Identifying and Accounting for Task-Dependent Bias in Crowdsourcing

Ece Kamar, Ashish Kapoor, Eric Horvitz
Microsoft Research
Redmond, WA USA
{eckamar, akapoor, horvitz}@microsoft.com

Abstract

Models for aggregating contributions by crowd workers have been shown to be challenged by the rise of task-specific biases or errors. Task-dependent errors in assessment may shift the majority opinion of even large numbers of workers to an incorrect answer. We introduce and evaluate probabilistic models that can detect and correct task-dependent bias automatically. First, we show how to build and use probabilistic graphical models for jointly modeling task features, workers' biases, worker contributions and ground truth answers of tasks so that task-dependent bias can be corrected. Second, we show how the approach can perform a type of transfer learning among workers to address the issue of annotation sparsity. We evaluate the models with varying complexity on a large data set collected from a citizen science project and show that the models are effective at correcting the task-dependent worker bias. Finally, we investigate the use of active learning to guide the acquisition of expert assessments to enable automatic detection and correction of worker bias.

Introduction

Crowdsourcing platforms such as Amazon Mechanical Turk provide access to a large pool of people for solving human intelligence tasks. Examples of tasks successfully applied to crowdsourcing include image labeling, product categorization, speech transcription, and handwriting recognition (Ipeirotis 2010). Studies of crowdsourcing marketplaces have shown that contributions from individual workers are noisy (Ipeirotis, Provost, and Wang 2010). Accuracy has been sought via statistical methods that leverage the *wisdom of the crowd*. For objective tasks (e.g., labeling), task owners have relied upon an assumption that aggregation of large numbers of contributions from the crowd will cancel out noise coming from individuals (Bachrach et al. 2012; Kamar, Hacker, and Horvitz 2012). The Condorcet Jury Theorem provides formal support for this assumption for binary decision-making problems (Condorcet 1785). Consequently, a well-studied problem in the field of human computation has been computational models for aggregating the contributions (annotations for labeling tasks) of individual workers to identify the correct answers (labels) of tasks. Under the assumption of the wisdom of the crowd, researchers have focused on modeling individuals workers' quality, bias, and expertise, and the relationship of their contributions to ground truth answers and task difficulty (Dawid and Skene 1979; Raykar et al. 2010; Whitehill et al. 2009).

Although the wisdom of the crowd assumption may hold in many cases, it may be invalid for specific tasks. Characteristics of tasks may induce a population wide or subgroup specific bias in worker contributions such that the majority opinion of a large number of individuals may be incorrect. Such systematic errors can be viewed as being analogous to erroneous, universally perceived optical illusions that are induced by certain visual patterns. Task-dependent biases can be based upon visual or other properties of tasks. Worker bias, usually represented with a confusion matrix, captures the likelihoods of a worker making specific types of mistake for tasks. With task-specific bias, workers may be more likely to make mistakes, such as confusing a specific class with another, when a task has certain characteristics.

The effect of task characteristics on worker bias has been studied on the Galaxy Zoo 2 project, a popular citizen science application for categorizing imagery of galaxies and other celestial objects, as captured by the Sloan Digital Sky Survey (SDSS). Researchers have identified image features associated with significant biases in worker annotations (Willett et al. 2013). Our analysis on Galaxy Zoo 2 data has shown that for 23% of the tasks, the majority opinion of 20 individuals disagrees with the ground truth answer. Existing aggregation models, assuming the correctness of the majority opinion of a large group and overlooking more sophisticated models of performance that consider task characteristics, cannot correct this bias. As a result, experts worked to manually craft debiasing procedures aimed at correcting the discovered bias (Willett et al. 2013). When data collection has a smaller scale than Galaxy Zoo 2 or resources for expert analysis is scarce, such task-dependent worker biases can go unrecognized. We describe methods that can automatically recognize and correct for task-specific biases. We believe that developing such methods will help to advance the practice of human computation on consensus tasks.

We investigate a family of probabilistic graphical models that can represent and infer multiple types of variables within the same model, handle missing data and capture uncertainty over all variables of the model. We show how the approach can detect and correct task-dependent worker bias so that accurate task labels can be accurately inferred from worker annotations. The methods can be used to infer the relationships among task features, workers' biases, annotations and ground truth task answers (labels). Given a set of ground-truth labels provided by experts, the models can detect and learn about task-dependent biases by observing when majority opinion disagrees with ground truth. Inferences about the relationship among specific task features and bias can be generalized to provide insights about tasks with unknown labels. We furthermore show how the methods can be used to selectively acquire a set of ground-truth labels from experts that work to debias in an efficient manner.

We develop models of varying complexity for identifying and correcting task-dependent worker bias. First, we build on the Bayesian classifier combination model (BCC) (Kim and Ghahramani 2012), to develop the Worker Bias Model (WBM), which models each workers' bias separately with a confusion matrix. Next, we move to leverage information from a population of users. Experiments on crowdsourcing marketplaces have identified the issue of annotation sparsity, where a majority of workers provide a small number of contributions. To overcome sparsity in learning bias models, we propose population wide and hybrid models for transfer learning among workers. The Population Bias Model (PBM) models bias with a confusion matrix shared by all workers. Finally, we present the Hybrid Bias Model (HBM), which learns a personalized mixture of population and individualistic confusion matrices to represent each worker's bias. A key contribution of this work is the introduction of task-dependent extensions of PBM and HBM for representing the relationship between task features and bias. This relationship is represented with the addition of logistic regression models, which are used to learn a weighted combination of task features to predict each row of the confusion matrix.

We evaluate our models on a large-scale data set collected from volunteer citizen scientists who annotated celestial objects on the Galaxy Zoo 2 system. The data set contains 91 thousand tasks (celestial objects) with an average of 27 annotations (votes), where tasks are described by 52 task features that were automatically extracted via machine vision analysis. Experimental results demonstrate that the task-dependent models infer ground truth labels of tasks more accurately than models without task features when some ground truth labels are provided as observations. They also show that the predictive performance can be further improved by learning a mixture of population wide and personalized components. Finally, we experiment with active learning to guide the potentially costly acquisition of ground-truth labels so as to learn task-dependent biases with minimal expert labeling. The results show that such active learning methods can successfully guide the contributions from experts so as to ideally learn about task-dependent bias.

Related Work

Aggregating worker annotations to infer ground truth task labels has been an active area of research. Several methods have been proposed for representing the relationship among task labels, worker annotations, and worker ability (Raykar et al. 2010; Yan et al. 2011; Welinder et al. 2010; Kamar, Hacker, and Horvitz 2012). These models represent worker ability in a single dimension that distinguishes workers by performance. Such a one-dimensional representation is expanded in other works to explicitly model worker bias in the form of confusion matrices (Dawid and Skene 1979; Kim and Ghahramani 2012; Simpson et al. 2013; Zhou et al. 2012; Liu and Wang 2012). Ipeirotis *et al.* propose an algorithm for estimating the inherent quality of a worker by separating systematic worker bias from errors (Ipeirotis, Provost, and Wang 2010). These models can represent the types of mistakes made by individual workers but do not represent the relationship between task characteristics and worker mistakes. One approach for linking worker ability with task characteristics is adding task difficulty as a latent variable in crowdsourcing models (Whitehill et al. 2009; Bachrach et al. 2012; Dai et al. 2013; Simpson et al. 2013). Without explicitly modeling task characteristics, modeling the relationship between worker bias and task difficulty is not sufficient to learn about and task-dependent worker biases; task difficulty helps to infer whether a worker is likely to make a mistake but does not indicate the *type* of mistake.

In the work of Kamar *et al.*, task features are used in discriminative training to predict ground truth answers and worker responses when large training sets are available (Kamar, Hacker, and Horvitz 2012). The resulting models do not explicitly represent worker bias nor the relationships among task features, worker annotations, ground truth answers and worker biases. Wauthier and Jordan present a Bayesian latent feature model in which shared factors influence worker bias in crowdsourcing (Wauthier and Jordan 2011). Their investigations study how shared strategies may affect worker bias but they overlook the influence of task features on the labeling process. Welinder *et al.* propose a multi-dimensional model of worker competence, expertise and bias (Welinder et al. 2010). In this generative model, abstract task features and worker characteristics may jointly influence worker annotations but the bias representation is not conditioned on features that describe the task. Instead, the investigations are targeted at learning worker bias from crowd consensus. They do not study systematic errors rising population-wide task-specific biases.

Ipeirotis has shown that the number of annotations per worker in crowdsourcing tasks is distributed with a power law, meaning that most workers provide only a small number of annotations (Ipeirotis 2010). To address the sparsity, Liu and Wang propose models that learn worker specific bias and population wide bias, and a hybrid of the two (Liu and Wang 2012). We follow a similar investigation in this work. Alternatively, researchers have proposed models for representing worker communities (Venanzi et al. 2014; Simpson et al. 2013).

Researchers have studied active learning for acquiring worker annotations and ground truth labels in order to learn about workers and their biases (Yan et al. 2011; Bachrach et al. 2012; Wauthier and Jordan 2011). We shall demonstrate the use of active learning to collect ground truth task labels aimed at ideal learning about task-dependent biases.

Bias Models

In this section, we discuss a family of probabilistic graphical models for representing worker bias. The models are generative—they describe the process of observed worker annotations being created based on latent variables (e.g, worker confusion matrices) and observed variables (e.g., task features). The relationships between variables in the probabilistic graphical models encode our independence assumptions about model variables (Pearl 1989). We use *factor graphs* to describe the probabilistic graphical models, which represent the joint probability distribution of variables as the product of factors (Koller and Friedman 2009). We start with models for representing the relationships among worker bias, worker annotations and true task labels. Then, we expand these models to reason about task characteristics.

Problem Formulation

We focus on a class of crowdsourcing tasks called *consensus tasks* (Kamar, Hacker, and Horvitz 2012). The goal of a consensus task is to identify a correct answer which is unknown to the task owner, via the aggregation of predictions provided by workers. Consensus tasks are common in crowdsourcing and provide workers with labeling challenges. Examples include *games with a purpose* (e.g., image labeling in the ESP game) (von Ahn and Dabbish 2008), paid crowdsourcing systems (e.g., product categorization in Mechanical Turk) (Ipeirotis 2010), and citizen science projects (e.g., efforts to classify birds, plants, and celestial objects) (Lintott et al. 2008).

Formally, let T be the number of consensus tasks that W workers are providing annotations for. K is the set of possible answers for the task and C is the number of elements in K. For each task i, $t_i \in K$ is the true class (label) of the task. $a_i^w \in K$ is the annotation from worker w for task i. Each task has a set of F-dimensional task features x_i, which are automatically generated to represent task characteristics.

The inference problem is predicting t, the true labels of consensus tasks, based on A, the set of collected annotations, and X, the set of task features. To facilitate the learning of worker biases, a subset of task labels T can be acquired from experts as observations. For simplicity of formalization, our notation assumes that all workers provide annotations for all tasks resulting in a dense worker-task annotation matrix. However, this assumption is not realistic in crowdsourcing since most workers provide a small number of annotations. Neither our implementation nor our experiments make this assumption and they can work with more realistic sparsely annotated matrices.

Task-Independent Bias Models

Prior Work: Worker Bias Model The core of our approach is the Bayesian Classifier Combination (BCC) model that was proposed in (Kim and Ghahramani 2012). We first describe this model and then describe several key extensions that provide more expressive bias models that incorporate task-dependent features.

The BCC model has been used in the crowdsourcing domain for modeling the relationship between individual workers' biases, the annotations collected from them, and the true labels of tasks (Simpson et al. 2013). The Worker Bias Model (WBM) given in Figure 1(a) is a modification of the BCC model so that we can easily integrate task features as will be described in the next section [1].

The factor graph for the Worker Bias Model is presented in Figure 1(a) and it defines key relationships among the workers, tasks, and class labels. The figure has three plates for representing classes, workers, and tasks, representing the number instances of each variable in the model. For example, the variable a_i^w is at the intersection of worker and task plates and there exists $W \times T$ instances of that variable in the model. Here, p is a random variable denoting class frequencies in task labels. The model represents worker w's bias with a C by C confusion matrix π^w, such that each row π_c^w is a probability vector. π^w is generated from σ^w, a C by C probability score matrix. The Worker Bias Model assumes that each worker's bias is independent of other workers' biases, and each task label is independently generated from the distribution p.

Given the annotations A, the factor model defines a posterior distribution on the true labels t, the class frequencies p, the confusion matrices π and the probability score matrices σ for all W workers.

$$Pr(p, t, \pi, \sigma | A) = \frac{1}{Z_{\text{w}}} \prod_{i=1}^{T} \psi_t(p, t_i) \times$$
$$\prod_{w=1}^{W} \psi_\pi(\sigma_{t_i}^w, \pi_{t_i}^w)\, \psi_a(\pi_{t_i}^w, a_i^w)$$

here Z_{w} is the partition function (normalization term) for the worker bias model, and the potentials ψ_t, ψ_a, ψ_π take the following form:

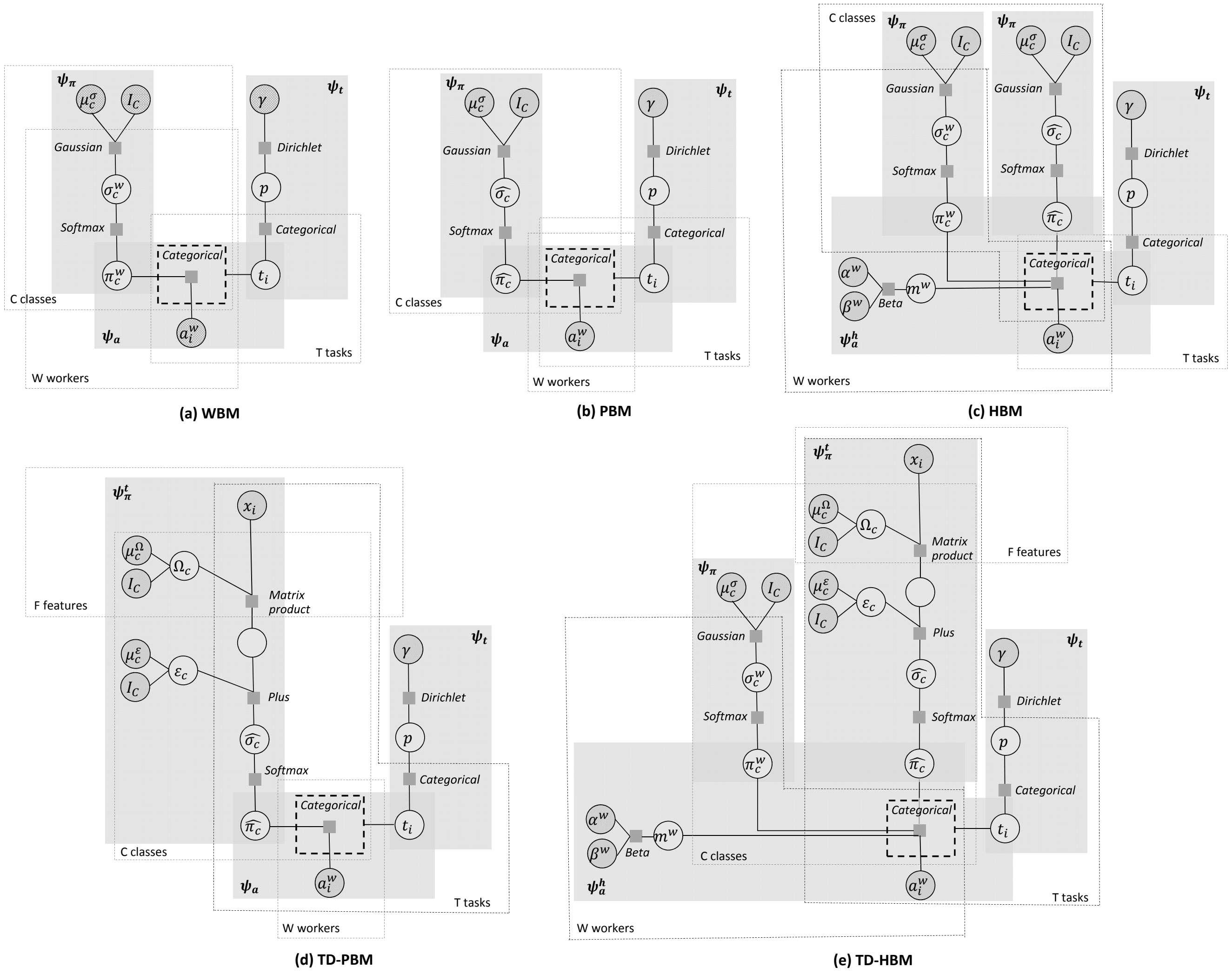
$$\begin{aligned}
\psi_t(p, t_i) &= Dir(p|\gamma)\, Cat(t_i|p) \\
\psi_\pi(\sigma_c^w, \pi_c^w) &= \mathcal{N}(\sigma_c^w | \mu_c^\sigma, I_C)\, \delta(\pi_c^w - softmax(\sigma_c^w)) \\
\psi_a(\pi_c^w, a_i^w) &= Cat(a_i^w | \pi_c^w)
\end{aligned}$$

Each potential is marked on Figure 1(a). Intuitively, potential ψ_t describes the generation of task labels. The true label t_i for task i is generated from a categorical distribution from parameter p, where p denotes frequencies of classes for all tasks and is generated from conjugate Dirichlet prior γ. γ is initialized with uniform values, giving equal prior likelihood to all classes.

The potential ψ_a describes the relationship between task label t_i, worker w's annotation for task i a_i^w and the worker's confusion matrix π^w. The random variable a_i^w is generated with a categorical distribution from $\pi_{t_i}^w$, where t_i identifies which row of the confusion matrix generates a_i^w. This relationship of t_i acting as an indicator of the categorical distribution between π^w and a_i^w is represented with a *gate* (dashed box) as introduced by (Minka and Winn 2008).

The potential ψ_π describes the process of generating worker confusion matrices from Gaussian priors using the

[1] In the BCC model, π_c^w is generated with a Dirichlet distribution. In the WBM model, it is generated by applying the softmax operator to σ_c^w, which is a multivariate Gaussian variable.



(a) WBM

(b) PBM

(c) HBM

(d) TD-PBM

(e) TD-HBM

Figure 1: Models for representing and inferring bias. In each model, shaded regions mark potentials defined for inference. Dashed lines represent plates for classes, workers, and tasks. Gates are represented with darker dashed squares and observed variables are represented with shaded circles.

softmax operator as proposed by Venanzi *et al.* (Venanzi et al. 2014). Each row of the probability score matrix σ_c^w is a multivariate Gaussian variable generated from Gaussian priors μ_c^σ and I_C. μ_c^σ is initialized to the c^{th} row of the identity matrix so that the prior of each confusion matrix assigns a higher likelihood to a worker annotation to be correct than incorrect. I_C is the identity matrix of dimension C. c^{th} row of the worker confusion matrix is generated by applying the softmax operator to σ_c^w to derive a proper probability vector. δ operator in ψ_π is the Dirac delta function ensuring that the posterior distribution is positive only when π_c^w is equal to $softmax(\sigma_c^w)$.

Extension 1: Population Bias Models The Worker Bias Model assumes that workers' biases are independent given common priors. Such an assumption of independence may often be incorrect: mistakes made by workers may depend in a systematic manner on human perception and the task itself. On the other hand, learning separate confusion matrices per worker that are accurate requires having sufficient annotation data per worker. Such data sufficiency does not hold in many crowdsourcing scenarios, including Galaxy Zoo, where a majority of workers perform only a few tasks (Kamar, Hacker, and Horvitz 2012) and the worker-task annotation matrix is sparse. Not having sufficient annotation data per worker can degrade the performance of the Worker Bias Model.

The Population Bias Model (PBM) makes the opposite assumption that all workers share the same confusion matrix, meaning that the ways that all workers make mistakes are identical. Instead of learning π^w, a confusion matrix per worker, it learns $\widehat{\pi}$, a single confusion matrix for the population. The model uses the annotation data collected from all workers to learn this single matrix. This assumption addresses the annotation sparsity problem but the population-wide model cannot learn about individual workers.

The factor model for the Population Bias Model is given in Figure 1(b). The model learns population score matrix $\widehat{\sigma}$ and confusion matrix $\widehat{\pi}$. The posterior distribution for variables $\widehat{\sigma}, \widehat{\pi}, t, \psi$ are inferred as follows using the potentials already defined, where Z_p is the partition function of PBM:

$$Pr(p, t, \widehat{\pi}, \widehat{\sigma}|A) = \frac{1}{Z_p} \prod_{i=1}^{T} \psi_t(p, t_i) \times \\ \psi_\pi(\widehat{\sigma_{t_i}}, \widehat{\pi_{t_i}}) \prod_{w=1}^{W} \psi_a(\widehat{\pi_{t_i}}, a_i^w)$$

Extension 2: Hybrid Bias Model Both PBM and WBM have limitations in that the former cannot learn about individual worker characteristics and the latter cannot generalize the data of a worker to other workers. The Hybrid Bias Model (HBM) bridges the gap between these models by modeling both worker specific and population-wide confusion matrices, and learning a mixture weight per worker describing which model to trust for that worker. By doing so, the model can learn to depend heavily on the population-wide data for workers with small number of annotations and it can learn an accurate worker confusion matrix for each worker with a large number of annotations.

HBM is presented in Figure 1(c). $\widehat{\pi}$ is the population confusion matrix and π^w is the confusion matrix for worker w. m^w is the mixture weight of worker w. The posterior distribution for variables $p, t, \widehat{\sigma}, \widehat{\pi}, \sigma, \pi, m$ are inferred as follows:

$$Pr(p, t, \widehat{\sigma}, \widehat{\pi}, \sigma, \pi, m|A) = \frac{1}{Z_h} \prod_{i=1}^{T} \psi_t(p, t_i) \times \\ \psi_\pi(\widehat{\sigma_{t_i}}, \widehat{\pi_{t_i}}) \prod_{w=1}^{W} \psi_\pi(\sigma_{t_i}^w, \pi_{t_i}^w)\, \psi_a^h(\widehat{\pi_{t_i}}, \pi_{t_i}^w, a_i^w, m^w)$$

where Z_h is the partition function (normalization term) for HBM. Potentials ψ_t, ψ_π were described previously and ψ_a^h take the following form:

$$\psi_a^h(\widehat{\pi}_c, \pi_c^w, a_i^w, m^w) = Beta(m^w|\alpha^w, \beta^w) \times \\ (m^w\, Cat(a_i^w|\widehat{\pi}_c) + (1 - m^w)\, Cat(a_i^w|\pi_c^w))$$

ψ_a^h is expanded from ψ_a to reason about the mixture of population and worker specific confusion matrices. m^w, mixture weight of worker w, determines the influence of $\widehat{\pi}$ and π^w in modeling worker w's bias; $\widehat{\pi}$ is weighted with m^w and π^w is weighted with $(1 - m^w)$. m^w is generated from a Beta distribution with parameters α^w and β^w. α^w is constant between workers. The value of β^w is set to be proportional to the number of annotations provided by worker w so that the prior of m^w is adjusted with respect the the amount of data available for worker w. For workers with a small number of annotations, the prior is skewed to have a higher value, which assigns a larger weight to $\widehat{\pi}$. For workers with a large number of annotations, the prior is skewed to have a lower value, which results in a higher dependency to π^w.

Task-Dependent Bias Models

The bias models presented in the previous section assume that workers' biases are conditionally independent of the task characteristics given the true labels of tasks. This assumption does not hold when task characteristics affect human perception and decision making (e.g., Galaxy Zoo 2 tasks). In this section, we expand the bias models presented so far to learn task-specific worker biases.

Extension 3: Task-Dependent Population Bias Model Figure 1(d) presents the factor model for the Task-Dependent Population Bias Model (TD-PBM). This model expands PBM by incorporating a generative process that describes the way task features influence worker mistakes. We model this process in the form of multinomial logistic regression (Bishop 2006), where the input is a vector of continuous task features x_i for task i and the output is $\widehat{\pi_c^i}$, the c^{th} row of the confusion matrix for task i. Using the logistic regression component, the model infers a confusion matrix for each task based on the task's features. $\widehat{\pi_c^i}$ describes the likelihood of a worker providing annotations when the true label of the task is c and it has features x_i. F is the dimension of the feature vector x_i. Ω_c is a F by C matrix of regression coefficients of the logistic regression model and ϵ_c is the mean vector to be added to the multiplication of x_i and Ω_c. The posterior distribution for variables $p, t, \widehat{\sigma}, \widehat{\pi}, \Omega, \epsilon$ are inferred as follows:

$$Pr(p, t, \widehat{\sigma}, \widehat{\pi}, \Omega, \epsilon|A, X) = \frac{1}{Z_t} \prod_{i=1}^{T} \psi_t(p, t_i) \times \\ \psi_\pi^t(\widehat{\sigma_{t_i}^i}, \widehat{\pi_{t_i}^i}, \Omega_{t_i}, \epsilon_{t_i}) \prod_{w=1}^{W} \psi_a(\widehat{\pi_{t_i}^i}, a_i^w)$$

where Z_t is the partition function (normalization term) for T-PBM. The new term, ψ_π^t take the following form:

$$\psi_\pi^t(\widehat{\sigma_c^i}, \widehat{\pi_c^i}, \Omega_c, \epsilon_c) = \mathcal{N}(\epsilon_c|\mu_c^\epsilon, I_C) \prod_{j=1}^{F} \mathcal{N}(\Omega_c^j|\mu_c^\Omega, I_C) \times \\ \delta(\widehat{\sigma_c^i} - (x_i\Omega_c + \epsilon_c))\, \delta(\widehat{\pi_c^i} - softmax(\widehat{\sigma_c^i}))$$

ψ_π^t denote the generation of the c^{th} row of the task-dependent confusion matrix $\widehat{\pi_c^i}$ with respect to the following equation of multinomial logistic regression:

$$\widehat{\pi_c^i} = softmax(x_i\Omega_c + \epsilon_c)$$

The logistic regression model links the task features with the matrix of regression coefficients (Ω_c) and the mean vector (ϵ_c) to predict $\widehat{\pi_c^i}$. Mean vector ϵ_c is a multivariate Gaussian generated with parameters μ_c^ϵ and identity matrix I_C. μ_c^ϵ is initialized to the c^{th} row of the identity matrix so that the prior of $\widehat{\pi_c^i}$ assigns a higher likelihood to producing a correct annotation than an incorrect annotation. Each row of the coefficients matrix Ω_c^j is a multivariate Gaussian generated with parameters μ_c^Ω and identity matrix I_C. μ_c^Ω is initialized to a vector of zeros. Finally, to make the model identifiable, we constrain the last element of Ω_c^j and μ_c^ϵ to be zero.

Extension 4: Task-Dependent Hybrid Bias Model The shortcoming of TD-PBM is that it assumes that all workers are identical and it does not model worker characteristics. One way to address this shortcoming is to expand the WBM to reason about task features similar to the construction of TD-PBM. This expansion results in an explosion in the number of variables in the model since the number of confusion matrices to be inferred increases with $T \times W$, where T is the number of tasks, and W is the number of workers. The sparsity of the worker-task annotation matrix makes learning task-dependent worker bias models infeasible.

The Task-Dependent Hybrid Bias Model (TD-HBM) uses the idea of learning a weighting of population and worker confusion matrices, as the Hybrid Bias Model does, to jointly model the relationships among task features, worker characteristics, and bias. The factor model for TD-HBM is given in Figure 1(e). The model decouples learning about task features from learning about worker characteristics. It has different generative processes for modeling population bias and for modeling worker bias. The population bias is task dependent. $\widehat{\pi^i}$, population confusion matrix for task i, is generated with the logistic regression component as described for TD-PBM based on task features x_i. On the other hand, π^w, confusion matrix for worker w, is task independent. With this construction, we can use the annotation set collected from all workers to learn about the task-specific biases that affect the population (e.g., the systematic influence of task features on human perception). We can also use the set of annotations collected from each worker to personalize the bias model independently of task features. The number of confusion matrices to be inferred by this model grows linearly with T and W.

The posterior distribution for variables p, t, $\widehat{\sigma}$, $\widehat{\pi}$, σ, π, m, Ω, ϵ are inferred as follows:

$$Pr(p, t,\widehat{\sigma}, \widehat{\pi}, \sigma, \pi, m, \Omega, \epsilon | A, X) = \frac{1}{Z_{th}} \times$$
$$\prod_{i=1}^{T} \psi_t(p, t_i)\ \psi_\pi^t(\widehat{\sigma_{t_i}^i}, \widehat{\pi_{t_i}^i}, \Omega_{t_i}, \epsilon_{t_i}) \times$$
$$\prod_{w=1}^{W} \psi_\pi(\sigma_{t_i}^w, \pi_{t_i}^w)\ \psi_a^h(\widehat{\pi_{t_i}}, \pi_{t_i}^w, a_i^w, m^w)$$

where Z_{th} is the partition function (normalization term) for TD-HBM. The factor model does not have any new components. The potentials ψ_t, ψ_π^t, ψ_π and ψ_a^h were introduced in the construction of WBM, TD-PBM and HBM.

Empirical Analysis

We evaluate the bias models on a dataset collected from the Galaxy Zoo 2 project. This project provides an ideal, large-scale data set to evaluate the models we have presented since expert annotations have verified that visual features of tasks bias worker annotations. Our experiments focus on automatic detection and correction of this bias with probabilistic modeling. We provide evaluations for both batch and incremental active learning settings.

Empirical Setup

Galaxy Zoo 2 is designed to collect classifications (annotations) for millions of galaxies from the Sloan Digital Sky Survey (SDSS) (Galaxy Zoo 2 Team 2009). Its interface presents an image of a galaxy and asks workers (a.k.a. citizen scientists) to provide classifications with increasing sophistication. The project has collected more than 16 million annotations for 304 thousand galaxies. A catalog containing aggregate statistics about worker labels and ground truth labels has been made available for a subset of galaxies studied in Galaxy Zoo 2 (Willett et al. 2013). We study the galaxies in the main sample with spectroscopic redshifts (Galaxy Zoo 2 data release 2013). In addition, we obtained two separate data sets from the Galaxy Zoo 2 team: a log file containing detailed information on worker annotations, and the SDSS feature file reporting 52 features describing the features of each task obtained via machine vision analysis of each celestial object. The final data is created by merging these data sets by filtering out galaxies with missing ground truth labels, SDSS features or that are missing in the annotation log files. In our studies, we focus on the top-level classification question of the Galaxy Zoo 2 task hierarchy, which asks workers to classify an image as one of the three classes (smooth, disk and artifact). Our data set includes 2.5 million annotations collected for 91 thousand galaxies from 46 thousand unique workers. The average, minimum and maximum number of annotations per task in the data set is 27, 8 and 76 respectively. The average number of annotations per worker is 51. The minimum and the maximum number of annotations by a worker are 1 and 13384. The distribution of the three classes (smooth, disk and artifact) in the resulting data set are 27.03%, 72.94% and 0.03%.

The availability of the SDSS features enables us to study the automatic detection of task-dependent worker bias for the Galaxy Zoo 2 tasks. Experts have compared worker annotations with expert annotations and identified 3 of the 52 SDSS features that are linked to significant systematic biases in worker annotations. The analysis was used to guide the manual authoring of a debiasing procedure for correcting worker bias to reach ground truth classifications of galaxies. In our studies, we evaluate the performance of various aggregation models in capturing and correcting this bias.

We evaluate the aggregation models on the complete data set created by merging SDSS, annotation and ground truth data sources. Most annotation sets collected for common crowdsourcing tasks are much smaller in size than the Galaxy Zoo 2 catalog. To evaluate the performance of our models on smaller scale data sets, we devised an algorithm to randomly sample a subset of the data. The algorithm takes two parameters, n–the number of tasks and k–the number of annotations per task (i.e., overlap). The algorithm randomly selects a worker and adds the complete annotation set of that worker until k many annotations are collected for n tasks. This sampling procedure reflects the long tail power-law property of the dataset in which the majority of workers provide a few annotations and a small number of workers provide large amounts of work.

For our experiments, we sample data sets of sizes 100, 1000 and 10000 in which each task has a minimum of 20

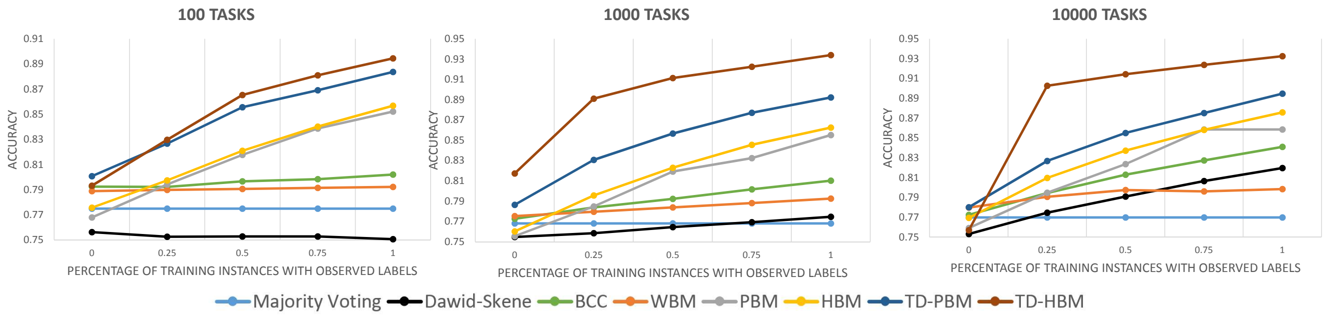


Figure 2: Accuracy of bias models with increasing number of ground truth label observations. The plots from left to right are for experiments with 100, 1000, and 10000 tasks respectively.

annotations. We subsample the annotations of each task to experiment with smaller values of overlap. In each experiment, we divide the data set in half; ground truth labels are provided as observations for only the tasks in the training set, and accuracy results are reported for the tasks in the testing portion. Experiments are repeated for 100, 20 and 10 times for data sets of sizes 100, 1000 and 10000. On average, data sets with 100 tasks and 20 overlap have 1246 judges with 1.64 annotations per worker and the worker with the maximum number of annotations has 30 maximum judgments. Data sets with 1000 tasks and 20 overlap have 5790 judges with 3.48 annotations per worker and the worker with the maximum number of annotations has 143 maximum judgments. Data sets with 10000 tasks and 20 overlap have 16858 judges with 11.86 annotations per worker and the worker with the maximum number of annotations has 2126 maximum judgments.

We compare the performance of the models introduced in earlier sections with the majority voting baseline, the Bayesian Classifier Combination (BCC) model (Kim and Ghahramani 2012), and the Dawid-Skene algorithm. The Dawid-Skene algorithm uses Expectation Maximization to simultaneously infer worker bias, worker annotations and ground truth answers, and is shown to perform consistently well across different data sets in comparison to other aggregation algorithms studied in the literature (Sheshadri and Lease 2013). We experiment with a public implementation of the algorithm that is available online (Ipeirotis and others. 2014). We use Infer.Net (Minka et al. 2014) for coding and for performing inference on BCC, WBM, PBM, HBM, TD-PBM, TD-HBM models. Inference is done using variational message passing (Koller and Friedman 2009) for a fixed number of 100 iterations. We use the Wilcoxon sign test to evaluate the significance of differences between the performance of models. We report the model performance with predictive accuracy for interpretability of the results. All trends and significant results hold when performance is reported based on the F1 metric.

Batch Learning

Figure 2 compares the accuracy of bias models when 0%, 25%, 50%, 75% and 100% of the tasks in the training set have observed ground truth labels. The tasks have 20 overlap. TD-PBM and TD-HBM are given all 52 task features as input. The experiment is repeated for data sizes with 100, 1000 and 10000 tasks and plots summarizing the results are presented from left to right. The figure shows that the task-dependent models perform better than other models when some ground truth labels are provided as observations ($p < 0.05$). The performance of these models improves with increasing numbers of ground truth labels as these observations help the models to identify which features are linked to biases in worker annotations ($p < 0.01$). The results also show that TD-HBM, having both population and worker-specific components, helps to learn more accurate models; TD-HBM outperforms TD-PBM for data sizes 1000 and 10000 for varying percentages of training instances with ground truth labels ($p < 0.05$). The significance of the differences between the accuracy of TD-PBM and TD-HBM increase with large data sets as these data sets contain larger amounts of annotations per worker, which help in learning about individual worker's characteristics.

Table 1 reports the time required to perform inference on various models for data sizes of 100, 1000 and 10000 tasks when tasks have 20 overlap and when all training instances are given ground truth labels. For all models except Dawid-Skene, model generation and inference are performed on the Infer.Net platform. The running times are recorded on a machine with 2.50 GHz CPU and 64 GB RAM. In terms of modeling worker confusion matrices, modeling population-wide confusion matrices results in the fastest inference and the hybrid components are the most expensive. Reasoning about task-dependent bias also increase the complexity of inference.

Figure 3 shows the effect of annotation overlap on the performance of models when each data set has 1000 tasks and ground truth labels are observed for all tasks in the training set. The figure shows that increasing overlap from three to larger values has a positive effect on the models reasoning about task features ($p < 0.01$). It shows that larger overlap helps to learn more accurate task-dependent models as tasks with larger overlap have more information about how task features affect worker annotations. For models without task features, we observe an opposite trend. The performances

	100 tasks	1000 tasks	10000 tasks
Dawid-Skene	0.79	70.96	2449.35
BCC	1.30	7.75	71.38
WBM	4.79	24.61	163.10
PBM	0.69	2.85	24.88
HBM	6.95	38.10	216.81
TD-PBM	2.59	18.38	173.39
TD-HBM	8.42	60.02	376.96

Table 1: Running times (in seconds) of inference on various models for increasing data set sizes.

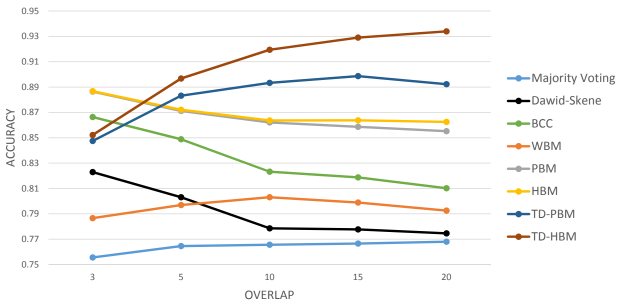


Figure 3: Accuracy of bias models with variation in annotation overlap.

of Dawid-Skene, BCC, PBM and HBM degrade when the overlap is increased from 3 to larger numbers ($p < 0.05$) since the generative processes expressed in these models are not expressive enough to represent the different facets of the data. Overall, the results highlight the value of machine learned aggregation models for the Galaxy Zoo tasks. For this data set, the value to be gained from more expressive machine learned models is greater than the improvement achieved from increasing the overlap from 3 to 20 when majority voting is the chosen method for aggregation. Understanding the effect of annotation overlap on model performance can be a guiding factor in deciding how to allocate worker resources in crowdsourcing tasks. Models that can achieve the performance of majority voting with fewer annotation overlap can solve more crowdsourcing tasks with the same worker resources.

Next we take a closer look at the way task-dependent bias models detect and correct worker bias for a run on 1000 randomly selected tasks. On this run, majority voting has 0.76 accuracy and TD-HBM has an accuracy of 0.94 when all training instances have observed ground truth labels. Among the tasks that majority opinion predicted the ground truth answer incorrectly, TD-HBM is able to correctly infer the answer 75% of the time. For only one of the tasks, TD-HBM makes a mistake when the majority opinion is correct. The task-dependent bias models are able to correct worker bias without knowing about the structure of the bias, by only observing worker annotations and ground truth labels. The way we choose to represent the relationship between task features and worker bias with a logistic regression component does not have the same form as the expert representation in the hand-crafted debiasing procedure. In fact, we seek to understand how well the automated approaches can detect and correct worker bias without knowing about the structure of this relationship.

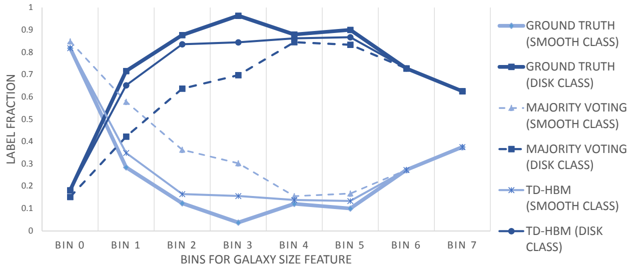


Figure 4: Label fractions for two most likely classes as a function of galaxy size. The figure displays ground truth fractions (thick line), predictions by majority voting (dashed line), and by TD-HBM (thin line).

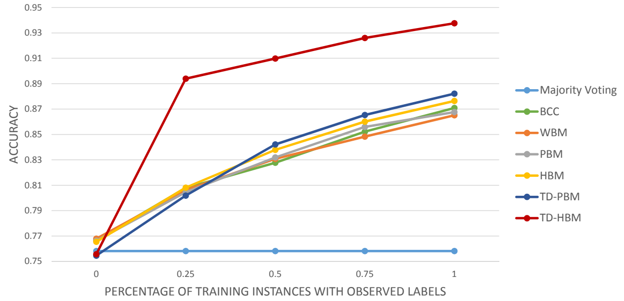


Figure 5: Accuracy of bias models when evaluated on the complete Galaxy Zoo 2 data set.

To analyze the extent that the task-dependent bias models can correct worker bias, we focus on the three task features that are identified by experts as influencing worker annotations and observe how the predictions of majority voting and TD-HBM change over the different values of these features. Figure 4 presents a visualization of the way one of the expert identified features (galaxy size) influences worker annotations and how this bias is corrected by TD-HBM (1000 tasks, 20 overlap, ratio of training instances with labels 1.0). In the figure, the gap between the thick and dashed lines quantifies how much majority opinion diverges from ground truth. As shown by the figure, majority voting underestimates the fraction of smooth galaxies between bins 1 and 5. The label distribution inferred by TD-HBM is closer to ground truth. We observe patterns similar to the ones reported in Figure 4 for the two other features.

Finally, Figure 5 reports the performance of the models when they are evaluated on the complete data set. Each task in the data set is randomly assigned to either to the training or the testing set. The implementation we use for the Dawid-Skene approach does not scale. The results highlight a large improvement from the TD-HBM model, combining the hybrid modeling of confusion matrices with the modeling of task features.

Overall, the results show that when task-dependent worker bias is present, the effort for generating task features and reasoning about them in aggregation leads to improved model performance. Results given in Figures 2 and 5 show that the percentage of training instances with observed labels have a similar effect on performance over data sets with different number of tasks. This observation reflects a characteristic of the Bayesian joint inference framework. Learning and inference takes into account both instances with observed and unobserved labels. The ratio of these instances determines how much weight to put into explaining observations on ground truth labels versus labels collected from workers. When the ratio of instances with ground truth labels is small, the weights of these instances can be adjusted to promote learning from gold set questions for better predictive performance.

Active Learning

The batch learning experiments show that task-dependent models are effective in correcting worker bias when ground truth labels can be provided by experts. We can apply active learning techniques to guide decisions about which expert labels to acquire so as to build accurate models while minimizing the quantity of potentially costly expert annotations.

We implement and study two active learning methods. The guidance provided by these methods depends on inferences performed with existing observations. Once a new label is collected, and revised posterior probabilities are used to generate recommendations on the next annotations to seek. The *entropy* method computes the entropy of label prediction, $H(t_i)$, for each task i and chooses the task with the highest entropy to be labeled by the expert. The *information gain* method maximizes the expected reduction of entropy over the label predictions of all tasks (IG) (MacKay 1992). IG for task i is computed as:

$$IG_i = \sum_{j=1}^{C} Pr(t_i = j) \sum_{k=1}^{T} (H(t_k) - H(t_k|t_i = j))$$

where $H(t_k|t_i = j)$ is the label entropy of k^{th} task when the inference is repeated after observing the true label of task i as j. This calculation is expensive; it requires repeating model inference for each possible value of each t_i.

We implement the *random* method as a baseline, which chooses a remaining task randomly at each round. We average the results of the random method 10 times for each data set.

We compare random, entropy and information gain methods when they are applied to task-dependent and task-independent models. Given the complexity of running the information gain method, we run the experiments on data sets with 100 tasks and 20 overlap and apply PBM and TD-PBM models, which have smaller running times but reach accuracies comparable to their hybrid counter parts for this data set size. Figure 6 shows that the task selection using the active learning methods improves model accuracy. Both the *entropy* and *information gain* methods are able to choose tasks

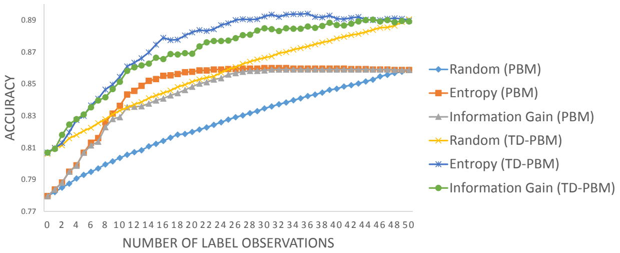


Figure 6: Accuracy of active learning methods.

that help the models improve more quickly with additional expert annotations than when tasks are chosen randomly for expert annotation. The results also highlight the advantage of task-dependent models on accuracy in the active learning setting. The results do not report a superior performance for the *information gain* method as compared to *entropy*. In our setting, the observations available from the start (annotations and task features) may contradict expert labels. As a result, the entropy of the model may not decrease monotonically. Thus the myopic nature of information gain calculation may fail to capture the true value of a label.

Conclusion and Future Work

We reviewed methods aimed at automating the identification and correction of task-dependent worker biases in crowdsourcing. We leverage probabilistic graphical models to perform joint learning and inference about the relationships among workers' biases, their annotations, ground truth task labels, and task features. Our evaluations showed that models with explicit representations of task features and their influences are effective in correcting worker bias when they are given some task labels as observations. The experiments also show that the models can successfully generalize or transfer observations among workers so as to facilitate learning under annotation sparsity.

The proposed models can be improved in several ways. We believe we can enhance active learning by moving beyond greedy, single-step computation of information gain. Speed-ups in model inference can enable the use of richer computation of the value of information of expert annotations that would perform richer lookahead. We are also interested in pursuing richer transfer learning among workers. We observed transfer learning via the use of a mixture of personalized and population-wide confusion matrices. Future extensions of transfer learning are possible with the integration of community-based models as proposed by Venanzi *et al.* (Venanzi et al. 2014). We also see possibilities for moving beyond models that capture the relationships among task features and bias to include other factors such as incentives and market dynamics. Finally, the probabilistic graphical models representing uncertainty over model variables can enable algorithms that can be used to route tasks among experts and workers based on a consideration of the likelihood of the presence different kinds of biases.

Acknowledgments

We thank the Galaxy Zoo team for data access, John Guiver for assistance with using Infer.Net and the anonymous reviewers for feedback. Finally, we thank the Zooniverse project and the citizen scientists who have contributed to Galaxy Zoo for making this study possible.

References

Bachrach, Y.; Graepel, T.; Minka, T.; and Guiver, J. 2012. How to grade a test without knowing the answers - A bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *International Conference on Machine Learning, ICML*.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc.

Condorcet, M. d. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale.

Dai, P.; Lin, C. H.; Mausam; and Weld, D. S. 2013. POMDP-based control of workflows for crowdsourcing. *Artificial Intelligence* 202:52–85.

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics* 20–28.

Galaxy Zoo 2 data release. 2013. Galaxy zoo 2 data sets. http://data.galaxyzoo.org/.

Galaxy Zoo 2 Team. 2009. Galaxy Zoo 2. http://zoo2.galaxyzoo.org/.

Ipeirotis, P., and others. 2014. Dawid-Skene Algorithm Implementation. https://github.com/ipeirotis/get-another-label.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *ACM SIGKDD workshop on human computation*, 64–67. ACM.

Ipeirotis, P. G. 2010. Analyzing the Amazon Mechanical Turk marketplace. *ACM Crossroads* 17(2):16–21.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *International Conference on Autonomous Agents and Multiagent Systems, AAMAS*.

Kim, H., and Ghahramani, Z. 2012. Bayesian classifier combination. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, 619–627.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press.

Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; Murray, P.; and Vandenberg, J. 2008. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389(3):1179–1189.

Liu, C., and Wang, Y. 2012. Truelabel + confusions: A spectrum of probabilistic models in analyzing multiple ratings. In *International Conference on Machine Learning, ICML*.

MacKay, D. J. C. 1992. Information-based objective functions for active data selection. *Neural Computation* 4(4):590–604.

Minka, T., and Winn, J. 2008. Gates: A graphical notation for mixture models. In *Advances in Neural Information Processing Systems*, 1073–1080.

Minka, T.; Winn, J.; Guiver, J.; Webster, S.; Zaykov, Y.; Yangel, B.; Spengler, A.; and Bronskill, J. 2014. Infer.NET 2.6. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Pearl, J. 1989. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* 11:1297–1322.

Sheshadri, A., and Lease, M. 2013. SQUARE: A benchmark for research on computing crowd consensus. In *AAAI Conference on Human Computation and Crowdsourcing, HCOMP*.

Simpson, E.; Roberts, S. J.; Psorakis, I.; and Smith, A. 2013. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*. 1–35.

Venanzi, M.; Guiver, J.; Kazai, G.; Kohli, P.; and Shokouhi, M. 2014. Community-based bayesian aggregation models for crowdsourcing. In *International World Wide Web Conference, WWW*, 155–164.

von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Commun. ACM* 51(8):58–67.

Wauthier, F. L., and Jordan, M. I. 2011. Bayesian bias mitigation for crowdsourcing. In *Conference on Neural Information Processing Systems*, 1800–1808.

Welinder, P.; Branson, S.; Belongie, S.; and Perona, P. 2010. The multidimensional wisdom of crowds. In *Conference on Neural Information Processing Systems*, 2424–2432.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. R. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Conference on Neural Information Processing Systems 2009*, 2035–2043.

Willett, K. W.; Lintott, C. J.; Bamford, S. P.; Masters, K. L.; Simmons, B. D.; Casteels, K. R. V.; Edmondson, E. M.; Fortson, L. F.; Kaviraj, S.; Keel, W. C.; Melvin, T.; Nichol, R. C.; Raddick, M. J.; Schawinski, K.; Simpson, R. J.; Skibba, R. A.; Smith, A. M.; and Thomas, D. 2013. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* stt1458.

Yan, Y.; Rosales, R.; Fung, G.; and Dy, J. G. 2011. Active learning from crowds. In *International Conference on Machine Learning, ICML*, 1161–1168.

Zhou, D.; Platt, J. C.; Basu, S.; and Mao, Y. 2012. Learning from the wisdom of crowds by minimax entropy. In *Conference on Neural Information Processing Systems*, 2204–2212.