Bi-convex Optimization to Learn Classifiers from Multiple Biomedical Annotations

Xin Wang and Jinbo Bi

Abstract—The problem of constructing classifiers from multiple annotators who provide inconsistent training labels is important and occurs in many application domains. Many existing methods focus on the understanding and learning of the crowd behaviors. Several probabilistic algorithms consider the construction of classifiers for specific tasks using consensus of multiple labelers annotations. These methods impose a prior on the consensus and develop an expectation-maximization algorithm based on logistic regression loss. We extend the discussion to the hinge loss commonly used by support vector machines. Our formulations form bi-convex programs that construct classifiers and estimate the reliability of each labeler simultaneously. Each labeler is associated with a reliability parameter, which can be a constant, or class-dependent, or varies for different examples. The hinge loss is modified by replacing the true labels by the weighted combination of labelers' labels with reliabilities as weights. Statistical justification is discussed to motivate the use of linear combination of labels. In parallel to the expectation-maximization algorithm for logistic based methods, efficient alternating algorithms are developed to solve the proposed bi-convex programs. Experimental results on benchmark datasets and three real-world biomedical problems demonstrate that the proposed methods either outperform or are competitive to the state of the art.

Index Terms—Biconvex optimization, classifier training, multiple annotators, learning from crowds

1 Introduction

Learning from multiple labelers who provide inconsistent annotations to the training data is an emerging machine learning problem. Recent technological innovations have created easily accessible crowdsourcing platforms such as Amazon Mechanical Turk (AMT)[1] and Crowdflower[2], through which tasks such as text or image annotations can be assigned to enormous online annotators at affordable prices. These crowdsourcing methods largely reduce the economic and time costs associated with massive annotating tasks. However, they have imposed great technical challenges in deriving and modeling ground truth in many cases. For instance, to diagnose cancer, although a biopsy provides ground truth, the procedure is complicated and combines with discomforts. Hence, series of X-ray images can instead be read and annotated by multiple radiologists to facilitate the diagnosis. An early work in cancer research reported the problem that different radiologists have different reliabilities of recognizing a lesion in the same diagnostic image [1]. Since one expert's expertise may be biased and/or incomplete, integration of knowledge from multiple experts is necessary to build accountable and reliable cancer informatics systems. Learning from multiple annotators has become necessary and beneficial in a variety of fields. In the bioinformatics field, nature language processing (NLP) tasks have been performed by AMT workers to extract knowledge from biological and medical documents [2], [3], [4]. A recent work [5] recruited non-expert AMT workers to annotate CT images with little prior training and then had to use integration such as majority voting strategies to detect polyps (colon cancer).

The traditional learning task constructs a classifier mapping from input features to ground truth labels, which becomes difficult in the scenarios where ground truth labels are unknown and need to be estimated from multiple annotators of different expertise. Fig. 1 illustrates the learning problem using an example of heart motion analysis. Multiple radiologists annotate a set of echocardiograms in terms of whether an image shows abnormal heart motion. The labels from these radiologists do not agree. The goal is to train a classifier that utilizes the inconsistent radiologist annotations to predict unseen images. The echocardiograms are very difficult to interpret even for the best physicians [6]. Inter-observer studies showed that even world-class experts only agreed on 80% of their diagnoses. The learning problem described in Fig. 1 is especially difficult when radiologists' expertise and reliability are unknown.

Several methods have been proposed in the recent machine learning literature to learn models from crowds, or more precisely, from crowdsourced labels [7]. These methods typically impose a probabilistic model on the labeling process, such as Bernoulli model or Gaussian model on the true labels [8], or two-coin model for annotators [9], [10], and then use an expectation-maximization (EM) process to build logistic regression classifiers. Two recent works [11], [12] also propose convex formulations based on logistic regression, but the true classifier is estimated by taking an average effect of the classifiers trained with each labeler, which may be impacted significantly by malicious labelers or spammers. There has been limited effort in extending support vector machines (SVM) to build classifiers from crowd-annotated data. It has been shown that SVM may bear some advantages over logistic regression when data

• *Xin Wang and Dr. Jinbo Bi are with the Department of Computer Science and Engineering, University of Connecticut, Connecticut, CT, 06279. Correspondence should be addressed to Jinbo Bi. E-mail: xin.wang@uconn.edu, jinbo.bi@uconn.edu*

1. https://www.mturk.com/
2. http://www.crowdflower.com/

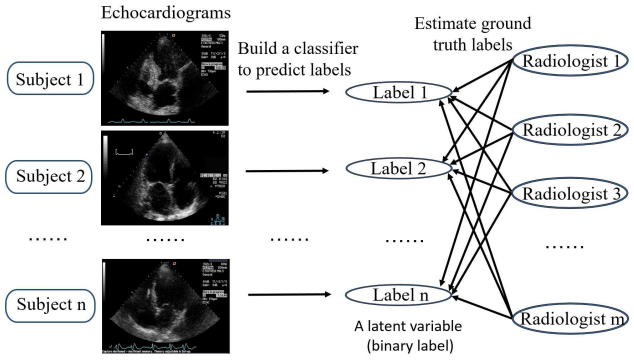


Fig. 1: Classifier training from multiple annotators. Echocardiograms of n subjects are annotated by m radiologists, and the ground truth for each image is unknown. A classifier is constructed to map an image to its ground truth label that is estimated from the different radiologist labels.

follows certain distribution such as multivariate or mixture of distributions, and SVM methods may require less features than logistic regression to achieve a better or equivalent classification accuracy [13], [14], [15].

In this paper, we propose a bi-convex optimization approach that performs simultaneously three tasks: (1) assess how good each labeler is, (2) estimate the true labels, and (3) build a classifier using approximate true labels estimated from the multiple labels. The key step is to modify the hinge loss used in the SVM where the unknown true labels are replaced by their estimates. In the proposed approach, we associate each labeler with a reliability factor. Three learning models, each forming a bi-convex program, are derived by making the hinge loss reflect three different kinds of assumptions on the labeler reliability. The proposed methods follow a general principle that the labels from a more reliable labeler should contribute more to the construction of the classifier. If a labeler has a constant reliability factor, it represents an overall performance of the labeler for the task. For binary classification tasks, if a labeler has a predisposition to one class than the other, his/her reliability differs between the distinct classes, which brings a more complex reliability structure. The most complex one assumes that the labeler reliability varies on individual examples if the labeler is not equally competent to annotate different examples.

2 Related works

Many existing methods for *learning from crowds* focus on modeling of an annotation process and estimating error rates for the labelers independent of any classifiers. The early statistical methods [1], [16], [17] on error rate estimation for repeated but conflicting test results, and the recent work on learning crowd behaviors [18], [19], [20], are good examples. The latest work in this direction ranks annotators to identify spammers [21], uses Multinomial probabilistic models to quantify the competency of each labeler [22], and parameterizes labeler expertise or reliabilities as well as the difficulty of an annotation task to model human annotation more accurately [23], [24], [25]. Moreover, in the work of [26] and [27], reliabilities are estimated from gold standard tasks and then used in a weighted combination for new labeling tasks. Another method in [28] models the labelers using a stochastic model and select examples to teach the labelers via a greedy algorithm. These methods study the problem of optimizing the task assignment in a crowdsourcing system. We adopt the similar strategy as [26] and [27] to aggregate the inconsistent labels assigned by multiple labelers, but the labeler reliabilities are jointly estimated with a classifier.

Recently the interest of learning from crowds has increased to directly build classifiers from multi-labeler data. Repeated labeling methods [29], [30] identify the labels that should be reacquired from some labelers in order to improve classification performance or data quality. A recent theoretical work [31], however, argues that the repeated labeling negatively impacts the relative size of the training sample. Another set of approaches [32], [33] assume the existence of prior knowledge relating the different labelers, and the prior is used to identify the samples for each labeler that are appropriate to be used in the classifier estimation. Several methods [7], [8], [9], [10], [11], [12], [34], [35], however, neither assume that labels can be reacquired, nor assume existence of any prior on labeler relations. These approaches rely on certain data distribution, such as Bernoulli model on the true binary labels or Gaussian model on the true continuous labels [8] or two-coin model on the process of how an annotator provides a label [9], [10], and then develop a posterior solution with logistic regression and use an EM algorithm to estimate the model parameters.

Among the methods that build a classifier and estimate labelers' error rates simultaneously, the models of [11], [12] and [8] are the most similar to our work. In [11], [12], a linear classifier with coefficients $\mathbf{w}_j$ is built for each individual labeler j based on his/her own annotation using logistic regression and the final classifier with a coefficient vector $\mathbf{w}$ is obtained by enforcing a regularization term, that is either $\sum_j ||\mathbf{w}_j - \mathbf{w}||^2$ in [11] or $\sum_{j,k} ||\mathbf{w}_j - \mathbf{w}_k||^2$ where j, k denotes the indexes of the classifiers constructed from an individual labeler's annotation [12]. The final classifier ($\mathbf{w}$) is hence constructed by taking an average effect of individual labeler's classifiers rather than by minimizing the final classifier's own loss on the training data. This classifier may collapse if there are many malicious labelers due to the kind of majority voting effect. In [8], it is assumed that a labeler's competence may vary when annotating different sample points, so a classifier is built for each labeler to parameterize his/her reliability on an example. Then the final classifier is built by modeling the reliabilities of the different labelers in a logistic regression based EM algorithm. Unlike this method, we impose no specific distributions but more general and intuitive assumptions on the labelers' reliabilities.

In one of our recent works [36], we propose a strategy to model simultaneously two sources of labeling ambiguity: (1) the one caused by the inconsistent labels from multiple annotators; (2) and the ambiguity that a class label is associated with a bag of input instances rather than each instance. The second labeling ambiguity is often referred to as the multiple instance learning problem where a bag is labeled as positive as long as one of its instances shows evidence to be positive. We first modify the hinge loss to

take a consensus of different labelers' labels, which is, in spirit, the same as the first model in this current paper - the constant reliability model. Then the work in [36] focuses on the extension of that model to multiple instance learning. In this paper, however, we focus on solving the multiple labeler learning problem by reforming the early constant reliability model and proposing two other more advanced new formulations.

3 The Proposed Formulations

We derive the learning formulations in this section. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ comprise the n examples, where $\mathbf{x}_i \in R^d$, and is annotated with multiple versions of the label $\{y_i^1, y_i^2, \cdots, y_i^m\}$. We focus on the case of binary classification where $y_i^j \in \{-1, 1\}$, $j \in \{1, 2, ..., m\}$. Suppose that the true label of $\mathbf{x}_i$ is y_i and we consider linear models of the form $\mathbf{x}^\top \mathbf{w} + b$ where $\mathbf{w}$ is the weight vector and b is the offset to be determined for the classifier. We derive our models by modifying the hinge loss $[1 - y_i(\mathbf{x}_i^\top \mathbf{w} + b)]_+ = \max\{0, 1 - y_i(\mathbf{x}_i^\top \mathbf{w} + b)\}$ where we replace the unknown true label y_i by a linear combination of y_i^j.

The use of a linear combination of y_i^j as an approximate of y_i is rooted from a probabilistic understanding of the learning problem. For instance, in the model with constant labeler reliability derived in the next section, the essential motivation is that the true (unobserved) label y_i is a linear combination of y_i^j's that are *i.i.d.* sampled from the hidden true y_i, taking a Gaussian form $y_i^j \sim \mathcal{N}(y_i, \sigma_j^2)$ where y_i is the mean and σ_j^2 is the precision. Then the *a posteriori* distribution of y_i given all the observed labels follows $\mathcal{N}(\mu_i, \sigma_i^2)$, where $\mu_i = \sum_j \sigma_j^2 y_i^j / \sum_j \sigma_j^2$, and $\sigma_i^2 = \sum_j \sigma_j^2$. So one can see *a posteriori* mean is a weighted linear combination of all observed labels, and the weights sum to 1.

3.1 The model with constant labeler reliability

We approximate an example's true label y_i by a weighted combination of each labeler's labels, e.g., $y_i \simeq \sum_{j=1}^m r_j y_i^j$ and each labeler j is associated with a reliability factor r_j where $0 \leq r_j \leq 1$. If the reliability factors of all labelers are equal, this combination amounts to the majority voting. If we require additionally $\sum_j r_j = 1$, we approximate y_i by a convex combination of labelers' opinions. We believe these combinations may all be reasonable, and the most appropriate one may be problem-specific. If the weighted consensus of all labelers $\sum_j r_j y_i^j > 0$, the example i is more likely to be in the class of $y = 1$; or otherwise, it likely has a true label of $y = -1$.

We modify the hinge loss by replacing the true labels y_i by the weighted consensus, which yields a bi-convex function $[1 - (\sum_j r_j y_i^j)(\mathbf{x}_i^\top \mathbf{w} + b)]_+$ (convex with respect to $(\mathbf{w}, b)$ for fixed $\mathbf{r}$ and convex with respect to $\mathbf{r}$ for fixed $(\mathbf{w}, b)$). When the consistency is high among the labels given by different labelers, especially by reliable labelers, the magnitude of $\sum_j r_j y_i^j$ tends to be large regardless of its sign, showing high annotation confidence for $\mathbf{x}_i$. Minimizing the modified loss leads to a classifier that works hard to correctly classify $\mathbf{x}_i$. When the labeling consistency is low among reliable labelers for some examples, the assignment of them to either class can be a vague guess. The linear combination of labels will lead to a small value in magnitude due to the cancellation effect of the mixed $+1$ and -1 labels. The modified loss then reports a low value on such cases, which hence does not emphasize the classification performance on these examples. This justifies the validity of the modified loss.

By adding a regularization term $||\mathbf{w}||^2$ to the empirical loss, we minimize the following optimization problem

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\mathbf{r}} & \lambda||\mathbf{w}||^2 + \sum_i [1 - (\sum_j r_j y_i^j)(\mathbf{x}_i^\top \mathbf{w} + b)]_+ \\ \text{s.t.} & \sum_j r_j = 1, \quad r_j \geq 0, \\ & i = 1, 2, ..., n, \quad j = 1, 2, ..., m. \end{array} \tag{1}$$

The constraints on $\mathbf{r}$ are affine, which formulate the convex combinations of labelers' opinions and enforce competition among the labelers by limiting the sum of their reliabilities to a constant 1. It is easy to verify that Problem (1) is a case of bi-convex optimization because the objective function is bi-convex and constraints are affine. To translate the problem into a canonical form, the modified loss is translated into constraints $(\sum_j r_j y_i^j)(\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 - \xi_i$ for each example i where the slack variables $\xi_i \geq 0$, and both $\mathbf{r}$ and $(\mathbf{w}, b)$ are variables to be determined in the following optimization problem

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\boldsymbol{\xi},\mathbf{r}} & \lambda||\mathbf{w}||^2 + \sum_i \xi_i \\ \text{s.t.} & (\sum_j r_j y_i^j)(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \sum_j r_j = 1, \quad r_j \geq 0, \quad \xi_i \geq 0, \\ & i = 1, 2, ..., n, \quad j = 1, 2, ..., m. \end{array} \tag{2}$$

Problem (2) is also a quadratically constrained quadratic optimization problem but with one of its constraints bi-convex.

For binary classification, if the training data is balanced in the distribution of either class (e.g., close to even numbers of positive and negative examples), the proposed model with constant reliability is often sufficient to estimate a labeler's overall reliability. This is sometimes referred to as a one-mode model. The labelers with higher reliabilities are expected to be assigned with larger weights by solving Problem (2). However, this one-mode model will hardly take care of the situation when labelers have different labeling accuracies with respect to the positive or negative class labels. In practice, when the problem data is very unbalanced, the true positive rate and true negative rate will be important factors to reflect the real performance. A model considering a labeler's class-dependent reliability will be needed.

3.2 The model with class-dependent labeler reliability

A labeler's reliability may naturally be class dependent. For instance, online annotators may have different accuracies in labeling documents with respect to different topics relying on whether they are more familiar with some topics than others. If a labeler tends to always label examples to the $+1$ class, his positive predictive value (PPV) (the percentage of examples labeled by the labeler as positive that are actually positive) may be low but his negative predictive value (NPV) (the percentage of examples labeled by the labeler as negative that are actually negative) can be high.

We extend the model discussed in Section 3.1 to class-dependent reliability factors. The model still estimates the true labels by the weighted combination of

each labeler's labels. However, two parameters $\alpha_j \geq 0$ and $\beta_j \geq 0$ are needed to estimate the j-th labeler's PPV and NPV, respectively. We set the true labels $y_i \simeq \sum_{j=1}^{m} \alpha_j^{(1+y_i^j)/2} \beta_j^{(1-y_i^j)/2} y_i^j$. If labeler j gives $y_i^j = +1$, α_j should be used as the corresponding weight for y_i^j, and $\beta_j^{(1-y_i^j)/2}$ is degraded to 1. If $y_i^j = -1$, β_j should be used in the combination. Unlike the constant reliability model, we now require $\sum_j \alpha_j = 1$ and $\sum_j \beta_j = 1$, that can enforce competition among labelers. When the two parameters are used in the modified hinge loss, we optimize the following optimization problem for the best class-dependent reliabilities and classifier

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\mathbf{r}} \; & \lambda||\mathbf{w}||^2 + \sum_i [1 - (\sum_{j=1}^{m} \alpha_j^{\frac{1+y_i^j}{2}} \beta_j^{\frac{1-y_i^j}{2}} y_i^j)(\mathbf{w}^\top \mathbf{x}_i + b)]_+ \\ \text{s.t.} \quad & \sum_j \alpha_j = 1, \quad \sum_j \beta_j = 1, \\ & \alpha_j \geq 0, \quad \beta_j \geq 0, \\ & i = 1, 2, ..., n, \quad j = 1, 2, ..., m. \end{aligned} \tag{3}$$

Bi-convexity still holds for Problem (3) since α_j and β_j are not used at the same time for each y_i^j in the constraints given the values of y_i^j are already known. The same optimization algorithm used to solve Problem (1) can be applied to solve Problem (3). Problem (3) can also be translated into a canonical form by utilizing slack variables ξ to represent the hinge losses, and the resultant optimization problem is written as follows:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\mathbf{r}} \; & \lambda||\mathbf{w}||^2 + \sum_i \xi_i \\ \text{s.t.} \quad & (\sum_{j=1}^{m} \alpha_j^{(1+y_i^j)/2} \beta_j^{(1-y_i^j)/2} y_i^j)(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \sum_j \alpha_j = 1, \quad \sum_j \beta_j = 1, \\ & \xi_i \geq 0, \quad \alpha_j \geq 0, \quad \beta_j \geq 0, \\ & i = 1, 2, ..., n, \quad j = 1, 2, ..., m. \end{aligned} \tag{4}$$

This two-mode model, is similar in spirit, to the two-coin model used in [9], [10], where a labeler's expertise was also described by two factors, sensitivity and specificity. In [9], [10], labelers' expertise, true labels and the classifier were learnt with an EM algorithm based on logistic regression. However, we observe that this prior model can become numerically unstable when a large number of laberlers are present [36]. The two-coin model updates the estimated ground truth denoted by μ, a probability of the true label being $+1$, based on the multiplications of sensitivities and specificities, denoted by $0 \leq \alpha_j \leq 1$ and $0 \leq \beta_j \leq 1$ for labeler j respectively (with a little mixed use of notation). The products $\Pi_{j=1}^{m} \alpha_j$ and $\Pi_{j=1}^{m} \beta_j$, assuming there are m labelers, become extremely small with large m. Consequently, μ becomes oscillating between 0 and 1 since the two products are used in the numerator and denominator of the updating formula for μ.

The proposed model is instead scalable and reliable to deal with a large number of labelers. By selecting high-quality labelers for use in the combination, redundant labelers may be excluded and sparsity has been observed in the estimated reliabilities when a large number of labelers are included. Our empirical results show that, for both the models with constant reliabilities and class-dependent reliabilities, the true labels can be sufficiently estimated from few labelers and information from other labelers might be redundant. Our models could automatically select labelers whose labels are valid to make accurate estimates of the ground truth and exclude correlated or redundant labelers.

3.3 The model with sample-specific labeler reliability

If a labeler is not equally competent to annotate all sample subjects, his/her reliability r will become a factor relying on individual samples $\mathbf{x}$ and hence becomes a function of $\mathbf{x}$ as $r(\mathbf{x})$. Such an issue often takes place in real life applications. Radiologists may not have the equal reliability dealing with high-quality images versus images of different kinds of noise. Few previous studies examined this practical difficulty. The methods in [8], [34] built a classifier $\mathbf{x}^\top \mathbf{w}_j + b_j$ for each labeler j based on his/her own annotation, and defined $r_j(\mathbf{x})$ as a sigmoid translation $(1 + \exp(-(\mathbf{x}^\top \mathbf{w}_j + b_j)))^{-1}$ of the linear classifier. The probability of observing y_i^j was $p(y_i^j) = (1 - r_j(\mathbf{x})^{|y_i^j - y_i|} r_j(\mathbf{x})^{1-|y_i^j - y_i|}$. Due to the use of sigmoid functions and absolute values in the exponent, it has created complex optimization problems.

We will use the functional distance from each $\mathbf{x}_i$ to the separation boundary ($\mathbf{x}^\top \mathbf{w}_j + b_j = 0$) that is computed from a labeler j's annotation to determine the labeler's confidence on labeling $\mathbf{x}_i$ relative to other labelers. The farther from the separation boundary, the more confident the labeler annotates the example. Hence, the reliability function $r_j(\mathbf{x}_i) = |\mathbf{x}_i^\top \mathbf{w}_j + b_j| / \sum_j |\mathbf{x}_i^\top \mathbf{w}_j + b_j|$. The denominator is used to compute the jth labeler's confidence relative to other labelers' confidence. The individual labelers' classifiers can be built by minimizing standard hinge loss defined as $\eta_i^j = [1 - y_i^j (\mathbf{x}_i^\top \mathbf{w}_j + b_j)]_+$. Since these classifiers are used to determine reliabilities, they should be constructed more or less accurately (i.e., close to the final classifier determined by $\mathbf{w}$), which motivates to impose an additional regularizer $R(\mathbf{w}, \mathbf{w}_j) = \sum_j ||\mathbf{w} - \mathbf{w}_j||^2$ assuming the final classifier is a more accurate estimate of the true classifier. Importantly, this regularizer will enforce individual labeler's classifiers to have similar $||\mathbf{w}_j||$. Because $r_j(\mathbf{x}_i)$ is defined through a functional distance from $\mathbf{x}_i$ to the boundary, the similarity among different $||\mathbf{w}_j||$ will render that $r_j(\mathbf{x}_i)$ is largely proportional to the geometric (Euclidean) distance from the point to the boundary.

We define the modified hinge loss $\xi_i = [1 - (\sum_j \frac{(|\mathbf{x}_i^\top \mathbf{w}_j + b_j|}{\sum_j (|\mathbf{x}_i^\top \mathbf{w}_j + b_j|} y_i^j)(\mathbf{x}_i^\top \mathbf{w} + b)]_+$ and the additional constraint for ξ_i would be $(\sum_j \frac{(|\mathbf{x}_i^\top \mathbf{w}_j + b_j|}{\sum_j (|\mathbf{x}_i^\top \mathbf{w}_j + b_j|} y_i^j)(\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 - \xi_i$. The modified hinge loss appears complex. However, it can be re-organized through simple algebraic calculations by moving the denominator in $r_j(\mathbf{x}_i)$ to the right-hand side. After re-organization, this constraint becomes $\sum_j |\mathbf{x}_i^\top \mathbf{w}_j + b_j| (y_i^j (\mathbf{x}_i^\top \mathbf{w} + b) - (1 - \xi_i)) \geq 0$. The use of the absolute value of $\mathbf{x}_i^\top \mathbf{w}_j + b_j$ can complicate the optimization problem. Hence, we replace the absolute value by the upper bound $u_i^j \geq 0$ that satisfies the constraints: $-u_i^j \leq \mathbf{x}_i^\top \mathbf{w}_j + b_j \leq u_i^j$ for all $i = 1, \cdots, n$ and $j = 1, \cdots, m$. Overall, we search for the best classifier $(\mathbf{w}, b)$

and the most accurate reliability estimate based on $(\mathbf{w}_j, b_j)$ by optimizing the following problem:

$$\begin{aligned}\min_{\mathbf{w},b,\boldsymbol{\xi},\mathbf{w}_j,b_j,\boldsymbol{\eta},\mathbf{u}} \quad & \lambda_1||\mathbf{w}||^2 + \lambda_2 \sum_j ||\mathbf{w}-\mathbf{w}_j||^2 \\ & + \sum_i \xi_i + \sum_i \sum_j \eta_i^j + \sum_i \sum_j u_i^j \\ \text{s.t.} \quad & \sum_j u_i^j (y_i^j(\mathbf{x}_i^\top \mathbf{w} + b) - (1-\xi_i)) \geq 0, \\ & -u_i^j \leq \mathbf{x}_i^\top \mathbf{w}_j + b_j \leq u_i^j, \\ & y_i^j(\mathbf{x}_i^\top \mathbf{w}_j + b_j) \geq 1 - \eta_i^j, \\ & \xi_i \geq 0, \quad \eta_i^j \geq 0, \quad u_i^j \geq 0, \\ & i = 1, 2, ..., n, \quad j = 1, 2, ..., m.\end{aligned} \tag{5}$$

Problem (5) has a convex objective function, a bi-affine constraint (the first constraint) and all other constraints are affine. This problem is also a bi-convex program. We can group the variables into two groups: one group is related to the final classifier including variables $\mathbf{w}, b, \boldsymbol{\xi}$, and the other group is related to individual classifiers including variables $\mathbf{w}_j, b_j, \boldsymbol{\eta}_j, \mathbf{u}$. When fixing one group of the variables, Problem (5) becomes a convex quadratic program in terms of the other group of variables.

Besides the regularizer that enforces the similarity between individual $\mathbf{w}_j$'s and the final classifier's $\mathbf{w}$, an additional regularizer $||\mathbf{w}_j||^2$ can be directly applied to individual $\mathbf{w}_j$. We will also evaluate an alternative formulation by revising the objective function in Problem (5) to

$$\begin{aligned}\min_{\mathbf{w},b,\boldsymbol{\xi},\mathbf{w}_j,b_j,\boldsymbol{\eta},\mathbf{u}} \quad & \lambda_1||\mathbf{w}||^2 + \lambda_2 \sum_j ||\mathbf{w}-\mathbf{w}_j||^2 \\ & + \lambda_3 \sum_j ||\mathbf{w}_j||^2 + \sum_i \xi_i + \sum_i \sum_j \eta_i^j \\ & + \sum_i \sum_j u_i^j\end{aligned} \tag{6}$$

where the same constaints in Problem (5) apply.

Besides the prior methods in [8], [34], the methods in [11], [12] also estimate a labeler's reliability by building a classifier $(\mathbf{w}_j, b_j)$ from the labeler's own annotations. However, the final classifier $(\mathbf{w}, b)$ was built by minimizing $\sum_j ||\mathbf{w} - \mathbf{w}_j||$, and hence $\mathbf{w}$ is estimated as the centroid of all $\mathbf{w}_j$'s, and thus suffering from significant outlier labelers.

4 The Optimization Algorithm

In this section, we present an effective algorithm to solve the proposed Problems (1), (3), (5) and (6) respectively. We adopt the alternating optimization approach for each problem where we alternate between solving for the final classifier and solving for the parameters related to reliability iteratively until the algorithm converges. Because the three proposed problems are all bi-convex as discussed in the last section, the subproblems formulated for solving each group of variables is convex. To solve the subproblems in (1) and (3), we used their equivalent formulations (2) and (4) by introducing slack variables. Algorithm 1 summarizes the algorithmic procedure for each of the problems. For illustration convenience, we list the procedure for each of the problems together in Algorithm 1.

The initialization choices listed in Algorithm 1 are believed to follow the most common sense. For instance, without prior knowledge, we may assume that all labelers are equally competent (with equal initial $\mathbf{r}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) and let the algorithm determine and update the reliability factors based on sample data. We also empirically notice that the algorithm is insensitive to initial values in the sense that it gives the same solution when we perturb the listed initial values by random white noise.

Algorithm 1 Alternating Optimization Algorithm for the Proposed Bi-convex Programs

Input: λ's, $\mathbf{X}$, $\mathbf{y}^j$, $j = 1, \cdots, m$, and a tolerance ϵ.
Initialize: let k denote the current number of iterations, and k is initialized to 0. Let Θ denote the set of all variables needed to be optimized.
For the constant reliability model, set $\mathbf{r}^{(0)} = 1/m$. For the class-dependent reliability model, set $\boldsymbol{\alpha}^{(0)} = 1/m$ and $\boldsymbol{\beta}^{(0)} = 1/m$. For the sample-specific reliability model, set $(\mathbf{w}_j^{(0)}, b_j^{(0)})$ to the SVM classifiers that are built from each labeler's labels, $j = 1, \cdots, m$.
repeat
  Step 1:
  For Problem (1), solve Problem (2) for $(\mathbf{w}^{(k)}, b^{(k)})$ and $\boldsymbol{\xi}$ with fixed $\mathbf{r}^{(k-1)}$.
  For Problem (3), solve Problem (4) for $(\mathbf{w}^{(k)}, b^{(k)})$ and $\boldsymbol{\xi}$ with fixed $\boldsymbol{\alpha}^{(k-1)}$ and $\boldsymbol{\beta}^{(k-1)}$.
  For Problem (5) and (6), solve for $(\mathbf{w}^{(k)}, b^{(k)})$ and $\boldsymbol{\xi}$ with fixed $(\mathbf{w}_j^{(k-1)}, b_j^{(k-1)})$, $\boldsymbol{\eta}_j^{(k-1)}$, and $\mathbf{u}_j^{(k-1)}$.
  Step 2:
  For Problem (1), solve Problem (2) for $\mathbf{r}^{(k)}$ and $\boldsymbol{\xi}$ with fixed $(\mathbf{w}^{(k)}, b^{(k)})$.
  For Problem (3), solve Problem (4) for $\boldsymbol{\alpha}^{(k)}$ and $\boldsymbol{\beta}^{(k)}$ with fixed $(\mathbf{w}^{(k)}, b^{(k)})$.
  For Problem (5) and (6), solve for $(\mathbf{w}_j^{(k)}, b_j^{(k)})$, $\boldsymbol{\eta}_j^{(k)}$, and $\mathbf{u}_j^{(k)}$ with fixed $(\mathbf{w}^{(k)}, b^{(k)})$ and $\boldsymbol{\xi}^{(k)}$.
until $||\Theta^{(k)} - \Theta^{(k-1)}||^2 \leq \epsilon$.
Output: $(\mathbf{w}, b)$, and $\mathbf{r}$, (or $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, or $(\mathbf{w}_j, b_j)$).

We point out a small derivation difference in solving the three problems. Problems (1) and (3) can be solved using the same split of working variables, that is the algorithm optimizes either $(\mathbf{w}, b)$ or $\mathbf{r}$ (or $(\boldsymbol{\alpha}, \boldsymbol{\beta})$) in an alternating step. The slack variables ξ_i in Problems (2) and (4) are only used to update the hinge loss in Problems (1) and (3), respectively, at each step and hence they are included in both the working groups of variables. For the sample-specific reliability model, the modified hinge loss is not bi-convex by its literal form. We have reformulated the problem using the first constraint in Problems (5) and (6). In this situation, we group the variables ξ_i with the final classifier parameters $(\mathbf{w}, b)$.

According to the convergence analysis in [37], [38], the alternating Algorithm 1 used for solving our programs (1), (3), (5) and (6) converges to a set of fixed points which in general includes global minimizers, local minimizers and the saddle points. Due to the bi-convexity, the fixed points of our programs do not include saddle points [38].

We give a brief discussion on the complexity of Algorithm 1 using Problems (1) and (3). To optimize for $(\mathbf{w}, b)$, Algorithm 1 solves a quadratic program similar to SVM. To solve for $\mathbf{r}$ (or $(\boldsymbol{\alpha}, \boldsymbol{\beta})$), Algorithm 1 solves a linear program.

Effective algorithms such as Dantzig's simplex method, or later interior point methods have been developed for these programs. In our implementation, we used the CPLEX software to solve them with choices of simplex-based methods. Rigorous bounds on the number of operations required by these methods have been established. For instance, the complexity of solving the linear program is in order of $d^2\ell$ with a constant where d is the number of variables and ℓ is the number of constraints in the program. We observe that Algorithm 1 typically stops within 10 iterations, so the overall complexity of Algorithm 1 is a constant (e.g., 10) multipying the sum of the complexity of solving the linear and quadratic programs.

5 Experiments

The proposed methods were tested on five benchmark datasets at first. Four of them are commonly used in evaluating machine learning algorithms. These datasets are all for binary classification and come with ground truth labels, but they are not labeled by multiple annotators. We created synthetic labelers for these datasets. The fifth benchmark dataset is the facial expression dataset where each face shot image was labeled by multiple real online workers. Besides the benchmark datasets, we also tested the proposed methods on three real-world problems of analyzing biomedical images. The first problem was to detect breast cancer in digitized mammographic images. The other two problems aimed respectively to detect Heart Wall Motion Abnormality (HWMA) using features extracted from echocardiogrmas and to diagnose Alzheimer's disease using features extracted from magnetic resonance imaging (MRI).

Our methods were compared against four recently-published methods, all of which construct classifiers: Two-coin model [10], EMGaussian model and EMBernoulli model [8], and a convex model [11]. The classifier trained with ground truth labels was supposed to achieve the best performance whereas the majority voting approach served as our baseline. The proposed methods *Model with Constant Reliability, Model with Class-Dependent Reliability and Model with Sample-Dependent Reliability* were respectively referred to as MCR, MCDR and MSDR.

Ten-fold cross validation (CV) was used to run all algorithms on each of the datasets with the same stratified CV split. For the proposed methods and the convex method in [11], we tuned their regularization parameters within the training data in the first CV fold using another internal three-fold CV for each dataset and then fixed the parameters for the nine remaining CV folds. We selected the parameters that obtained the best performance from the range of $[10^{-3}, 10^{-2}, \cdots, 10^{3}]$.

5.1 Benchmark datasets

In this section, we provide the details on the experimental procedure and results obtained on the benchmark datasets.

5.1.1 UCI datasets with synthetic annotators

We used four datasets including Cleveland, Glass, Ionosphere and Pima, which were downloaded from UCI Machine Learning Repository[3]. Table 1 has the details about these datasets. The datasets were preprocessed for performing binary classification. Although all these datasets had no multiple versions of labels from real labelers, they were frequently used by many previous multi-labeler learning methods, including the three methods in [9], [10], [8] that we compared in this paper. The rational was to have ground truth labels and known labeler reliabilities to test against the algorithms.

TABLE 1: Details of the UCI benchmark datasets with synthetic labelers

Dataset	Cases(positive)	Features
Cleveland	303 (139)	13
Glass	214 (163)	9
Ionosphere	351 (126)	33
Pima	768 (268)	8

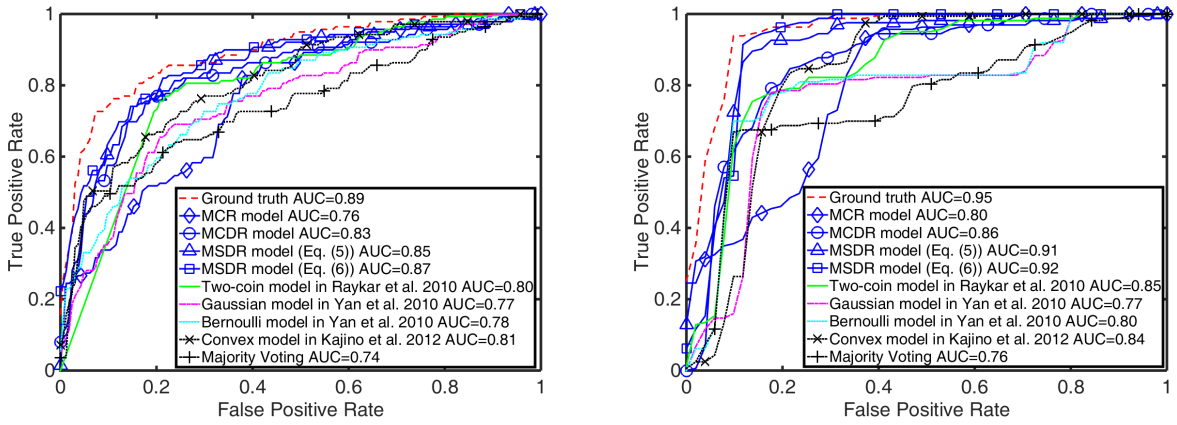
Since it was not straightforward to create synthesized labelers according to the pre-specified PPV's and NPV's, we created the labelers based on the pre-fixed sensitivities and specificities. The labelers were created following the same procedure used in [10]. We first specified two parameters for each labeler, the sensitivity α and specificity β. Five synthetic labelers were created for each of the four datasets. Their sensitivities and specificities were pre-defined as [0.6,0.6, 0.5, 0.7, 0.2] and [0.6, 0.6, 0.5, 0.2, 0.7], respectively. The third labeler's performance was close to a random guess. The first two labelers were given equal sensitivity and specificity, while the last two labelers were prejudicial in the sense that one of them had higher sensitivity and the other one had the exactly opposite parameter values.

Once the parameters were specified for a labeler, a random number was generated uniformly from $[0, 1]$ for each example. When the true label was $+1$ (or -1), if the random number was not bigger than the labeler's α (or β), this labeler chose the original label; or otherwise, (s)he flipped the sign of the label. After the labelers were created, their PPV's and NPV's were calculated.

In order to simulate the case where labelers have different levels of reliability on different examples, we randomly selected 50% of the data samples and ran k-means cluster analysis to group them into five subgroups. Then we made each of the five simulated labelers particularly accurate in annotating one of the subgroups, respectively, with no overlapping. Their labels coincided with the golden truth on the subgroup which they were assigned to. For the rest of data samples not belonging to the subgroup that was assigned to a labeler, the labeler will presume the same sensitivity and specificity levels used in the early experiments.

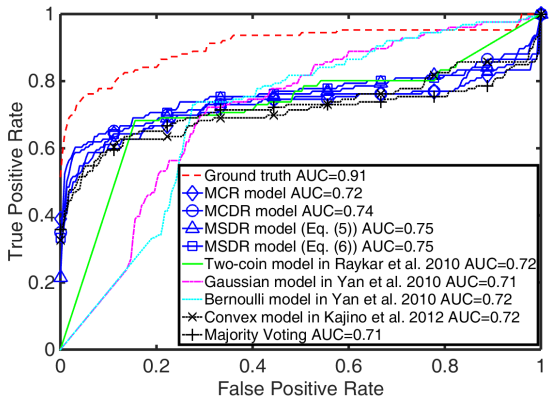
Fig. 2 shows the Receiver Operating Characteristic (ROC) curves achieved by all the methods in comparison on the four datasets with five simulated labelers. The ROC was plotted by merging all the validation data from the 10 folds of the CV. From the ROC plots, we found that the MSDR model with Eq.(6) generally achieved superior performance over the other models by learning the varying expertise jointly with estimating the true labels. Among the other

3. https://archive.ics.uci.edu/ml/datasets.html

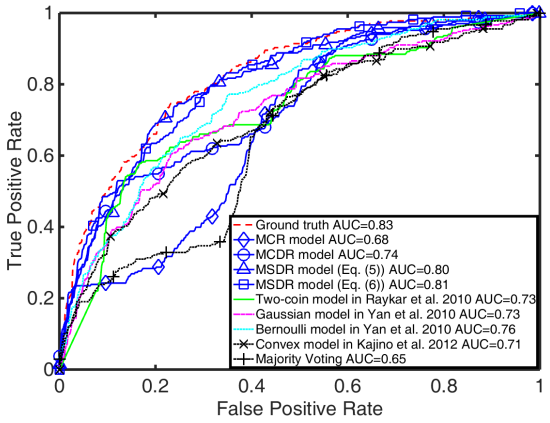


(a) Cleveland

(b) Glass



(c) Ionosphere



(d) Pima

Fig. 2: ROC curves on Cleveland, Glass, Ionosphere and Pima datasets with five simulated labelers (where two of them were simulated as good labelers, the third labeler was close to a random guess, the forth one was more accurate in labeling positive examples than negative ones and the last labeler was on the opposite of the forth one.).

models, MCDR as an extension of the MCR model, which estimated labelers' weights based on the PPV and NPV, consistently achieved better performance than the MCR model that only used one parameter to capture the labeling accuracy. Compared to the method in [9], [10] which also built a two-coin model (with two parameters), the MCDR model could also achieve a slightly better performance in general.

5.1.2 *Facial expression recognition dataset*

The facial expression dataset was previously used to study crowdsourcing behavior [39]. The original dataset contained 585 head-shots of 20 users. For each user, images were collected in which the user could be looking at 4 directions: straight, left, right and up, and the user could present 4 different kinds of facial expression: neutral, happy, sad and angry. The images were labeled with respect to the 4 types of facial expression by totally 27 online labelers at the Amazon Mechanical Turk. Because not all labelers labeled each image, on average, each image was labeled by 9 labelers. The previous study reported a labeling accuracy of only 63.3% using majority voting among labelers. Hence, the task of building a good feature-based classifier is very challenging.

We selected 220 images with the users looking straight, left and right without wearing sunglasses and performed experiments to classify, based on the image features, if an image contained a happy face. Twenty-four labelers were involved in labeling the 220 images. True positive labels were associated with 55 of the images, and the rest were labeled by -1. We set the missing labels to 0, which would automatically be ignored by any of the comparison methods. We segmented the region of an image containing a human face into 6×6 blocks. Local Binary Pattern (LBP) features [40] were extracted from each block and we aligned all these features together (2088 of them) to represent an image. We applied principal component analysis to reduce the dimensions to 120 that explained $\geq$ 95 % of the total data variance.

The area under the ROC curve (AUC) of each classifier was reported and summarized into Table 2 (the first column), where we used the actual labels of each image collected from online workers. Due to the difficulty of the problem itself and the significant amount of missing labels, all methods achieved modest AUC values. Our models MCDR, MSDR (both Eqs.(5) and (6)) and the Bernoulli model were among the best methods with MSDR models performing slightly better. All multi-labeler methods outperformed the

majority voting baseline.

To test how the compared methods perform as the number of annotators increased, we also created more synthetic labelers following the same procedure as mentioned in Section 5.1.1. We set 30% of the labelers to have sensitivities and specificities around [0.6, 0.6], 30% around [0.5, 0.5] (random guess), 20% around [0.8, 0.2] while the rest 20% with [0.2, 0.8]. The results were also reported in Table 2 (from the second to the last columns) which clearly show that the difference in performance was magnified. The two MSDR models improved the performance by 3% to 13% over the other multi-labeler learning methods in these experiments. We also ran an experiment with 1000 synthesized labelers (results not shown in Table 2). We observed that the two-coin model of [10] had extremely worse performance (AUC = 0.55) than other models (e.g., the best MCDR AUC = 0.73), which shows that this model may perform poorly with a large number of annotators. Besides, the convex model of [11] did not work well either since this model suffered a lot from the synthesized annotators with lower accuracies (AUC = 0.57).

Fig. 3 shows the average run time for an iteration of each method versus the number of annotators. All methods required longer time as the number of annotators increased. The two proposed models MCR and MCDR were more scalable since their run time curves were flatter than others and time costs were lower. It is partially because increasing the number of annotators only affects the optimization of the sub-problem, i.e., solving Problem (1) and (2) for $\mathbf{r}^{(k)}$ and $\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}$ when the classifier parameters are fixed. This sub-problem is a simple linear program and easily scalable with a large number of labelers. Given the two formulations of MSDR had similar run time, Fig. 3 reports the run time for MSDR with Eq.(5) only. The MSDR model was timely consuming in comparison with other models that also built individual labelers' classifiers, which may require the development of a more efficient optimization algorithm and we leave it for future work.

TABLE 2: AUC comparision on the facial expression dataset when the number of labelers increases.

Methods	$24^{\dagger}$	$40^{\ddagger}$	$60^{\ddagger}$	$80^{\ddagger}$	$100^{\ddagger}$	$200^{\ddagger}$
MCR	0.66	0.58	0.58	0.57	0.59	0.66
MCDR	0.68	**0.68**	0.60	0.60	0.62	0.70
MSDR (Eq. (5))	0.70	**0.68**	**0.68**	**0.63**	0.67	**0.73**
MSDR (Eq. (6))	**0.71**	**0.68**	0.67	**0.63**	**0.70**	**0.73**
Two-coin model	0.68	0.64	0.59	0.59	0.64	0.63
Gaussian model	0.66	0.61	0.61	0.56	0.61	0.59
Bernoulli model	0.67	0.65	0.65	**0.63**	0.61	0.66
Convex model	0.66	0.63	0.61	0.60	0.61	0.61
Majority voting	0.62	0.60	0.56	0.58	0.57	0.59

† Real annotators

‡ Synthetic annotators

5.2 Biomedical Image Analysis

To diagnose a complex disease, a diagnostic image is often interpreted by multiple radiologists to enhance the diagnostic accuracy. In this section, we describe how the proposed methods can help with cancer detection, heart abnormality detection and Alzheimer's disease analysis based on features that were extracted from a variety of imaging modalities, including mammographic images, ultrasound clips or MRI scans of brain. The mammographic images and echocardiograms were annotated by multiple radiologists. The MRI dataset of Alzheimer's disease contained records for multiple visits and a doctor's annotation was supplied in each visit. The final reading of the images was also provided and served as the ground truth labels for a patient. We used the diagnoses in the different visits as multiple annotations or created synthetic labelers to provide multiple versions of annotations.

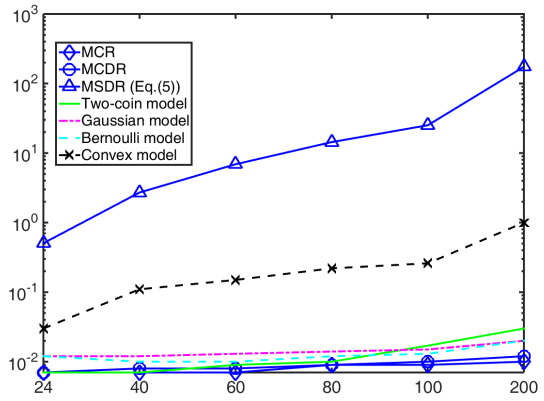


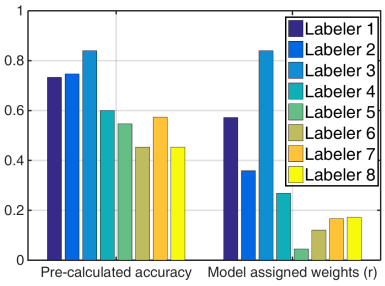
Fig. 3: Average runtime per iteration for every method on facial expression datasets.

5.2.1 *Detecting breast cancer in mammographic images*

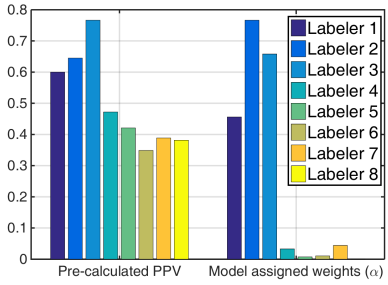
In this dataset, 75 mammograms were collected from real patients, of which the ground truth labels were obtained from biopsy which annotated whether the mammographic image contained a lesion. There were 28 positive samples (having a lesion) and 47 negative samples. Each sample image was represented by 8 attributes and was associated with the labels assigned by three radiologists. We created 5 more synthetic labelers by leveraging the ground truth labels. The labelers were synthesized with sensitivities [0.60, 0.50, 0.50, 0.20, 0.70] and specificities [0.60, 0.50, 0.50, 0.70, 0.20], which controlled the accuracies of the labelers in terms of annotating either a positive or negative example.

We drew the ROC curves of the classifiers constructed by the different methods in comparison together with the AUC statistic in Fig. 4. According to the AUC values, MSDR model (Eq.(6)) performed better than all the other multi-labeler learning algorithms. MCR achieved the lowest performance among the multi-labeler models but still outperformed the majority voting baseline. The two-coin model and MCDR performed similarly probably because both used two reliability parameters. Among the three models that used sample-specific reliabilities, our model was the best (beter than EMGaussian, EMBernoulli, and the early convex model).

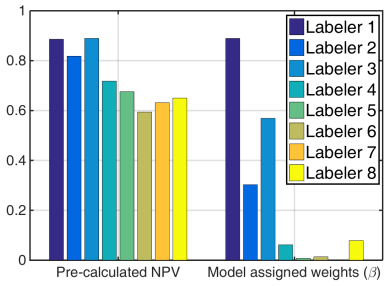
Fig. 5 (5a, 5b and 5c) shows the estimated reliability factors and compares them against the true labels or the simulated labeler performance. The first three labelers represent the radiologists. From Figs. 5a, 5b and 5c, we observed



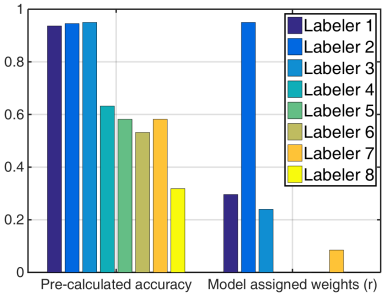
(a) Estimate $\mathbf{r}$ (Mammography data)



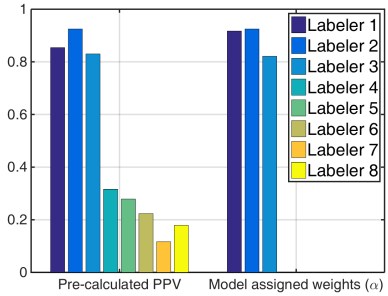
(b) Estimate $\boldsymbol{\alpha}$ (Mammography data)



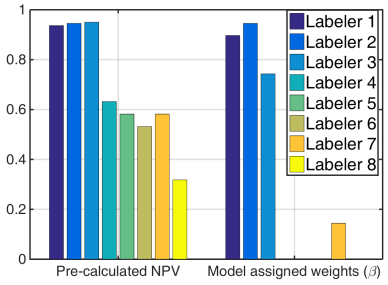
(c) Estimate $\boldsymbol{\beta}$ (Mammography data)



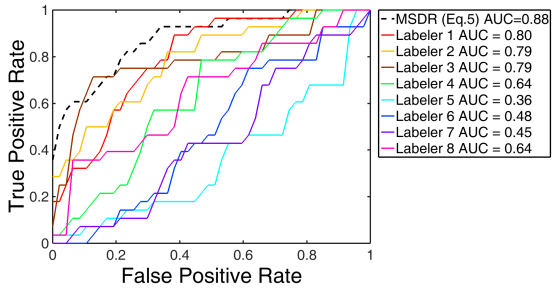
(d) Estimate $\mathbf{r}$ (HWMA data)



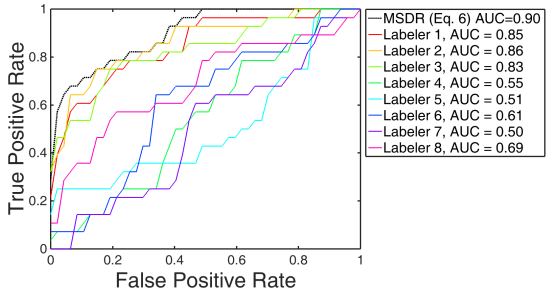
(e) Estimate $\boldsymbol{\alpha}$ (HWMA data)



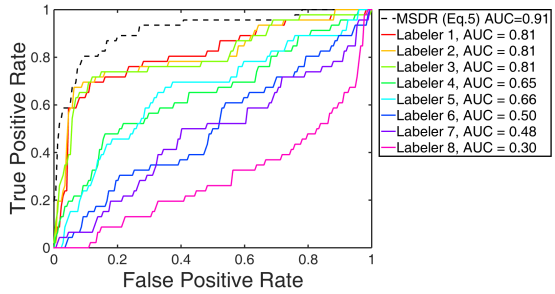
(f) Estimate $\boldsymbol{\beta}$ (HWMA data)



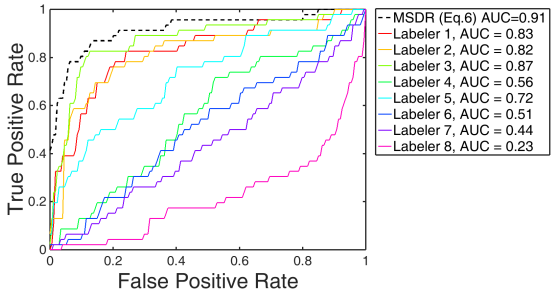
(g) $\{\mathbf{w}, b\}$ vs $\{\mathbf{w}_j, b_j\}$ (Mammographic images, MSDR (Eq. 5))



(h) $\{\mathbf{w}, b\}$ vs $\{\mathbf{w}_j, b_j\}$ (Mammographic images, MSDR (Eq. 6))



(i) $\{\mathbf{w}, b\}$ vs $\{\mathbf{w}_j, b_j\}$ (HWMA data, MSDR (Eq. 5))



(j) $\{\mathbf{w}, b\}$ vs $\{\mathbf{w}_j, b_j\}$ (HWMA data, MSDR (Eq. 6))

Fig. 5: The figure shows the various parameters learned on Mammography and HWMA datasets. Sub-figures (a), (b), (c) and (g) are drawn for Mammography dataset, while (d), (e), (f) and (h) belong to HWMA dataset. Further, (a) and (d) show the estimated reliabilities by MCR against the true labeler accuracies; (b), (c) and (e), (f) show the estimated $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by MCDR against the synthesized PPV and NPV; (g), (h), (i) and (j) show the ROC curves of the two final classifiers and each labeler's classifier obtained by MSDR.

that the MCR and MCDR models are able to sketch a general picture of the varying labeler expertise that is close to the true values/trend. For the MSDR model that builds a final classifier jointly with individual labelers' classifiers, the ROC plot in Figs. 5g and 5h show performance for the classifiers constructed by the two MSDR models. The final classifier clearly outperformed the classifiers built from any individual labeler's data. Our models created shrinkage effects that produced sparse $\mathbf{r}$ (or sparse $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$). As discussed early on, this shrinkage effect shows that true labels can be estimated from few reliable labelers for the tested datasets.

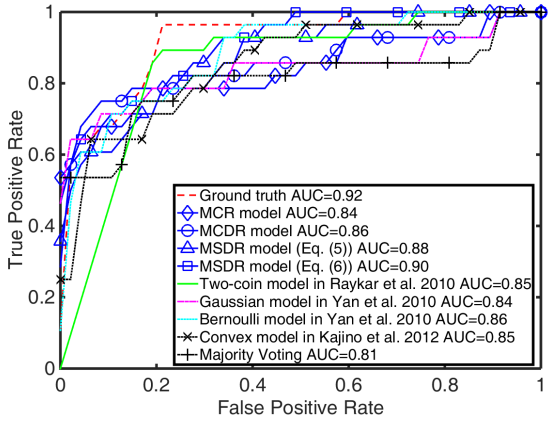


Fig. 4: ROC comparison on the mammography dataset.

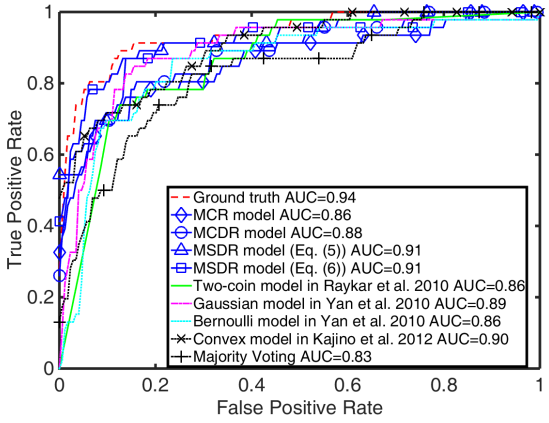


Fig. 7: ROC comparison on HWMA dataset

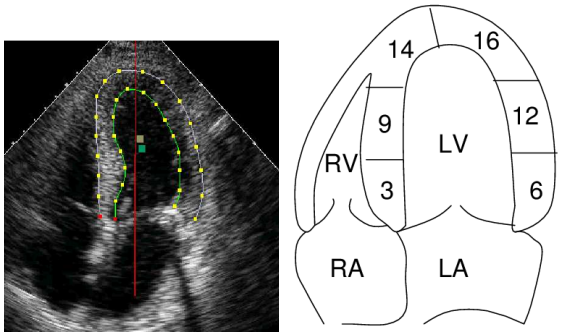


Fig. 6: Left: an ultrasound image of Apical 4 Chamber (A4C) view; right: the 6 heart segments seen from the A4C view.

5.2.2 Heart wall motion analysis

The Heart Wall Motion Abnormality (HWMA) detection dataset contained the features extracted from the images of the wall motion of left ventricles in 222 heart cases. The wall of left ventricle is medically segmented into 16 segments. Fig. 6 shows 6 of the 16 wall segments seen from the apical 4 chamber (A4C) view of an ultrasound clip. For each segment, 25 features were extracted. The feature extraction process was described in more detail in [6]. For each heart case and each segment, the ratings are provided by 5 doctors as the level of severity ranking from 1 to 5, besides, 0 would stand for the missing ratings. We assume that if the ratings are greater or equal to 2 then the label can be set as +1, which means there exists abnormality, otherwise the label is -1. Additionally, at the heart level, if two or more segments of one heart have been claimed as abnormal, the heart-level label would be +1, which means the heart overall has abnormality, otherwise it is -1.

In this experiment, among all these 16 segments, the data extracted from segment 14 were more balanced than the other ones. We used this set of data to test our methods. Because only two cases from this dataset missed radiologists' ratings, there were total 220 examples. Since there was no ground truth available for the data, it is reasonable to make the majority voted labels from the 5 real doctors be the ground truth, and then we randomly selected three real doctors and created 5 synthetic labelers using the same settings for varying sensitivities and specificities as in Section 5.1.1. Fig. 7 shows that the two proposed MSDR methods achieved the superior performance over the other methods.

Similarly, we also illustrated the reliability factors reported by the proposed models MCR and MCDR, which were included in Figs. 5 (5d, 5e and 5f). The real radiologists were shown as the first three annotators. We observed that the proposed models excluded most of the synthetic annotators whose labels were not in good quality. The labels from three radiologists were already sufficient to train the classifier well. Figs. 5i and 5j show the classifiers trained by the MSDR model and each annotator's labels, where we can see that the three radiologists had similar labeling expertise and they were much better than synthetic labelers. The MSDR model combined the expertise of good labelers and thus achieved the best performance.

5.2.3 MRI-based Alzheimer's disease analysis

We tested the proposed models on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset[4]. In the ADNI project, the collected data such as MRI and PET images of participants are used as the predictors to predict the progression of disease. The data we used contained 3063 MRI images taken from 882 participants including Alzheimer's disease (AD) patients, mild cognitive impairment (MCI) subjects and elderly controls. The participants were included in two ADNI study phases, ADNI GO and ADNI$_2$. Fig. 8 shows an example of a participant's brain MRI image in the axial view. A participant had multiple MRI scans collected as he/she had several follow-up visits and the MRI scans were taken at each visit.

We used each MRI image as an example and constructed the classifier to predict the diagnosis of AD or MCI based on the features extracted from MRI images. Among all the 3063 MRI images, there were 833 normal cases (labeled by -1) whereas the remaining images are for AD or MCI patients (labeled by +1). Each MRI image was preprocessed by FreeSurfer[5] and represented by 307 features. The features can be categorized into 5 types: cortical thickness average,

4. The ADNI website: http://adni.loni.usc.edu/
5. http://adni.loni.usc.edu/methods/mri-analysis/

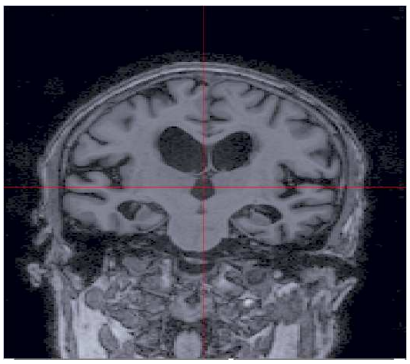


Fig. 8: An example image of an MRI scan along the axial view

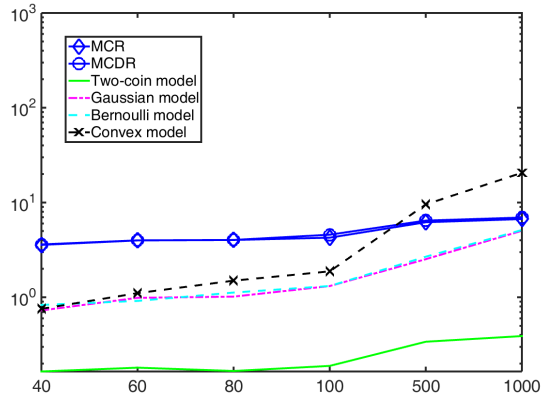


Fig. 9: Average runtime per iteration for every method on ADNI dataset. The x-axis indicates the number of labelers used in the experiments.

cortical thickness standard deviation, volume of cortical parcellation, volume of white matter parcellation and surface area.

In our first set of experiments, we extracted the data for 147 patients who completed (all) four visits at the month 3, 6, 12 and 24, respectively, and then we used the diagnoses for the first three visits as annotated labels so we had three different versions of the label. We used the diagnoses for the forth visit as the ground truth as it gave the latest stage of AD and MCI. Among all the compared methods, the classifier trained with the ground truth served the oracle model with the best performance of an AUC value of 0.64. The classifier trained with majority voted labels served as the baseline (AUC=0.58). The other methods achieved similar performance in general. The MCDR and MSDR models performed slightly better than others. (However, the difference became more significant when we increased the number of labelers as shown below). The three annotations at the month 3, 6, 12) served as good labelers with accuracies of [0.9, 0.93, 0.97] where the last labeler was the best. The relative labeling accuracy was reflected in the estimated reliabilities by the MCR model. For instance, $\mathbf{r}$=[0.2, 0, 0.8] indicated that the last labeler itself plays significant role in predicting the final diagnosis. We observed the similar labeler selection in the MCDR model given $\boldsymbol{\alpha}$=[0, 0.1, 0.9] and $\boldsymbol{\beta}$=[0, 0, 1].

TABLE 3: AUC comparision on the ADNI dataset when the number of synthetic labelers increases.

Methods	40	60	80	100	500	1000
MCR	0.63	0.68	0.71	0.74	0.73	0.78
MCDR	**0.66**	**0.72**	**0.74**	**0.76**	**0.76**	**0.81**
Two-coin model	**0.66**	0.70	0.70	0.72	0.71	0.70
Gaussian model	0.61	0.69	0.65	0.67	0.74	0.76
Bernoulli model	0.61	0.63	0.64	0.69	0.75	0.78
Convex model	0.65	0.66	0.65	0.65	0.68	0.70
Majority voting	0.61	0.63	0.63	0.60	0.65	0.67

In the second set of experiments, we created [40, 60, 80, 100, 500, 1000] synthetic labelers in the same way as the description in the experiments with the facial expression dataset. Because the MCR and MCDR models were more scalable to large datasets than the MSDR models, we further tested the MCR and MCDR models on the ADNI dataset using all 3063 images. The AUC values for the different methods were summarized into Table 3. The results in the table show that the MCDR model had achieved a superior performance when we increased the number of labelers. We also recorded the averaged run time of one iteration for these methods. The comparison of the run time in Fig. 9 shows that the two-coin model required the lowest run time across all the experiments with different numbers of labelers. When the number of labelers was relatively small, the MCR and MCDR had larger time costs than the other models. However, the two proposed models were more scalable to the number of labelers as we can see the run time curves were more flat. The convex model of [11] became more time consuming than the MCR and MCDR models when the number of labeler increased to 500 and 1000.

6 Conclusion

We have studied the multi-labeler learning problem that constructs classifiers from crowdsourcing labels and proposed three novel and unique formulations that all form bi-convex programs. By approximating the true labels with a weighted consensus of all labelers' opinions with the labeler reliabilities as the weights, we are able to modify the hinge loss function to become bi-affine with respect to the classifier parameters and the reliability factors of labelers. We employed three very general assumptions on the labeler reliability, including constant, class-dependent, and example-specific labeler reliability. The bi-convex programs can be effectively optimized by the widely-used alternating optimization algorithm, and outperform the state of the art in empirical tests.

Future extension of this work can examine the bi-convexity of the models more thoroughly, and explore some global optimization algorithms such as the one in [2] that can find a global minimizer for a bi-convex program although these algorithms are significantly more complex. It is also worthy examining the varying reliability scores estimated by the third model, which may prove potential utility in real-world applications, for example, to group

the crowdsourcing labelers according to their labeling reliabilities and behaviours. The datasets used in our experiments are relatively small. For many large datasets having inconsistent labels collected from crowdsourcing platforms such as AMT, they provide no input features but raw data examples, such as plain texts or images. Extracting meaningful features (input variables) from those datasets needs significant efforts, which goes beyond our goal of study in this work. Moreover, many of these datasets may provide no ground truth that can be used in model evaluation. Our future work will also include searching for larger datasets that are suitable for objectively and systematicly testing the proposed models.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their insightful comments. This work was supported by the US National Science Foundation grants IIS-1320586, DBI-1356655, and CCF-1514357, and an National Institutes of Health grant R01DA037349. Jinbo Bi was also supported by US National Science Foundation grants IIS-1407205 and IIS-1447711. Jinbo Bi is the corresponding author.

REFERENCES

[1] S. L. Hui and X. H. Zhou, "Evaluation of diagnostic tests without gold standards," *Statistical methods in medical research*, vol. 7, no. 4, pp. 354–70, 1998.

[2] M. Yetisgen-Yildiz, I. Solti, F. Xia, and S. R. Halgrim, "Preliminary experience with amazon's mechanical turk for annotating medical named entities," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 180–183.

[3] J. D. Burger, E. Doughty, S. Bayer, D. Tresner-Kirsch, B. Wellner, J. Aberdeen, K. Lee, M. G. Kann, and L. Hirschman, "Validating candidate gene-mutation relations in medline abstracts via crowdsourcing," in *Data Integration in the Life Sciences*. Springer, 2012, pp. 83–91.

[4] B. M. Good and A. I. Su, "Crowdsourcing for bioinformatics," *Bioinformatics*, vol. 29, no. 16, pp. 1925–1933, 2013.

[5] T. B. Nguyen, S. Wang, V. Anugu, N. Rose, M. McKenna, N. Petrick, J. E. Burns, and R. M. Summers, "Distributed human intelligence for colonic polyp classification in computer-aided detection for ct colonography," *Radiology*, 2012.

[6] M. Qazi, G. Fung, S. Krishnan, J. Bi, B. Rao, and A. Katz, "Automated heart abnormality detection using sparse linear classifiers," *IEEE Engineering in Medicine and Biology*, vol. 26, no. 2, pp. 56–63, 2007.

[7] R. Jin and Z. Ghahramani, "Learning with multiple labels," in *Advances in Neural Information Processing Systems*, S. T. S. Becker and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 897–904.

[8] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, and J. Dy, "Modeling annotator expertise: learning when everybody knows a bit of something," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 932–939.

[9] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: who to trust when everyone lies a bit," *Proceedings of the 26th International Conference on Machine Learning*, pp. 96–103, 2009.

[10] V. C. Raykar, S. Yu, G. H. Valadez, C. Florin, L. Bogoni, L. H. Zhao, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[11] H. Kajino and H. Kashima, "A convex formulation for learning from crowds," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012, pp. 73–79.

[12] H. Kajino, Y. Tsuboi, and H. Kashima, "Clustering crowds," in *Proceedings of the AAAI conference on Artificial Intelligence*, 2013, pp. 1120–1127.

[13] N. Pochet and J. Suykens, "Support vector machines versus logistic regression: improving prospective performance in clinical decision-making," *Ultrasound in Obstetrics & Gynecology*, vol. 27, no. 6, pp. 607–608, 2006.

[14] D. A. Salazar, J. I. Vélez, and J. C. Salazar, "Comparison between svm and logistic regression: Which one is better to discriminate?" *Revista Colombiana de Estadística*, vol. 35, no. 2, pp. 223–237, 2012.

[15] T. Verplancke, S. Van Looy, D. Benoit, S. Vansteelandt, P. Depuydt, F. De Turck, and J. Decruyenaere, "Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, p. 56, 2008.

[16] A. P. Dawid and A. M. Skeene, "Maximum likelihood estimation of observed error-rates using the EM algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 20–28, 1979.

[17] P. S. Albert and L. E. Dodd, "A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard," *Biometrics*, vol. 60, no. 2, pp. 427–435, 2004.

[18] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2008, pp. 254–263.

[19] D. Zhou, J. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, 2012, pp. 2204–2212.

[20] D. Zhou, Q. Liu, J. C. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 262–270.

[21] V. C. Raykar and S. Yu, "Ranking annotators for crowdsourced labeling tasks," in *Advances in Neural Information Processing Systems 20*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Cambridge, MA: MIT Press, 2011, pp. 1809–1817.

[22] C. Liu and Y.-M. Wang, "Truelabel+confusion: A spectrum of probabilistic models in analyzing multiple ratings," in *Proceedings of the International Conference on Machine Learning*, 2012, pp. 225–232.

[23] S. B. P. Welinder, S. Branson and P. Perona, "The multidimensional wisdom of crowds," in *Proceedings of the 2010 Neural Information Processing Systems (NIPS) Conference*, 2010, pp. 2424–2432.

[24] Y. Tian and J. Zhu, "Learning from crowds in the presence of schools of thought," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 226–234.

[25] T. W. J. B. J. Whitehill, P. Ruvolo and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proceedings of the 2009 Neural Information Processing Systems (NIPS) Conference.*, 2009, pp. 2035–2043.

[26] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in Neural Information Processing Systems*, 2011, pp. 1953–1961.

[27] C.-J. Ho, S. Jabbari, and J. W. Vaughan, "Adaptive task assignment for crowdsourced classification," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 534–542.

[28] A. Singla, I. Bogunovic, G. Bartók, A. Karbasi, and A. Krause, "Near-optimally teaching the crowd to classify," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 154–162.

[29] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labeling of Venus images," 1995, pp. 1085–1092.

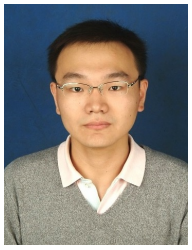
[30] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 2008, pp. 614–622.

[31] O. Dekel, O. Shamir, and I. th Annual International Conference on Machine Learning, "Good learners for evil teachers," in *Proceedings of the International Conference on Machine Learning*, 2009, pp. 233–240.

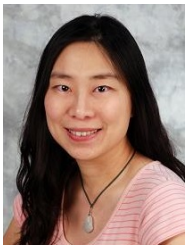
[32] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 129–136.

[33] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *Journal of Machine Learning Research*, vol. 9, no. 2, pp. 1757–1774, 2009.

[34] Y. Yan, R. Rosales, G. Fung, and J. Dy, "Modeling multiple annotator expertise in the semi-supervised learning scenario," *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pp. 241–248, 2010.
[35] M. Fang, J. Yin, and D. Tao, "Active learning for crowdsourcing using knowledge transfer," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 1809–1815.
[36] J. Bi and X. Wang, "Learning classifiers from dual annotation ambiguity via a minmax framework," *Neurocomputing*, vol. 151, Part 2, pp. 891 – 904, 2015.
[37] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Advances in Soft Computing, Lecture Notes in Computer Sciences*, vol. 2275, 2002, pp. 288–300.
[38] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, vol. 66, no. 3, pp. 373–407, 2007.
[39] B. Mozafari, P. Sarkar, M. J. Franklin, M. I. Jordan, and S. Madden, "Active learning for crowd-sourced databases," *Computing Research Repository*, vol. abs/1209.3686, 2012. [Online]. Available: http://arxiv.org/abs/1209.3686
[40] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of the 12th International Conference on Pattern Recognition (ICPR), 1994*, vol. 1, 1994, pp. 582–585.



Xin Wang Xin Wang, M.Sc received a master degree in Control Theory and Control Engineering from Dalian University of Technology, China. He is currently studying at the University of Connecticut toward his Ph.D. degree in Computer Science. His research interests include machine learning, data mining and intelligent systems.



Jinbo Bi Jinbo Bi, Ph.D. received a Ph.D. degree in mathematics from Rensselaer Polytechnic Institute, USA, and a master degree in Electrical Engineering and Automatic Control from Beijing Institute of Technology, China. She is an associate professor of Computer Science and Engineering at the University of Connecticut. Prior to her current appointment, she worked with Siemens Medical Solutions on computer aided diagnosis research and Partners Healthcare on clinical decision support systems, respectively. Her research interests include machine learning, data mining, bioinformatics and biomedical informatics.