Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

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 crowdsoucing 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. Empirical 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 decomposition and unfolded matrix factorization. However the theoretical guarantee of these heuristic methods is still an open question and the comparison experiment results from [Liu *et al.*, 2012a] show that LRTC has a more stable performance on both synthetic data and real-world data.

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 $\mathbf{x}$ 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 $\mathcal{X}_{ijk}$. A slice of a three-way tensor $\mathcal{X}$ is denoted as $\mathcal{X}_{i::}, \mathcal{X}_{:j:}$ or $\mathcal{X}_{::k}$. A fiber of a three-way tensor is denoted as $\mathcal{X}_{:jk}, \mathcal{X}_{i:k}$ or $\mathcal{X}_{ij:}$. The norm of a tensor is analogous to the matrix Frobenius norm: $||\mathcal{X}||_F = (\sum_{i,j,k} |\mathcal{X}_{ijk}|^2)^{1/2}$. The trace norm of a matrix M is defined as: $||M||_* = \sum_i \sigma_i(M)$ and $\sigma_i(M)$ denotes the i^{th} singular value in descending order. Let Ω denote the index set of a tensor, and $|\Omega|$ denote the cardinality of Ω. One important operation of a tensor $\mathcal{X}$ is called *matricization* or *unfold*, which reorders a n-way tensor into a matrix. We denote $\mathcal{X}_{(k)}$ as the output of *unfold* operation along the k-th dimension of a tensor $\mathcal{X}$, i.e., $\mathcal{X}_{(k)} = \mathit{unfold}_k(\mathcal{X})$. Similarly, the $\mathit{fold}_k(\mathcal{X}_{(k)})$ 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 $\mathcal{T}^0 \in \mathbb{R}^{N_w \times N_i \times N_c}$ and an index set Ω. 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 $\mathcal{T}^0$ and index set Ω are built as follows: If a worker i has labeled an item j with label k, the corresponding fiber $\mathcal{T}^0_{ij:}$ is initialized with an unit vector, which has value of 1 in kth entry and value of 0's in the rest. Meanwhile the corresponding index triplets of fiber $\mathcal{T}^0_{ij:}$ are added into the index set Ω. However a worker does not necessarily have to label all items. If worker i does not label item j, fiber $\mathcal{T}^0_{ij:}$ is initialized with a zero vector and Ω remains unchanged. If that is the case, the label tensor $\mathcal{T}^0$ 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 matrix completion is to minimize the matrix trace norm by solving the following convex optimization problem [Candès and Recht, 2009].

$$\min_X : ||X||_*, \quad s.t. : X_\Omega = M_\Omega \tag{1}$$

where X and M are matrices of the same size. Ω 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 Ω are missing. For tensor completion, [Liu *et al.*, 2012a] followed the same formulation with the following trace norm definition of an n-way tensor $\mathcal{X}$:

$$||\mathcal{X}||_* = \sum_{l=1}^{n} \alpha_l ||\mathcal{X}_{(l)}||_* \\ s.t. : \sum_{l=1}^{n} \alpha_l = 1, \alpha_l \geq 0,\, l = 1, ..., n \tag{2}$$

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:

$$\min_{\mathcal{X}} : \sum_{l=1}^{n} \alpha_l ||\mathcal{X}_{(l)}||_*, \quad s.t. : \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{3}$$

In the formulation, $\mathcal{X}$ is the target tensor that needs to be completed. However, the unfolded matrices $\mathcal{X}_{(l)}, l = 1, ..., n$, 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:

$$\min_{\mathcal{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathcal{X}_{(l)} - M_l||_F^2, \\ s.t. : \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{4}$$

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 $\mathcal{T} \in \mathbb{R}^{(N_w+1) \times N_i \times N_c}$ and index set Ω, our optimization problem becomes:

$$\min_{\mathcal{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathcal{X}_{(l)} - M_l||_F^2 + R(\mathcal{X}_{i_g::}) \\ s.t. : \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{5}$$

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 the 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 γ. Our key motivation of this regularization is to updating the item labels by combining the prior statistics and tensor structure information. The formulation becomes:

$$\min_{\mathcal{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathcal{X}_{(l)} - M_l||_F^2 + \frac{\gamma}{2} ||\mathcal{X}_{i_g::} - S||_F^2 \\ s.t. : \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{6}$$

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

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 $\mathcal{X}_{i_g j:}$ is posed to be 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 γ:

$$\min_{\mathcal{X}, M_l} : \sum_{l=1}^{n} \alpha_l ||M_l||_* + \frac{\beta_l}{2} ||\mathcal{X}_{(l)} - M_l||_F^2 + \frac{\gamma}{2} \sum_{j=1}^{N_i} \xi_j^2 \\ s.t. : \sum_{k=1}^{N_c} \mathcal{X}_{ijk} - 1 = \xi_j, i = i_g, \forall j = 1, ..., N_i \\ \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{7}$$

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: $\mathcal{X}, M_1, M_2$ 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 $\mathcal{X}$; 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:

$$\min_{M_l} : \frac{\alpha_l}{\beta_l} ||M_l||_* + \frac{1}{2} ||\mathcal{X}_{(l)} - M_l||_F^2 \tag{8}$$

The close-form solution of this problem has been given by [Cai *et al.*, 2010] as $D_\tau(\mathcal{X}_{(l)}) = U \Sigma_\tau V^T$. We first compute singular value decomposition of matrix $\mathcal{X}_{(l)} = U \Sigma V^T$, then replace Σ with its shrinkage version: $\Sigma_\tau = diag(\{\sigma_i - \tau\}_+)$. Here $a_+ = \max(a, 0)$ and τ 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 $\mathcal{X}$

Prior guided formulation:

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

$$\begin{aligned} &\min_{\mathcal{X}} : \sum_{l=1}^{n} \frac{\beta_l}{2} ||\mathcal{X}_{(l)} - M_l||_F^2 + \frac{\gamma}{2} ||\mathcal{X}_{i_g::} - S||_F^2 \\ &s.t. : \mathcal{X}_\Omega = \mathcal{T}_\Omega \end{aligned} \tag{9}$$

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

$$\begin{aligned} L = &\sum_{l=1}^{n} \frac{\beta_l}{2} \sum_{i,j,k} (\mathcal{X}_{ijk} - (fold_l(M_l))_{ijk})^2 \\ &+ \frac{\gamma}{2} ||\mathcal{X}_{i_g::} - S||_F^2 + \sum_{(i,j,k)\in\Omega}^{|\Omega|} \lambda_{ijk}(\mathcal{X}_{ijk} - \mathcal{T}_{ijk}) \end{aligned} \tag{10}$$

Elements of tensor $\mathcal{X}$ can be divided into three sets. First set $\mathcal{C}_1$ has its elements belong to the index set: $(i, j, k) \in \Omega$. The elements of the second set $\mathcal{C}_2$ 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 $\mathcal{C}_3$ 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 $\mathcal{C}_1$ do not appear in the second term of the Lagrangian. Easily we know that the solution is:

$$\mathcal{X}_{ijk} = \mathcal{T}_{ijk} \tag{11}$$

The elements in set $\mathcal{C}_2$ do not appear in the second and third terms of the Lagrangian. We take the derivative of the Lagrangian w.r.t. $\mathcal{X}_{ijk}$ and set it to 0, then we get:

$$\mathcal{X}_{ijk} = \left(\frac{\sum_{l=1}^{n} \beta_l fold_l(M_l)}{\sum_{l=1}^{n} \beta_l} \right)_{ijk} \tag{12}$$

The elements in set $\mathcal{C}_3$ do not appear in the third term of the Lagrangian. We take the derivative of the Lagrangian w.r.t. $\mathcal{X}_{ijk}$ and set it to 0, then we get:

$$\mathcal{X}_{ijk} = \left(\frac{\sum_{l=1}^{n} \beta_l fold_l(M_l)}{\sum_{l=1}^{n} \beta_l + \gamma} \right)_{ijk} + \left(\frac{\gamma S}{\sum_{l=1}^{n} \beta_l + \gamma} \right)_{jk} \tag{13}$$

Relaxed simplex formulation:

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

$$\begin{aligned} &\min_{\mathcal{X}} : \sum_{l=1}^{n} \frac{\beta_l}{2} ||\mathcal{X}_{(l)} - M_l||_F^2 + \frac{\gamma}{2} \sum_{j=1}^{N_i} \xi_j^2 \\ &s.t. : \sum_{k=1}^{N_c} \mathcal{X}_{ijk} - 1 = \xi_j, i = i_g, \forall j = 1, ..., N_i \\ &\quad\quad \mathcal{X}_\Omega = \mathcal{T}_\Omega \end{aligned} \tag{14}$$

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

$$\begin{aligned} L = &\sum_{l=1}^{n} \frac{\beta_l}{2} \sum_{i,j,k} (\mathcal{X}_{ijk} - (fold_l(M_l))_{ijk})^2 + \frac{\gamma}{2} \sum_{j=1}^{N_i} \xi_j^2 \\ &+ \sum_{j=1}^{N_i} \tau_j (\sum_{k=1}^{N_c} \mathcal{X}_{ijk} - 1 - \xi_j) + \sum_{(i,j,k)\in\Omega}^{|\Omega|} \lambda_{ijk}(\mathcal{X}_{ijk} - \mathcal{T}_{ijk}) \end{aligned} \tag{15}$$

Similar as the prior guided formulation, here the elements of tensor $\mathcal{X}$ are also divided into three sets $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$. The solutions for the elements in sets $\mathcal{C}_1, \mathcal{C}_2$ stay the same as shown in equations (11) and (12). The elements in set $\mathcal{C}_3$ do not appear in the third term of Lagrangian. We take the derivative of the Lagrangian w.r.t. $\mathcal{X}_{ijk}$ and ξ_j and set them to 0, then we get:

$$\mathcal{X}_{ijk} = \frac{\sum_{l=1}^{n} \beta_l (fold_l(M_l))_{ijk} - \tau_j}{\sum_{l=1}^{n} \beta_l} \tag{16}$$

$$\xi_j = \frac{\tau_j}{\gamma} \tag{17}$$

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

$$\tau_j = \frac{\sum_{l=1}^{n} \beta_l (\sum_{k=1}^{N_c} fold_l(M_l))_{ijk} - 1)}{N_c + \frac{1}{\gamma} \sum_{l=1}^{n} \beta_l} \tag{18}$$

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

$$\begin{aligned} \mathcal{X}_{ijk} = &\left(\frac{\sum_{l=1}^{n} \beta_l fold_l(M_l)}{\sum_{l=1}^{n} \beta_l} \right)_{ijk} \\ &+ \frac{\gamma \sum_{l=1}^{n} \beta_l (1 - \sum_{k=1}^{N_c} (fold_l(M_l))_{ijk})}{(\gamma N_c + \sum_{l=1}^{n} \beta_l) \sum_{l=1}^{n} \beta_l} \end{aligned} \tag{19}$$

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 $N_c \times N_c$ worker quality confusion matrix as follows: the diagonal entries are independently and

Algorithm 1: PG-TAC

Data: Augmented tensor $\mathcal{T}$, prior statistics S, α, β, γ, ϵ.
Result: Completed tensor $\mathcal{X}$.

while $||\mathcal{X} - \mathcal{T}||_F / ||\mathcal{T}||_F \geq \epsilon$ **do**
  for *l = 1:n* **do**
    Updating M_l based on equation (8).
  end
  Updating elements in set $\mathcal{C}_1$ based on equation (11);
  Updating elements in set $\mathcal{C}_2$ based on equation (12);
  Updating elements in set $\mathcal{C}_3$ based on equation (13);
end

uniformly sampled from a certain probability range. Empirically, the labeling difficulty of each item should rise with the increasing number of labels N_c, therefore it would be inappropriate to sample the diagonal probability entries in a fixed range for different N_c values. In order to simulate the real-world situations, we assign each diagonal entry with a probability which is the product of random guess and a scale factor κ. We empirically draw κ from a uniform distribution of range $[1.5, 1.99]$. For instance, if the scale factor is drawn as 1.8, then for a 3-labeling task, the diagonal element will have accuracy value of 0.6. The non-diagonal entries are randomly assigned with positive probabilities under the constraint that the sum of each row of the confusion matrix is equal to 1. Since each worker does not have to label all items, we draw a labeling decision for each worker from a Bernoulli distribution with the probability of q. In addition to prior guided tensor augmentation and completion (PG-TAC) and relaxed simplex tensor augmentation and completion (RS-TAC), we also use no constraint tensor augmentation and completion (NC-TAC) as a simple baseline method for comparison. The evaluation metric is the error rate of label prediction. All the outcomes of our proposed methods are result of 10 independent runnings and the performance is shown in figure 1. The initial configuration is $N_w = 50, N_i = 400, N_c = 4$ and $q = 0.7$. Based on this, we report the performance of our proposed methods on synthetic data sets generated with four groups of configurations: (a). N_w varies in range of $[20, 90]$ by a step size of 10. (b). N_c varies in range of $[2, 8]$ by a step size of 1. (c). N_i varies in range of $[50, 1000]$ by a step size of 50. (d). q varies in range of $[0, 0.95]$ by a step size of 0.05. Under each configuration, other data set parameters remain consistent with the initial configuration.

On all synthetic data sets we generated, the PG-TAC method achieves the lowest error rate on all configurations. The RS-TAC method does not necessarily improve the performance. In many configurations, RS-TAC and NC-TAC have almost the same performance. We also observe when the data sets are not sufficiently labeled, for instance, if N_w and N_i are very small or q is relatively large, the performance of our proposed methods are almost the same. When more data is used, the PG-TAC and RS-TAC have better performance.

5.2 Real data

We also evaluate our methods with various other methods on six real world crowdsourcing data sets. There are three data sets have binary labels and three data sets have multiple labels. The binary labeling data sets include Temp data set [Snow *et al.*, 2008]), RTE data set [Snow *et al.*, 2008] and Spam data set [Zhou *et al.*, 2015]. The multi-class labeling data sets include Dog data set [Zhou *et al.*, 2012], Web data set [Zhou *et al.*, 2012] and Age data set [Han *et al.*, 2015].

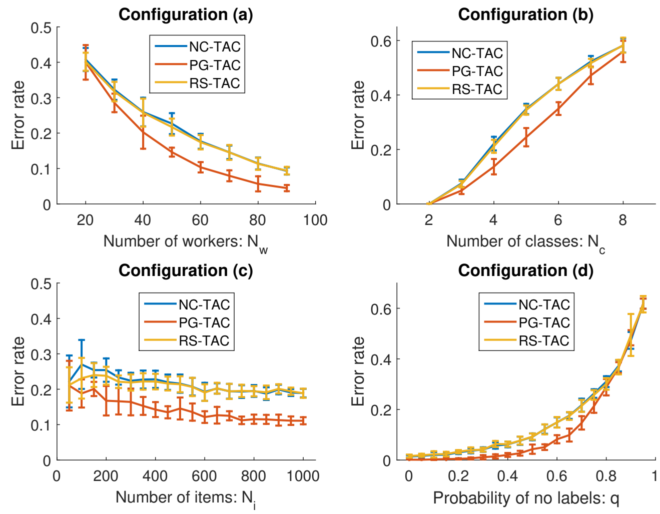

Figure 1: Comparison results of our proposed methods on various synthetic data sets configurations.

	# classes	# items	# workers	# total labels
RTE	2	800	164	8000
Temp	2	462	76	4620
Web	5	2653	177	15567
Dog	4	807	109	7354
Spam	2	149	18	1901
Age	7	1002	165	10020

Table 1: For Dog data set, the unqualified workers, who have only labeled a small amount of images, are remained. 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 filed method and this model is proposed by [Liu *et al.*, 2012b]. 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,

	MV	DS-EM	DS-MF	GLAD	MMCE	NC-TAC	PG-TAC	RS-TAC
RTE	10.31	7.25	**6.63**	7.00	7.50	7.13	7.00	7.25
Temp	6.39	5.84	5.84	5.63	5.63	5.84	**5.41**	5.84
Web	26.93	16.92	18.24	19.30	11.12	11.16	**10.82**	11.23
Dog	17.91	15.86	**15.74**	–	16.23	15.86	**15.74**	**15.74**
Spam	19.80	13.42	**12.75**	18.12	**12.75**	14.10	**12.75**	13.42
Age	34.88	39.62	36.33	35.73	**31.14**	31.24	32.44	**31.14**

Table 2: Comparison results of all methods on six real data sets in error rate (in percentage)

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 γ is very small, PG-TAC is approximately equal to NC-TAC; when γ is sufficiently large, PG-TAC reduces to its prior statistics. The objective of RS-TAC has a regularization term, which is parameterized by γ, 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: α_l, β_l and γ. Here $l = 1, ..., 3$ and 3 is the mode of the tensor. The values of α_l are assigned with value of $1/3$ and we let $\delta_l = \frac{\alpha_l}{\beta_l}$. Given δ_l, the value of β_l is also determined. Therefore it is straightforward to verify that we only need to tune δ_l in BCD iterations no matter it is in the step of computing M_l or in the step of computing $\mathcal{X}$. For simplicity, we let all δ_l be the same for all three modes. Eventually we can apply the grid search on two regularization parameters δ_l and γ, 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 $||\mathcal{X} - \mathcal{T}||_F / ||\mathcal{T}||_F$ as the stopping criteria and it is set to 10^{-5}. The final label prediction is performed as follows: in each fiber $\mathcal{X}_{i_g j:}$ 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

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20:1956–1982, 2010.

[Candès and Recht, 2009] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, pages 717–772, 2009.

[Candès and Tao, 2010] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transaction of Information Theory.*, pages 2053–2080, 2010.

[Dawid *et al.*, 1979] P. Dawid, A. M. Skene, A. P. Dawidt, and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pages 20–28, 1979.

[Gandy *et al.*, 2011] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 2011.

[Han *et al.*, 2015] Hu Han, Charles Otto, Xiaoming Liu, and Anil K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1148–1161, 2015.

[Huberman *et al.*, 2009] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Crowdsourcing, attention and productivity. *Journal of Information Science*, 35:758–765, 2009.

[Karger *et al.*, 2011] David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *CoRR*, abs/1110.3564, 2011.

[Kittur *et al.*, 2008] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[Kolda and Bader, 2009] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51:455–500, 2009.

[Liu *et al.*, 2012a] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:208–220, 2012.

[Liu *et al.*, 2012b] Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems*, 2012.

[Mineiro, 2011] P. Mineiro. http://www.machinedlearnings.com/2011/08/low-rank-confusion-modeling-of.html, 2011.

[Raykar *et al.*, 2010] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[Recht, 2011] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research.*, 12:3413–3430, 2011.

[Snow *et al.*, 2008] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 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*, pages 254–263, 2008.

[Tseng, 2001] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, pages 475–494, 2001.

[Welinder *et al.*, 2010] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, 2010.

[Whitehill *et al.*, 2009] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, 2009.

[Xu *et al.*, 2013] Yangyang Xu, Ruru Hao, Wotao Yin, and Zhixun Su. Parallel matrix factorization for low-rank tensor completion. *CoRR*, 2013.

[Zhang *et al.*, 2014] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, 2014.

[Zhou *et al.*, 2012] Dengyong Zhou, John C. Platt, Sumit Basu, and Yi Mao. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, 2012.

[Zhou *et al.*, 2015] Dengyong Zhou, Qiang Liu, John C. Platt, Christopher Meek, and Nihar B. Shah. Regularized minimax conditional entropy for crowdsourcing. *CoRR*, abs/1503.07240, 2015.