Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Dengyong Zhou DENZHO@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

Qiang Liu QLIU1@UCI.EDU
University of California, Irvine, CA 92697

John C. Platt JPLATT@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

Christopher Meek MEEK@MICROSOFT.COM
Microsoft Research, Redmond, WA 98052

Abstract

We propose a method to aggregate noisy ordinal labels collected from a crowd of workers or annotators. Eliciting ordinal labels is important in tasks such as judging web search quality and rating products. Our method is motivated by the observation that workers usually have difficulty distinguishing between two adjacent ordinal classes whereas distinguishing between two classes which are far away from each other is much easier. We formulate our method as minimax conditional entropy subject to constraints which encode this observation. Empirical evaluations on real datasets demonstrate significant improvements over existing methods.

1. Introduction

There has been considerable amount of work on learning when labeling is expensive, such as techniques on transductive inference and active learning. With the emergence of crowdsourcing services, like Amazon Mechanical Turk, labeling costs in many applications have dropped dramatically. Large amounts of labeled data can now be gathered at low price. Due to a lack of domain expertise and misaligned incentives, however, labels provided by crowdsourcing workers are often noisy. To overcome the quality issue, each item is usually simultaneously labeled by several workers, and then we aggregate the multiple labels with some manner, for instance, majority voting.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

An advanced approach for label aggregation is suggested by Dawid & Skene (1979). They assume that each worker has a latent confusion matrix for labeling. The off-diagonal elements represent the probabilities that a worker mislabels an arbitrary item from one class to another while the diagonal elements correspond to her accuracy in each class. Worker confusion matrices and true labels are jointly estimated by maximizing the likelihood of observed labels. One may further assume a prior distribution over worker confusion matrices and perform Bayesian inference (Raykar et al., 2010; Liu et al., 2012; Chen et al., 2013).

The method of Dawid & Skene (1979) implicitly assumes that a worker performs equally well across all items in a common class. In practice, however, it is often the case that one item is more difficult to label than another. To address this heterogeneous issue, Zhou et al. (2012) propose a minimax entropy principle for crowdsourcing. It results in that each item is associated with a latent confusion vector besides a latent confusion matrix for each worker. Observed labels are determined jointly by worker confusion matrices and item confusion vectors through an exponential family model. Moreover, it turns out that the probabilistic labeling model can be equivalently derived from a natural assumption of objective measurements of worker ability and item difficulty. Such kinds of objectivity arguments have been widely discussed in the literature of mental test theory (Rasch, 1961; Lord & Novick, 1968).

All the above approaches are for aggregating multiclass labels. In many scenarios, the labels are ordinal. To be concrete, let us consider the example of screening mammograms. A mammogram is an x-ray picture used to check for breast cancer in women. Radiologists often rate mammograms on a scale such as no cancer, benign cancer, possible malignancy, or malignancy. Since ordinal labels are

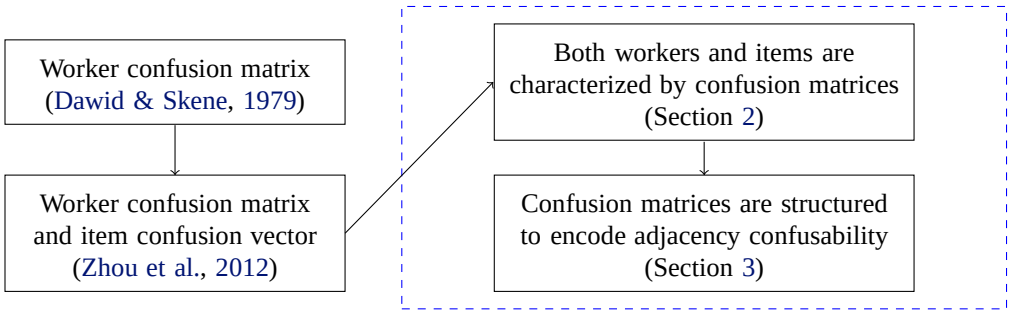

Figure 1. Roadmap for the development of our method for aggregating crowdsourced ordinal labels.

special cases of multiclass labels, one may apply the previous general multiclass approaches to aggregate ordinal labels. However, workers have different error patterns in ordinal labeling. We observe that they usually have difficulty distinguishing between two adjacent ordinal classes whereas distinguishing between two classes which are far away from each other is much easier. We refer to this observation as *adjacency confusability*. In the example of screening mammograms, a radiologist may rate a mammogram which indicates possible malignancy as malignancy, but it is less likely that she rates a mammogram which indicates no cancer as malignancy.

We propose a method to aggregate ordinal labels by taking adjacency confusability into account. The roadmap for the development of our method is illustrated in Figure 1. We first develop a general minimax conditional entropy approach for aggregating multiclass labels (Section 2). It extends the work of Zhou et al. (2012), and generates a labeling model in which both item difficulty and worker ability are characterized by confusion matrices. In contrast, Zhou et al. (2012) model item difficulty using confusion vectors. Then, as the main contribution in this paper, we adapt the general multiclass minimax conditional entropy approach to ordinal labels (Section 3). It is minimax conditional entropy subject to a different set of worker and item constraints which encode adjacency confusability observed in ordinal labeling. The formulation gives rise to an ordinal labeling model parameterized with *structured* worker and item confusion matrices. Due to the introduced structure, the ordinal labeling model has fewer parameters than the multiclass labeling model if there are more than two classes. In the case in which there are only two classes, the two models coincide as expected. In Section 4, we further introduce two kinds of regularization into the minimax conditional entropy scheme to address two practical issues. One is for preventing overfitting, and the other is for obtaining probabilistic labels. In practice, probabilistic labels are more useful than deterministic ones. When the label estimate of an item is close to uniform over several classes, we may want to either ask for more labels for the item or forward the item to an external expert. In Section 5, we present a dual coordinate ascent method to solve the minimax conditional entropy program as well as an efficient model selection technique. In Section 6, we empirically compare our method with existing methods that aggregate multiclass or ordinal labels. Finally, we conclude the paper with a discussion of future directions in Section 7.

2. Multiclass Minimax Conditional Entropy

In this section, we develop a method for aggregating multiclass labels from crowds. It extends the work of Zhou et al. (2012) such that item difficulty is represented by a two-dimensional confusion matrix instead of one-dimensional confusion vector. Matrix forms are more flexible for encoding domain knowledge in different types of labeling tasks. The flexibility is demonstrated when this general multiclass method is adapted to ordinal labels in Section 3.

Let us first introduce some notations. Assume that there are a group of workers indexed by i, a set of items indexed by j, and a number of classes indexed by k or c. Let x_{ij} be the observed label that worker i assigns to item j, and X_{ij} be the corresponding random variable. Denote by Y_j the true label of item j and $Q(Y_j = c)$ the probability that item j belong to class c. A special case is that $Q(Y_j = c)$ is simply an indicator function, that is, deterministic labels.

Now we introduce two four-dimensional tensors with each dimension corresponding to workers i, items j, observed labels k, or true labels c. One is the *empirical tensor* in which each element is given by

$$\widehat{\phi}_{ij}(c,k) = Q(Y_j = c)\mathbb{I}(x_{ij} = k)$$

to represent an observed confusion from class c to class k by worker i on item j, and the other is the *expected tensor* in which each element is given by

$$\phi_{ij}(c,k) = Q(Y_j = c)P(X_{ij} = k|Y_j = c)$$

to represent an expected confusion from class c to class k by worker i on item j. In the development of our method, these two tensors play the fundamental roles.

Denote the sum of the entropies of the observed labels conditioned on the true labels by

$$H(X|Y) = -\sum_{j,c} Q(Y_j = c) \sum_{i,k} P(X_{ij} = k|Y_j = c) \times \log P(X_{ij} = k|Y_j = c).$$

To estimate the true labels, we minimize-maximize the conditional entropy sum, that is,

$$\min_{Q} \max_{P} \quad H(X|Y) \tag{1}$$

subject to the worker and item constraints

$$\sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \ \forall i,k,c, \tag{2a}$$

$$\sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \ \forall j,k,c, \tag{2b}$$

plus the probability constraints

$$\sum_{k} P(X_{ij} = k|Y_j = c) = 1, \ \forall i,j,c, \tag{3a}$$

$$\sum_{c} Q(Y_j = c) = 1, \ \forall j, \ Q(Y_j = c) \geq 0, \ \forall j,c. \tag{3b}$$

The constraints in Equation (2a) enforce the expected confusion counts in the worker dimension to match their empirical counterparts. Symmetrically, the constraints in Equation (2b) enforce the expected confusion counts in the item dimension to match their empirical counterparts.

The Lagrangian of the inner maximization problem in (1) can be written as

$$L = H(X|Y) + L_\sigma + L_\tau + L_\lambda \tag{4}$$

with

$$L_\sigma = \sum_{i,c,k} \sigma_i(c,k) \sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right],$$

$$L_\tau = \sum_{j,c,k} \tau_j(c,k) \sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right],$$

$$L_\lambda = \sum_{i,j,c} \lambda_{ijc} \left[\sum_{k} P(X_{ij} = k|Y_j = c) - 1\right],$$

where $\sigma_i(c,k), \tau_j(c,k)$ and λ_{ijc} are introduced as the Lagrange multipliers. By the Karush-Kuhn-Tucker (KKT) conditions (Boyd & Vandenberghe, 2004),

$$\frac{\partial L}{\partial P(X_{ij} = k|Y_j = c)} = 0,$$

which implies

$$\log P(X_{ij} = k|Y_j = c) = \lambda_{ijc} - 1 - \sigma_i(c,k) - \tau_j(c,k).$$

Combining the above equation and the probability constraints in (3a) eliminates λ and yields

$$P(X_{ij} = k|Y_j = c) = \frac{1}{Z_{ij}} \exp[\sigma_i(c,k) + \tau_j(c,k)], \tag{5}$$

where Z_{ij} is the normalization factor given by

$$Z_{ij} = \sum_{k} \exp[\sigma_i(c,k) + \tau_j(c,k)].$$

Although the matrices $[\sigma_i(c,k)]$ and $[\tau_j(c,k)]$ in Equation (5) are obtained as the consequence of maximum conditional entropy, they can be understood rather intuitively. We can regard the matrix $[\sigma_i(c,k)]$ as the measure of the intrinsic ability of worker i. The (c,k)-th entry represents how likely worker i labels a randomly chosen item in class c as class k. Similarly, we can regard the matrix $[\tau_i(c,k)]$ as the measure of the intrinsic difficulty of item j. The (c,k)-th entry represents how likely item j in class c is labeled as class k by a randomly chosen worker. In what follows, we refer to $[\sigma_i(c,k)]$ as worker confusion matrices and $[\tau_i(c,k)]$ as item confusion matrices.

Minimum conditional entropy can be both intuitively and theoretically understood. Intuitively, it means that we believe that the observed labels are the least random given the true labels (Zhou et al., 2012). Theoretically, minimum conditional entropy can be understood as maximum likelihood. Substituting the labeling model in Equation (5) into the Lagrangian in Equation (4), we can obtain the dual form of the minimax program (1) as

$$\max_{\sigma,\tau,Q} \quad \sum_{j,c} Q(Y_j = c) \sum_{i} \log P(X_{ij} = x_{ij}|Y_j = c).$$

It is easy to see that, to be optimal, the true label distribution has to be deterministic. Consequently, the dual Lagrangian can be written as

$$\log \left\{ \prod_{j} \sum_{c} Q(Y_j = c) \prod_{i} P(X_{ij} = x_{ij}|Y_j = c) \right\},$$

which is nothing else but the log complete likelihood. In Section 4, we show how to reformulate the objective function in (1) to obtain probabilistic labels.

The main difference between the presented multiclass label aggregation method and the work of Zhou et al. (2012) is that the item constraints in the latter are formulated as

$$\sum_{i,c} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \ \forall j,k. \tag{6}$$

Such a formulation results in a one-dimensional confusion vector $[\tau_j(k)]$ representing the difficulty of item j instead of a confusion matrix. The connection between Equation (6) and (2b) is straightforward. Equation (6) can be recovered by summing Equation (2b) over all possible values of the true labels. So the constraints in Equation (6) are less restricted than the constraints in Equation (2b).

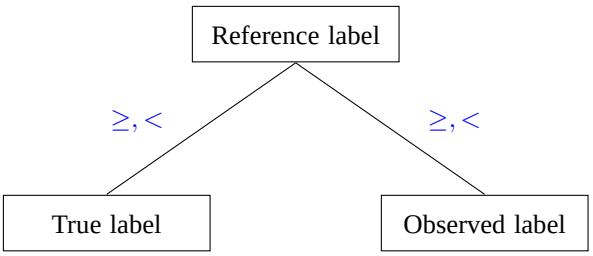

Figure 2. Indirect comparison between a true label and an observed label via comparing both to a reference label which varies through all possible values in a given ordinal label set.

3. Ordinal Minimax Conditional Entropy

In this section, we adapt the general multiclass label aggregation method developed in Section 2 to ordinal labels. Overall, we construct a different set of worker and item constraints to encode adjacency confusability observed in ordinal labeling that we discussed in Section 1. The formulation leads to an ordinal labeling model parameterized with structured worker and item confusion matrices.

Let us first introduce two symbols Δ and ∇ which take on arbitrary binary relations in $\mathcal{R} = \{\geq, <\}$. To estimate the true ordinal labels, we consider

$$\min_{Q} \max_{P} \; H(X|Y) \tag{7}$$

subject to the ordinal-based worker and item constraints

$$\sum_{c\Delta s} \sum_{k\nabla s} \sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \; \forall i, s, \tag{8a}$$

$$\sum_{c\Delta s} \sum_{k\nabla s} \sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = 0, \; \forall j, s, \tag{8b}$$

for all $\Delta, \nabla \in \mathcal{R}$, and the probability constraints in (3).

Let us explain the meaning of the constraints in Equation (8). To construct ordinal based constraints, the first issue that we have to address is how to compare two ordinal labels which are respectively observed label k and true label c in our scenario. For multiclass labels, as we have seen in Section 2, the comparison problem is trivial. We only need to check whether two given labels are the same or not. For ordinal labels, such a problem becomes tricky. Here, we suggest an indirect comparison between two ordinal labels by comparing both to a *reference label* s which varies through all possible values in a given ordinal label set (Figure 1). Then, for each chosen s, we partition the Cartesian product of the label set into four disjoint regions

$$\{(c,k)|c < s, k < s\}, \; \{(c,k)|c < s, k \geq s\},$$
$$\{(c,k)|c \geq s, k < s\}, \; \{(c,k)|c \geq s, k \geq s\}.$$

A partition example is shown in Table 1 where the given label set is $\{0, 1, 2, 3\}$. Equation (8a) defines a set of constraints for workers by summing Equation (2a) over each region. Similarly, Equation (8b) defines a set of constraints for items by summing Equation (2b) over each region.

From the above discussion, when there are more than two ordinal classes, the constraints in Equation (8) are less restricted than those in Equation (2). Consequently, the labeling model resulted from Equation (8) has fewer parameters, as we will see in Equation (11). In the case in which there are only two ordinal classes, those disjoint regions degenerate to single cells, and, thus, the two sets of constraints in Equation (8) and (2) are identical.

To understand the motivation underlying the constraints in Equation (8), let us write

$$\begin{aligned}\sum_{c\Delta s} \sum_{k\nabla s} \widehat{\phi}_{ij}(c,k) &= \sum_{c\Delta s} \sum_{k\nabla s} Q(Y_j = c)\mathbb{I}(x_{ij} = s) \\ &= \sum_{c\Delta s} Q(Y_j = c) \sum_{k\nabla s} \mathbb{I}(x_{ij} = s) \\ &= Q(Y_j \Delta s)\mathbb{I}(x_{ij} \nabla s). \end{aligned} \tag{9}$$

For example, when $\Delta = <$ and $\nabla = \geq$, the above equation becomes

$$\sum_{c<s} \sum_{k\geq s} \widehat{\phi}_{ij}(c,k) = Q(Y_j < s)\mathbb{I}(x_{ij} \geq s).$$

For a comparison between an observed label and a reference label, there are two possible outcomes: the observed label is larger or equal to the reference label; or the observed label is smaller than the reference label. Similarly, for a comparison between a true label and a reference label, there are also two possible outcomes. Putting together the above two comparisons, we have four possible outcomes in total. From Equation (9), the constraints in Equation (8a) enforce expected counts of all the four kinds of outcomes in the worker dimension to match their empirical counterparts. Symmetrically, the constraints in Equation (8b) enforce expected counts of all the four kinds of outcomes in the item dimension to match their empirical counterparts.

The Lagrangian of the maximization problem in (7) can be written as

$$L = H(X|Y) + L_\sigma + L_\tau + L_\lambda,$$

with

$$L_\sigma = \sum_{i,s} \sum_{\Delta,\nabla} \sigma_{is}^{\Delta,\nabla} \sum_{c\Delta s} \sum_{k\nabla s} \sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right],$$

$$L_\tau = \sum_{j,s} \sum_{\Delta,\nabla} \tau_{js}^{\Delta,\nabla} \sum_{c\Delta s} \sum_{k\nabla s} \sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right],$$

$$L_\lambda = \sum_{i,j,c} \lambda_{ijc} \left[\sum_{k} P(X_{ij} = k|Y_j = c) - 1\right],$$

where $\sigma_{is}^{\Delta,\nabla}, \tau_{js}^{\Delta,\nabla}$ and λ_{ijc} are the introduced Lagrange multipliers. By the similar process as in Section 2, we can

(a) Partitioning with $s = 1$

$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$
$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(2,0)$	$(2,1)$	$(2,2)$	$(2,3)$
$(3,0)$	$(3,1)$	$(3,2)$	$(3,3)$

(b) Partitioning with $s = 2$

$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$
$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(2,0)$	$(2,1)$	$(2,2)$	$(2,3)$
$(3,0)$	$(3,1)$	$(3,2)$	$(3,3)$

(c) Partitioning with $s = 3$

$(0,0)$	$(0,1)$	$(0,2)$	$(0,3)$
$(1,0)$	$(1,1)$	$(1,2)$	$(1,3)$
$(2,0)$	$(2,1)$	$(2,2)$	$(2,3)$
$(3,0)$	$(3,1)$	$(3,2)$	$(3,3)$

Table 1. Partitioning the Cartesian product of an ordinal label set $\{0, 1, 2, 3\}$. Each table shows a partition including four disjoint regions with respect to a possible reference label.

obtain a probabilistic ordinal labeling model

$$P(X_{ij} = k|Y_j = c) = \frac{1}{Z_{ij}} \exp[\sigma_i(c,k) + \tau_j(c,k)], \quad (10)$$

where

$$\sigma_i(c,k) = \sum_{s\geq 1} \sum_{\Delta,\nabla} \sigma_{is}^{\Delta,\nabla} \mathbb{I}(c\Delta s, k\nabla s), \quad (11a)$$

$$\tau_j(c,k) = \sum_{s\geq 1} \sum_{\Delta,\nabla} \tau_{js}^{\Delta,\nabla} \mathbb{I}(c\Delta s, k\nabla s), \quad (11b)$$

and Z_{ij} is the normalization factor obtained by summing the numerator over all possible labels.

The ordinal labeling model in Equation (10) is actually the same as the multiclass labeling model in Equation (5) except the worker and item confusion matrices in Equation (10) are now subtly structured through Equation (11). Consequently, whenever there are more than two classes, the ordinal labeling model has fewer parameters than the multiclass labeling model. In the case in which there are only two classes, the ordinal labeling model and the multiclass labeling model coincide.

4. Regularized Minimax Conditional Entropy

In this section, we develop regularized minimax conditional entropy to address two practical issues—preventing overfitting and generating probabilistic labels.

Let us first look at the overfitting issue. Given a finite number of observed labels, the empirical counts in Equations (2) or (8) may not exactly match their expected values. It is more likely that they fluctuate around their expected values. On the other hand, those fluctuations should not be too large. Hence, to be more realistic, we have to move from exact matching to approximate matching while penalizing large fluctuations as in (Zhou et al., 2012).

Now let us look at the issue of probabilistic labels. In practice, probabilistic labels are usually more helpful than deterministic labels. When the estimated label distribution for an item is close to uniform over multiple labels, we may want to either ask for more labels for the item or forward the item to an external expert. Unfortunately, the minimax conditional entropy methods in Sections 2 and 3 can only generate deterministic labels. To remedy this issue, as shown below, we put an extra entropy regularization over the unknown true label distributions.

Let us denote by

$$H(Y) = -\sum_{j,c} Q(Y_j = c) \log Q(Y_j = c).$$

To estimate the true labels, we consider

$$\min_Q \max_P \; H(X|Y) - H(Y) - \frac{1}{\alpha}\Omega(\xi) - \frac{1}{\beta}\Psi(\zeta) \quad (12)$$

subject to the relaxed worker and item constraints

$$\sum_{c\Delta s} \sum_{k\nabla s} \sum_{j} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = \xi_{is}^{\Delta,\nabla}, \forall i, s, \quad (13a)$$

$$\sum_{c\Delta s} \sum_{k\nabla s} \sum_{i} \left[\phi_{ij}(c,k) - \widehat{\phi}_{ij}(c,k)\right] = \zeta_{js}^{\Delta,\nabla}, \forall j, s, \quad (13b)$$

for all $\Delta, \nabla \in \mathcal{R}$, and the probability constraints in Equation (3).

The introduced slack variables $\xi_{is}^{\Delta,\nabla}$ and $\zeta_{js}^{\Delta,\nabla}$ in Equation (13) model the fluctuations, which are not restricted to be positive and could be rather arbitrary. However, when there are a sufficiently large number of observations, the fluctuations should be approximately normally distributed, due to the central limit theorem. Such observation motivates us to choose the regularization functions

$$\Omega(\xi) = \sum_{i,s} \sum_{\Delta,\nabla} \left(\xi_{is}^{\Delta,\nabla}\right)^2, \; \Psi(\zeta) = \sum_{j,s} \sum_{\Delta,\nabla} \left(\zeta_{js}^{\Delta,\nabla}\right)^2$$

to penalize large fluctuations. The introduced entropy term $H(Y)$ in (12) can be considered as penalizing a large deviation from uniform distribution.

Substituting the labeling model in Equation (10) into the Lagrangian of (12), we can obtain the dual form

$$\max_{\sigma,\tau,Q} \quad \sum_{j,c} Q(Y_j = c) \sum_i \log P(X_{ij} = x_{ij}|Y_j = c) + H(Y) - \alpha\Omega(\sigma) - \beta\Psi(\tau). \quad (14)$$

Algorithm 1 Regularized Minimax Conditional Entropy

input: $\{x_{ij}\}, \alpha, \beta$

initialize:

$$Q(Y_j = c) \propto \sum_i \mathbb{I}(x_{ij} = c) \tag{15}$$

repeat:

$$\{\sigma, \tau\} = \arg\min_{\sigma,\tau} \alpha\Omega(\sigma) + \beta\Psi(\tau) - \sum_{j,c} Q(Y_j = c) \sum_i \log P(X_{ij} = x_{ij} | Y_j = c) \tag{16a}$$

$$Q(Y_j = c) \propto \prod_i P(X_{ij} = x_{ij} | Y_j = c) \tag{16b}$$

output: Q

When $\alpha = 0$ and $\beta = 0$, the objective function in (14) turns out to be a lower bound of the log marginal likelihood

$$\begin{aligned}
&\log \left\{ \prod_j \sum_c \prod_i P(X_{ij} = x_{ij} | Y_j = c) \right\} \\
=&\log \left\{ \prod_j \sum_c \frac{Q(Y_j = c)}{Q(Y_j = c)} \prod_i P(X_{ij} = x_{ij} | Y_j = c) \right\} \\
\geq& \sum_{j,c} Q(Y_j = c) \sum_i \log P(X_{ij} = x_{ij} | Y_j = c) + H(Y).
\end{aligned}$$

The last step is based on Jensen's inequality. Maximizing marginal likelihood is supposed to be more appropriate than maximizing complete likelihood since only the observed labels essentially matter in our inference. The regularized form of multiclass minimax conditional entropy in Section 2 can be established in the same way.

5. Optimization and Model Selection

The dual problem of regularized minimax conditional entropy for either multiclass or ordinal labels is nonconvex. A stationary point can be obtained via coordinate ascent (Algorithm 1), which is essentially Expectation-Maximization (EM) (Dempster et al., 1977; Neal & Hinton, 1998). We first initialize the label estimate via aggregating votes in Equation (15). Then, in each iteration step, given the current estimate of labels, update the estimate of worker and item confusion matrices by solving the optimization problem in (16a); and, given the current estimate of worker and item confusion matrices, update the estimate of labels through the closed-form formula in (16b). The program in (16a) is strongly convex and smooth. So it can be solved with linear convergence rates (Nesterov, 2004). The closed-form formula in Equation (16b) is identical to applying Bayes' rule with a uniform prior.

Next we discuss how to select the regularization parameters α and β. If the true labels of a subset of items are known—such subsets are usually referred to as validation sets, we may choose (α, β) such that those known true labels can be best predicted. Otherwise, we suggest to choose (α, β) via k-fold likelihood-based cross-validation. Specifically, for each possible (α, β) chosen from a candidate set:

1. Randomly partition the set of observed labels into k equal-size subsamples;
2. Leave out a single subsample and use the remaining $k-1$ subsamples to estimate worker and item confusion matrices using Algorithm 1;
3. Compute the marginal likelihood of the observed labels in the left-out subsample using the estimated worker and item confusion matrices;
4. Repeat steps (2)−(3) till each subsample is left out once and only once;
5. Compute the average of the obtained marginal likelihoods over all k subsamples.

The (α, β) resulting in the largest average marginal likelihood is considered to be optimal. Finally, we run Algorithm 1 over the full dataset using the optimal (α, β). The cross-validation parameter k is typically set to 5 or 10.

To make the model selection process less time consuming, we would like to further suggest to set

$$\begin{aligned}
\alpha &= \gamma \times (\text{number of classes})^2, \\
\beta &= \frac{\text{number of labels per worker}}{\text{number of labels per item}} \times \alpha.
\end{aligned} \tag{17}$$

In our experiments, we choose γ from $\{1/4, 1/2, 1, 2, 4\}$. From our limited empirical studies, larger candidate sets for γ do not cause more gains. There are two observations that motivate us to consider using the square of the number of classes in Equation (17). First, the square of the number of classes has the same magnitude as the number of parameters in a confusion matrix. Second, label noise dramatically increases when the number of classes increases, requiring more than linearly scaled regularization.

6. Experiments

In this section, we report empirical results of our method on real crowdsourced data in comparison with state-of-the-art methods that aggregate multiclass or ordinal labels. Three error metrics are considered. Let us denote by y the true rating and $\widehat{y}$ the estimate. The error metrics are defined as: (1) L0 $= \mathbb{I}(y \neq \widehat{y})$; (2) L1 $= |y - \widehat{y}|$; and (3) L2 $= |y - \widehat{y}|^2$.

For convenience, in what follows, the regularized minimax conditional entropy method for multiclass labels is referred to as `entropy multiclass` or `entropy(M)`,

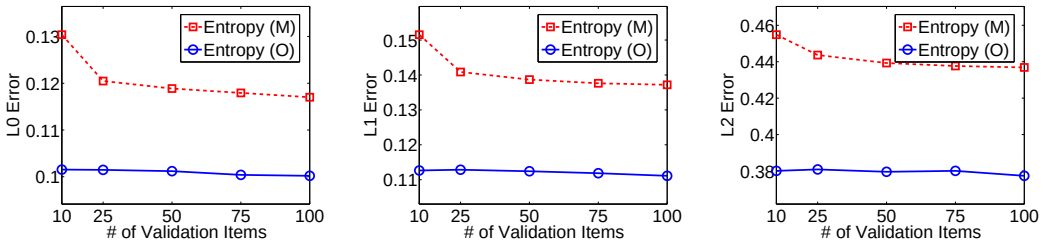

Figure 3. Error rates of the two entropy methods on the web dataset. The regularization parameter γ is chosen by using validation sets.

	Majority Vote	Dawid & Skene	Latent Trait	Entropy (M)	Entropy (O)
L0 Error	0.269	0.170	0.201	0.111	**0.104**
L1 Error	0.428	0.205	0.211	0.131	**0.118**
L2 Error	0.930	0.539	0.481	0.419	**0.384**

Table 2. Error rates of different methods on the web dataset. The regularization parameter γ for both the entropy methods is chosen via likelihood-based cross-validation without using any ground truth labels.

Probability Bin	$(0, 0.5)$	$(0.5, 0.6)$	$(0.6, 0.7)$	$(0.7, 0.8)$	$(0.8, 0.9)$	$(0.9, 1)$
# Items	173	291	292	313	406	1178
L0 Error	0.416	0.381	0.199	0.080	0.020	0.001
L1 Error	0.543	0.395	0.216	0.093	0.025	0.001
L2 Error	0.832	0.423	0.250	0.118	0.035	0.001

Table 3. Positive correlation between probabilistic labels and error rates. The results are from `entropy ordinal` on the web dataset. The labels estimated with larger probabilities are more likely to be correct. We have a similar observation for `entropy multiclass`.

and the regularized minimax conditional entropy method for ordinal labels is referred to as `entropy ordinal` or `entropy(O)`. Both are implemented by Algorithm 1.

Majority voting and the method of Dawid & Skene (1979) are considered as baselines. We also compare our method with latent trait analysis (Andrich, 1978; Master, 1982; Uebersax & Grove, 1993), which are the only ordinal label aggregation methods that we have seen in the literature. Roughly speaking, in such kind of scheme, each item is assumed to have a latent real-valued score, and each worker is assumed to have her personal class thresholds which characterize her class definitions. Given an item, an observed label from a worker is assumed to be generated through a probabilistic model which is a logistic (or normal) ogive (Lord & Novick, 1968) of the item score, the worker's class thresholds, and the measurement error parameters for the worker and item. In our empirical studies, we take an open source implementation of latent trait analysis by Mineiro (2011). We observe that the method of Zhou et al. (2012) performs almost the same as `entropy multiclass` so its results are not additionally reported.

Web search relevance rating. The web search relevance rating dataset contains 2665 query-URL pairs and 177 workers (Zhou et al., 2012). Give a query-URL pair, a worker is required to provide a rating to measure how the URL is relevant to the query. The rating scale is 5-level: perfect, excellent, good, fair, or bad. On average, each pair was labeled by 6 different workers, and each worker labeled 90 pairs. More than 10 workers labeled only one pair. The average L0 error rate of workers is 62.95%.

We first compare the two entropy methods with the parameter γ chosen by validation sets. We randomly select 1500 pairs to form a test set, and then select 10 to 100 pairs from the remaining pairs to form validation sets. The error rates on the test set are summarized in Figure 3. Each data point is obtained via averaging over 100 random sampling trials. `Entropy ordinal` outperforms `entropy multiclass` on all the three error metrics.

We then compare the entropy methods with the existing methods that we have discussed. All the methods except the entropy methods have no parameter to tune. It would be unfair if we require additional validation sets to tune the parameter γ in the entropy methods. Thus, we choose γ through 5-fold data likelihood based cross-validation. The error rates of different methods are summarized in Table 2. From the results, `entropy ordinal` outperforms `entropy multiclass` which outperforms all the others. The results also show that marginal likelihood based cross-validation works pretty well. It is completely comparable with validation sets based model selection.

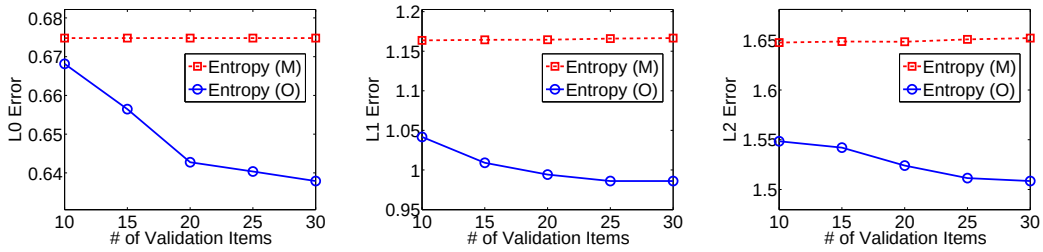

Figure 4. Error rates of the two entropy methods on the price dataset. The regularization parameter γ is chosen by using validation sets.

	Majority Vote	Dawid & Skene	Latent Trait	Entropy (M)	Entropy (O)
L0 Error	0.675	0.650	0.688	0.675	**0.613**
L1 Error	1.125	1.050	1.063	1.150	**0.975**
L2 Error	1.605	1.517	1.504	1.643	**1.492**

Table 4. Error rates of different methods on the price dataset. The regularization parameter γ for both entropy methods is chosen via likelihood based cross-validation without using any ground truth labels.

To investigate the correlation between probabilistic labels and error rates, we collect the obtained label probabilities into bins from $(0, 0.5)$ to $(0.9, 1)$ and check the error rates in each bin (Table 3). We observe that the labels estimated with larger probabilities are more likely to be correct.

Household item price estimation. The price dataset consists of 80 household items collected from stores such as Amazon and Costco. The prices of those items are estimated by 155 undergraduate students (Liu et al., 2013). There is a total of seven price bins: \$0−\$50, \$51−\$100, \$101−\$250, \$251−\$500, \$501−\$1000, \$1001−\$2000, and \$2001−\$5000. Given an item, a student has to decide which bin its price falls in. All the items are estimated by all the students, that is, we have a full data matrix. The average L0 error rate of students is 69.47%, compared to the L0 error rate of random guessing at 85.71%.

We first compare the two entropy methods with the parameter γ chosen through validation sets. We randomly select 50 items to form a test set, and then select 10 to 30 items from the remaining items to form validation sets. The error rates on the test set are summarized in Figure 4. Each data point is obtained via averaging over 100 random sampling trials. `Entropy ordinal` performs substantially better than `entropy multiclass` on all the three error metrics. We then compare the entropy methods with the state-of-the-art methods that we have discussed. The parameter γ in the entropy methods is chosen via 5-fold data likelihood based cross-validation. We summarize the error rates of different methods in Table 4. `Entropy ordinal` again performs the best among all the methods.

Although it is proposed as an ordinal approach, latent trait analysis performs even worse than the multiclass method by Dawid & Skene (1979) on both datasets in terms of L0 and L1 errors. It is perhaps partially because latent trait analysis violates the Main Principle for inference suggested by Vapnik (1998). As its intermediate step, latent trait analysis attempts to estimate a real-valued score for each item. This is a more ambitious problem than estimating ordinal labels. To estimate ordinal labels, we only have to know the ranges which real-valued scores fall in. The data that we have may be sufficient for estimating ordinal labels, but they may be insufficient for estimating continuous scores.

7. Conclusion and Discussion

We have presented a novel method for aggregating ordinal labels from a crowd of workers. The key component in our method is an ordinal labeling model in which worker ability and item difficulty are illustrated with structured confusion matrices. The matrix forms are uniquely determined through minimax conditional entropy subject to worker and item constraints which encode adjacency confusability observed in ordinal labeling—workers usually have difficulty distinguishing between two adjacent ordinal classes whereas distinguishing between two classes which are far away from each other is much easier. Empirical results on real crowdsourced data show that our method performs substantially better than existing methods.

Our minimax conditional entropy scheme is general and it can be extended to many other labeling tasks which involve structured labels, such as protein folding (Khatib et al., 2011), machine translation (Zaidan & Callison-Burch, 2011), and speech captioning (Murphy et al., 2013). To achieve these extensions, we need to formulate domain-specific worker and item constraints which may result in differently structured confusion matrices to parameterize the probabilistic labeling models for those tasks.

References

Andrich, D. A rating formulation for ordered response categories. *Psychometrika*, 43:561–73, 1978.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Chen, X., Lin, Q., and Zhou, D. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *Proceedings of the 30th International Conferences on Machine Learning*, 2013.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1220–1229, 2013.

Dawid, A. P. and Skene, A. M. Maximum likeihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society*, 28(1):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*, 39(1): 1–38, 1977.

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, 2011.

Karger, D. R., Oh, S., and Shah, D. Budget-optimal task allocation for reliable crowdsourcing systems. arXiv:1110.3564, 2013.

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., Baker, D., and Players, F. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47):18949–18953, 2011.

Liu, Q., Peng, J., and Ihler, A. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems 25*, pp. 701–709, 2012.

Liu, Q., Steyvers, M., and Ihler, A. Scoring workers in crowdsourcing: How many control questions are enough? In *Advances in Neural Information Processing Systems 26*, pp. 1914–1922, 2013.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Master, G. N. A Rasch model for partial credit scoring. *Psychometrika*, 47:149–174, 1982.

Mineiro, P. Ordered values and mechanical turk. http://www.machinedlearnings.com, 2011.

Murphy, M., Miller, C. D., Lasecki, W. S., and Bigham, J. P. Adaptive time windows for real-time crowd captioning. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 13–18, 2013.

Neal, R. M. and Hinton, G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I. (ed.), *Learning in Graphical Models*, pp. 355–368. Kluwer Academic, Dordrecht, MA, 1998.

Nesterov, Yu. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic, 2004.

Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pp. 321–333, Berkeley, CA, 1961.

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*, 11:1297–1322, 2010.

Tian, Y. and Zhu, J. Learning from crowds in the presence of schools of thought. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 226–234, 2012.

Uebersax, J. S. and Grove, W. M. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, 49:823–835, 1993.

Vapnik, V. N. *Statistical learning theory*. Wiley, NY, 1998.

Welinder, P., Branson, S., Belongie, S., and Perona, P. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, pp. 2424–2432, 2010.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pp. 2035–2043, 2009.

Zaidan, O. F. and Callison-Burch, C. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 1220–1229, 2011.

Zhou, D., Platt, J. C., Basu, S., and Mao, Y. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25*, pp. 2204–2212, 2012.