Discover Multiple Novel Labels in Multi-Instance Multi-Label Learning*

Yue Zhu [1,2] **and Kai Ming Ting** [3] **and Zhi-Hua Zhou** [1,2]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210023, China
[3] Federation University, Victoria 3842, Australia
{zhuy, zhouzh}@lamda.nju.edu.cn, kaiming.ting@federation.edu.au

Abstract

Multi-instance multi-label learning (MIML) is a learning paradigm where an object is represented by a bag of instances and each bag is associated with multiple labels. Ordinary MIML setting assumes a fixed target label set. In real applications, multiple novel labels may exist outside this set, but hidden in the training data and unknown to the MIML learner. Existing MIML approaches are unable to discover the hidden novel labels, let alone predicting these labels in the previously unseen test data. In this paper, we propose the first approach to discover multiple novel labels in MIML problem using an efficient augmented lagrangian optimization, which has a bag-dependent loss term and a bag-independent clustering regularization term, enabling the known labels and multiple novel labels to be modeled simultaneously. The effectiveness of the proposed approach is validated in experiments.

Introduction

In traditional supervised learning, an object is represented by a single instance associated with a single label (SISL). In many real applications, however, an object can be simultaneously associated with multiple labels, whereas the object can be described by multiple instances. An image in the image classification task, for example, can be divided into multiple patches (as multiple instances); and the image can be tagged with multiple semantic labels (Zhou and Zhang 2007). The Multi-instance multi-label learning (MIML) framework has been established to handle this kind of complex objects (Zhou *et al.* 2012).

Many MIML approaches have been proposed recently (Nguyen 2010; Zhang 2010; Briggs *et al.* 2012; Huang *et al.* 2014). Almost all of them assume that all bags belong to a fixed target label set. However, this assumption is frequently violated. Some novel labels may exist but hidden in training data, even though they do not exist in the initial target label set for some reasons. It is important to build a model which can attune itself to novel labels (Zhou 2016): not only to identify them in the training data, but enable the learned model to predict new objects with these novel labels.

*This research was supported by the NSFC (61333014) and the 111 Program (B14020).

Pham *et al.* (2015) has proposed a probabilistic model to identify novel instances in MIML setting. They have assumed that all novel instances have the same label. The novel instances, however, may actually belong to different labels. In many applications, a learned model is not only required to detect novel instances, but also able to classify these instances into one of the multiple novel labels. For example, it is desirable to distinguish multiple new voices in bird song recognition, which may correspond to different birds which are the new arrivals in the area.

In this paper, we aim to discover multiple novel labels in MIML learning, and propose a discriminative approach called `DMNL`. The contributions of this paper are: (1) formalize the multiple novel labels discovering problem in MIML learning; (2) propose the first approach to effectively discover novel labels by formulating the problem as a non-negative orthogonal constrained optimization, minimizing a bag-level loss plus a clustering regularization; (3) propose two new evaluation metrics for novel labels discovery.

In contrast to Pham *et. al.* (2015)'s probabilistic approach, our discriminative approach has the following advantages: (a) `DMNL` discovers multiple novel labels, rather than dealing with one novel label only. (b) `DMNL`'s computational cost increases linearly, rather than increasing exponentially w.r.t. the number of bag labels.

There are two other lines of related works: (i) class-incremental learning (Da *et al.* 2014; Kuzborskij *et al.* 2013) mainly focuses on one novel class under a SISL setting; and (ii) MIML learning with weak labels (Yang *et al.* 2013) tries to recover the missing known labels for each bag. They could not be applied to discover multiple novel labels in MIML learning, where both known labels and novel labels must be modeled simultaneously.

The rest of this paper is organized as follows. We provide the conceptual overview of the proposed approach, the preliminaries, the problem formalization, the proposed approach, and the optimization method in the next five sections. Experimental results are presented before conclusion.

Conceptual Overview

In our work, we discover multiple novel labels in both instance-level annotation and bag-level prediction tasks in MIML setting. Specifically, we design a new discriminative approach called `DMNL` that optimizes the bag-level loss on

known labels (in a loss term), and it takes into consideration the cluster structure among all instances (in a regularization term), independent of bags.

In the regularization term, we exploit the structure among instances to regularize the hypothesis space. Specifically, we derive our regularization based on an assumption about the following structure: instances, independent of bags, with the same label are close to each other in the feature space, thus are grouped into a single cluster. The intuition is as follows. Suppose there are two bags $\{\boldsymbol{x}_{1,1},\cdots,\boldsymbol{x}_{1,n_1}\}$ and $\{\boldsymbol{x}_{2,1},\cdots,\boldsymbol{x}_{2,n_2}\}$, both with bag labels $\{l_1,l_2\}$. If no structure is considered, there are $(2^{n_1}-2)(2^{n_2}-2)$ possible combinations of instance labels for instance annotation. In contrast, if we know the instance groupings, say $\{\boldsymbol{x}_{1,1},\boldsymbol{x}_{1,2},\boldsymbol{x}_{2,1},\boldsymbol{x}_{2,2}\}$ are in one group and the rest are in the other group, the number of possible combinations will be reduced to 2.

The loss term is designed based on the bag-level performance on known labels. We propose to utilize all instances in a bag for the purpose of loss computation. This loss is different from that of existing discriminative MIML approaches. They usually take the instance with the largest predictive value in a bag as the bag representative for each label (Zhang and Zhou 2008; Briggs *et al.* 2012; Huang *et al.* 2014). The bag label, however, is usually depended on some structure relation among the instances rather than a single instance (Zhou *et al.* 2009). As the crucial structure is unknown, we choose to consider the contribution of all instances to the bag label. Besides, each bag label is of equal importance, even if it may be associated with different number of instances. So we introduce a rescaling (Zhou and Liu 2010) strategy, which is popular in class imbalance learning, to balance the contribution of each label.

The closest related work (Pham *et al.* 2015) on novel instance detection in MIML learning treats all novel instances as belonging to a single novel label only. It is a degenerated version of the multiple novel labels problem. The approach learns a probabilistic model by maximizing the log likelihood. It utilizes a dynamic programming method to reduce the likelihood estimation complexity for each bag i from $O((c_i+1)^{z_i})$ to $O((c_i+1)2^{(c_i+1)}z_i)$, where c_i is the number of observed bag labels, and z_i is the bag size.

It is non-trivial to extend Pham *et al.* (2015)'s work to model multiple novel labels. In their estimation of the likelihood for each bag with the only one possible novel label, 2 cases are considered: (a) the known labels only; (b) both the known labels and the novel label. It is more complicated for $k>1$ novel labels since all of the subsets of k novel labels should be taken into account to compute the likelihood, which yields 2^k cases in total.

Preliminaries

Let $\mathcal{X}$ be the instance feature space, and $L=\{l_1,\cdots,l_c\}$ be the target label set of size c. Further we define $D=\{(X_1,\boldsymbol{y}_1),\cdots,(X_m,\boldsymbol{y}_m)\}$ as the training set of size m, where $X_i=\{\boldsymbol{x}_{i,1};\cdots;\boldsymbol{x}_{i,z_i}\}$ is a bag of z_i instances with each instance $\boldsymbol{x}_{i,j}\in\mathcal{X}$, and $\boldsymbol{y}_i=[y_{i,1},\cdots,y_{i,c}]\in\{0,1\}^c$ is the observed bag label vector. If bag i belongs to l_j, then $y_{i,j}=1$; otherwise $y_{i,j}=0$.

Table 1: Some commonly used notations.

I	identity matrix	$\mathbf{1}$	all-one vector
$\circ$	element-wise product	$\div$	element-wise divide
$\mathrm{Tr}(.)$	matrix trace	$\vee$	pairwise OR operator
$\bigvee$	$\bigvee_{i=1}^{n}\boldsymbol{a}_i=\boldsymbol{a}_1\vee\boldsymbol{a}_2\vee\cdots\vee\boldsymbol{a}_n$		
$\Omega_c(.)$	the first c columns of a matrix		
$\mathbb{I}(.)$	1 if the argument is true, 0 otherwise		
$\mathrm{diag}(.)$	a diagonal matrix with the arguments on the diagonal		

Suppose there are k novel labels; the combined label set becomes $\hat{L}=L\cup\overline{L}$, where $\overline{L}=\{l_{c+1},\cdots,l_{c+k}\}$ represents the k novel labels. Let $\hat{\boldsymbol{y}}_{i,j}=[\hat{y}_{i,j,1},\cdots\hat{y}_{i,j,c+k}]\in\{0,1\}^{c+k}$ be the unknown instance label vector for instance j in bag i.

We follow a common assumption in the MIML setting, i.e., each instance belongs to a single label only (Briggs *et al.* 2012; Pham *et al.* 2014; 2015). Hence, we have $\sum_{l=1}^{c+k}\hat{y}_{i,j,l}=1$ for each instance $\boldsymbol{x}_{i,j}$. Note that novel labels and instance labels are not available for training.

Let $A=[X_1;\cdots;X_m]$ be the all-instance matrix which is the concatenation of instances from all bags, and $\boldsymbol{a}_i, i\in\{1,\cdots,n\}$ be the i-th row of A, where $n=\sum_{i=1}^{m}z_i$ is the total number of instances. Define $\hat{Y}_i=[\hat{\boldsymbol{y}}_{i,1};\cdots;\hat{\boldsymbol{y}}_{i,z_i}]$ as the instance label matrix of bag i, and $\tilde{Y}=[\hat{Y}_1;\cdots;\hat{Y}_m]$ as the concatenation of all instance label vectors. Each row of A and $\tilde{Y}$ is an instance and its label vector, respectively.

Other commonly used notations are listed in Table 1.

Problem Formalization

Problem Definition: Given a training set D, consists of bags with known labels, the problem of discovering multiple novel labels in MIML learning is to detect previously unknown labels (i.e., novel labels) in each bag in the training set, and build a model that can predict bag labels from the set of known and novel labels for a previously unseen bag.

The problem can be tackled in two levels given D: (i) the instance-level annotation task is to learn a mapping from an instance to a label (in the set of known labels and novel labels) $f:\mathcal{X}\to\hat{L}$; (ii) the bag-level prediction task is to learn a mapping from a bag to a set of labels $\Psi:2^{\mathcal{X}}\to 2^{\hat{L}}$.

We assume that the labels of a bag include all instance labels in that bag. As a consequence, when task (i) is solved and each instance label $\hat{\boldsymbol{y}}_{i,j}$ in bag i has been predicted, task (ii) will be solved by predicting $\bigvee_{j=1}^{z_i}\hat{\boldsymbol{y}}_{i,j}$ as the bag label for bag i.

Proposed Approach: DMNL

In this section, we provide the details of the loss term and the clustering regularization term mentioned in the section on conceptual overview. We first introduce two properties of instance labels. To simplify notations, we illustrate the properties using known labels only. Those properties also hold in the multiple novel labels setting with minor change in notations. The properties are presented by the following propositions.

Proposition 1 *$\boldsymbol{y}_i = \boldsymbol{\beta}_i^\top \hat{Y}_i$, where $\boldsymbol{\beta}_i = [\beta_{i,1}; \cdots; \beta_{i,z_i}]$, with each $\beta_{i,j} = 1/\sum_{l=1}^{c}(\mathbb{I}(\hat{y}_{i,j,l}=1)\sum_{q=1}^{z_i}\hat{y}_{i,q,l})$.*

In Proposition 1, each instance label contributes to the bag label, and $\boldsymbol{\beta}_i$ corresponds to the contribution weights. In order to balance the importance of each bag label (they may be associated with different number of instances), we introduce a rescaling (Zhou and Liu 2010) strategy. Suppose there are n_l instances with the l-th label in the bag, then the weight for each of them will be $1/n_l$ (i.e., by assuming that every instance of the same label has equal importance). $\beta_{i,j}$ is the weight of the j-th instance in the i-th bag. This proposition also satisfies $\boldsymbol{y}_i = \bigvee_{j=1}^{z_i}\hat{\boldsymbol{y}}_{i,j}$. For example, given a bag $\{\boldsymbol{x}_1; \boldsymbol{x}_2; \boldsymbol{x}_3; \boldsymbol{x}_4\}$ [1], if $\boldsymbol{x}_1$ belongs to l_1, $\boldsymbol{x}_2$ and $\boldsymbol{x}_4$ belong to l_2, $\boldsymbol{x}_3$ belong to l_3, then $\boldsymbol{\beta}_i = [1, 0.5, 1, 0.5]^\top$ according to Proposition 1, where $\boldsymbol{x}_2$ and $\boldsymbol{x}_4$ share the same label and make the equal contribution to l_2. The set of $\boldsymbol{\beta}_i$, $i \in \{1, 2, \cdots, m\}$ is denoted as $\tilde{\boldsymbol{\beta}} = \{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m\}$.

We assume that, independent of bags, there exists a prototype for each label, and instances with the same label are close to the label prototype, and far away from the prototypes of other labels. This is the cluster structure assumption.

Proposition 2 *Given the prototype $\boldsymbol{p}_l$ of label l, instance label matrix $\tilde{Y}$ will be a solution (Ding* et al. *2006) to*

$$\min_{G} \sum_{i=1}^{n}\sum_{l=1}^{c} G_{i,l}\|\boldsymbol{a}_i - \boldsymbol{p}_l\|^2 : G \in \{0,1\}^{n\times c}, G^\top G = S, \quad (1)$$

where $S = \mathrm{diag}(\mathbf{1}^\top G)$.

Recall that every instance holds a single label only. Thus, we have the orthogonal constraints in Eqn. (1): $G^\top G = S$, and the non-zero elements in S represents the total number of positive instances for each label.

Proposition 1 enables us to design a bag-level loss, considering the contribution of each instance. Specifically, we use the squared misclassification loss on the known bag labels, and derive the loss as: $\sum_{i=1}^{m}\|\boldsymbol{y}_i - \boldsymbol{\beta}_i^\top \Omega_c(\hat{Y}_i)\|^2$.

Proposition 2 enables us to design a regularization considering the cluster structure. Note that prototype $\boldsymbol{p}$ is unknown, and it is related to the ground truth distribution of each label. Thus, the optimization must avoid calculating $\boldsymbol{p}$. Having defined $H = \tilde{Y}S^{-\frac{1}{2}}$, optimization in Eqn. (1) is transformed (Zha *et al.* 2001) into

$$\max_{H} \ \mathrm{Tr}(H^\top AA^\top H) : H^\top H = I, H \geq 0,$$

where I is an identity matrix. Thus, the regularization term we used is defined as $-\mathrm{Tr}((\tilde{Y}S^{-\frac{1}{2}})^\top AA^\top \tilde{Y}S^{-\frac{1}{2}})$.

Combining the loss and the regularization term together under the non-negative orthogonal constraints, we obtain the following optimization task:

$$\begin{aligned} \min_{\tilde{Y}} \ & \sum_{i=1}^{n}\|\boldsymbol{y}_i - \boldsymbol{\beta}_i^\top \Omega_c(\hat{Y}_i)\|^2 - \lambda\mathrm{Tr}((\tilde{Y}S^{-\frac{1}{2}})^\top AA^\top(\tilde{Y}S^{-\frac{1}{2}})), \\ \text{s.t.} \ & (\tilde{Y}S^{-\frac{1}{2}})^\top(\tilde{Y}S^{-\frac{1}{2}}) = I, \ \tilde{Y} \in \{0,1\}^{n\times(c+k)}, \end{aligned} \quad (2)$$

where λ is a trade-off parameter to be tuned.

[1] To simplify notation, we have dropped the bag subscript i here.

Note that we consider an inductive setting, and reduce Eqn. (2) from a difficult integer optimization to an easier continuous optimization in $[0,1]^{n\times(c+k)}$. Hence, we denote by $W = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{c+k}]$ the parameter to be learned, and define $g_l(\boldsymbol{x}, W) = \exp(\boldsymbol{x}\boldsymbol{w}_l)/\sum_{l'=1}^{c+k}\exp(\boldsymbol{x}\boldsymbol{w}_{l'})$ as the predictive function for instance $\boldsymbol{x}$ on label l, whose output value will lie in range $[0,1]$. Let $\boldsymbol{g} = [g_1, \cdots, g_{c+k}]$, thus we have $\boldsymbol{g}(\boldsymbol{x}, W) = [g_1(\boldsymbol{x}, W), \cdots, g_{c+k}(\boldsymbol{x}, W)]$ and $\boldsymbol{g}(X, W) = [g(\boldsymbol{x}_1, W); \cdots; g(\boldsymbol{x}_z, W)], \boldsymbol{x}_i \in X$. Then $\tilde{Y}$ can be modeled by $\boldsymbol{g}(A, W)$. Substituting $\tilde{Y}$ in Eqn. (2), the final optimization for `DMNL` is given as follows:

$$\begin{aligned} \min_{W} \ & \sum_{i=1}^{m}\|\boldsymbol{y}_i - \boldsymbol{\beta}_i^\top \Omega_c(\boldsymbol{g}(X_i, W))\|^2 \\ & -\lambda\mathrm{Tr}((\boldsymbol{g}(A,W)S^{-\frac{1}{2}})^\top AA^\top(\boldsymbol{g}(A,W)S^{-\frac{1}{2}})), \\ \text{s.t.} \ & (\boldsymbol{g}(A,W)S^{-\frac{1}{2}})^\top(\boldsymbol{g}(A,W)S^{-\frac{1}{2}}) = I. \end{aligned} \quad (3)$$

Remark 1 If a manifold structure (Belkin *et al.* 2006) is used instead of the cluster structure, we can simply replace $-\lambda\mathrm{Tr}((\boldsymbol{g}(A,W)S^{-\frac{1}{2}})^\top AA^\top(\boldsymbol{g}(A,W)S^{-\frac{1}{2}}))$ in Eqn. (3) with $+\lambda\mathrm{Tr}((\boldsymbol{g}(A,W)^\top L\boldsymbol{g}(A,W))$ where L is the laplacian matrix, so as to encourage similar instances to have similar predictive values. The optimization is similar.

Instance-level annotation: Having learned W, instance $\boldsymbol{x}_{i,j}$ is assigned the label with the maximum predictive value, i.e.,

$$\hat{y}_{i,j,l} = \begin{cases} 1, & l = \arg\max_{l'} g_{l'}(\boldsymbol{x}_{i,j}, W), l' \in \{1, \cdots, c+k\}; \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Bag-level prediction for bag i is obtained as: $\bigvee_{j=1}^{z_i}\hat{\boldsymbol{y}}_{i,j}$.

Optimization

The optimization of Eqn. (3) could not be done directly, because: (i) $\tilde{\boldsymbol{\beta}}$ and S are based on $\tilde{Y}$, i.e., the discrete 0-1 matrix derived from Eqn. (4); (ii) a complex function w.r.t W is involved in the non-negative orthogonal constraints.

To handle (i), an alternating optimization strategy is applied: we fix $\tilde{\boldsymbol{\beta}}$ and S and update W; then we use the new W to derive $\tilde{\boldsymbol{\beta}}$ and S. Specifically, given W, $\tilde{Y}$ is derived based on Eqn. (4); then $\tilde{\boldsymbol{\beta}}$ is calculated according to Proposition 1 and S is calculated as $S = \mathrm{diag}(\mathbf{1}^\top \tilde{Y})$.

To deal with (ii), we define a new variable $\hat{H} = g(A, W)S^{-\frac{1}{2}}$. Thus, the constraints are concisely expressed as $\hat{H}^\top\hat{H} = I, \hat{H} \geq 0$ to simplify the optimization.

Let $\hat{H} = [H_1; \cdots; H_m]$, where $H_i = g(X_i, W)S^{-\frac{1}{2}}$. Then, Eqn. (3) can be rewritten as:

$$\min_{W,\hat{H}} \phi(W) + \psi(\hat{H}) : \hat{H}^\top\hat{H} = I, \hat{H} = g(A,W)S^{-\frac{1}{2}}, \quad (5)$$

where $\phi(W) = \frac{1}{2}\sum_{i=1}^{m}\|\boldsymbol{y}_i - \boldsymbol{\beta}_i^\top\Omega_c(g(X_i,W))\|^2$, and $\psi(\hat{H}) = \frac{1}{2}\sum_{i=1}^{m}\|\boldsymbol{y}_i - \boldsymbol{\beta}_i^\top\Omega_c(H_iS^{\frac{1}{2}})\|^2 - \lambda\mathrm{Tr}(\hat{H}^\top AA^\top\hat{H})$. Notice that we consider prediction loss in both $\phi(W)$ and $\psi(\hat{H})$, so as to obtain a better result and a faster convergence.

Algorithm 1 DMNL

Input: D, λ, ρ
Output: W
Process:

Initialize W, $\hat{H}$, Λ, S and $\boldsymbol{\beta}_i, i = 1, \cdots, m$;
repeat:
  Update W via SGD; \\Solve Eqn. (7)
  Update $\hat{H}$ via Eqn. (10); \\ Solve Eqn. (8)
  Update Λ via Eqn. (9);
  Predict $\tilde{Y}$ according to Eqn. (4);
  Calculate $\boldsymbol{\beta}_i, i = 1, \cdots, m$, according to Proposition. 1;
  Calculate $S = \mathrm{diag}(\mathbf{1}^\top \tilde{Y})$;
until convergence or the maximum iteration is reached.

To solve Eqn. (5), we follow an augmented lagrangian optimization framework (Boyd *et al.* 2010). Specifically, the augmented lagrangian of Eqn. (5) is given by:

$$\mathcal{L}(W,\hat{H},\Lambda) = \phi(W) + \psi(\hat{H}) + \frac{\rho}{2}\|\hat{H} - g(A,W)S^{-\frac{1}{2}} + \Lambda\|_F^2 + \zeta,$$

where $\|.\|_F$ is the Frobenius norm, Λ is the dual variable, ρ is the penalty parameter, and ζ is a constant which can be dropped during the optimization. Then, solving Eqn. (5) is equivalent to solving the following optimization problem:

$$\min_{W,\hat{H},\Lambda} \mathcal{L}(W,\hat{H},\Lambda) : \hat{H}^\top \hat{H} = I, \hat{H} \geq 0. \quad (6)$$

We optimize Eqn. (6) w.r.t. W, $\hat{H}$ and Λ in an alternating manner. Let $W^{(t)}$, $\hat{H}^{(t)}$ and $\Lambda^{(t)}$ be the solution of the t-th iteration. We have the update rules as follows:

$$W^{(t+1)} = \arg\min_W \mathcal{L}(W, \hat{H}^{(t)}, \Lambda^{(t)}); \quad (7)$$

$$\hat{H}^{(t+1)} = \arg\min_{\hat{H} \geq 0} \mathcal{L}(W^{(t+1)}, \hat{H}, \Lambda^{(t)}) : \hat{H}^\top \hat{H} = I; \quad (8)$$

$$\Lambda^{(t+1)} = \Lambda^{(t)} + \hat{H}^{(t+1)} - g(A, W^{(t+1)})S^{-\frac{1}{2}}. \quad (9)$$

Algorithm 1 summarizes the procedure.

Update W. In order to efficiently obtain the solution of Eqn. (7), stochastic gradient descent (SGD) is applied. Specifically, solving Eqn. (7) is equivalent to minimizing

$$\mathcal{L}_W = \phi(W) + \frac{\rho}{2}\|\hat{H} - g(A,W)S^{-\frac{1}{2}} + \Lambda\|_F^2.$$

Then we decompose $\mathcal{L}_W = \sum_{i=1}^m \mathcal{L}_W^{(i)}$ according to bags. Specifically, $\phi(W)$ can be naturally written as $\phi(W) = \sum_{i=1}^m \phi_i(W)$ with each $\phi_i(W) = \|\boldsymbol{y}_i - \boldsymbol{\beta}_i^\top \Omega_c(g(X_i, W))\|^2$. Recall that $A = [X_1; \cdots; X_m]$, $\hat{H} = [H_1; \cdots; H_m]$, and we decompose $\Lambda = [\Lambda_1; \cdots; \Lambda_m]$, where Λ_i is the counterpart of X_i. Then $\frac{\rho}{2}\|\hat{H} - g(A,W)S^{-\frac{1}{2}} + \Lambda\|_F^2$ can be decomposed as $\sum_{i=1}^m \frac{\rho}{2}\|H_i - g(X_i,W)S^{-\frac{1}{2}} + \Lambda_i\|_F^2$. Based on the above decompositions, we have

$$\mathcal{L}_W^{(i)} = \phi_i(W) + \frac{\rho}{2}\|H_i - g(X_i,W)S^{-\frac{1}{2}} + \Lambda_i\|_F^2.$$

Let $F_i = g(X_i, W)$. Given $(X_i, \boldsymbol{y}_i, H_i, \Lambda_i)$, the gradient of $\mathcal{L}_W^{(i)}$ w.r.t. W is given by

$$\nabla_W = X_i^\top\big((\boldsymbol{\beta}_i \Delta_{W1}) \circ G_i \circ J_i\big) + \rho X_i^\top((\Delta_{W2} S^{-\frac{1}{2}}) \circ G_i),$$

where $\Delta_{W1} = [\boldsymbol{y}_i, 0^{1\times k}] - \boldsymbol{\beta}_i^\top F_i$, $\Delta_{W2} = H_i + \Lambda_i - F_i S^{-\frac{1}{2}}$, $G_i = F_i \circ (F_i - 1)$ and $J_i = [1^{1\times c}, 0^{1\times k}]$.

We initialize W with $W^{(t)}$, then update W in each iteration $i, (i \in \{1, \cdots, m\})$ via $W \leftarrow W - \eta_i \nabla_W$, where η_i is the step size. Finally, we output the averaged W on all iterations in SGD algorithm as $W^{(t+1)}$.

Update $\hat{H}$. For the optimization task of Eqn. (8), there is an non-negative orthogonal constraints $\hat{H}^\top \hat{H} = I$, $\hat{H} \geq 0$. $\hat{H}^\top \hat{H} = I$ is known as the Stiefel manifold (Boumal *et al.* 2014), thus we try to solve Eqn. (8) on the Stiefel manifold. To simplify the description, we drop the superscript of $\hat{H}^{(t)}$ below. Based on Choi (2008)'s work, we derive the update rule for $\hat{H}$ as

$$\hat{H}^{(t+1)} = \hat{H} \circ \frac{[\nabla_{\hat{H}}]^+ + \hat{H}([\nabla_{\hat{H}}]^-)^\top \hat{H}}{[\nabla_{\hat{H}}]^- + \hat{H}([\nabla_{\hat{H}}]^+)^\top \hat{H}}, \quad (10)$$

where $\nabla_{\hat{H}}$ is the gradient of $\mathcal{L}(W^{(k+1)}, \hat{H}, \Lambda^{(k)})$ w.r.t. $\hat{H}$; $[\nabla_{\hat{H}}]^+$ and $[\nabla_{\hat{H}}]^-$ satisfy $[\nabla_{\hat{H}}]^+ > 0$, $[\nabla_{\hat{H}}]^- > 0$, and $[\nabla_{\hat{H}}] = [\nabla_{\hat{H}}]^+ - [\nabla_{\hat{H}}]^-$.

Note that all of the operations involved in Eqn. (10) are element-wise, i.e., element-wise product and element-wise divide. As a result, it can be divided into blocks according to bags. Specifically, $\nabla_{\hat{H}_i}$ is given by

$$\nabla_{\hat{H}_i} = \boldsymbol{\beta}_i\big(\Delta_{H_i 1} \circ J_i\big) S^{\frac{1}{2}} - \lambda X_i \sum_{j=1}^m X_j^\top H_j + \rho \Delta_{H_i 2},$$

where $\Delta_{H_i 1} = \boldsymbol{\beta}_i^\top H_i S^{\frac{1}{2}} - [\boldsymbol{y}_i, 0^{1\times k}]$, $\Delta_{H_i 2} = H_i - F_i S^{-\frac{1}{2}} + \Lambda_i$, and $J_i = [1^{1\times c}, 0^{1\times k}]$.

To simplify description, let $M_i = X_i \sum_{j=1}^m X_j^\top H_j$. Because some elements in M_i and Λ_i may be negative, we define $\Phi^+(M) = (\mathrm{abs}(M) + \mathrm{abs}(M))/2$ and $\Phi^-(M) = (\mathrm{abs}(M) - \mathrm{abs}(M))/2$, where $\mathrm{abs}(.)$ returns the absolute value on each element, so as to obtain $\Phi^+(M) \geq 0$ and $\Phi^-(M) \geq 0$ satisfying $M = \Phi^+(M) - \Phi^-(M)$. Then $[\nabla_{\hat{H}_i}]^+$ and $[\nabla_{\hat{H}_i}]^-$ is given by

$$[\nabla_{\hat{H}_i}]^+ = \boldsymbol{\beta}_i\big((\boldsymbol{\beta}_i^\top \hat{H} S^{\frac{1}{2}}) \circ J_i\big) S^{\frac{1}{2}} + \lambda \Phi^-(M_i) + \rho(H_i + \Phi^+(\Lambda_i)),$$

$$[\nabla_{\hat{H}_i}]^- = \boldsymbol{\beta}_i\big([\boldsymbol{y}_i, 0^{1\times k}] \circ J_i\big) S^{\frac{1}{2}} + \lambda \Phi^+(M_i) + \rho(F_i S^{-\frac{1}{2}} + \Phi^-(\Lambda_i)).$$

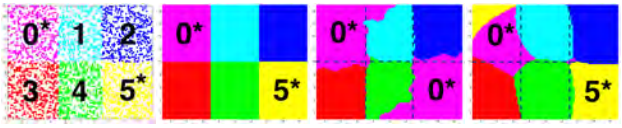
Initialization. Each element in W is set as a random value within $[0, 1]$. Then, we cluster all instances via k-means with $(c + k)$ clusters, and initialize $\tilde{Y}$ with the obtained cluster indicator matrix, whose element of row i column j suggests the ith instance belongs to jth cluster. After that, we set $S = \mathrm{diag}(\mathbf{1}^\top \tilde{Y})$, and calculate $\boldsymbol{\beta}$ according to Proposition 1. Finally, $\hat{H}$ is initialized by $\hat{H} \leftarrow \tilde{Y} S^{-\frac{1}{2}}$.

Experiments on Toy Dataset

Experimental Setting. We randomly generate 300 bags for training from 6 different classes (i.e., 0-5), which corresponds to 6 different colored rectangles in Figure 1(a). Each bag contains 10 instances and possesses 2.47 labels on average. During the training, only class 1-4 are observed in the bag level as known labels, and class 0 and 5 are novel labels (with a star mark in Figure 1) which are unknown in the training data. For testing the instance level annotation

with novel labels, we uniformly sample 10,000 instances from all classes, shown in Figure 1(b). We compare our approach with MIML-NC (Pham *et al.* 2015) which detects all novel instances with a single novel label. The convergence has been validated in experiments.

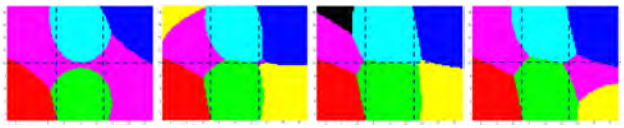
Results. Figures 1(c) and 1(d) show the main results of MIML-NC and DMNL respectively, where dash lines correspond to the boundary of the ground truth. As can be observed, MIML-NC predicts all novel instances as the same label "0", whereas our proposed approach is able to discover multiple novel labels (i.e., both "0" and "5").



(a) Training data (b) Test data (c) MIML-NC (d) DMNL

Figure 1: Toy data experiments

Influence of the Number of Novel Labels k. In practice, the number of novel labels is unknown, thus a number k has to be specified by a user. Figure 2 shows the results of DMNL on the toy dataset by setting different k values.



(a) k=1 (b) k=2 (c) k=4 (d) k=8

Figure 2: Influence of the number of novel labels k

When k matches the ground truth, i.e., $k = 2$, the best detection performance are achieved on both the known labels and the novel labels. This observation suggests that we are able to select k via cross validation according to the detection performance on the known labels.

When k is larger than the ground truth, some novel instances from the same class may be separated into different categories: see the black part in Figure 2(c). Note that the algorithm does not always produce the user-specified k novel classes. This is because the optimization process takes into consideration both the bag-level loss as well as cluster structure. Those detected labels with very few instances (due to the orthogonal constraints) will be regarded as the noise rather than novel labels. Figure 2 shows that $k=4$ produced 3 novel labels; and $k=8$ produced 2 novel labels.

For $k = 1$, the detection performance for multiple novel labels is less than ideal because instances from two different labels are forced to be considered with the same label. The basic assumption is violated: instances with the same label are in the same cluster.

Experiments on Real Datasets

Datasets. The datasets include *MSRCv*$_2$ image dataset (Winn *et al.* 2005), two letter datasets (Briggs *et al.* 2012) (i.e., *Letter Carroll* and *Letter Frost*), and the *MNIST* handwritten dataset (LeCun *et al.* 1998) [2]. Note that *MNIST* is a single-instance single-label dataset taken from 10 digits. In order to transform it to a MIML format, we randomly sample 200 bags from the 10 digits, resulting each bag having 27.6 instances and 3.09 labels on average.

Experimental Setting. We follow the same setting as used in Pham *et al.* (2015): the class labels are split into known and unknown labels (i.e. novel labels) in each dataset. Specifically, we consider 3 different types of splits: the 1st-4th, the 1st-8th and the 1st-16th labels are taken as novel labels for the first 3 datasets; and the 1st-2nd, the 1st-4th and the 1st-6th labels are treated as novel labels for MINIST, since this dataset contains only 10 labels. Then the novel labels are removed for training.

The evaluation is conducted on following aspects: (A) instance-level annotation, including (A1) discovering different novel labels; (A2) detecting instances with novel labels; and (A3) instance annotation on known labels; (B) bag-level prediction, including (B1) prediction for multiple novel labels; and (B2) prediction for known labels. Note that, there are no existing evaluation metrics for multiple novel labels discovery, thus we have to design a new measure for instance-level annotation and bag-level prediction.

Specifically, in order to evaluate the performance in A1, we define a new metric $\mathrm{F_{INL}}$. Let h_i and t_i be the predicted label index and ground truth label index, respectively, for instance $\boldsymbol{x}_i$. By ordering known labels before novel labels, $\mathrm{F_{INL}}$ measure is defined as:

$$\mathrm{F_{INL}} = 2\mathrm{Prec_{INL}}\mathrm{Rec_{INL}}/(\mathrm{Prec_{INL}} + \mathrm{Rec_{INL}});$$

$$\mathrm{Prec_{INL}} = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(h_i=h_j)\mathbb{I}(t_i=t_j)\mathbb{I}(t_i>c)\mathbb{I}(h_i>c)}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(h_i=h_j)\mathbb{I}(h_i>c)};$$

$$\mathrm{Rec_{INL}} = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(h_i=h_j)\mathbb{I}(t_i=t_j)\mathbb{I}(t_i>c)\mathbb{I}(h_i>c)}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{I}(t_i=t_j)\mathbb{I}(t_i>c)},$$

where c is the number of known labels. $\mathrm{Prec_{INL}}$ measures the fraction that pairs of instances from the same discovered novel label are the real pairs from the same novel label; $\mathrm{Rec_{INL}}$ measures the fraction of pairs, that match with the same novel label, have been discovered; $\mathrm{F_{INL}}$ is the combination of $\mathrm{Prec_{INL}}$ and $\mathrm{Rec_{INL}}$.

For B1, F_{INL} measure cannot be applied, because a bag may possess multiple novel labels whereas an instance at most holds one. Instead, we define $\mathrm{F_{BNL}}$ for bag-level evaluation of multiple novel labels. Let $G \in \{0,1\}^{n\times(c+k)}$ denote the predicted label matrix for bags, $Y \in \{0,1\}^{n\times(c+k')}$ be the groundtruth, and $G_{:,l}$ is the lth column of G. Define $\mathcal{F}(\boldsymbol{y},\boldsymbol{g})$ as f-measure function, where $\boldsymbol{y}$ is a predicted label vector, and $\boldsymbol{g}$ is the groundtruth vector. By ordering known labels before novel labels, $\mathrm{F_{BNL}}$ is given by:

$$\mathrm{F_{BNL}} = \frac{1}{k}\sum_{i=1}^{k}\max(\{\mathcal{F}(G_{:,c+i}, Y_{:,c+j}), j\in\{1,\cdots,k\}\}),$$

which measures the average performance on detected multiple novel labels on the ground-truth label that best matches.

[2] We use *MS*, *LC*, *LF* and *MN* as short names for *MSRCv*$_2$, *Letter Carroll*, *Letter Frost* and *MNIST*, respectively.

Both F_{INL} and F_{BNL} are the higher the better.

For the evaluation in other aspects, existing metrics can be applied. A2 is exactly the same task as that in (Pham *et al.* 2015), thus we take the same metric (AUC) for evaluation. For A3, accuracy is applied to evaluate the performance on instance annotation of known labels. For B2, hamming loss is used to evaluate bag-level performance on known labels.

Three state-of-the-art MIML approaches are used as baselines: ORLR (Pham *et al.* 2014) , MIML_{fast} (Huang *et al.* 2014) and MIML-NC (Pham *et al.* 2015). The first two baselines, which do not directly handle novel labels, assign an instance to a novel label if all the predictive values are lower than a user-specified threshold (Lou *et al.* 2013). Because all baselines are unable to discover multiple novel labels directly, kmeans is employed to cluster the detected novel instances into k groups in the post-processing. Besides, we implement a variant of DMNL based on manifold assumption (see Remark 1) [3].

All parameters in each approach are tuned via 5-fold cross validation on the training set on the known labels in bag level, except the number of novel labels k in the baselines. As training data has no information about novel labels, we set the same k for all baselines as that tuned in DMNL, for a fair comparison.

Results on Instance-Level Annotation. Performance from three aspects (A1, A2 & A3) are evaluated.

For A1, i.e, discovery of multiple novel labels, the F_{INL} results are exhibited in Table 2. "•" indicates that our approach is significantly better than the baseline (paired t-tests at 95% significance level), and the best average performance is represented in bold. Without exception, DMNL achieves the best performance on all datasets.

DMNL performs better than all baselines because they must take additional procedures outside the optimization process to discover multiple novel labels (i.e., clustering for all three baselines, and thresholding for ORLR and MIML_{fast}). These procedures do not take bag information into consideration. In contrast, DMNL establishes a unified model for both known labels and multiple novel labels, and considers both the cluster structure among all instances and the contribution of each instance in a bag. Note that F_{INL} is conservative measure, i.e., a positive contribution is counted only if both instances in a pair are correctly predicted to be novel instances of the same label; otherwise a negative contribution is counted. As a result, a small difference in F_{INL} suggests a big gap between the approaches. This is also the reason why F_{INL} is small in Table 2.

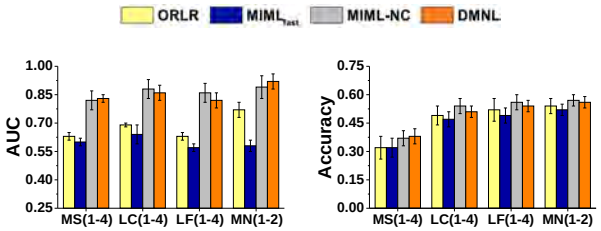
In terms of the ability to detect instances with any novel label (A2), Figure 3(a) summarizes the AUC results. As observed, DMNL achieves significant better performance than ORLR and MIML_{fast} and a comparable performance to MIML-NC, which is the state-of-the-art novel instance detection approach for MIML setting.

In terms of instance annotation on known labels (A3), Figure 3(b) shows that DMNL performs comparably to MIML-NC, and sightly better than ORLR and MIML_{fast}.

[3]It achieves comparable performance to DMNL. Details will be presented in a longer version due to the space limit.

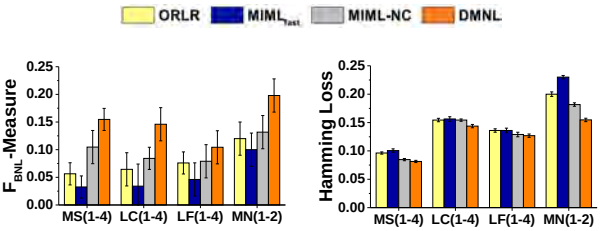
Table 2: F_{INL} results for discovering multiple novel labels

	ORLR	MIML_{fast}	MIML-NC	DMNL
MS(1-4)	.13±.02•	.11±.01•	.19±.02•	**.22±.02**
MS(1-8)	.12±.01•	.10±.02•	.16±.03•	**.20±.02**
MS(1-16)	.12±.02•	.08±.04•	.14±.03•	**.20±.03**
LC(1-4)	.15±.02•	.14±.04•	.24±.02•	**.31±.04**
LC(1-8)	.15±.02•	.14±.04•	.23±.04	**.28±.04**
LC(1-16)	.15±.02•	.14±.05•	.21±.03•	**.26±.03**
LF(1-4)	.15±.02•	.13±.05•	.19±.04•	**.26±.03**
LF(1-8)	.13±.04•	.12±.05•	.21±.03•	**.27±.04**
LF(1-16)	.13±.04•	.13±.05•	.19±.05	**.22±.05**
MN(1-2)	.13±.04•	.11±.05•	.15±.05	**.19±.04**
MN(1-4)	.19±.05•	.17±.05•	.24±.05	**.26±.04**
MN(1-6)	.16±.05•	.15±.03•	.19±.05•	**.25±.03**



(a) AUC for A2 (b) Accuracy for A3

Figure 3: Results on instance-level annotation (A2 and A3)



(a) F_{BNL} for B1 (b) Hamming Loss for B2

Figure 4: Results on bag-level prediction (B1 and B2)

This shows that DMNL, though considering multiple novel labels, does not degrade the performance on known labels.

Results on Bag-Level Prediction. Figure 4 summarizes the bag-level prediction results on B1 and B2. As expected, DMNL surpasses all baselines in terms of the ability to discover multiple novel labels on bag level (B1). In terms of bag-level prediction on known labels (B2), DMNL has better results than ORLR and MIML_{fast}, which support the argument of Pham *et al.* (2015) that modeling novel instances in MIML can improve the prediction results of known labels. Due to DMNL's ability to simultaneously model both known and novel labels, it achieves even better performance than MIML-NC.

Table 3: Runtime comparison results (in second)

	MS	LC	LF	MN
MIML-NC	206	166	86	176
DMNL	17	10	9	17

Runtime Comparison. We compare DMNL with MIML-NC in terms of runtime. The results are shown in Table 3. It can be observed that DMNL (in MATLAB) achieves about 10 times faster than MIML-NC (with C implementations in some parts).

Conclusion

We presented the first model for discovering and predicting multiple novel labels in MIML learning. The proposed discriminative model has the following unique feature: the problem is formulated as a non-negative orthogonal constrained optimization problem that has a bag-dependent loss term and a clustering regularization term which is bag-independent. This enables both the known labels and the multiple novel labels to be simultaneously modeled. Experiments results validate the effectiveness and the efficiency of our approach on discovering and predicting multiple novel labels in MIML learning.

References

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(11):2399–2434, 2006.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Machine Learning*, 3(1):1–122, 2010.

F. Briggs, X. Z. Fern, and R. Raich. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 534–542, 2012.

S. Choi. Algorithms for orthogonal nonnegative matrix factorization. In *Proceedings of 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1828–1832. IEEE, 2008.

Q. Da, Y. Yu, and Z.-H. Zhou. Learning with augmented cass by exploiting unlabeled data. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1760–1766, 2014.

C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 126–135, 2006.

S.-J. Huang, W. Gao, and Z.-H. Zhou. Fast multi-instance multi-label learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1868–1874, 2014.

I. Kuzborskij, F. Orabona, and B. Caputo. From n to n+1: multiclass transfer incremental learning. In *Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3358–3365, 2013.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Q. Lou, R. Raich, F. Briggs, and X. Z. Fern. Novelty detection under multi-label multi-instance framework. In *Proceedings of the 23rd IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2013.

N. Nguyen. A new svm approach to multi-instance multi-label learning. In *Proceedings of the 10th IEEE International Conference on Data Mining*, pages 384–392, 2010.

A. T. Pham, R. Raich, and X. Z. Fern. Dynamic programming for instance annotation in multi-instance multi-label learning. *arXiv preprint arXiv:1411.4068*, 2014.

A. T. Pham, R. Raich, X. Z. Fern, and J. P. Arriaga. Multi-instance multi-label learning in the presence of novel class instances. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2427–2435, 2015.

J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, pages 1800–1807, 2005.

S.-J. Yang, Y. Jiang, and Z.-H. Zhou. Multi-instance multi-label learning with weak label. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1862–1868, 2013.

H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon. Spectral relaxation for k-means clustering. In *Proceedings of Advances in Neural Information Processing Systems 13*, pages 1057–1064. 2001.

M.-L. Zhang and Z.-H. Zhou. M3miml: A maximum margin method for multi-instance multi-label learning. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 688–697, 2008.

M.-L. Zhang. A k-nearest neighbor based multi-instance multi-label learning algorithm. In *Proceedings of 22nd IEEE International Conference on Tools with Artificial Intelligence*, pages 207–212, 2010.

Z.-H. Zhou and X.-Y. Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *Proceedings of Advances in Neural Information Processing Systems 19*, pages 1609–1616. 2007.

Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedigns of the 26th International Conference on Machine Learning*, pages 1249–1256, 2009.

Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.

Z.-H. Zhou. Learnware: On the future of machine learning. *Frontiers of Computer Science*, 10(4):589–590, 2016.