Multi-label Crowdsourcing Learning with Incomplete Annotations

Shao-Yuan Li and Yuan Jiang(✉)

National Key Lab for Novel Software Technology,
Nanjing University, Nanjing 210023, China
{lisy,jiangy}@lamda.nju.edu.cn

Abstract. In this paper we consider multi-label crowdsourcing (MLC) learning with labeling information from non-expert crowds. Previous crowdsourcing works typically care about single-label tasks, which ignore the label correlations. While the preliminary MLC studies concern themselves with the label correlations, they focus on *local* correlations whose estimation relies heavily on the annotations' quality and requires complete annotations for the labeled instances. However, annotations' quality in MLC is often various and incomplete annotations are common. For example, the crowds may just tag a few labels and leave the other labels untouched due to the heavy workload or labeling uncertainty. In this paper, we deal with the incomplete annotation issue. We propose a two stage approach considering the *global* low-rank structure correlation between the labels and crowds. Being able to learn with incomplete annotations, we also extend the proposed model to active annotation collection which significantly reduces the labeling cost. Experiments validate the effectiveness of our proposals.

Keywords: Multi-label · Crowdsourcing · Incomplete annotation
Low-rank structure · Tensor completion · Active query

1 Introduction

A rising challenge faced by machine learning algorithms is the scenario of multiple labels associated with data, e.g., one image may be tagged with labels *'urban'* and *'road'*, and one document may involve topics like *'economics'* and *'politics'*. To handle such tasks, multi-label learning (MLL) has received significant attention [30]. Typical MLL requires the groundtruth labels, which are rather expensive resources. Through distributing the task to multiple workers and estimating the true labels via some aggregation schemes, crowdsourcing [9] provides an economic alternative to collect labels.

Previous crowdsourcing works mainly care about single-label tasks [6,11,17,20,26,27,31], whereas using crowdsourcing for MLL is still in the preliminary stage. [1,7,22] extended the single-label methods by considering the *local*

Supported by NSF of China (61673201).

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11012, pp. 232–245, 2018.
https://doi.org/10.1007/978-3-319-97304-3_18

label correlations estimated from the crowds' annotations, which rely heavily on the annotations' quality. Furthermore, they require complete annotations for the labeled instances, i.e., they assume the positive labels are tagged and the untagged labels are negative. In practice, the crowds may just tag a few labels and leave the others untouched due to the heavy workload of examining the whole label set or labeling uncertainty. Simply regarding the untagged labels as negative would deteriorate the label correlation estimation and subsequent learning.

In this paper, we consider the incomplete annotation issue. We propose a two-stage approach **CRIA** (CRowdsourcing with Incomplete Annotations) to first estimate the incomplete annotations and then estimate the groundtruth. Considering that the labels are correlated and determined by a few factors, we assume a low-rank structure between labels; besides, with the basic crowdsourcing assumption that most workers are willing to provide good annotations, the tagged labels are very likely to be correct. As the workers are labeling the same task, their annotations should be closely related. Regarding the annotations as a three-mode (instance, label, worker) tensor, we assume a *global* low-rank structure of the tensor and propose an optimization objective to estimate the missing annotations, and then infer the groundtruth labels using ensemble over crowds.

Besides, motivated by the low-rank structure which means the full annotations can be approximated by only a subset of them, we also propose strategies to actively select the few most helpful annotations to reduce the labeling burden and cost. Based on the groundtruth prediction of the proposed **CRIA** model, we define criteria to measure the uncertainty, informativeness and reliability of instances, labels and workers, and select the most informative labels of the most uncertain instances to collect annotations from the most reliable workers.

2 Related Work

Dealing with examples associated with multiple labels, multi-label learning has received significant attention. The simplest way is to deal with each label independently. To exploit the label correlations, various advanced approaches were proposed, see [30] and references therein. Typical multi-label algorithms require the groundtruth labels for learning, which are expensive.

With the advent of crowdsourcing platforms such as Amazon Mechanical Turk (AMT), crowdsourcing [9] provides an alternative to collect supervised information by distributing the task to multiple easy to access workers. As workers may make mistakes, the common wisdom is to estimate the higher quality labels via some aggregation schemes. Numbers of studies for single-label tasks have been proposed, mainly by using probabilistic models and estimating the workers' expertise from different perspectives, such as measures with explicit explanation like accuracy [11,27], confusion matrix [16,17,31], and more complex multidimensional vectors [26]. In the multi-label learning field, using crowdsourcing is in the preliminary stage. To estimate the groundtruth labels, methods extending the single-label methods by exploiting label correlations were explored.

[7] proposed three implementations P-DS, D-DS, ND-DS respectively extending DS [6] by incorporating dependency relationships among all label power set, label set of two labels, and conditional label dependency. [22] captured the co-occurrence dependencies between labels exploiting the notion of latent label clusters. But they assume that the workers annotate the positive labels and the rest are negative, i.e., the annotations are complete, which may not be true in practice.

We also note works constructing the taxonomy of labels using crowdsourcing [1,5,21]. Among them, [1,5] collected annotations for items, the similar setting as ours. [1] also implemented an approach considering the label co-occurrence to estimate the 'true' labels of items. Whereas the methodology assumes that all annotations are equally reliable, thus the same parameter value for all workers are used, whose inferiority is demonstrated in our experiments.

Our technique of low-rank tensor completion is related to the field of matrix, tensor completion [3,4,8,15,18,25]. The trace norm has been shown to be the tightest convex approximation for the rank of matrices, and efficient algorithms for matrix completion using trace norm were proposed [3,4,18]. As there is no direct way to determine the rank for tensors, heuristic models such as Tucker model based tensor factorization [25], parallel factor analysis model [8], and tensor trace norm definition using matrix trace norm [15] were proposed. Similar to [32] which exploited tensor for multi-class problem, considering that the well developed tools of matrix completion can be exploited, we build our approach based on the model of [15]. Different from [32] which inserted a groundtruth layer into the tensor, and relied on prior information of results of other crowdsourcing methods, we estimate the groundtruth through ensemble over the completed annotations, which is more efficient, robust and stable.

Other related work may contain partial label learning (PLL) and multi-label active learning (MLA). Our awareness that workers may not tag all labels shares the similar concern with PLL, which learns from a partial set of groundtruth labels [2,28]. Our active annotation collection idea is inspired by MLA, which reduces the labeling cost by collecting the most valuable labels, using measures like uncertainty and informativeness [10,14,19]. Our previous work [13] considered active multi-label crowdsourcing by incorporating the local neighborhoods' label correlations, which are solicited from the initial set of groundtruth labels.

3 Crowdsourcing with Incomplete Annotations

We use bold capital letters such as $\mathbf{X}$ to denote matrix, $\|\cdot\|_*$, $\|\cdot\|_F$ the trace and Frobenious norm of one matrix. Calligraphic letters, such as $\mathcal{X}$ denote tensors. For a 3-mode tensor $\mathcal{X}$, its (i, j, t)-th element is represented as $\mathcal{X}_{ijt}$. $\mathcal{X}_{(k)}$ denotes the output matrix of the *unfold* operation along the k-th dimension on $\mathcal{X}$. The opposite operation *fold* is the inverse of *unfold* and returns the tensor. The Frobenius norm of a 3-mode tensor is defined as $\|\mathcal{X}\|_F = (\sum_{i,j,t} |\mathcal{X}_{ijt}|^2)^{1/2}$.

We represent the annotations for a set of multi-label instances $\mathbf{X} \in \mathbb{R}^{N\times d}$ as a 3-mode tensor $\mathcal{T} \in \{+1, -1, 0\}^{N\times L\times M}$, where N, L, M, d respectively denotes the number of instances, labels, workers and feature dimension. $\mathcal{T}_{ijt}$ denotes worker t's annotation result on label j for instance i. $\mathcal{T}_{ijt} = 1(-1)$ denotes worker t tags label j as positive (negative) for instance i, and $\mathcal{T}_{ijt} = 0$ denotes worker t doesn't tag label j. Note that the column $\mathcal{T}_{i:t} = 0^{L\times 1}$ means that worker t doesn't tag instance i. Our target is to estimate groundtruth labels from $\mathcal{T}$.

Previous MLC assumes that positive labels are tagged and the remaining are negative, i.e., each column $\mathcal{T}_{i:t}$ is either $\{+1,-1\}^{L\times 1}$ or $0^{L\times 1}$. This is not true in practice. To deal with this issue, we propose to first estimate the untagged annotations by exploiting the low rank structure between labels and annotations, and then estimate the groundtruth using ensemble.

3.1 Missing Annotation Estimation

Following typical crowdsourcing learning, we assume that most workers are acting with *good will*, i.e., they are willing to provide good annotations. Considering that the labels are correlated and determined by a few factors, and the annotation results of different workers should be closely related, the full annotation tensor is expected to be low rank. We adopt the tensor completion model in [15] to recover the untagged annotations. Formally, we wish to estimate an annotation tensor $\mathcal{X} \in \mathbb{R}^{N\times L\times M}$ whose entries in the observed positions should be close to the observed annotations, and at the same time, it should be low rank:

$$\min_{\mathcal{X}} \quad \textstyle\sum_{k=1}^{3} \alpha_k \|\mathcal{X}_{(k)}\|_* \quad s.t. \;\; \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{1}$$

Here $\Omega = \{(i,j,t)|\mathcal{T}_{ijt} \neq 0\}$ denotes the index set of observed annotations. $\mathcal{X}_{(1)}, \mathcal{X}_{(2)}, \mathcal{X}_{(3)}$ are respectively the unfolded annotation matrix along the instance, label and worker dimension. α_k are predefined scalars with $\sum_{k=1}^{3} \alpha_k = 1$, $\alpha_k \geq 0$. From Eq. 1, we can see that the tensor completion is essentially fulfilled by conducting low rank matrix completion on each of the unfolded annotation matrix along the instance, label and worker dimension. By using this model, we can easily exploit well developed techniques such as side information utilization and efficient optimization methods for low rank matrix completion.

Since the trace norms in Eq. 1 are not independent, to conduct optimization, matrices $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ are introduced and the optimization is converted as:

$$\min_{\mathcal{X},\mathbf{M}_k} \quad \textstyle\sum_{k=1}^{3} \alpha_k \|\mathbf{M}_k\|_* + \frac{\beta_k}{2}\|\mathcal{X}_{(k)} - \mathbf{M}_k\|_F^2 \quad s.t. \;\; \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{2}$$

To exploit the instance features as side information to augment the learning, we assume a linear relationship between the crowds' annotations and the instances' features, and the unfolded matrix $\mathcal{X}_{(1)}$ can be represented by $\mathcal{X}_{(1)} = \mathbf{X}\mathbf{W}$, where $\mathbf{W}$ is the coefficients. Thus our learning objective becomes:

$$\min_{\mathcal{X},\mathbf{W},\mathbf{M}_2,\mathbf{M}_3} \alpha_1\|\mathbf{X}\mathbf{W}\|_* + \frac{\beta_1}{2}\|\mathcal{X}_{(1)} - \mathbf{X}\mathbf{W}\|_F^2 + \textstyle\sum_{k=2}^{3} \alpha_k \|\mathbf{M}_k\|_* + \frac{\beta_k}{2}\|\mathcal{X}_{(k)} - \mathbf{M}_k\|_F^2$$
$$s.t. \;\; \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{3}$$

Algorithm 1. $\mathcal{X}$ Estimation

```
1: **Initialization:** $\mathcal{X} = \mathcal{T}$,$\Omega = \{(i,j,k)|\mathcal{T}_{ijk} \neq 0\}$, $\beta_k$, stopping criterion $\epsilon$
2: **while** stopping criterion is not satisfied **do**
3:     Calculate $\mathcal{X}$ by Eq. 5
4:     Calculate $\mathbf{M}_2, \mathbf{M}_3$ by Eq. 6
5:     Calculate $\mathbf{W}$ by Eq. 7 using APGD
6: **end while**
```

Incorporating the features into learning, we achieve two advantages: (1) it allows us to predict labels for novel unseen examples; (2) in cases where the feature dimension d is much smaller than the number of instances N, the matrix $\mathbf{W} \in \mathbb{R}^{d\times LM}$ mode would be much smaller than the matrix $\mathbf{M}_1 \in \mathbb{R}^{N\times LM}$, we reduce the number of parameters to learn and speed up the computation.

To estimate $\mathcal{X}, \mathbf{W}, \mathbf{M}_2, \mathbf{M}_3$, we employ block coordinate descent (BCD) [23] to estimate them iteratively which converges at rate $O(1/T)$.

Computing $\mathcal{X}$: With the other variables fixed, the optimization with respect to $\mathcal{X}$ is given by the following subproblem:

$$\min_{\mathcal{X}} \tfrac{\beta_1}{2}\|\mathcal{X}_{(1)} - \mathbf{XW}\|_F^2 + \textstyle\sum_{k=2}^{3} \tfrac{\beta_k}{2}\|\mathcal{X}_{(k)} - \mathbf{M}_k\|_F^2 \quad s.t. \quad \mathcal{X}_\Omega = \mathcal{T}_\Omega \tag{4}$$

The closed form solution is induced:

$$\mathcal{X}_{ijt} = \begin{cases} \mathcal{T}_{ijt} & (i,j,t) \in \Omega \\ (\frac{\beta_1 fold_1(\mathbf{XW}) + \sum_{k=2}^{3} \beta_k fold_k(\mathbf{M}_k)}{\sum_{k=1}^{3} \beta_k})_{ijt} & (i,j,t) \notin \Omega \end{cases} \tag{5}$$

Computing $\mathbf{M}_k$: With the other variables fixed, the optimization with respect to $\mathbf{M}_k$ is given by:

$$\min_{\mathbf{M}_k} \frac{1}{2}\|\mathcal{X}_{(k)} - \mathbf{M}_k\|_F^2 + \frac{\alpha_k}{\beta_k}\|\mathbf{M}_k\|_* \tag{6}$$

which has closed form solution by SVT [3].

Computing $\mathbf{W}$: With the other variables fixed, the optimization with respect to $\mathbf{W}$ is given by:

$$\min_{\mathbf{W}} \frac{\alpha_1}{\beta_1}\|\mathbf{XW}\|_* + \frac{1}{2}\|\mathcal{X}_{(1)} - \mathbf{XW}\|_F^2 \tag{7}$$

Following [28], we exploit Accelerated Proximal Gradient Descend (APGD) [24] to optimize Eq. 7 iteratively, which converges at rate $O(1/T^2)$.

3.2 Groundtruth Inference

After the tensor completion, we get one full annotation estimation tensor $\mathcal{X} \in \mathbb{R}^{N\times L\times M}$. To infer the groundtruth $\{\hat{\mathbf{z}}_{ij}\}$, we conduct ensemble over the recovered annotations. Compared to the observed discrete $\{1, -1\}$ labeling results, the recovered annotations are signed real valued. We test two voting strategies:

(1) **Signed Voting:** we use the sign $\{1, -1\}$ of the estimated annotations as the hard label annotation, and conduct voting over the workers:

$$\hat{\mathbf{z}}_{ij} = \sum_t sign(\mathcal{X}_{ijt})/M \tag{8}$$

This scheme is named as $\mathbf{CRIA}_S$ (CRowdsourcing with Incomplete Annotations).

(2) **Valued Voting:** we believe that the estimation values can represent some confidence level for the workers, for example, given the annotation estimation of two workers t, t' on label j of instance i $\mathcal{X}_{ijt} = 0.9$, $\mathcal{X}_{ijt'} = -0.1$, label j is more likely to be positive on instance i.

$$\hat{\mathbf{z}}_{ij} = \sum_t (\mathcal{X}_{ijt})/M \tag{9}$$

This scheme is named as $\mathbf{CRIA}_V$.

For a set of novel instances $\mathbf{X}_s$, to predict their groundtruth, we first estimate the crowds' annotations $\mathcal{X}_s = fold_1(\mathbf{X}_s\mathbf{W})$, and then conduct the above voting.

4 Active Annotation Collection

Considering that the labeling budget is often limited, whereas collecting annotations without control may lead to not only unnecessary information redundancy, but also possible out of control labeling errors (e.g., incomplete annotation in this paper), we propose to actively collect the most valuable annotations.

Instance Selection. To select the most informative instance, typical active methods defined some uncertainty measure for the instances considering the prediction uncertainty over labels. As the proposed **CRIA** model gives real valued prediction without hard label prediction, we define an uncertain measure for the instance, computed as the gap between its largest and smallest prediction of labels. The smaller the gap, it means the more difficult to split the positive and negative labels for the instance. The instance with the smallest gap is selected:

$$i^* = \arg\min_i U(\mathbf{x}_i) := \max_j \hat{\mathbf{z}}_{ij} - \min_j \hat{\mathbf{z}}_{ij} \tag{10}$$

Label Selection. Different from traditional active methods querying the most uncertain labels, we consider the *label sparsity* property of multi-label tasks, i.e., while the label size can be large, the number of positive labels is usually very small, whose information however is critical to learning. E.g., for the two image tasks concerning 6 and 16 labels in our experiment, the average number of positive labels are respectively 1.24 ± 0.45 and 1.80 ± 0.90. This *label sparsity* phenomenon is also observed on numerous multi-label benchmark data.[1] Driven by this, we propose to query the most possibly *positive* labels for annotation:

$$L_j^* = \{j^* \mid \hat{\mathbf{z}}_{i^*j} \text{ ranks top } l \text{ among label predictions}\} \tag{11}$$

[1] http://mulan.sourceforge.net/datasets-mlc.html.

Algorithm 2. Active Crowdsourcing Procedure

```
Input: instances $\mathbf{X}$, initial annotations $\mathcal{T}$
Repeat:
    Estimate the full annotations $\mathcal{X}$ and groundtruth for $\mathbf{X}$ respectively by Alg. 1 and Eq. 9
    Select instance, labels, worker $(\mathbf{x}_{i^*}, L_j^*, t^*)$ respectively by Eq. 10, 11, 12
    Query annotations for $(\mathbf{x}_{i^*}, L_j^*, t^*)$ and add into $\mathcal{T}$
Until The maximum number of queries is reached
```

The number l can be determined based on reality, e.g., if the cardinality of positive labels is large (small), we can set l large (small). We test l with varying values in the experiments.

Worker Selection. Given the selected instance and label (i^*, L_j^*) by Eqs. 10 and 11, we select the worker which most possibly gives the positive annotations:

$$t^* = \arg\max_t \textstyle\sum_{j^* \in L_j^*} \mathcal{X}_{i^* j^* t} \tag{12}$$

After the instance, label, worker indices (i^*, L_j^*, t^*) are selected, the corresponding annotations are collected and added into $\mathcal{T}_{i^* j^* t^*}$ to update the full annotation tensor and groundtruth label estimation using the CRIA model. The overall process is summarized in Algorithm 2. We compare with a few baselines in the experiment, which shows significant annotation savings.

5 Running Time Analysis

For $\mathbf{CRIA}_S$ ($\mathbf{CRIA}_V$), the computation mainly comes from the SVT while computing $\mathbf{M}_k$ for Eq. 6 and $\mathbf{W}$ for Eq. 7, which can be implemented very efficiently using readily available high-quality packages, see [3] for details. As approximate solution is good enough, early stop can be employed for the iterative BCD and APGD, e.g., their maximum iteration number are set as 200, 100 in our experiment. Thus compared to the probabilistic crowdsourcing learning methods which rely on the EM procedure over all labels and workers, $\mathbf{CRIA}_S$ ($\mathbf{CRIA}_V$) is very efficient, especially in the case of not small number of workers and labels.

6 Experiment

6.1 Data Sets

We distributed two multi-label image annotation tasks with different instance and label size on AMT, and ask the workers to tag the positive label for the image they see, the same way as existing multi-label crowdsourcing works do. Thus the annotations are expected to be a fair comparison benchmark.

Scene. *Scene* contains 700 images concerning 6 labels. On average each image has 1.24 ± 0.45 labels. 18 workers annotating the most images (each no less than

70) are kept for experiment. On average each worker annotated 267 ± 201 images, each image was annotated by 6.9 ± 2.3 workers.

Image. *Image* contains 1495 images concerning 16 labels. On average each image has 1.80 ± 0.90 labels. 15 workers annotating the most images (each no less than 100) are kept. On average each worker annotated 397 ± 453 images, each image was annotated by 10.1 ± 1.41 workers.

The groundtruth labels are annotated by human volunteers, and a 1248 dimension fisher vector is extracted as the feature. We conduct some rough analysis to get some idea about the data quality. For each worker, we compute its MacroF1 score on its annotated instances. MacroF1 [30] is the macro average of the F1 score over all labels, with value in [0, 1], the larger, the better. Respectively 15 and 14 workers' macroF1 range in $[0.70, 0.85]$ for *Scene* and *Image*, indicating that annotations from most workers are reliable. In the following, we first conduct experiments concerning the incomplete annotation issue, then test the effect of active annotation collection strategies.

6.2 Crowdsourcing Learning

We test two settings for the incomplete annotation learning. In the *transductive* setting, we make annotations of the whole data uniformly random missing, and our target is to estimate the groundtruth labels for the data set. We vary the observed fraction of annotations p from 100% to 10%. In the *inductive* setting, we randomly split the data into 10% test data with no annotations and 90% training data with annotations, with the observation varies also from 100% to 10%. Results on the *test data* are reported. Each experiment is repeated for 10 times and the average and standard deviation results are recorded.

We compare with four multi-label baselines **D-DS**, **P-DS**, **ND-DS** [7], **MLNB** [1], and four state of the art single-label baselines **MV**, **DS** [6], **Yutc** [17], **MaxEn** [31]. The single-label baselines can deal with the incomplete annotation issue by learning on each label separately, but they ignore the label correlations thus are expected to work worse than ours. The multi-label baselines treat the untagged labels for the annotated instances as negative. For the proposed method, parameters are set as $\alpha_k = 1/3$, $\beta_k = 0.1$, $\epsilon = 10^{-2}$. The maximum iteration for BCD, APGD are set as 200, 100. The codes of the baselines are provided by their authors and the default parameter suggested there are used. Except for DS, the two coin model implemented by [16] is used. Besides, we also incorporate another baselines **LinEn**, which builds a linear classifier for each label on each worker using its corresponding instances, and the prediction is the ensembles of workers. To prevent overfitting, l2 regularization with trade-off parameter 1 is used. As our method uses linear classifier, and also considers the missing annotation issue and the label/worker correlations, we choose LinEn as a fair baseline to demonstrate the importance of the above two factors.

As our method currently learns no threshold to separate the positive and negative predictions, to evaluated the classification performance, we treat the top k ranked labels as positive and rest as negative. Here k is the number of true positive labels of each example. For comparison fairness, the best result among this strategy and the heuristic thresholding strategy (0.5 for crowdsourcing baselines and 0 for LinEns) for baselines are used. We report the macroF1 (macF.) results, performance on other measures such as hamming loss, microF1, average arecision and ranking loss are similar and we omit them due to space limitation.

Transductive Results. The results for the *transductive* setting are shown in Table 1. We can see that CRIA$_V$ outperforms other methods significantly in most time, whereas the signed voting scheme CRIA$_S$ is inferior. When $p = 100\%$ which is the same scenario of previous multi-label baselines, our approach also achieves significantly better performance. It's notable that when the annotations are small, e.g., 10% for *Scene* and 20% for *Image*, LinEn performs the best. This may be due to the information insufficiency for the crowdsourcing methods to learn reliable parameters. When the annotations increase, LinEn is not able to gain as much benefit as the crowdsourcing baselines. Among the multi-label baselines, MLNB ranks the worst. Two reasons may explain this, first, MLNB considering label co-occurrence is designed for learning the label hierarchy, but for problems lacking such hierarchies, label co-occurrence is not as common; more importantly, MLNB uses the same parameter for all workers, which ignores their expertise variance. Comparing D-DS, P-DS, and ND-DS which extend DS [6] by considering label relationship among all label powersets, label set of two labels, ND-DS performs most effectively by learning the conditional label dependency.

Comparing the single-label baselines, when the annotations are no less than 60%, MaxEn which models each worker's confusion on each example performs the best. Whereas when annotations are less, Yutc utilizing the instances' feature performs better. Besides, the multi-label baselines are not necessarily always better than the single-label ones, which is reasonable in crowdsourcing. Since their local label correlations are solicited solely from the collected annotations, whose quality would rely heavily on the annotations' quality and quantity.

Inductive Results. The results for the *inductive* setting are shown in Table 2. Except for Yutc and LinEn, the other baselines didn't consider the feature information and learned no classifier, thus no results are reported for them. The comparison is similar with the transductive setting, but with much lower performance, indicating that inductive crowdsourcing learning is more challenging.

6.3 Active Results

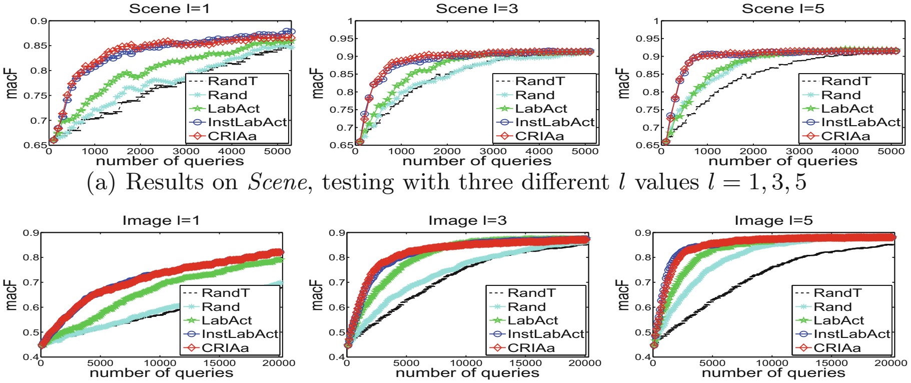
In the above, we have validated the superiority of the proposed **CRIA** model, in this section, we test the proposed active strategies. Using the real valued voting scheme **CRIA**$_V$ as learning model, we test: (1) **Rand** which selects instance, labels, worker randomly; (2) **LabAct** which selects instance and worker randomly, selects labels using the proposed strategy; (3) **InstLabAct** which selects worker randomly, selects instance and labels using the proposed strategy; (4) **CRIA**$_a$ which selects instance, labels, worker using the proposed strategy. Each time one instance is first selected, then l labels for this instance are selected, and the worker is selected. We also compare with one totally random strategy (5) **RandT** which selects l random entries of the tensor at each time.

Table 1. The *MacF1* results for the *transductive* setting. The randomly observed annotations p varies from 100% to 10%. CRIA$_S$ and CRIA$_V$ are the proposed approach with signed and real valued voting. mean ± std results over 10 times random repetitions are recorded. The best results on each row are bolded, with the comparable ones (pairwise single-tailed t-test at 95% confidence level) marked by •.

	p	100%	80%	60%	40%	20%	10%
Scene	D-DS	.869 ± .000	.854 ± .006	.812 ± .006	.744 ± .012	.598 ± .019	.419 ± .019
	P-DS	.421 ± .000	.411 ± .003	.394 ± .003	.357 ± .008	.271 ± .011	.181 ± .011
	ND-DS	.883 ± .000	.863 ± .008	.813 ± .007	.751 ± .009	.607 ± .014	.422 ± .020
	MLNB	.186 ± .000	.207 ± .014	.207 ± .013	.196 ± .008	.176 ± .011	.151 ± .006
	MV	.895 ± .000	.871 ± .006	.840 ± .005	.780 ± .008	.620 ± .015	.441 ± .018
	DS	.853 ± .003	.833 ± .010	.799 ± .010	.775 ± .014	.692 ± .014	.533 ± .019
	MaxEn	.893 ± .001	.884 ± .007	.863 ± .006	.828 ± .010	.716 ± .013	.537 ± .019
	YuTc	.854 ± .004	.847 ± .010	.833 ± .013	.835 ± .017	.764 ± .039	.571 ± .102
	LinEn	.809 ± .000	.798 ± .005	.788 ± .006	.777 ± .009	.754 ± .009	**.726 ± .009**
	CRIA$_S$	.907 ± .000	.900 ± .004	.882 ± .007	.847 ± .012	.743 ± .015	.617 ± .013
	CRIA$_V$	**.918 ± .000**	**.914 ± .005**	**.900 ± .009**	**.871 ± .010**	**.783 ± .013**	.647 ± .012
Image	D-DS	.794 ± .000	.757 ± .006	.703 ± .007	.623 ± .013	.462 ± .012	.293 ± .002
	P-DS	.161 ± .000	.156 ± .005	.148 ± .007	.129 ± .010	.086 ± .009	.049 ± .007
	ND-DS	.803 ± .000	.752 ± .002	.686 ± .001	.607 ± .000	.454 ± .014	.294 ± .001
	MLNB	.120 ± .000	.115 ± .007	.113 ± .005	.114 ± .009	.103 ± .005	.093 ± .004
	MV	.850 ± .000	.811 ± .005	.760 ± .004	.662 ± .005	.476 ± .008	.311 ± .006
	DS	.769 ± .004	.753 ± .010	.747 ± .011	.699 ± .009	.516 ± .010	.339 ± .008
	MaxEn	.845 ± .000	.827 ± .006	.796 ± .004	.717 ± .008	.522 ± .011	.337 ± .004
	YuTc	.813 ± .003	.797 ± .012	.781 ± .008	.738 ± .022	.568 ± .040	.366 ± .057
	LinEn	.707 ± .000	.693 ± .003	.675 ± .011	.652 ± .012	**.621 ± .020**	**.558 ± .040**
	CRIA$_S$	.877 ± .000•	.852 ± .003	.805 ± .006	.702 ± .007	.508 ± .005	.363 ± .009
	CRIA$_V$	**.877 ± .000**	**.864 ± .005**	**.832 ± .005**	**.760 ± .006**	.599 ± .007	.471 ± .006

Table 2. The *MacF1* results for the *inductive* setting. The randomly observed annotations p varies from 100% to 10%. CRIA$_S$ and CRIA$_V$ are the proposed approach with signed and real valued voting. mean ± std results over 10 times random repetitions are recorded. The best results on each row are bolded, with the comparable ones (pairwise single-tailed t-test at 95% confidence level) marked by •.

	p	100%	80%	60%	40%	20%	10%
Scene	YuTc	.286 ± .018	.291 ± .028	.270 ± .020	.246 ± .041	.060 ± .022	.114 ± .062
	LinEn	**.293 ± .000**	**.303 ± .021**	.286 ± .018	.287 ± .023	.270 ± .025	.241 ± .038
	CRIA$_S$	.279 ± .000	.286 ± .033	.286 ± .023	.283 ± .019	.293 ± .030	.286 ± .029
	CRIA$_V$	.286 ± .000	.301 ± .033•	**.290 ± .029**	**.291 ± .031**	**.303 ± .012**	**.300 ± .044**
Image	YuTc	.368 ± .000	.331 ± .000	.311 ± .000	.274 ± .000	.195 ± .000	.094 ± .000
	LinEn	.339 ± .000	.314 ± .014	.272 ± .022	.237 ± .019	.186 ± .022	.163 ± .019
	CRIA$_S$	.369 ± .000	.337 ± .008	.299 ± .018	.264 ± .019	.224 ± .019	.199 ± .021
	CRIA$_V$	**.399 ± .000**	**.374 ± .015**	**.348 ± .021**	**.324 ± .020**	**.278 ± .017**	**.235 ± .015**



(a) Results on *Scene*, testing with three different l values $l = 1, 3, 5$

(b) Results on *Image*, testing with three different l values $l = 1, 3, 5$

Fig. 1. The MacF1 results of different annotation collection strategies.

Here we don't compare with active matrix/tensor completion method like adaptive column subset selection [12], because with the target of recovering the matrix, they require a whole column each time which corresponds to all the labels of an instance, whereas our concern is just that the worker may not annotate all labels. We do not compare with traditional multi-label active learning because they learn with groundtruth labels. We also do not compare with related but different work [29] for single-label crowdsourcing tasks and [13] which relies on groundtruth for label correlation exploitation.

For each data, we randomly select 10% of the annotations as the initial data, and iteratively select (instance, label, worker) to query annotations and update the groundtruth estimation. The average MacF1 results over 10 times repetition in the *transductive* setting are shown in Fig. 1, $l = 1, 3, 5$ are tested. It can be seen that both the instance and label active query strategies (InstLabAct vs LabAct; LabAct vs Rand) play significant role in finding the most helpful annotations. Besides, comparing Rand and RandT, we can see some subtle advantage of Rand at the early stage, indicating that focusing on querying labels of specific instances is more preferred. For worker selection, no much difference between the random and active strategy is observed. This may be explained by the overall uniform annotation quality of the workers. On the two experiment data, $l = 3$ is a moderate number for $\mathbf{CRIA}_a$, which converges much faster than $l = 1$ and no slower than $l = 5$. This is due to that, the average number of positive labels for *Scene* and *Image* are respectively 1.24 and 1.80, for which 3 is large enough.

6.4 Parameter Study

In the experiment, parameters α_k and β_k are fixed. α_k is conventionally set as 1/3 (3 is the mode number). β_k trade-off between the approximation to the observed annotations and the low-rank property of annotations. We study the effect of β_k to our multi-label learning. Let $\gamma = \alpha_k/\beta_k$, we vary γ in $[10^{-3}, \dots, 10^{3}]$, and plot in Fig. 2 the results with observed annotation rate $10\%, 40\%, 70\%$ in the transductive setting. From Fig. 2 we can see that when γ is no less than 10, the learning performance is fairly stable, which is consistent with [15].

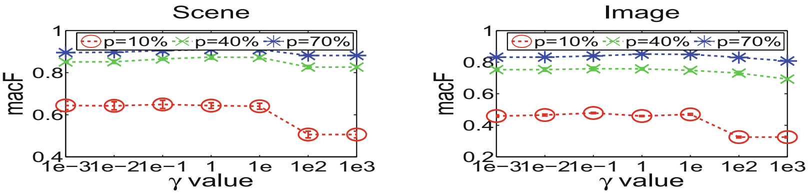


Fig. 2. The influence of β with three different observation rate

7 Conclusion

In this paper, we deal with multi-label crowdsourcing learning where annotations for the tagged instances are incomplete. We exploit the low-rank structure between labels and crowds to estimate the unobserved annotations and infer the groundtruth. Experiments show the superiority of the proposed approach, even in the complete annotation case. We also propose active annotation collection strategies to effectively reduce the labeling workload and cost. Currently, we do not pay special attention to spammer workers which would provide no beneficial annotations, for future work, we would like to deal with such problem.

References

1. Bragg, J., Mausam, Weld, D.: Crowdsourcing multi-label classification for taxonomy creation. In: First AAAI Conference on Human Computation and Crowdsourcing (2013)
2. Bucak, S.S., Jin, R., Jain, A.K.: Multi-label learning with incomplete class assignments. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, pp. 2801–2808 (2011)
3. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
4. Candes, E., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inform. Theory **56**(5), 2053–2080 (2009)
5. Chilton, L., Little, G., Edge, D., Weld, D., Landay, J.: Cascade: crowdsourcing taxonomy creation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1999–2008 (2013)
6. Dawid, A.P., Skene, A.M.: Maximum likeihood estimation of observer error-rates using the em algorithm. J. Roy. Stat. Soc. **28**(1), 20–28 (1979)
7. Duan, L., Oyama, S., Sato, H., Kurihara, M.: Separate or joint? Estimation of multiple labels from crowdsourced annotations. Expert Syst. Appl. **41**(13), 5723–5732 (2014)

8. Harshman, R.A.: Foundations of the PARAFAC procedure: models and conditions for an explanatory multi-modal factor analysis. In: UCLA Working Papers in Phonetics, vol. 16, pp. 1–84 (1970)
9. Horvitz, E.: Reflections on challenges and promises of mixed-initiative interaction. AI Mag. **28**(2), 13–22 (2007)
10. Huang, S.J., Zhou, Z.H.: Active query driven by uncertainty and diversity for incremental multi-label learning. In: Proceedings of the 13th IEEE International Conference on Data Mining, pp. 1079–1084 (2013)
11. Karger, D., Oh, S., Shah, D.: Iterative learning for reliable crowdsourcing systems. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 24, pp. 1953–1961 (2011)
12. Krishnamurthy, A., Singh, A.: Low-rank matrix and tensor completion via adaptive sampling. In: Advances in Neural Information Processing Systems 26, Lake Tahoe, Nevada, pp. 836–844 (2013)
13. Li, S.Y., Jiang, Y., Zhou, Z.H.: Multi-label active learning from crowds. CoRR abs/1508.00722 (2015)
14. Li, X., Guo, Y.: Active learning with multi-label SVM classification. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, pp. 1479–1485 (2013)
15. Liu, J., Musialski, P., Wonka, P., Ye, J.: Tensor completion for estimating missing values in visual data. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 208–220 (2011)
16. Liu, Q., Peng, J., Ihler, A.: Variational inference for crowdsourcing. In: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 692–700 (2012)
17. Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. J. Mach. Learn. Res. **11**, 1297–1322 (2010)
18. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
19. Singh, M., Curran, E., Cunningham, P.: Active learning for multi-label image annotation. In: Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science (2008)
20. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.: 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, Honolulu, Hawaii, pp. 254–263 (2008)
21. Sun, Y., Singla, A., Fox, D., Krause, A.: Building hierarchies of concepts via crowdsourcing. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, pp. 844–853 (2015)
22. Tam, N.T., Viet, H.H., Hung, N.Q.V., Weidlich, M., Yin, H., Zhou, X.: Multi-label answer aggregation for crowdsourcing. Technique report (2016)
23. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**, 475–494 (2001)
24. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, Seattle (2008)
25. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**, 279–311 (1966)
26. Welinder, P., Branson, S., Belongie, S., Perona, P.: The multidimensional wisdom of crowds. In: Lafferty, J., Williams, C.I., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems 23, pp. 2024–2432 (2010)

27. Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J.: Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22, pp. 2035–2043 (2009)
28. Xu, M., Jin, R., Zhou, Z.H.: Speedup matrix completion with side information: application to multi-label learning. In: Burges, C., Bottou, L., Ghahramani, Z., Weinberger: K. (eds.) Advances in Neural Information Processing Systems 26, pp. 2301–2309 (2013)
29. Yan, Y., Rosales, R., Fung, G., Dy, J.: Active learning from crowds. In: Proceedings of the 28th International Conference on Machine Learning, pp. 1161–1168 (2011)
30. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. **26**(8), 1819–1837 (2014)
31. Zhou, D., Basu, S., Mao, Y., Platt, J.: Learning from the wisdom of crowds by minimax entropy. In: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 2195–2203 (2012)
32. Zhou, Y., He, J.: Crowdsourcing via tensor augmentation and completion. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, pp. 2435–2441 (2016)