Crowdsourcing via Tensor Augmentation and Completion

Yao Zhou, Jingrui He
Arizona State University, Tempe, Arizona
yzhou174@asu.edu, jingrui.he@asu.edu

Abstract
Nowadays, the rapid proliferation of data makes it possible to build complex models for many real applications. Such models, however, usually require large amount of labeled data, and the labeling process can be both expensive and tedious for domain experts. To address this problem, researchers have resorted to crowdsourcing to collect labels from non-experts with much less cost. The key challenge here is how to infer the true labels from the large number of noisy labels provided by non-experts.

Different from most existing work on crowdsourcing, which ignore the structure information in the labeling data provided by non-experts, in this paper, we propose a novel structured approach based on tensor augmentation and completion. It uses tensor representation for the labeled data, augments it with a ground truth layer, and explores two methods to estimate the ground truth layer via low rank tensor completion. Experimental results on 6 real data sets demonstrate the superior performance of the proposed approach over state-of-the-art techniques.

1 Introduction
Recent years have seen explosive growth of data being collected from a variety of domains. Such unprecedented amount of data makes it possible to build complex models for prediction and inference. On the other hand, building such models requires accurate label information, the collection of which from domain experts is typically both expensive and tedious. Alternatively, crowdsourcing has been proposed to collect large amount of label information from non-experts, which is much less expensive [Kittur et al., 2008; Huberman et al., 2009]. However, due to the noisy nature of the labels provided by non-experts, a key challenge in crowdsourcing is how to infer the true labels from the large number of noisy labels.

To address this problem, a variety of techniques have been proposed in the past decades. Among others, the most straightforward method is majority voting, which is based on the assumption that all labels are equally reliable. However, this assumption may not hold in practice, and majority voting has been proven sub-optimal [Karger et al., 2011]. More recently, [Dawid et al., 1979] proposed an iterative algorithm based on Expectation Maximization (EM) to estimate worker quality and infer the item true label at the same time. Its performance is further improved by a variety of recent algorithms [Zhou et al., 2012; Liu et al., 2012b; Zhang et al., 2014; Raykar et al., 2010]. Section 2 provides a brief review of these algorithms.

In this paper, for the first time, we approach the crowdsourcing problem using tools and concepts from tensor augmentation and completion (TAC). Compared with existing techniques, we are able to effectively leverage the structured information in the labeled data. First of all, we represent the set of labels provided by non-experts (workers) as a three-way tensor, and then augment it with an extra tensor slice named the ground truth layer. Second, to infer the true labels in the ground truth layer, we leverage the low rank property of the augmented tensor, and introduce two optimization problems named PG-TAC (prior guided) and RS-TAC (relaxed simplex). Finally, we propose various algorithms for solving these problems using block coordinate descent. Experimental results on 6 real data sets demonstrate the effectiveness of the proposed methods in both binary and multi-class labeling tasks, outperforming several state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we briefly review existing working on crowdsourcing and tensor completion. Then in Sections 3 and 4, we present our proposed model and optimization algorithms, followed by experimental results on both synthetic and real data sets in Section 5. Finally, we conclude the paper in Section 6.

2 Related work
In this section, we briefly review the related work on crowdsourcing and missing value completion.

One of the earliest works on crowdsourcing is [Dawid et al., 1979], which proposes an iterative algorithm based on Expectation Maximization (EM) to estimate worker quality and infer the item true label at the same time. They assume each worker is associated with a probabilistic confusion matrix for item labeling. Each diagonal entry of the confusion matrix represents the labeling accuracy in each labeling class and the off-diagonal entries of each row represent the mislabeling probabilities. However their model implicitly ignores the item variations in the same class and assumes all items, which
have the same true labels, will have the same degree of difficulties. That assumption does not hold in many real-world situations, then [Zhou et al., 2012] improved upon their work by proposing a minimax entropy principle to infer the true label, the labeling difficulty of the item, and the quality of the worker. Besides the worker quality, their method assumes that each item has its own intrinsic difficulty of being mislabeled. When the item difficulty is ignored, their model is reduced to the EM method proposed by [Dawid et al., 1979]. Another flaw of the EM method, proposed by [Dawid et al., 1979], is that their likelihood function is nonconvex, therefore its performance is initialization sensitive because the EM iterations can possibly converge at a local optimum. To address this issue, [Zhang et al., 2014] proposed a two-staged algorithm in which the initial worker confusion matrix is estimated using the spectral method, and then their algorithm turns to EM iterations. Their model has been proved to be able to achieve the minimax rates of convergence up to a logarithmic factor. [Liu et al., 2012b] also proposed a graphical model that performs variational inference method using belief propagation and mean field (MF) algorithms. Another probabilistic model named GLAD, which can simultaneously estimate the ground truth, item difficulty and worker ability, has also been proposed by [Whitehill et al., 2009]. However the GLAD model can only work on binary tasks and it does not model the worker bias, its performance can get worse when the bias variation of different workers is high [Welinder et al., 2010]. Later on, the GLAD model is generalized to work multi-class labelling tasks by [Mineiro, 2011].

Missing values are commonly seen in many real-world applications, such as recommendation systems, which motivates the study of missing value completion. This problem is initially proposed by [Candès and Recht, 2009] in order to recover the missing entries in matrices. Theoretically it has already been proved that most low rank matrices can be recovered from a small fraction of entries by formatting a rank minimization problem. However this rank minimization problem is NP-hard and non-convex, which results in the optimization problem that uses trace norm as the objective. This is addressed and mentioned in a variety of works [Candès and Tao, 2010; Recht, 2011]. The advantage is that trace norm is the tightest convex envelop for matrix rank. In many practical situations, higher dimensional data is more desired and it requires to generalize the completion methods on tensors. Similar to matrix completion, it is straightforward to think of formulating the tensor completion as a rank minimization problem. However, unlike the matrix rank, there is no direct algorithm that can decide the rank of a tensor [Kolda and Bader, 2009]. To overcome this issue, similarly, [Liu et al., 2012a] proposed to approximate the rank minimization problem as a trace norm minimization problem. They introduce one type of the definition for tensor trace norm, while there exists many other definitions [Gandy et al., 2011]. [Liu et al., 2012a] also relaxes the objective function so that the optimization problem becomes convex. Their final low rank tensor completion method (LRTC) shows the broad capability to recover data in various format. Meanwhile there are many other heuristic methods [Xu et al., 2013] that can be applied to do tensor completions by employing tensor decom-

3 Problem formulation
3.1 Notation
In this article, we use calligraphic letters, such as $\mathcal{X}$, to denote tensors. We use upper case letters, such as $M$, to denote matrices. Vectors and scalars are denoted by the bold lower case letters and a lower case letters such as $x$ and $x$. A $n$-way tensor is denoted as $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times \ldots \times N_n}$. The $(i,j,k)$th element of a three-way tensor $\mathcal{X}$ is represented by $X_{ijk}$. A slice of a three-way tensor $\mathcal{X}$ is denoted as $X_{i\cdot:j\cdot:k}$ or $X_{i\cdot}$.

A fiber of a three-way tensor is denoted as $X_{j\cdot:k}$ or $X_{i\cdot:j\cdot}$.

The norm of a tensor is analogous to the matrix Frobenius norm: $||\mathcal{X}\||_F = (\sum_{i,j,k} |X_{ijk}|^2)^{1/2}$. The trace norm of a matrix $M$ is defined as: $||M||_T = \sum_{i} \sigma_i(M)$ and $\sigma_i(M)$ denotes the $i$th singular value in descending order. Let $\Omega$ denote the index set of a tensor, and $|\Omega|$ denote the cardinality of $\Omega$. One important operation of a tensor $\mathcal{X}$ is called matricization or unfold, which reorders a $n$-way tensor into a matrix. We denote $X_{(k)}$ as the output of unfold operation along the $k$-th dimension of a tensor $\mathcal{X}$, i.e., $X_{(k)} = \text{unfold}_k(\mathcal{X})$. Similarly, the $\text{fold}_k(X_{(k)})$ is the inverse operation of unfold and it returns the tensor $\mathcal{X}$. The details of operations fold and unfold can be found at [Kolda and Bader, 2009].

3.2 Tensor augmentation and completion
We propose to reorganize the worker labels from crowdsourcing as a three-way label tensor $T^0 \in \mathbb{R}^{N_w \times N_i \times N_c}$ and an index set $\Omega$. Here $N_w$, $N_i$ and $N_c$ are denoted as number of the workers, number of the items and number of the classes respectively. Each worker gives each item either exactly one label or no label, then label tensor $T^0$ and index set $\Omega$ are built as follows: If a worker $i$ has labeled an item $j$ with label $k$, the corresponding fiber $T^0_{ijk}$ is initialized with an unit vector, which has value of 1 in $k$th entry and value of 0’s in the rest. Meanwhile the corresponding index triplets of fiber $T^0_{ijk}$ are added into the index set $\Omega$. However a worker does not necessarily have to label all items. If worker $i$ does not label item $j$, fiber $T^0_{ijk}$ is initialized with a zero vector and $\Omega$ remains unchanged. If that is the case, the label tensor $T^0$ will have missing entries. In our approach, we propose to augment the label tensor with an extra tensor slice of size $N_i \times N_c$, called the ground truth layer, on the worker dimension. All entries of the ground truth layer are assumed to be missing and our objective is to infer the true labels of items.

Recall that the common approach for tensor completion is to minimize the matrix trace norm by solving the following convex optimization problem [Candès and Recht, 2009].

$$\min_{\mathcal{X}} : ||\mathcal{X}||_T, \quad s.t.: \mathcal{X}_{\Omega} = M_{\Omega}$$

(1)

where $X$ and $M$ are matrices of the same size. $\Omega$ is the index set of matrix $M$, and $X$ is the matrix such that its rank should be minimized while completing procedure. The entries that
do not belong to $\Omega$ are missing. For tensor completion, [Liu et al., 2012a] followed the same formulation with the following trace norm definition of an $n$-way tensor $\mathbf{X}$:

$$||\mathbf{X}||_* = \sum_{l=1}^{n} \alpha_l ||\mathbf{X}(l)||_*$$

$$s.t. : \sum_{l=1}^{n} \alpha_l = 1, \alpha_l \geq 0, l = 1, ..., n$$

where $\alpha_l, l = 1, ..., n$ are pre-defined scalars of tensor trace norm. Analogous to the matrix completion formulation, the tensor completion problem can be written as follows:

$$\min_{\mathbf{X}} : \sum_{l=1}^{n} \alpha_l ||\mathbf{X}(l)||_* , \quad s.t. : \mathbf{X}_{\Omega} = \mathbf{T}_{\Omega}$$

In the formulation, $\mathbf{X}$ is the target tensor that needs to be completed. However, the unfolded matrices $\mathbf{X}(l), l = 1, ..., n$, are not independent with each other. In order to split them and solve them independently, same number of intermediate matrices $M_l, l = 1, ..., n$ are introduced in this problem. Then this optimization problem can be relaxed and formulated as:

$$\min_{\mathbf{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathbf{X}(l) - M_l||_F^2, \quad s.t. : \mathbf{X}_{\Omega} = \mathbf{T}_{\Omega}$$

We propose to formulate the crowdsourcing problem as an augmented tensor completion problem with certain regularization on the ground truth layer. Since our task only requires a three-way tensor, from now on, without other specifications, all tensors in our equations have an order of three, namely $n = 3$. Given the augmented label tensor $\mathbf{T} \in \mathbb{R}^{(N_w+1) \times N_i \times N_c}$ and index set $\Omega$, our optimization problem becomes:

$$\min_{\mathbf{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathbf{X}(l) - M_l||_F^2 + R(\mathbf{X}_{\Omega})$$

$$s.t. : \mathbf{X}_{\Omega} = \mathbf{T}_{\Omega}$$

Here $i_g$ denotes the ground truth layer index on the worker dimension of the tensor.

### 3.3 Two formulations for inferring the ground truth layer

In the formulations, we propose to regularize the ground truth layer in two different ways: One is to regularize the discrepancy between the ground truth layer of the tensor and a given prior statistics of the items. Another one is to constraint each tensor fiber of the ground truth layer in a simplex. Under these two regularizations, the inferred ground truth layer can have distinct interpretations.

**Prior guided ground truth inference**

The objective function of the first formulation has a regularization term w.r.t. the discrepancy between ground truth layer and prior statistics matrix, and the regularization is parameterized with a positive value $\gamma$. Our key motivation of this regularization is to updating the item labels by combining the prior statistics and tensor structure information. The formulation becomes:

$$\min_{\mathbf{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathbf{X}(l) - M_l||_F^2 + \frac{\gamma}{2} ||\mathbf{X}_{\Omega} - S||_F^2$$

$$s.t. : \mathbf{X}_{\Omega} = \mathbf{T}_{\Omega}$$

Here $S \in \mathbb{R}^{N_w \times N_c}$ represents the item prior statistics matrix of the tensor $\mathbf{T}$.

**Relaxed simplex ground truth inference**

The second formulation has regularization terms w.r.t. the tensor fibers of the ground truth layer. Originally each item fiber $\mathbf{X}_{i_gj}$ is posed to constraint in a simplex. However the amount of labels collected for each item is usually limited to a small number in empirical experiment. It is likely that the labels of these items fluctuate around their expected values. In order to prevent overfitting, we formulate our objective with relaxed simplex constraint and penalize the large fluctuations according to the value of parameter $\gamma$:

$$\min_{\mathbf{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathbf{X}(l) - M_l||_F^2 + \frac{\gamma}{2} \sum_{j=1}^{N_c} \xi_j^2$$

$$s.t. : \sum_{k=1}^{N_c} \mathbf{X}_{ijk} = 1 = \xi_j, i = i_g, \forall j = 1, ..., N_i$$

$$\mathbf{X}_{\Omega} = \mathbf{T}_{\Omega}$$

### 4 Algorithm

All the terms in the objective function are convex, therefore we can employ the block coordinate descent (BCD) for the optimization problems (6) and (7). BCD is guaranteed to converge [Tseng, 2001] and is computational easier and cheaper than the batch update. Then we apply the coordinate descent to optimize one target variable while fixing others. In our case, we have four blocks: $\mathbf{X}$, $M_1$, $M_2$ and $M_3$, because the observed tensor has only three dimensions: the worker, the item and the label. There are two major iteration steps in BCD: First iteration updates one intermediate matrix $M_l$ while fixing the other intermediate matrices and the tensor $\mathbf{X}$; Second iteration updates the tensor and fixing all intermediate matrices.

#### 4.1 Updating $M_l$

Under certain simplification, the optimization problem of first BCD iteration becomes:

$$\min_{M_l} : \frac{\alpha_l}{\beta_l} ||M_l||_* + \frac{1}{2} ||\mathbf{X}(l) - M_l||_F^2$$

The close-form solution of this problem has been given by [Cai et al., 2010] as $D_\tau(\mathbf{X}(l)) = U \Sigma \tau V^T$. We first compute singular value decomposition of matrix $\mathbf{X}(l) = U \Sigma V^T$, then replace $\Sigma$ with its shrinkage version: $\Sigma_\tau = \text{diag}(|\sigma_i - \tau+|)$. Here $a_+ = \max(a, 0)$ and $\tau$ is the threshold of shrinkage SVD. No matter under prior guided formulation or relaxed simplex formulation, both problems will have problem (8) as the sub-problem in their BCD iterations.
4.2 Updating $\mathbf{X}$

Prior guided formulation:

With intermediate matrices $M_1$, $M_2$ and $M_3$ fixed in this iteration, the optimization problem becomes:

$$\min_{\mathbf{X}} \sum_{l=1}^{n} \frac{\beta_l}{2} ||x_l - M_l||^2_F + \frac{\gamma}{2} ||x_{i_g} - S||^2_F$$

s.t. : $\mathbf{X}_\Omega = T_\Omega$

(9)

This problem is convex and the objective can be rewritten in elementary manner and then the Lagrangian of the optimization problem is given as:

$$L = \sum_{l=1}^{n} \frac{\beta_l}{2} \sum_{i,j,k} (X_{ijk} - (fold_i(M_l))_{ijk})^2$$

$$+ \frac{\gamma}{2} ||x_{i_g} - S||^2_F + \sum_{(i,j,k) \in \Omega} \lambda_{ijk}(\mathbf{X}_{ijk} - T_{ijk})$$

(10)

Elements of tensor $\mathbf{X}$ can be divided into three sets. First set $C_1$ has its elements belong to the index set: $(i, j, k) \in \Omega$. The elements of the second set $C_2$ neither belong to the index set nor the ground truth layer: $(i, j, k) \notin \Omega$ and $i \neq i_g$. The elements of the third set $C_3$ do not belong to the index set but belong to the ground truth layer: $(i, j, k) \notin \Omega$ and $i = i_g$. The elements in set $C_1$ do not appear in the second term of the Lagrangian. Easily we know that the solution is:

$$X_{ijk} = T_{ijk}$$

(11)

The elements in set $C_2$ do not appear in the second and third terms of the Lagrangian. We take the derivative of the Lagrangian w.r.t. $X_{ijk}$ and set it to 0, then we get:

$$X_{ijk} = \left(\sum_{l=1}^{n} \frac{\beta_l}{l} fold_l(M_l) \right)_{ijk} \left(\sum_{l=1}^{n} \frac{\beta_l}{l} \right)_{ijk}$$

(12)

The elements in set $C_3$ do not appear in the third term of the Lagrangian. We take the derivative of the Lagrangian w.r.t. $X_{ijk}$ and set it to 0, then we get:

$$X_{ijk} = \left(\sum_{l=1}^{n} \frac{\beta_l}{l} fold_l(M_l) \right)_{ijk} \left(\sum_{l=1}^{n} \frac{\beta_l}{l} \right)_{ijk} + \frac{\gamma S}{\sum_{l=1}^{n} \beta_l + \gamma}$$

(13)

Relaxed simplex formulation:

The intermediate matrices $M_1$, $M_2$ and $M_3$ are fixed, the optimization problem becomes:

$$\min_{\mathbf{X}} \sum_{l=1}^{n} \frac{\beta_l}{2} ||x_l - M_l||^2_F + \frac{\gamma}{2} \sum_{j=1}^{N_i} \xi_j^2$$

s.t. : $\sum_{k=1}^{N_i} X_{ijk} - 1 = \xi_j, i = i_g, \forall j = 1, ..., N_i$

$$\mathbf{X}_\Omega = T_\Omega$$

(14)

Similarly, we rewrite the objective element wise, and the Lagrangian of the optimization problem becomes:

$$L = \sum_{l=1}^{n} \frac{\beta_l}{2} \sum_{i,j,k} (x_{ijk} - (fold_i(M_l))_{ijk})^2$$

$$+ \frac{\gamma}{2} \sum_{j=1}^{N_i} \xi_j^2$$

$$+ \sum_{j=1}^{N_i} \lambda_j (x_{ijk} - T_{ijk})$$

(15)

Similar as the prior guided formulation, here the elements of tensor $\mathbf{X}$ are also divided into three sets $C_1, C_2$ and $C_3$. The solutions for the elements in sets $C_1, C_2$ stay the same as shown in equations (11) and (12). The elements in set $C_3$ do not appear in the third term of Lagrangian. We take the derivative of the Lagrangian w.r.t. $X_{ijk}$ and $\xi_j$ and set them to 0, then we get:

$$X_{ijk} = \frac{\sum_{l=1}^{n} \beta_l (fold_i(M_l))_{ijk} - \tau_j}{\sum_{l=1}^{n} \beta_l}$$

$$\xi_j = \frac{\tau_j}{\gamma}$$

(16)

(17)

Substituting Equations (16) and (17) into the relaxed constraint $\sum_{k=1}^{N_c} X_{ijk} - 1 = \xi_j$, we get:

$$\tau_j = \frac{\sum_{l=1}^{n} \beta_l (\sum_{k=1}^{N_c} fold_i(M_l))_{ijk} - 1}{N_c + \gamma \sum_{l=1}^{n} \beta_l}$$

(18)

Substituting Equation (18) in Equation (16), we get:

$$X_{ijk} = \frac{\sum_{l=1}^{n} \beta_l fold_i(M_l)_{ijk}}{\sum_{l=1}^{n} \beta_l}$$

$$+ \frac{\gamma \sum_{l=1}^{n} \beta_l (1 - \sum_{k=1}^{N_c} (fold_i(M_l))_{ijk})}{\gamma N_c + \sum_{l=1}^{n} \beta_l \sum_{l=1}^{n} \beta_l}$$

(19)

Our proposed PG-TAC method is described in Algorithm 1. The algorithm of RS-TAC is omitted due to space limit.

5 Experiments

In this section, we report the results of our proposed methods on four groups of synthetic data sets. The purpose of this is to study the behavior of our methods under various data set configurations. Moreover, we compare our methods with a variety of state-of-the-art algorithms on six real data sets.

5.1 Synthetic data

Generation of synthetic data sets is based on four parameters: number of worker $N_w$, number of items $N_i$, number of classes $N_c$ and probability of no labels $q$. Given $N_c$ and $N_i$, the true labels of these items are sampled from a multinomial distribution with probabilities $p_1, p_2, ..., p_{N_c}$. In order to have balanced data, these probabilities should be the same. However we add random noise to the probabilities without breaking the rule of the sum being to 1. Now, the data set is unbalanced and is more analogous to a real data set. Then for each worker, we generate a $N_c \times N_i$ worker quality confusion matrix as follows: the diagonal entries are independently and
5.2 Real data

Real data are very small or sets are not sufficiently labeled, for instance, if almost the same performance. We also observe when the data performance. In many configurations, RS-TAC and NC-TAC have the lowest error rate on all configurations. The RS-TAC method does not necessarily improve the performance. In some configurations, PG-TAC achieves the lowest error rate on all configurations. Under each configuration, other data set parameters remain consistent with the initial configuration.

For Dog data set, the unqualified workers, who have only labeled a small amount of images, are removed. For web data set, 12 items have been removed due to lack of true labels. For age data set, data has been discretized into 7 bins: [0, 9], [10, 19], [20, 29], [30, 39], [40, 49], [50, 59], [60, 100].

5.3 Methods

In our experiment, we employed eight methods for the purpose of comparison: Majority Voting (MV) is the most straightforward method to implement and we use it as one of our baseline methods. Dawid-Skene Expectation Maximization (DS-EM), proposed by [Dawid et al., 1979], is a generative model which jointly infers the item true labels and worker qualities. Dawid-Skene Mean Field (DS-MF) employs variational inference using mean field method and this model is proposed by [Liu et al., 2012]. Generative model of Labels, Abilities and Difficulties (GLAD), proposed by [Whitehill et al., 2009], is a probabilistic framework that can simultaneously infer worker quality, item difficulty and item true labels. Here we use its variant, implemented by [Mineiro,
layer is initialized using histogram. If we do not have any of the augmented tensor will not increase if the ground truth linear combination of label tensor slices, therefore the rank of worker labels. Normalized histogram is the mean value of the tensor and normalized by using the open source implementation provided by Zhou et al., 2011], which can work on multi-class data. Minimax Conditional Entropy (MMCE) uses the minimax entropy principle [Zhou et al., 2012] to infer items ground truth from noisy labels. When the item difficult is ignored, the MMCE model is reduced to DS-EM method. NC-TAC is another simple baseline of our proposed method without the constraint on ground truth layer. PG-TAC employs an tensor slice as its prior statistics. When regularization parameter $\gamma$ is very small, PG-TAC is approximately equal to NC-TAC; when $\gamma$ is sufficiently large, PG-TAC reduces to its prior statistics. The objective of RS-TAC has a regularization term, which is parameterized by $\gamma$, to control the strength of relaxation on ground truth layer.

### 5.4 Parameter selection

PG-TAC and RS-TAC both have three parameters in objective: $\alpha_l, \beta_l$ and $\gamma$. Here $l = 1, ..., 3$ and 3 is the mode of the tensor. The values of $\alpha_l$ are assigned with value of 1/3 and we let $\delta_l = \frac{\alpha_l}{\sum_l \alpha_l}$. Given $\delta_l$, the value of $\beta_l$ is also determined. Therefore it is straightforward to verify that we only need to tune $\delta_l$ in BCD iterations no matter it is in the step of computing $M_l$ or in the step of computing $X^\gamma$. For simplicity, we let all $\delta_l$ be the same for all three modes. Eventually we can apply the grid search on two regularization parameters $\delta_l$ and $\gamma$, and the procedure is described as follows: all data sets we used are publicly available online and they all come with ground truth labels. We run our proposed algorithms on a 2-D grid parameter space. For each possible parameter pair on the searching grid, a subset of worker labels is randomly chosen from current data set without replacement. In practice, we empirically choose 90 percent of worker labels as a subset, run our methods, and evaluate the performance. Then we repeat the same procedure ten times for each possible parameter pair on the grid. Eventually the regularization parameter pair is chosen as the one that have lowest average error rate.

### 5.5 Implementation details

The results of MV, DS-EM and MMCE methods are verified by using the open source implementation provided by Zhou et al., 2015]. Our PG-TAC method uses the DS-EM as prior statistic and the ground truth layer is initialized using histogram of worker labels. In our empirical studies, we have tried initializing ground truth layer with majority voting of worker labels, mean value of the tensor and normalized histogram of worker labels. Normalized histogram is the linear combination of label tensor slices, therefore the rank of the augmented tensor will not increase if the ground truth layer is initialized using histogram. If we do not have any information about the ground truth layer, there is no hope to recover the unknown ground truth layer with meaningful returning values. This has been verified by initialize ground truth layer with all 0’s and the final completed values on it are meaningless. The ground truth layer of RS-TAC method is initialized with DS-EM. We use $||X - T||_F / ||T||_F$ as the stopping criteria and it is set to $10^{-5}$. The final label prediction is performed as follows: in each fiber $X^\gamma_{i,j}$ of the completed ground truth layer, the entry with larger values are more likely to be correctly predicted.

### 5.6 Results

Table 2 summarizes the error rates of various methods on six real data sets. For fairly comparison, all methods have been fed with the same format of input data. Our proposed methods PG-TAC and RS-TAC have consistently lower error rate than other state-of-the-art methods in most data sets. The RS-TAC method has the best performance on Age data set, which is the most difficult one we employed. Among all other data sets, the RS-TAC method has similar performance as NC-TAC. We observe the PG-TAC has outperformed all state-of-the-arts methods in most real data sets. The performance shown by PS-TAC is within our anticipation because PG-TAC combines the prior information and structural information inferred from tensor. From Equation (13), we know that inferred layer is actually the linear combination of prior statistics and NC-TAC. The only exception is on Age data set, which has severely unbalanced label distributions. Interestingly, DS-EM has the worst performance on Age data set among all methods. Even though the prior statistics is severely biased, PG-TAC still can achieve competitive results.

### 6 Conclusion

In this paper, we propose two novel methods (PG-TAC and RS-TAC) to infer the true labels of items in both binary and multi-class crowdsourcing settings. These methods capture the structure information in the data by representing the noisy labels provided by workers with tensors. Furthermore, we propose to augment the data tensor with an extra ground truth layer, and explore various tensor completion techniques to infer the true labels in the ground truth layer. Our experiment results on 6 real data set demonstrate that our proposed methods outperform state-of-the-art techniques.

### Acknowledgements

The research was partially supported by NSF research grant IIS-1552654; and by an IBM Faculty Award. The views and
conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the U.S. Government.

References


