He is currently the Director of the Laboratory for Human Machine Control and a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests include computer vision, robotics, machine intelligence, and control.



**Zhengtao Yu** received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005.

He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include natural language process, image processing, and machine learning.



**Kun Yue** received the B.Sc. degree from Yunnan University, Kunming, China, in 2001, the M.Sc. degree from Fudan University, Shanghai, China, in 2004, and the Ph.D. degree from Yunnan University in 2009, all in computer science.

He is currently a Professor with Yunnan University. His current research interests include massive data analysis and knowledge engineering.



**Lizhen Wang** (M'16) received the B.S. and M.Sc. degrees in computational mathematics from Yunnan University, Kunming, China, in 1983 and 1988, respectively, and the Ph.D. degree in computer science from the University of Huddersfield, Huddersfield, U.K., in 2008.

She is currently a Professor with the Department of Computer Science and Engineering, Yunnan University. She has authored or co-authored over 50 scientific articles. Her research interests include data mining, data warehouses, and computer algorithms.