Learning to Aggregate Ordinal Labels by Maximizing Separating Width

Guangyong Chen [1] **Shengyu Zhang** [1] **Di Lin** [2] **Hui Huang** [2] **Pheng Ann Heng** [1]

Abstract

While crowdsourcing has been a cost and time efficient method to label massive samples, one critical issue is quality control, for which the key challenge is to infer the ground truth from noisy or even adversarial data by various users. A large class of crowdsourcing problems, such as those involving age, grade, level, or stage, have an ordinal structure in their labels. Based on a technique of sampling estimated label from the posterior distribution, we define a novel separating width among the labeled observations to characterize the quality of sampled labels, and develop an efficient algorithm to optimize it through solving multiple linear decision boundaries and adjusting prior distributions. Our algorithm is empirically evaluated on several real world datasets, and demonstrates its supremacy over state-of-the-art methods.

1. Introduction

Crowdsourcing has drawn increasing popularity in the field of machine learning by annotating millions of items in a short time with relatively low cost (Howe, 2006; Welinder & Perona; Deng et al., 2013; Jiang et al., 2015). This provides a great opportunity to build up large-scale training sets for complex models, such as deep neural networks (Krizhevsky et al., 2012), and to reach consensus among non-experts, such as peer grading in today's popular massive open online course (MOOC) systems. However, the quality of the collected results is often unreliable and diverse, and there are spammers who give random labels to make easy money, or even adversaries who deliberately give wrong answers. To address this issue, most crowdsourcing systems resort to distributing each item to a number of redundant workers. This raises a challenging question of how to aggregate such noisy and redundant labels.

An intuitive and baseline approach for crowdsourcing is to identify each item following the majority voting (MV) result of workers. Unfortunately, this approach is error-prone since it treats each worker equally, and the accuracy severely deteriorates with the fraction of less qualified workers, spammers or adversaries. Weighted majority voting (WMV) (Karger et al., 2011) method tries to address this issue by associating each worker with a weight to characterize his expertise. Specially, max-margin majority voting (M^{3}V) method (Tian & Zhu, 2015) optimizes the associated variables in WMV by maximizing the minimal difference between the aggregated score of the potential true label and the aggregated scores of others.

In a different approach, Dawid-Skene (DS) model (Dawid & Skene, 1979) represents each worker's expertise by a confusion matrix and uses a latent variable model to generate collected labels, which implicitly assumes a worker to perform equally well across all items in a common class. This model can be iteratively inferred by the famous Expectation-Maximization (EM) method (Dempster et al., 1977), and works well in practice. In particular, (Zhang et al., 2016) employs the spectral method (Anandkumar et al., 2012) to initialize the DS model, and obtains an optimal convergence rate up to a logarithmic factor by EM method. Recently, (Zhou et al., 2012; Tian & Zhu, 2015) proposed to improve the aggregating performance by integrating the merits of MV method with DS model. The performance of DS model and its variants often relies on the specially conceived priors with some manually configured hyperparameters.

All the above mentioned approaches are for aggregating general multiclass labels. In many practical applications, however, the labels have a natural ordinal structure. For instance, in MOOCs, students are often required to grade their own assignments on an ordinal scale of *5 (excellent), 4 (good), 3 (fair), 2 (pass)* and *1 (failure)*. In medical imaging, doctors are often required to mark images on an ordinal scale of *stage 1*, *stage 2*, *stage 3*, and *stage 4*. Ordinal label faces an issue of diverse standards. For example, given four assignments whose true grades are *5, 4, 3* and *2*, a strict marker may rate them as *4, 3, 2* and *1*. If we ignore the ordinal structure, we may identify this marker as an adversary because all his answer labels are considered wrong. But actually this marker grades all assignments in a correct

[1] The Chinese University of Hong Kong, Hong Kong, China. [2] Shenzhen University, China. Correspondence to: Shengyu Zhang <syzhang@cse.cuhk.edu.hk>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

order, which should be incorporated to improve the crowdsourcing performance.

This inspires us to transform the K-class ordinal labeling to $K-1$ binary classifications. That is, instead of directly using a label answer $k \in [K] = \{1, 2, \dots, K\}$, we use it to answer $K-1$ questions "Is the label greater than i" for all $i \in [K-1]$. In this way, the harsh marker's answer 4 for the first assignment would give three correct answers (on thresholds $i = 1, 2, 3$) and only one wrong answer (on threshold $i = 4$), making the marker's answers highly useful in the aggregation.

For each binary problem, we can employ a Gibbs sampler to generate label estimations from the generative model of crowdsourcing. However, these label estimations could be error-prone especially for the difficult items, whose labels may be sampled according to the uniform distribution, and it is well known that the performance of a generative model heavily relies on the specially conceived priors. To address these issues in binary crowdsourcing tasks, we define a separating width to characterize the quality of label estimations, and solve it by optimizing a linear decision boundary. The similar idea has been previously explored in (Cortes & Vapnik, 1995) and found a lot of success for supervising learning problems. By optimizing the separating width among two classes, we can improve the sampling accuracy and update the prior distributions automatically during the learning process. To characterize the quality of aggregating ordinal labels from K classes, we introduce $K-1$ decision boundaries to help optimize the separating width. As demonstrated empirically, our method achieves the best performance on the real-world datasets compared to other state-of-the-art methods.

The rest of this paper is organized as follows. Sec. 2 introduces some preliminary works for crowdsourcing tasks. Sec. 3 presents the generative model employed in this paper. Sec. 4.1 derives the objective function for binary aggregating problem, which is extended for the ordinal case in Sec. 4.2. The derivations of inference method are discussed in Sec. 5. Sec. 6 evaluates the performance of our method on some real-world datasets, and Sec. 7 concludes this paper.

2. Problem Setting and Preliminary Work

In this section, we formalize the problem and survey some preliminary methods. Suppose that there are M workers and N items taken from a total of K classes. For item i, define an $M \times K$ matrix R^i by putting $R^i_{jk} = 1$ if worker j labels the item as k, and $R^i_{jk} = 0$ otherwise. Note that R^i is a highly sparse matrix since each item is usually assigned to a small number of workers. The objective of a crowdsourcing problem is *to identify the true label z_i of item i based on the sparse matrices $\{R^1, \dots, R^N\}$.*

2.1. Majority Voting Method and its Variants

Majority Voting (MV) has been widely used to find the most likely label for item i by solving the following problem,

$$z_i = \arg_k \max \mathbf{1}_M^T R^i e_k, \qquad (1)$$

where $\mathbf{1}_M$ is a all-one column vector of dimension M and e_k is the k-th standard basis. Weighted majority voting (WMV) (Karger et al., 2011) generalizes MV by assigning weight vector $\eta \in \mathbb{R}^{M\times 1}$ to the workers and solving the following problem

$$z_i = \arg_k \max \eta^T R^i e_k. \qquad (2)$$

Specially, max-margin majority voting (M^{3}V) (Tian & Zhu, 2015) defines the crowdsourcing margin as the minimal difference between the aggregated score of the potential true label and aggregated scores of other labels, and solves η by maximizing the sum of the crowdsourcing margins of all items.

2.2. Dawid-Skene Model and its Variants

Dawid-Skene (DS) model has been another popular way to aggregate collected labels by capturing the uncertainties of labeling behaviors in a generative model. Compared with WMV and M^{3}V, both of which characterize the expertise of worker j by a scaler variable, DS model characterizes the expertise of worker j with an individual confusion matrix $A^j \in \mathbb{R}^{K\times K}$, where the (k, d)-th entry denotes the probability that worker j labels a class k sample as class d. Denote $\boldsymbol{A} = \{A^j\}_{j=1}^M$. DS model aims to maximize the likelihood of observed samples $\boldsymbol{R} = \{R^i\}_{i=1}^N$ as follows,

$$\max_{\boldsymbol{A}} \mathcal{L} = \sum_{i=1}^N \ln \int p(R^i|z_i, \boldsymbol{A}) p(z_i) dz_i, \qquad (3)$$

where $p(R^i|z_i, \boldsymbol{A}) = \prod_{j=1}^M \prod_{d=1}^K (A^j_{z_i d})^{R^i_{jd}}$ and z_i is a latent variable with $p(z_i) = \frac{1}{K}, \forall i \in [N]$. This likelihood function can be optimized iteratively by EM method (Dempster et al., 1977) as,

$$\begin{aligned} &\textbf{E-step:} \quad q(z_i = k) \propto \exp \sum_{j=1}^M \sum_{d=1}^K R^i_{jd} \ln A^j_{kd}, \\ &\textbf{M-step:} \quad A^j_{kd} \propto \sum_{i=1}^N q(z_i = k) R^i_{jd}. \end{aligned} \qquad (4)$$

Thus, the collected labels are aggregated following the rule $z_i = \arg_k \max \exp \sum_{j=1}^M \sum_{d=1}^K R^i_{jd} \ln A^j_{kd}$, where the unknown parameters $\boldsymbol{A}$ can be updated in **M-step** through maximum likelihood estimation (MLE) principle.

Recently, spectral methods have been applied to obtain a better initialization of the DS model (Zhang et al., 2016), which achieves an optimal convergence rate up to a logarithmic factor. By assuming some special structures of the confusion matrices $\boldsymbol{A}$, (Raykar et al., 2010) studies homogeneous DS model, and (Moreno et al., 2015) studies the existence of clusters of workers.

2.3. Recent Achievements

Some recent improvements have been achieved by combining MV related methods and DS related methods. (Zhou et al., 2012) assumes labels are generated according to a distribution over workers, items and labels, which can be inferred by minimizing its entropy with constraints developed from MV method and DS model. (Tian & Zhu, 2015) incorporates M^{3}V method with DS model in a regularized Bayesian framework (Zhu et al., 2014), and approximates the posterior distribution over the true labels with a Gibbs sampler. Nowadays, binary and general multi-class crowdsourcing problems have been widely studied in the literature, but the ordinal sibling has not received nearly as much attention yet. The work (Zhou et al., 2014) tries to use the ordinal structure and makes an assumption that workers have difficulty distinguishing between two adjacent ordinal classes whereas it is much easier to distinguish between two far-away classes. In this paper, we will develop a novel objective function to aggregate the ordinal labels, and achieve the best performance on the real-world datasets.

3. Generative Model of Crowdsourcing

In this section, we present a fully Bayesian model to generate observed matrices $\boldsymbol{R}$. First we note that some items may be intrinsically hard to label even for experts (which is not uncommon in, for example, medical imaging). To model such difficulty, we introduce a K-dimensional vector ω^i to denote the prior distribution of true label of the item i even for experts. (For items clearly from category k, the vector ω^i would be just the standard basis e_k.) Denote $\boldsymbol{\omega} = \{\omega^i\}_{i=1}^N$. We can obtain a joint distribution as follows.

$$p(\boldsymbol{R}, \boldsymbol{z}|\boldsymbol{A}, \boldsymbol{\omega}) = \prod_{i,j,d,k} (A^j_{kd})^{R^i_{jd}\mathbb{I}(z_i=k)} {\omega^i_k}^{\mathbb{I}(z_i=k)}, \quad (5)$$

where $\boldsymbol{A}$ contains the confusion matrices of all workers like DS model, $\boldsymbol{z}$ is the label vector to be solved and $\mathbb{I}(\cdot)$ is an indicator function. Since usually most workers just annotate a few items, we may not have sufficient samples to infer $\boldsymbol{z}$, $\boldsymbol{A}$ and $\boldsymbol{\omega}$. To overcome this, we formulate a fully Bayesian framework over $\boldsymbol{A}$ and $\boldsymbol{\omega}$ with prior from Dirichlet distributions (Minka, 2000), a family that has found numerous successful applications (such as topic models) to generate prior distributions. We assume that both workers' expertise $\boldsymbol{A}$ and items' difficulty $\boldsymbol{\omega}$ are random variables from the family of Dirichlet distributions

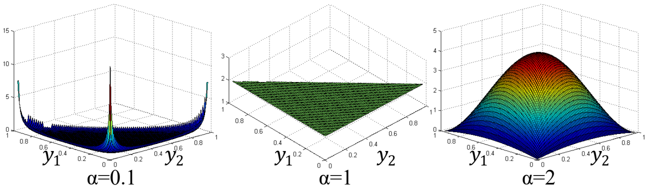


Figure 1. Illustrating how the Dirichlet density changes with respect to the scalar value α.

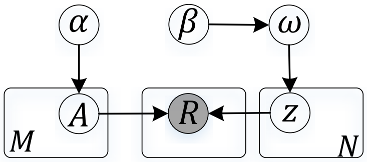


Figure 2. Graphical model of our generative model.

$$D(x|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{t=1}^{K} x_t^{\alpha-1}, \quad (6)$$

where $\Gamma(\cdot)$ is the gamma function. As illustrated in Figure 1, the concentration parameter α controls the sparsity preference of random vector. A precisely described item i should have ω^i be associated with a small concentration parameter, resulting in a sparse prior vector, while a vaguely described item should be associated with a large concentration parameter. To model the expertise A^j of the worker j, we also prefer that it has a small concentration parameter. We formulate the prior distributions over $\boldsymbol{A}$ and $\boldsymbol{\omega}$ as follows.

$$p(\boldsymbol{A}|\boldsymbol{\alpha}) = \prod_{j,k} D(A^j_{k:}|\alpha_j), p(\boldsymbol{\omega}|\boldsymbol{\beta}) = \prod_i D(\omega^i|\beta_i), \quad (7)$$

where $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^M$ and $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^N$. Combining Eq. (5) and (7) gives the following joint distribution,

$$p(\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{R}, \boldsymbol{z}|\boldsymbol{A}, \boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{\beta})p(\boldsymbol{A}|\boldsymbol{\alpha}).$$

The graphical model can be found in Fig. 2.

Given the matrices $\boldsymbol{R}$, we can get the posterior distribution over $\boldsymbol{A}, \boldsymbol{z}$ and $\boldsymbol{\omega}$, which can be formulated as

$$p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega}|\boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{\int p(\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{A}\boldsymbol{z}\boldsymbol{\omega}}. \quad (8)$$

We can obtain a classifier as $\arg_{\boldsymbol{z}} \max p(\boldsymbol{z}|\boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \arg_{\boldsymbol{z}} \max \int p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega}|\boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{A}\boldsymbol{\omega}$ to label the item, parametrized by the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Conventionally, researchers mainly focus on how to approximate the posterior distribution with better accuracy and running-time performance with the fixed prior distributions, or updating the prior distributions by introducing new priors over $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (Kim & Ghahramani, 2012; Moreno et al., 2015). However, the performance of the above generative model heavily relies on the specially conceived priors to incorporate domain knowledge, which transmits affects on the posterior estimations through Bayes' rules. Given a family of prior choices, we prefer the classifier with the more powerful discriminative capability to achieve better generalization performance.

4. Maximizing the Separating Distance

4.1. Binary Crowdsourcing Problem

Before we present the objective function to aggregate ordinal labels, we firstly consider a simple case, the binary crowdsourcing problem with $K = 2$. As shown in Eq. (8), by varying the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we can obtain a series of posterior approximations to identify unlabeled items via Bayes rule. Moreover, by fixing the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we can get multiple estimations of the true label of one item, which are randomly sampled from the posterior distribution over its true label. Thus, we are motivated to find the most favored set of label estimations over all items.

For a better generalization performance, we try to maximize the separation width between two classes. As shown in Fig. 3, the label set 2 is preferred to the set 1, because the set 2 has a larger separation width between two classes. To evaluate the separating width of samples with the label set $\boldsymbol{z} = \{z_i\}_{i=1}^N$, with $z_i \in \{-1, 1\}, \forall i \in [N]$, we introduce a linear decision boundary $f(R^i) = a^T R^i b$ with $a \in \mathbb{R}^{M\times 1}$ and $b \in \mathbb{R}^{K\times 1}$. Our decision boundary is formulated refer to the formulas in Eq. (1) and (2), where a denote the worker expertise and b transforms worker's label into a scale variable. Thus, we define an optimization problem as,

$$\begin{aligned} \min_{a,b} \quad & \mathcal{L}(a,b) = \|a\|_2^2 \|b\|_2^2, \\ s.t. \quad & z_i a^T R^i b \geq 1, \quad \forall i \in [N], \end{aligned} \tag{9}$$

where $\|x\|_2^2 = x^T x$ and the minimal value $\mathcal{L}(a^*, b^*)$ characterizes the separating width of the label set $\boldsymbol{z}$. This optimization problem can be understood from the objective function used in support vector machine (SVM) (Cortes & Vapnik, 1995), where the objective function is to maximize the margin width $(\|ba^T\|_F)^{-1} = (\|a\|_2^2\|b\|_2^2)^{-1}$ and the constrains state that all samples lie on the correct side of the margin. (The constraint in the above optimization problem can be viewed as the inner product of R^i and a rank-1 matrix ba^T. One may wonder why confining to rank-1 measurements. Note that MV (Eq.(1)) and WMV (Eq.(2)) are also of the rank-1 form, and our experiments also show that using higher rank measurements actually makes the generalization performance worse; see experiments in Appendix.) Since $\boldsymbol{z}$ is a random variable generated from the posterior distribution (8), we need to reformulate the objective function (9) as follows,

$$\begin{aligned} \min_{a,b,\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \mathcal{L}(a,b,\boldsymbol{\alpha},\boldsymbol{\beta}) = \|a\|_2^2 \|b\|_2^2, \\ s.t. \quad & \mathbb{E}_{p(z_i|R^i,\boldsymbol{\alpha},\boldsymbol{\beta})} z_i a^T R^i b \geq 1, \quad \forall i \in [N], \\ & p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega} | \boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto p(\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega} | \boldsymbol{\alpha}, \boldsymbol{\beta}). \end{aligned} \tag{10}$$

Practically, the labeled samples are often linearly inseparable by a single hyperplane; see Set 3 in Fig. 3. To cope with this issue, we relax the hard constrains by introducing non-negative slack variables ξ_i, one for each sample, and obtain a "soft" model as follows.

$$\begin{aligned} \min_{a,b,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}} \quad & \mathcal{L}(a,b,\boldsymbol{\alpha},\boldsymbol{\beta}) = \|a\|_2^2 \|b\|_2^2 + \frac{\lambda_2}{\lambda_1}\sum_{i=1}^N \xi_i, \\ s.t. \quad & \mathbb{E}_{p(z_i|R^i,\boldsymbol{\alpha},\boldsymbol{\beta})} z_i a^T R^i b \geq 1 - \xi_i, \quad \forall i \in [N], \\ & p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega} | \boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto p(\boldsymbol{R}, \boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega} | \boldsymbol{\alpha}, \boldsymbol{\beta}), \end{aligned} \tag{11}$$

where $\frac{\lambda_2}{\lambda_1}$ is used as a positive regularization parameter for later convenience, and $1 - \xi_i$ is the soft-margin for item i. If R^i lies on the correct side of the margin, $\xi_i = 0$. For sample on the wrong side, ξ_i is proportional to the distance to the margin. Thus, the value of ξ_i reflects the difficulty of identifying item i, or the error allowed to misclassify the item i. The calculation of $p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega} | \boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is intractable because it involves that of the marginal distribution $p(\boldsymbol{R}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. To address this issue, we introduce a redundant distribution $q(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega})$ and rewrite the optimization problem as follows.

$$\begin{aligned} \min_{a,b,\boldsymbol{\alpha},\boldsymbol{\beta},q,\boldsymbol{\xi}} \quad & \mathcal{L}(a,b,\alpha,\beta,\boldsymbol{\xi}) = \|a\|_2^2 \|b\|_2^2 + \frac{\lambda_2}{\lambda_1}\sum_{i=1}^N \xi_i, \\ s.t. \quad & \mathbb{E}_{q(z_i)} z_i a^T R^i b \geq 1 - \xi_i, \quad \forall i \in [N], \\ & \mathrm{KL}(q\|p) = 0. \end{aligned} \tag{12}$$

where q and p are shorthand for $q(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega})$ and $p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega} | \boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, respectively, to simplify the presentations in the rest of the paper. Let $\zeta_i = 1 - z_i a^T R^i b$, we can turn the optimization problem into the following one with two regularization terms: $\min_{a,b,\boldsymbol{\alpha},\boldsymbol{\beta},q,\boldsymbol{\xi}} \ \mathcal{L}(a,b,\boldsymbol{\alpha},\boldsymbol{\beta},q)$, where $\mathcal{L}$ is defined by

$$\begin{aligned} \mathcal{L}(a,b,\boldsymbol{\alpha},\boldsymbol{\beta},q) = {} & \mathrm{KL}(q\|p) + \lambda_1 \|a\|_2^2 \|b\|_2^2 \\ & + 2\lambda_2 \sum_{i=1}^N \sum_{z_i \in \{-1,1\}} q(z_i)(\zeta_i)_+, \end{aligned} \tag{13}$$

where $(\zeta_i)_+ = \max\{0, \zeta_i\}$ is the hinge loss function widely used in training classifiers . The factor $2\lambda_2$ is introduced here to simplify the derivations of inference methods

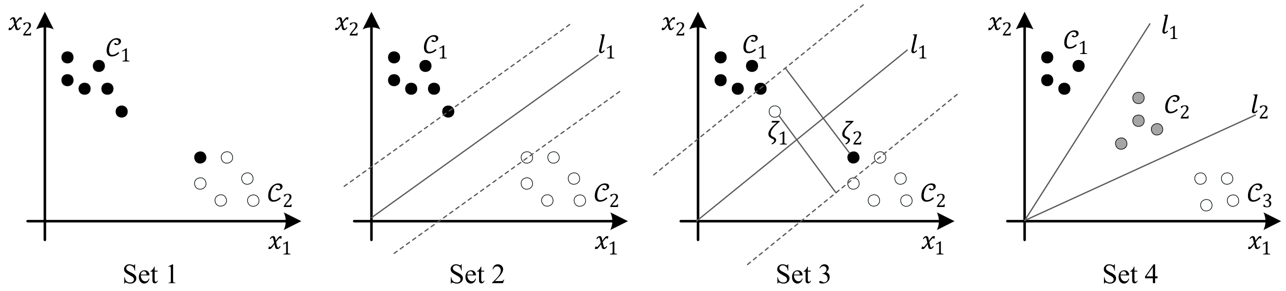


Figure 3. Four label sets. Set 1 contains one possible label set. Set 2 contains another label set, which is preferred to the set 1. Set 3 contains a label set, whose separating width can be estimated with the slack variables. Set 4 contains a set of ordinal labels, whose separating width can be estimated by transforming the ordinal problem into multiple binary ones.

later. By optimizing the unknown parameters in the objective function in Eq.(13), we can obtain the estimated labels with the largest separating width.

4.2. Ordinal Crowdsourcing Problem

As introduced in Section 1, transforming a K-class ordinal labeling problem ("what is the label of this item?") to $(K-1)$ binary classification problems ("Is the label of this item greater than k?" for $k \in [K-1]$) allows us to exploit more useful information from workers. The transform is illustrated in Set 4 of Fig. 3, where we have items coming from K ordered classes, $\mathcal{C}_1, \ldots, \mathcal{C}_K$. We look for $K-1$ decision boundaries, with boundary t discriminating classes $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_t$ and classes $\mathcal{C}_{t+1} \cup \cdots \cup \mathcal{C}_K$. For the t-th binary question, we introduce a linear decision boundary as $f_t(R^i) = a_t^T R^i b_t$. It is easily verified that all boundaries intersecting at the zero point. With $\boldsymbol{a} = \{a_t\}_{t=1}^{K-1}$ and $\boldsymbol{b} = \{b_t\}_{t=1}^{K-1}$, the loss function in Eq. (13) becomes

$$\begin{aligned}\mathcal{L}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha},\boldsymbol{\beta},q) =& \mathrm{KL}(q\|p) + \lambda_1 \sum_{t=1}^{K-1} \|a_t\|_2^2 \|b_t\|_2^2 \\ &+ 2\lambda_2 \sum_{i=1}^{N} \sum_{z_i=1}^{K} q(z_i) \sum_{t=1}^{K-1} (\zeta_{it})_+, \quad (14)\end{aligned}$$

where $\zeta_{it} = 1 - \mathrm{sgn}_t(z_i) a_t^T R^i b_t$ with $\mathrm{sgn}_t(z_i) = -1$ if $z_i \le t$ and $\mathrm{sgn}_t(z_i) = 1$ if $z_i > t$. It is obvious that our ordinal model will degenerate into binary one when $K = 2$. When $K \ge 3$, the ordinal label should be estimated by considering the predicted results from $K-1$ binary problems.

5. Inference Details

In this section, we present the implementation details to infer the true labels and all other unknown parameters involved in ordinal crowdsourcing problems. Our inference method consists of two parts. In the first part, we employ a Gibbs sampler to approximately sample from the posterior distribution $p = p(\boldsymbol{A}, \boldsymbol{z}, \boldsymbol{\omega} | \boldsymbol{R}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. In the second part, we update the hyperparameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and linear decision boundaries based on the gradient method to achieve the largest separating width.

To approximate the intractable posterior distribution p, there are two standard approaches, which are Variational Bayesian (VB) and Gibbs sampling. Compared with the Gibbs sampling method, VB is usually difficult in its functional optimization, especially hard in our case due to the hinge loss function. Moreover, VB often suffers from inaccuracy because of the potentially impractical assumption of independence of variables. Gibbs sampling is applicable here, because it provides numerical approximations to the integration problems in large dimensional spaces by generating an instance from the conditional distribution of each variable in turn. It can be shown that the sequence of samples constitutes a Markov chain, which finally converges to the targeted posterior distribution as the stationary distribution.

Since the sampling process of the confusion matrices $\boldsymbol{A}$ and the items' difficulties $\boldsymbol{\omega}$ can be developed in the standard way, we leave their derivations in Appendix and mainly discuss the sampling process of true labels $\boldsymbol{z}$ here. The difficulty of sampling $\boldsymbol{z}$ is mainly due to the hinge loss function $(\zeta_{it})_+$. We employ data augmented technique (Polson & Scott, 2011) to approximate the hinge loss function. According to the equality (Andrews & Mallows, 1974),

$$\exp(-2\lambda_2(\zeta_{it})_+) = \int \phi(z_i, \gamma_{it} | R^i) d\gamma_{it}, \quad (15)$$

with $\phi(z_i, \gamma_{it} | R^i) = (2\pi\gamma_{it})^{-\frac{1}{2}} \exp(\frac{-1}{2\gamma_{it}}(\gamma_{it} + \lambda_2\zeta_{it})^2)$ and γ_{it} as a non-negative augmented variable, we can reformulate the objective function in Eq.(14) as follows.

$$\begin{aligned}\mathcal{L}(\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha},\boldsymbol{\beta},q) \le\ & \mathrm{KL}(q\|p) + \lambda_1 \sum_t a_t^T a_t b_t^T b_t \quad (16) \\ &+ \sum_{i,z_i,t} \int q(z_i) q(\gamma_{it}) \ln \frac{q(\gamma_{it})}{\phi(z_i, \gamma_{it} | R^i)} d\gamma_{it},\end{aligned}$$

where the inequality comes from Jensen's inequality with a new distribution $q(\gamma_{it})$ to help to approximate the hinge loss function $\exp(-2\lambda_2(\zeta_{it})_+)$. Note that the right hand side of the inequality is tractable, minimizing which would give an upper bound of the original optimization problem. Before we sample the true labels of items, we need to first generate augmented variables $\boldsymbol{\gamma} = \{\gamma_{it}\}_{i=1,t=1}^{N,K-1}$. When fixing other random variables, we can generate the (i,t)-th augmented parameter according to the following generalized inverse Gaussian (GIG) distribution,

$$\gamma_{it} \sim \frac{1}{Z}\gamma_{it}^{-\frac{1}{2}} \exp[-\frac{1}{2}(\gamma_{it} + \frac{\lambda_2^2\zeta_{it}^2}{\gamma_{it}})], \tag{17}$$

where Z is the normalization term. It has been shown that $\frac{1}{\gamma_{it}}$ can be drawn efficiently with $\mathcal{O}(1)$ time complexity (Michael et al., 1976).

Here, we can sample the true labels of all items. Let $\phi(\boldsymbol{z}, \boldsymbol{\gamma}|\boldsymbol{R}) = \prod_{i=1}^{N}\prod_{t=1}^{K-1} \phi(z_i, \gamma_{it}|R^i)$. Rewrite the objective function shown in Eq. (16) with respect to $q(z_i)$ as follows,

$$\begin{aligned} &\mathcal{L}(q(z_i)) \\ &\leq \mathrm{KL}(q\|p) + \sum_{i,z_i,t}\int q(z_i)q(\gamma_{it}) \ln \frac{q(\gamma_{it})}{\phi(z_i,\gamma_{it}|R^i)} d\gamma_{it} \\ &= \mathrm{KL}(q \cdot q(\boldsymbol{\gamma})\|p \cdot \phi(\boldsymbol{z}, \boldsymbol{\gamma})), \end{aligned} \tag{18}$$

Thus, with all other parameters fixed, we can sample $z_i \in [K]$ according to the following distribution,

$$q(z_i) \propto p(R^i, \boldsymbol{A}, z_i, \omega_i|\boldsymbol{\alpha}, \boldsymbol{\beta}) \prod_{t=1}^{K-1} \phi(z_i, \gamma_{it}|R_t^i), \tag{19}$$

Let us examine the two terms on the right hand side. The first term comes from the generative model of crowdsourcing, while the second term maximizes the separating width of the estimated ordinal labels. For the binary crowdsourcing problem, we have only one decision boundary to measure the separating width, while for the ordinal crowdsourcing problem with K classes, we get $K-1$ intersected decision boundaries to measure the separating width.

After obtaining a set of random samples to approximate the joint posterior distribution q over all model parameters and augmented variables, the objective function shown in Eq. (14) becomes a parametric function with respect to $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Thus, we can intuitively update these parameters based on the gradient method. The derivations of the updating formulas over $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be found in Appendix.

Let $\langle f(x)\rangle = \int q(x)f(x)dx$ denote the expectation of $f(x)$ with respect to the distribution of $q(x)$. Our method is outlined in Algorithm 1, in which each **while** iteration consists of two **for** loops, and the source code with demo can be found on the website[1]. In the first **for** loop, we employ a Gibbs sampler to generate the random variables to approximate the posterior distribution. In the second part, we solve the separating width by optimizing $K-1$ decision boundaries, and update the prior distributions by gradient method. Compared with the traditional generative model of crowdsourcing, including DS model and its variants, our method introduces an augment variable γ_{it} to approximate the hinge loss function, which is further involved in the generation of true labels. This algorithm is iteratively implemented to reach a local optimum.

Algorithm 1 Our Ordinal Crowdsourcing Method

```
Input: $\boldsymbol{R} = \{R^i\}_{i=1}^N$, $\lambda_1$, $\lambda_2$ and the learning rates $\eta$.
Initializing $\boldsymbol{z} = \{z_i\}_{i=1}^N$ by MV, $\boldsymbol{a}, \boldsymbol{b}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$
while not convergence do
    for $i = 1 : N$ do
        $A_{kd}^j \sim D(A_{kd}^j|\alpha_j + \sum_{i=1}^N R_{jd}^i \mathbb{I}(z_i = k))$
        $\omega_k^i \sim D(\omega_k^i|\beta_i + \mathbb{I}(z_i = k))$
        $\gamma_{it} \sim \frac{1}{Z}\gamma_{it}^{-\frac{1}{2}} \exp[-\frac{1}{2}(\gamma_{it} + \frac{\lambda_2^2\zeta_{it}^2}{\gamma_{it}})]$
        $z_i \sim p(R^i, \boldsymbol{A}, z_i, \omega_i|\boldsymbol{\alpha}, \boldsymbol{\beta}) \prod_{t=1}^{K-1} \phi(z_i, \gamma_{it}|R_t^i)$
    end for
    for $t = 1 : K-1$ do
        $\Sigma_{a_t} = 2\lambda_1\|b_t\|_2^2\boldsymbol{I} + \sum_{i=1}^N \frac{\lambda_2^2}{\langle\gamma_{it}\rangle} R^i b_t b_t^T {R^i}^T$
        $a_t = \Sigma_{a_t}^{-1}(\sum_{i=1}^N(\lambda_2 + \frac{\lambda_2^2}{\langle\gamma_{it}\rangle})\langle\mathrm{sgn}_t(z_i)\rangle R^i b_t)$
        $\Sigma_{b_t} = 2\lambda_1\|a_t\|_2^2\boldsymbol{I} + \sum_{i=1}^N \frac{\lambda_2^2}{\langle\gamma_{it}\rangle} a_t^T R^i {R^i}^T a_t$
        $b_t = \Sigma_{b_t}^{-1}(\sum_{i=1}^N(\lambda_2 + \frac{\lambda_2^2}{\langle\gamma_{it}\rangle})\langle\mathrm{sgn}_t(z_i)\rangle a_t^T R^i)$
    end for
    $\alpha_j \leftarrow \alpha_j - \eta\frac{\partial\mathcal{L}(\alpha_j)}{\partial\alpha_j}, \forall j \in [M]$
    $\beta_i \leftarrow \beta_i - \eta\frac{\partial\mathcal{L}(\beta_i)}{\partial\beta_i}, \forall i \in [N]$
end while
```

6. Experiments and Discussions

To fully evaluate the ideas employed in this paper, we present empirical studies of our aggregating method in comparison with competitive ones not only on ordinal crowdsourcing tasks, but also binary crowdsourcing tasks. For our method, we configure $\lambda_1 = \lambda_2 = 1, \boldsymbol{\alpha} = \boldsymbol{1}_M, \boldsymbol{\beta} = \boldsymbol{1}_N, \eta = 1 \times 10^{-5}$ and initialize z_i by the majority voting result. In each run of our method, we generate 80 samples to approximate the posterior distribution and discard the first 10 samples as burn-in steps. The reported error rate of our method is averaged over 10 runs, and all experiments are conducted in a PC with Intel Core i7 1.8GHz CPU and 8.00GB RAM.

6.1. Binary Crowdsourcing Tasks

We first evaluate our method on three binary benchmark datasets shown in Table 1, include labeling bird species

[1] http://appsrv.cse.cuhk.edu.hk/~gychen/

Table 1. The summary of real-world datasets used in the comparison experiments.

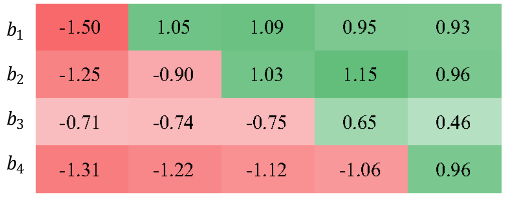
Name	Classes	Items	Workers	Labels/item
Bird	2	108	39	39.0
RTE	2	800	164	10.0
TREC	2	19,033	762	4.64
Web	5	2,665	177	5.84
Age	7	1,002	165	10

(Welinder et al., 2010) (Bird dataset), recognizing textual entailment (Snow et al., 2008) (RTE dataset) and accessing the relevance of topic-document pairs with a binary judgment in TREC 2011 crowdsourcing track (Gabriella & Matthew, 2011) (TREC dataset). The competitive methods include the pure majority voting estimator (refereed to as MV), the EM method for DS model initialized by majority voting (refereed to as MV-DS), the EM method for DS model initialized by spectral method (refereed to as Opt-DS) (Zhang et al., 2016), the Gibbs sampler for the Bayesian extension of M^{3}V (Tian & Zhu, 2015) (referred to as G-CrowdSVM), the SVD-based algorithm proposed in (Ghosh et al., 2011) (referred to as Gh-SVD), and the Eigenvalues of Ratio algorithm proposed in (Dalvi et al., 2013) (referred to as Eig-Ratio). The performance of all methods are evaluated by error as $l_0 = 1 - \frac{1}{|\boldsymbol{z}|}\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_0$, where $\boldsymbol{z}$ contains true labels of items (available in all these datasets) and $\hat{\boldsymbol{z}}$ contains estimations given by our algorithm (not using any information of $\boldsymbol{z}$). Noted that the reported error rates of G-CrowdSVM are the average over 10 random runs as our method.

As shown in Table 2, our method achieves the best performance among all methods on three benchmark datasets. Without regards to the prior knowledges over workers' expertise and items' difficulties, we can degenerate our model into MV-DS model by setting $\lambda_2 = \lambda_1 = 0$. Comparing with the performance of MV-DS model, especially Opt-DS method, we present a more complicated generative model, leading to better predictive results. Compared with G-CrowdSVM method, our method updates prior distributions and improves the sampling accuracy by optimizing the separating width during the learning process, which leads to the improvements of predictive performance on all datasets.

6.2. Ordinal Crowdsourcing Tasks

In this part, we report empirical results of our method on ordinal benchmark datasets in comparison with competitive ones. We consider MV, MV-DS, and G-CrowdSVM as baselines, and compare our method with Entropy(O) (Zhou et al., 2014), which consider the ordinal structures in labels. As shown in Table 1, we have two ordinal datasets. One is to judge the relevance of query-URL pairs with a 5-level rating score (Web dataset), and the other is to identify the age of each subject with a 7-level rating score (Age dataset). To evaluate the performance of aggregating ordinal labels, we define three following error measurements as: $l_0 = 1 - \frac{1}{|\boldsymbol{z}|}\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_0$, $l_1 = \frac{1}{|\boldsymbol{z}|}\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_1$ and $l_2 = \frac{1}{|\boldsymbol{z}|}\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2$. Compared with the error rate l_0, the measures l_1 and l_2 take precision into consideration, and may be preferred for aggregating ordinal labels when one cares about the severity of error.



b_1	-1.50	1.05	1.09	0.95	0.93
b_2	-1.25	-0.90	1.03	1.15	0.96
b_3	-0.71	-0.74	-0.75	0.65	0.46
b_4	-1.31	-1.22	-1.12	-1.06	0.96

Figure 4. The average value of 80 samples of $\{b_t\}_{t=1}^4$ in a random run.

Table 3 summarizes the performance of all methods on the ordinal datasets, and shows that our method consistently outperforms the others in predicting the ordinal labels of items. Similar with our method, G-CrowdSVM attempts to maximize the margin between the aggregated score of potential true label and the aggregated score of others, and achieves the better performance in comparison with the state-of-the-art method to aggregate ordinal labels, Entropy(O). Compared with G-CrowdSVM, we treat the problem of aggregating collected labels as the classification problem, and introduce $K - 1$ decision boundaries to consider the ordinal relationship among categories. As shown in Table 3, on the Web dataset, our method significantly reduces the average l_0 error rate from 7.99% to 3.22%. In addition, the average l_1 error of our method is 3.69%, which is only slightly larger than the l_0 error rate of 3.22%. It means that, even for the incorrect labels $\hat{z}_i$ outputted by our algorithm, our $\hat{z}_i$ is not far away from its true answer z_i, resulting in a relatively small damage.

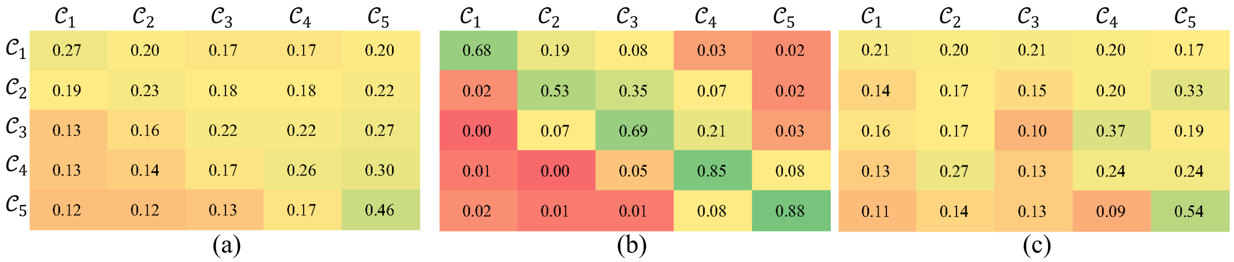
The Web dataset has been widely used in the evaluation of ordinal crowdsourcing problem. On this dataset, our method introduces 4 decision boundaries to measure the separating width of generated true labels. To help to understand the linear decision boundaries learned by our method, we illustrate the average value over 80 samples of $\{b_t\}_{t=1}^4$ as Fig. 4. It can be seen that the absolute value of all entries in b_t is approximated to 1. To be more concrete, let us consider the first decision boundary with b_1, which calculate $R^i b_1$ for the item i. Thus, $R^i b_1$ successfully reduces the ordinal problem into a binary one, the j-th entry in $R^i b_i$ would be -1 if worker j ranks item i as 1 and 1 if worker j rank item i as $2, 3, 4$ and 5. Note that a_t characterizes the expertise of all workers for the t-th binary problem. Fig. 5 contains three confusion matrices, including the averaged confusion matrix of all workers, the confusion matrices of

Table 2. l_0 error rate (%) in predicting the latent labels on three binary benchmark datasets.

Binary Dataset	Ours	G-CrowdSVM	Opt-DS	MV-DS	MV	Gh-SVD	Eig-Ratio
Bird	**9.25**±0.17	10.37±0.41	10.09	11.11	24.07	27.78	27.78
RTE	**7.00**±0.29	7.72±0.22	7.12	7.12	10.31	49.13	9.00
TREC	**29.30**±0.11	31.32±0.34	29.80	30.02	34.86	42.99	43.96

Table 3. Errors in predicting the latent labels on two ordinal benchmark datasets.

Ordinal Dataset		Ours	G-CrowdSVM	Entropy(O)	MV-DS	MV
Web	l_0	**0.0322**±0.0013	0.0799±0.0026	0.1040	0.1574	0.2693
	l_1	**0.0369**±0.0032	0.0940±0.0057	0.1173	0.2149	0.4251
	l_2	**0.2153**±0.0019	0.3629±0.0044	0.3816	0.5358	0.9247
Age	l_0	**0.3210**±0.0025	0.3298±0.0036	0.3732	0.3962	0.3488
	l_1	**0.3493**±0.0036	0.3737±0.0033	0.4541	0.5000	0.4083
	l_2	**0.6192**±0.0047	0.6592±0.0028	0.7936	0.8518	0.7297



(a)

	c_1	c_2	c_3	c_4	c_5
c_1	0.27	0.20	0.17	0.17	0.20
c_2	0.19	0.23	0.18	0.18	0.22
c_3	0.13	0.16	0.22	0.22	0.27
c_4	0.13	0.14	0.17	0.26	0.30
c_5	0.12	0.12	0.13	0.17	0.46

(b)

	c_1	c_2	c_3	c_4	c_5
c_1	0.68	0.19	0.08	0.03	0.02
c_2	0.02	0.53	0.35	0.07	0.02
c_3	0.00	0.07	0.69	0.21	0.03
c_4	0.01	0.00	0.05	0.85	0.08
c_5	0.02	0.01	0.01	0.08	0.88

(c)

	c_1	c_2	c_3	c_4	c_5
c_1	0.21	0.20	0.21	0.20	0.17
c_2	0.14	0.17	0.15	0.20	0.33
c_3	0.16	0.17	0.10	0.37	0.19
c_4	0.13	0.27	0.13	0.24	0.24
c_5	0.11	0.14	0.13	0.09	0.54

Figure 5. (a) the averaged confusion matrix over all worker, (b) the confusion matrix of an expert, (c) the confusion matrix of a spammer.

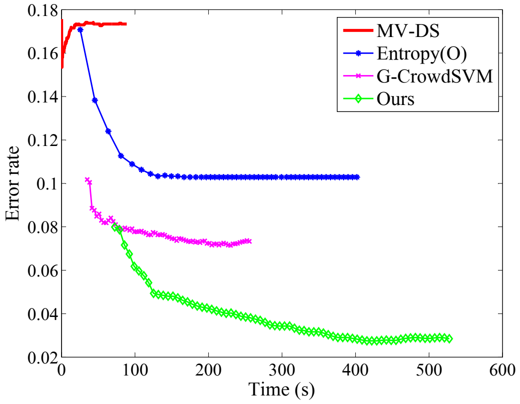


Figure 6. Error rates per iteration of various estimators on the Web dataset.

an expert and a spammer. It can be found that the spammer ranks all items randomly to make easy money, while the expert has a confusion matrix similar to the identical matrix. Our method can accurately estimate the confusion matrices of all workers even given the averaged confusion matrix acts like a spammer. Fig. 6 summarizes the training time and error rates after each iteration for all estimators on the Web dataset. It can be found that the proposed method coverages to a lower error rate and all other three methods have error convergence curves all above ours.

7. Conclusions

In this paper, we develop a novel method to aggregate ordinal labels by optimizing the separating width among classes. To measure the separating width among ordinal labels, we first investigate a binary case, and then extend our achievements to the ordinal one. With $K-1$ decision boundaries, we define an optimization problem for measuring the separating width among ordinal classes. The newly introduced boundaries not only help to optimize the hyperparameters, but also calibrate the estimated labels sampled from the generative model. A Gibbs sampler is adopted to approximate the posterior distribution, while the gradient method is used to calculate the separating width and optimize the hyperparameters.

As demonstrated on the ordinal datasets, which is the main focus of this paper, our method consistently achieves the best performance compared with competitive ones, and the improvements on Web dataset are significant. As demonstrated by the experimental results on the binary datasets, our algorithm works slightly better than any previous method. Thus, our algorithm provides a uniform method in both binary and ordinal cases.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments to improve the presentation of this paper. This work is supported by the China 973 Program (Project No. 2015CB351706) and a grant from the National Natural Science Foundation of China (Project No. 61233012). Shengyu Zhang was supported by Research Grants Council of the Hong Kong S.A.R. (Project no. CUHK14239416).

References

Anandkumar, A., Liu, Y. K., Hsu, D. J., Foster, D. P., and Kakade, S. M. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pp. 917–925, 2012.

Andrews, D. F. and Mallows, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 99–102, 1974.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dalvi, N., Dasgupta, A., and Kumar, R. .and Rastogi, V. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 285–294. ACM, 2013.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

Deng, J., Krause, J., and Li, F. F. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2013.

Gabriella, K. and Matthew, L. Overview of the trec 2011 crowdsourcing track. In *Proceedings of TREC 2011*, 2011.

Ghosh, A., Kale, S., and McAfee, P. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pp. 167–176. ACM, 2011.

Howe, J. The rise of crowdsourcing. *Wired magazine*, 14 (6):1–4, 2006.

Jiang, M., Huang, S., Duan, J., and Zhao, Q. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1072–1080. IEEE, 2015.

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

Kim, H. C. and Ghahramani, Z. Bayesian classifier combination. In *AISTATS*, pp. 619–627, 2012.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Michael, J. R., Schucany, W. R., and Haas, R. W. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2):88–90, 1976.

Minka, T. Estimating a dirichlet distribution, 2000.

Moreno, P. G., Artés-Rodríguez, A., Teh, Y. W., and Perez-Cruz, F. Bayesian nonparametric crowdsourcing. *Journal of Machine Learning Research*, 16:1607–1627, 2015.

Polson, N. G. and Scott, S. L. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

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

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 254–263. Association for Computational Linguistics, 2008.

Tian, T. and Zhu, J. Max-margin majority voting for learning from crowds. In *Advances in Neural Information Processing Systems*, pp. 1621–1629, 2015.

Welinder, P. and Perona, P. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*.

Welinder, P., Branson, S., Perona, P., and Belongie, S. J. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, pp. 2424–2432, 2010.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet em: a provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44, 2016.

Zhou, D., Basu, S., Mao, Y., and Platt, J. C. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, pp. 2195–2203, 2012.

Zhou, D., Liu, Q., Platt, J. C., and Meek, C. Aggregating ordinal labels from crowds by minimax conditional entropy. In *ICML*, pp. 262–270, 2014.

Zhu, J., Chen, N., and Xing, E. P. Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, 15(1): 1799–1847, 2014.