Multi-Instance Multi-Label Learning with Weak Label*

Shu-Jun Yang, Yuan Jiang, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{yangsj, jiangy, zhouzh}@lamda.nju.edu.cn

Abstract

Multi-Instance Multi-Label learning (MIML) deals with data objects that are represented by a bag of instances and associated with a set of class labels simultaneously. Previous studies typically assume that for every training example, all positive labels are tagged whereas the untagged labels are all negative. In many real applications such as image annotation, however, the learning problem often suffers from *weak label*; that is, users usually tag only a part of positive labels, and the untagged labels are not necessarily negative. In this paper, we propose the MIMLwel approach which works by assuming that highly relevant labels share some common instances, and the underlying class means of bags for each label are with a large margin. Experiments validate the effectiveness of MIMLwel in handling the weak label problem.

1 Introduction

Conventional supervised learning often assumes that an object is represented by a single-instance and associated with one class label. In a recent learning framework, Multi-Instance Multi-Label learning (MIML) [Zhou and Zhang, 2006; Zhou *et al.*, 2012], an object is allowed to be represented by a *bag* of instances and associated with multiple class labels simultaneously. This framework has been found useful in diverse tasks especially those involving complicated data objects, such as image annotation [Zha *et al.*, 2008], text categorization [Zhou *et al.*, 2012], video annotation [Xu *et al.*, 2012], acoustic classification [Briggs *et al.*, 2012], bioinformatics [Li *et al.*, 2009c], etc.

Many MIML algorithms have been developed during the past few years. To name a few, Zhou and Zhang [2006] proposed the MIMLSVM and the MIMLBoost by degenerating MIML to multi-label learning and multi-instance learning, respectively; Zha *et al.* [2008] addressed MIML problem with hidden conditional random field; Yang *et al.* [2009] tackled MIML with probabilistic generative models; Nguyen [2010] solved MIML problem through estimating instance labels; Briggs *et al.* [2012] optimized rankloss for each bag. MIML metric learning has also been studied in [Jin *et al.*, 2009]. All these studies assumed that the labels for training examples, both positive and negative, are all given explicitly in advance; in other words, the complete label assignments for each training example are available.

In many real applications, however, the label assignments are generally given by users, and it is often difficult to expect the users to specify all labels for every example. It is very likely that, for example, in image annotation, users only tag a subset of positive labels for an image, and therefore, the untagged labels are not necessarily negative. This setting is called *weak label*, which has only been studied in two pieces of work on multi-label learning: Sun *et al.* [2010] proposed the WELL approach by assuming that the similarities between instances are derived from a group of low-rank base similarities; Bucak et al. [2011] proposed the MLR-GL approach by exploiting group lasso to combine ranking errors. Note that these two studies focused on single-instance data, whereas to the best of our knowledge, the MIML weak label setting has never been touched.

MIML transductive learning and semi-supervised learning have been studied before [Feng and Xu, 2010; Xu *et al.*, 2012]. Both assumed that the labeled MIML examples are with complete label assignments, whereas unlabeled data are without any label information; this is related but different from the weak label setting because in these settings, once we get tagged labels for an example, we can conclude that all the untagged labels are negative.

In this paper, we propose the MIMLwel (MIML with WEak Label) approach for the weak label setting. Our basic assumption is that highly relevant labels generally share common instances, and the underlying class means of bags for each label are with a large margin. Specifically, we first explore a mapping from a bag of instances to a feature vector where each element measures the degree of the bag being associated with a group of similar instances. Then, we employ sparse predictors to learn the labels of bags such that the class means of bags for each label is maximized. We formulate the problem in a general framework and provide an efficient block coordinate descent solution. The effectiveness of the MIMLwel approach on handling the weak label problem is validated in experiments.

In the following we will present the MIMLwel approach

*This research was supported by NSFC (61273301, 61073097), 973 Program (2010CB327903) and Baidu Fund (181315P00441).

in Section 2, and then report our experiments in Section 3. Finally, we conclude the paper in Section 4.

2 The MIMLwel Approach

In the original MIML setting [Zhou and Zhang, 2006; Zhou *et al.*, 2012], we are given a training data set $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)\}$, where X_i is a bag containing n_i instances $\{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,n_i}\}$, $Y_i = [y_{i,1}, \ldots, y_{i,L}] \in \{0,1\}^L$ is a label vector containing L labels, where $y_{i,l} = +1$ if the lth label is positive for X_i, and 0 otherwise. Note that the labels of instances $\mathbf{x}_{i,j}$'s ($i = 1, \ldots, m; j = 1, \ldots, n_i$) are unknown.

In MIML weak label setting, however, only a subset of labels are tagged. Specifically, for X_i, we are given a label vector $\hat{Y}_i = [\hat{y}_{i,1}, \ldots, \hat{y}_{i,L}]$, where $\hat{y}_{i,l} = 1$ if the lth label is tagged for X_i, and 0 otherwise. The goal is to predict all the positive labels for unseen bags.

We first consider a mapping from a bag of instances to a feature vector, and then represent each bag X_i by $\phi^{\mathbf{C}}(X_i)$, where $\mathbf{C}$ are prototypes. With such a mapping, each bag is re-represented by a single feature vector, and thus classical single-instance learning algorithms can be applied. The prototypes $\mathbf{C}$ can be generated by a clustering process, e.g., k-means; however, such a process may result in a suboptimal performance. To address this problem, our proposed approach explicitly takes into account the learning of $\mathbf{C}$.

2.1 The Formulation

A straightforward strategy to deal with the weak label setting is to decompose the problem into L independent binary classification problems, each corresponding to a label; for the ith label, training data with the ith label tagged are regarded as positive training examples, whereas the untagged data are regarded as unlabeled data. Then, each of these problems can be addressed by PU-learning (Positive and Unlabeled learning) [Liu *et al.*, 2003]. Such a strategy, however, ignores useful information concealed in label relations and often leads to a suboptimal performance. It also requires a high computational workload, particularly when there are a large number of training bags. Furthermore, existing PU-learning algorithms were mostly designed for single-instance data, whereas the training data for each of these L binary classification problems are bags rather than single-instances; therefore, existing PU-learning algorithms could not be applied directly.

Considering that existing approaches could not be applied to MIML weak label setting directly, we propose the MIMLwel approach. In order to take label relations into account, we assume that the highly relevant labels often share common instances. Moreover, in order to improve efficiency, particularly when there are a large number of bags, we employ an efficient strategy by assuming that the underlying class means of bags for each label are separated with a large margin.

For simplicity, we employ L linear models one for each label, i.e.,$f_l(X) = \mathbf{w}_l^T \Phi^{\mathbf{C}}(X)$ where each $\mathbf{w}_l$ is a d- dimensional linear predictor $[w_{l,1}, \ldots, w_{l,d}]^T$ and $\mathbf{w}_l^T$ denotes the transpose of $\mathbf{w}_l$. To exploit label relations, we consider a label relation matrix $R \in [0,1]^{L \times L}$, where $R_{l,\tilde{l}} = 1$ if the labels l and $\tilde{l}$ are related, and 0 otherwise. Let $\mathbf{W}_{l,\tilde{l}}$ denote $[\mathbf{w}_l, \mathbf{w}_{\tilde{l}}]$ for the pair of related labels $(l, \tilde{l})$. Inspired by [Argyriou *et al.*, 2008], we assume that highly related labels often share common instances, implying that many rows of $\mathbf{W}_{l,\tilde{l}}$ should equal to zero; this can be characterized by a convexly relaxed term $\|\mathbf{w}(l,\tilde{l})\|_{(2,1)}$, which is a convex relaxation of $\|\mathbf{w}(l,\tilde{l})\|_{(2,0)}$. Thus, our goal is to find $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]$ and an output matrix $\bar{\mathbf{Y}}$ such that:

$$\begin{aligned} \min_{\mathbf{W}, \bar{\mathbf{Y}}} \; & -\eta \sum_{l=1}^{L} V(\{\bar{y}_{i,l}, X_i\}_{i=1}^m, \mathbf{w}_l) + \sum_{1 \le l, \tilde{l} \le L} R_{l,\tilde{l}} \|\mathbf{W}_{l,\tilde{l}}\|_{2,1}^2 \\ \text{s.t.} \; & |\bar{Y}_l - \hat{Y}_l|_1 / |\hat{Y}_l|_1 \le \epsilon; \\ & \bar{y}_{i,l} = \hat{y}_{i,l} \text{ if } \hat{y}_{i,l} = 1, \forall l = 1, \ldots, L. \end{aligned} \tag{1}$$

where V is a loss function for each label, $|\cdot|_1$ stands for the l_1-norm, ϵ controls the sparsity of $|\bar{Y}_l - \hat{Y}_l|_1$, and η trades off the empirical risk and model complexity. Typically, V can be defined as a sum of losses on each bag; this involves label estimation for each bag and may be computationally costly, especially when there are a large number of bags. Recently, Li *et al.* [2009b] indicated that the estimation of a simpler statistic, i.e., the underlying class means of bags for each label, can be used for an effective approximation. This motivates us to define $V(\{\bar{y}_{i,l}, X_i\}_{i=1}^m, \mathbf{w}_l)$ as

$$\frac{\sum_{i=1}^m \mathbf{w}_l^T \Phi^{\mathbf{C}}(X_i) \bar{y}_{i,l}}{\sum_{i=1}^m \bar{y}_{i,l}} - \frac{\sum_{i=1}^m \mathbf{w}_l^T \Phi^{\mathbf{C}}(X_i)(1 - \bar{y}_{i,l})}{\sum_{i=1}^m (1 - \bar{y}_{i,l})}, \tag{2}$$

which implies that the class means of bags for each label are separated with a large margin.

Here, inspired by [Zhou and Zhang, 2006; Zhou *et al.*, 2012], we consider to use label specified prototypes to initialize $\mathbf{C}$, and the definition of $\Phi^{\mathbf{C}}(X)$ can be:

$$\Phi^{\mathbf{C}}(X) = [s(X, \mathbf{c}_1^1), \ldots, s(X, \mathbf{c}_{r_1}^1), s(X, \mathbf{c}_1^2), \ldots, s(X, \mathbf{c}_{r_L}^L)],$$

where $\mathbf{C} = [\mathbf{c}_1^1, \mathbf{c}_2^1, \ldots, \mathbf{c}_{r_L}^L]$ are prototypes and $s(X, \mathbf{c})$ is a similarity function. Specifically, $[\mathbf{c}_1^l, \ldots, \mathbf{c}_{r_l}^l]$ are prototypes for the lth label, and r_l is set as $0.1|Y_l|$ in our experiments. Here, we measure the similarity by Gaussian Hausdorff function as in [Zhang and Wang, 2009], i.e., $s(X, \mathbf{c}) = \min_{\mathbf{x} \in X} \exp(-\|\mathbf{x} - \mathbf{c}\|^2 / \delta)$ where δ is set to the average distance between all pairs of instances. Notice that other alternative implementations are also feasible.

Learning the Prototypes C

So far we assume that the prototypes $\mathbf{C}$ are available; as aforementioned, this may lead to a suboptimal performance because it leaves the learning model out of account. One alternative way is to learn $\{\mathbf{W}, \bar{\mathbf{Y}}\}$ and $\mathbf{C}$ simultaneously, which can be formulated as

$$\begin{aligned} \min_{\mathbf{W}, \bar{\mathbf{Y}}, \mathbf{C}} \; & -\eta \sum_{l=1}^{L} V(\{\bar{y}_{i,l}, X_i\}_{i=1}^m, \mathbf{w}_l) + \sum_{1 \le l, \tilde{l} \le L} R_{l,\tilde{l}} \|\mathbf{W}_{l,\tilde{l}}\|_{2,1}^2 \\ & + \beta \Theta(\{X_i\}_{i=1}^m, \mathbf{C}) \\ \text{s.t.} \; & |\bar{Y}_l - \hat{Y}_l|_1 / |\hat{Y}_l|_1 \le \epsilon; \\ & \bar{y}_{i,l} = \hat{y}_{i,l} \text{ if } \hat{y}_{i,l} = 1, \forall l = 1, \ldots, L. \end{aligned} \tag{3}$$

Algorithm 1 MIMLwel

Input: $\{X_i, \hat{Y}_i\}_{i=1}^m, R, \eta, \beta, \epsilon$;
Output: $\mathbf{W}$, $\bar{\mathbf{Y}}$ and $\mathbf{C}$

```
Perform clustering for the positive bags on each label to initialize prototypes C;
while not converged do
    while not converged do
        Fix C and Ȳ, update W ← Eq. 5;
        Fix C and W, update Ȳ ← Eq. 7;
    end while
    Fix W and Ȳ, update C ← Eq. 8;
end while
```

Here, β is a parameter and $\Theta(\{X_i\}_{i=1}^m, \mathbf{C})$ can be realized by any clustering objective. For simplicity, we employ k-means and consider the positive bags to construct the prototypes as [Zhang and Wang, 2009], i.e.,

$$\Theta(\{X_i\}_{i=1}^m, \mathbf{C}) = \sum_{l=1}^{L} \sum_{i=1}^{m_l} \min_{j=1,\ldots,r_l} \{H(X_{i+}^l, \mathbf{c}_j^l)\}, \quad (4)$$

where $\{X_{i+}^l\}_{i=1}^{m_l}$ are the positive bags for the lth label, and $H(X_i, \mathbf{c}) = \sum_{j=1}^{n_i} \|\mathbf{x}_{i,j} - \mathbf{c}\|^2 / n_i$ is the average Euclidean distance.

2.2 Block Coordinate Descend Algorithm

The objective function in Eq. 3 involves $\mathbf{W}$, $\bar{\mathbf{Y}}$ and $\mathbf{C}$, and it is not easy to optimize with respect to all the variables simultaneously. Here we extend an efficient block coordinate descend [Tseng, 2001] algorithm. Specifically, we first optimize the objective function with respect to $\mathbf{W}$ when $\bar{\mathbf{Y}}$ and $\mathbf{C}$ are fixed, then optimize it with respect to $\bar{\mathbf{Y}}$ when $\mathbf{W}$ and $\mathbf{C}$ are fixed, and finally optimize it with respect to $\mathbf{C}$ when the first two are fixed. These three procedures are repeated until convergence. Algorithm 1 summarizes the pseudo-code of MIMLwel. In the following, we will present these three procedures.

Fix C and $\bar{\mathbf{Y}}$, Update W

When $\mathbf{C}$ and $\bar{\mathbf{Y}}$ are fixed, note that the term $\Theta(\{X_i\}_{i=1}^m, \mathbf{C})$ is not related to $\mathbf{W}$ and thus, we need to solve the following optimization problem:

$$\min_{\mathbf{W}} -\eta \sum_{l=1}^{L} V(\{\bar{y}_{i,l}, X_i\}_{i=1}^m, \mathbf{w}_l) + \sum_{1 \le l, \tilde{l} \le L} R_{l,\tilde{l}} \|\mathbf{W}_{l,\tilde{l}}\|_{2,1}^2, \quad (5)$$

which is convex for $\mathbf{W}$. To deal with the non-smoothness of $\|\mathbf{W}_{l,\tilde{l}}\|_{2,1}^2$, according to [Argyriou *et al.*, 2008], we have,

$$\|\mathbf{W}_{l,\tilde{l}}\|_{2,1}^2 = \min_{\boldsymbol{\lambda}^{(l,\tilde{l})} \in \mathcal{M}} \mathbf{w}_l^T D_{l,\tilde{l}}^+ \mathbf{w}_l + \mathbf{w}_{\tilde{l}}^T D_{l,\tilde{l}}^+ \mathbf{w}_{\tilde{l}}$$

where $\mathcal{M} = \{\boldsymbol{\lambda} | \lambda_i \ge 0, \sum_{i=1}^d \lambda_i \le 1\}$, $D_{l,\tilde{l}} = Diag(\boldsymbol{\lambda}^{(l,\tilde{l})})$ is a diagonal matrix and $D_{l,\tilde{l}}^+$ denotes the pseudoinverse of $D_{l,\tilde{l}}$. Eq. 5 can be rewritten as,

$$\begin{aligned} \min_{\mathbf{W}} \min_{\{\boldsymbol{\lambda}^{(l,\tilde{l})}\}_{1 \le l, \tilde{l} \le L}} & -\eta \sum_{l=1}^{L} V(\{\bar{y}_{i,l}, X_i\}_{i=1}^m, \mathbf{w}_l) + \varsigma D_{l,\tilde{l}}^+ \\ & + \sum_{1 \le l, \tilde{l} \le L} R_{l,\tilde{l}} \big(\mathbf{w}_l^T D_{l,\tilde{l}}^+ \mathbf{w}_l + \mathbf{w}_{\tilde{l}}^T D_{l,\tilde{l}}^+ \mathbf{w}_{\tilde{l}}\big). \\ \text{s. t.} \quad & \boldsymbol{\lambda}^{(l,\tilde{l})} \in \mathcal{M} \end{aligned} \quad (6)$$

Here, a small constant ς (e.g., 10^{-3} in our experiments) is introduced to ensure the convergence [Argyriou *et al.*, 2008]. Eq. 6 is jointly-convex for $\mathbf{W}$ and $\{\boldsymbol{\lambda}^{(l,\tilde{l})}\}_{1 \le l, \tilde{l} \le L}$, and thus, one can iteratively solve one of them by fixing the others as constants until convergence. Specifically, when $\{\boldsymbol{\lambda}^{(l,\tilde{l})}\}_{1 \le l, \tilde{l} \le L}$ is fixed and note that $\mathbf{w}_l$'s are decoupled in Eq. 6, Eq. 6 can be addressed by L independent subproblems each corresponding to one $\mathbf{w}_l$ via a simple quadratic programming (QP).

When $\mathbf{W}$ is fixed, according to [Argyriou *et al.*, 2008], $\{\boldsymbol{\lambda}^{(l,\tilde{l})}\}_{l,\tilde{l}=1}^L$ can be solved via a close-form solution, *i.e.*,

$$\boldsymbol{\lambda}^{(l,\tilde{l})} = [\sqrt{w_{l,1}^2 + w_{\tilde{l},1}^2 + \varsigma}/P, \ldots, \sqrt{w_{l,d}^2 + w_{\tilde{l},d}^2 + \varsigma}/P],$$

where $P = \|\mathbf{w}^{(l,\tilde{l})}\|_{2,1} + d\varsigma$. Above procedures are repeated iteratively until convergence.

Fix C and W, Update $\bar{\mathbf{Y}}$

When $\mathbf{C}$ and $\mathbf{W}$ are fixed, note that the second and third terms in Eq. 3 are not related to $\bar{\mathbf{Y}}$ and the $\bar{Y}_l$s are decoupled with respect to the objective and constraints in Eq. 3, Eq. 3 can be addressed by L independent subproblems each corresponding to one label:

$$\begin{aligned} \min_{\bar{Y}_l} & \frac{\sum_{i=1}^m \mathbf{w}_l^T \Phi^{\mathbf{C}}(X_i)(1 - \bar{y}_{i,l})}{\sum_{i=1}^m (1 - \bar{y}_{i,l})} - \frac{\sum_{i=1}^m \mathbf{w}_l^T \Phi^{\mathbf{C}}(X_i) \bar{y}_{i,l}}{\sum_{i=1}^m \bar{y}_{i,l}} \\ \text{s.t.} \ & |\bar{Y}_l - \hat{Y}_l|_1 / |\hat{Y}_l|_1 \le \epsilon; \\ & \bar{y}_{i,l} = \hat{y}_{i,l} \text{ if } \hat{y}_{i,l} = 1, \forall i = 1, \ldots, n. \end{aligned} \quad (7)$$

Let p_i denote $\mathbf{w}_l^T \Phi^{\mathbf{C}}(X_i)$, a_l denote $1/\sum_{i=1}^m (1 - \bar{y}_{i,l})$ and b_l denote $1/\sum_{i=1}^m \bar{y}_{i,l}$. Then, the objective function in Eq. 7 can be rewritten as: $\min_{\bar{Y}_l} \sum_{i=1}^m p_i \{a_l - (a_l + b_l)\bar{y}_{i,l}\}$. We have the following proposition:

Proposition 1 *At optimality, $p_i \ge p_j$ if $\bar{y}_{i,l} \ge \bar{y}_{j,l}$.*

Proof. Assume, to the contrary, that the optimal $\bar{Y}_l$ does not have the same sorted order as p. Then, there are two label elements $\bar{y}_{i,l}$ and $\bar{y}_{j,l}$, with $p_i \ge p_j$ but $\bar{y}_{i,l} \le \bar{y}_{j,l}$. Then $p_i\{a_l - (a_l + b_l)\bar{y}_{i,l}\} + p_j\{a_l - (a_l + b_l)\bar{y}_{j,l}\} \ge p_i\{a_l - (a_l + b_l)\bar{y}_{j,l}\} + p_j\{a_l - (a_l + b_l)\bar{y}_{i,l}\}$, as $(p_i - p_j)(\bar{y}_{i,l} - \bar{y}_{j,l}) \le 0$. Thus, $\bar{Y}_l$ is not optimal, a contradiction. □

According to Proposition 1, Eq. 7 can be solved by sorting. Specifically, let $\mathcal{B} = \{X_i | \hat{y}_{i,l} = 1\}$ and $\bar{\mathcal{B}} = \{X_i | \hat{y}_{i,l} = 0\}$. We first sort the p_i's with respect to the bags in $\bar{\mathcal{B}}$ in a descending order, and then add top-valued bag of $\bar{\mathcal{B}}$ into $\mathcal{B}$ iteratively, until the constraint $|\bar{Y}_l - \hat{Y}_l|_1 \le \epsilon |\hat{Y}_l|_1$ is violated; all these results are candidate solutions, and we finally output the best solution among them according to the minimal objective value in Eq 7.

Table 1: Experimental results (mean± std) on text data. ↑ (↓) indicates the larger (smaller), the better. • (○) indicates the compared method is significantly worse (better) than MIMLwel (pairwise t-tests at 95% significance level).

	W.L.R.	MIMLwel	MIMLwel-D	WELL+	KISAR	MIMLSVM	MIMLBoost	RANKLOSS
HL ↓	10%	.154 ± .001	.156 ± .002 •	.165 ± .005 •	.165 ± .050 •	.189 ± .095 •	.168 ± .028 •	.165 ± .005 •
	20%	.141 ± .002	.144 ± .001 •	.164 ± .004 •	.165 ± .054 •	.185 ± .001 •	.167 ± .003 •	.159 ± .002 •
	30%	.125 ± .002	.134 ± .002 •	.187 ± .008 •	.162 ± .061 •	.183 ± .063 •	.166 ± .002 •	.140 ± .003 •
	40%	.105 ± .002	.124 ± .002 •	.186 ± .007 •	.163 ± .099 •	.181 ± .001 •	.166 ± .003 •	.137 ± .008 •
maF1↑	10%	.060 ± .007	.057 ± .009 •	.002 ± .002 •	.001 ± .001 •	.123 ± .005 ○	.001 ± .002 •	.001 ± .002 •
	20%	.160 ± .012	.146 ± .009 •	.018 ± .005 •	.038 ± .077 •	.138 ± .006 •	.002 ± .005 •	.050 ± .003 •
	30%	.274 ± .018	.215 ± .012 •	.084 ± .006 •	.123 ± .142 •	.148 ± .007 •	.011 ± .010 •	.122 ± .001 •
	40%	.442 ± .017	.304 ± .014 •	.084 ± .007 •	.135 ± .004 •	.160 ± .006 •	.003 ± .005 •	.531 ± .047 ○
miF1↑	10%	.094 ± .007	.100 ± .014 ○	.003 ± .003 •	.002 ± .003 •	.386 ± .002 ○	.001 ± .001 •	.001 ± .001 •
	20%	.276 ± .012	.261 ± .014 •	.018 ± .015 •	.034 ± .069 •	.398 ± .004 ○	.001 ± .003 •	.051 ± .002 •
	30%	.432 ± .011	.370 ± .013 •	.352 ± .018 •	.140 ± .132 •	.404 ± .003 •	.013 ± .012 •	.244 ± .001 •
	40%	.581 ± .012	.462 ± .011 •	.382 ± .008 •	.247 ± .006 •	.411 ± .004 •	.006 ± .011 •	.614 ± .029 ○

Fix $\bar{\mathbf{Y}}$ and $\mathbf{W}$, Update $\mathbf{C}$

When $\bar{\mathbf{Y}}$ and $\mathbf{W}$ are fixed, note that the second term in Eq. 3 is not related to $\mathbf{C}$, we need to solve the following optimization problem:

$$\min_{\mathbf{C}} -\eta \sum_{l=1}^{L} V(\{\bar{y}_{i,l}, X_i\}_{i=1}^{m}, \mathbf{w}_l) + \beta\Theta(\{X_i\}_{i=1}^{m}, \mathbf{C}). \quad (8)$$

Eq. 8, however, is non-convex and non-differentiable. Let $g(\mathbf{C})$ denote the objective function of Eq. 8 and $\mathbf{C}^{(0)}$ denote the current solution. Here, we employ the subgradient method [Shor *et al.*, 1985] to find a refined solution $\bar{\mathbf{C}}$, i.e., $g(\bar{\mathbf{C}}) \leq g(\mathbf{C})$. Let $\nabla g(\mathbf{C})$ denote the subgradient of $\mathbf{C}$,we update $\mathbf{C}^{(u)}$ as, $\forall u = 0, 1, \ldots, v$

$$\mathbf{C}^{(u+1)} = \mathbf{C}^{(u)} - \alpha \bigtriangledown g(\mathbf{C}^{(u)}),$$

where α is the step size and v is the number of iterations; they are set as $10^{-4}/u$ and 10, respectively, in our experiment. Because the subgradient method is not a descent method, it is common to keep track of the best candidate found so far, i.e., the one with the smallest function value,

$$\bar{\mathbf{C}} = \underset{\mathbf{C}^{(u)},\ u=0,\ldots,v}{\arg\min} g(\mathbf{C}^{(u)}).$$

It is evident that we will have $g(\bar{\mathbf{C}}) \leq g(\mathbf{C})$. When $g(\bar{\mathbf{C}}) = g(\mathbf{C})$, the algorithm stops and outputs $\bar{\mathbf{C}}$ as the final result.

3 Experiments

In this section, we first compare MIMLwel with several state-of-the-art MIML algorithms on benchmark data sets, and then evaluate MIMLwel on a real-world image annotation task.

MIMLwel is compared with four state-of-the-art MIML algorithms including MIMLSVM [Zhou and Zhang, 2006], MIMLBoost [Zhou and Zhang, 2006], RANKLOSS [Briggs *et al.*, 2012] and KISAR [Li *et al.*, 2012]. MIMLwel is also compared with the weak label method WELL [Sun *et al.*, 2010]. Note that WELL works for multi-label learning, which could not be applied directly to MIML. For a fair comparison, we first employ the feature mapping learned by MIMLwel, and then apply the WELL approach. Furthermore, WELL works under transductive setting which requires to know the testing instances in advance. In our comparison, we feed the testing data without the labels as well as the training data to WELL , and evaluate the performance of WELL on the test data. Note that this setup is unfair to our proposal because our proposal does not see any test data in the training process. We call the extended version of WELL as WELL+. MIMLwel is further compared with MIMLwel-D, a degenerated version of MIMLwel that does not learn $\mathbf{C}$.

In our experiments we consider four *weak label ratio*s (W.L.R.), defined as $|\hat{\mathbf{Y}}_{.,l}|_1/|\mathbf{Y}_{.,l}|_1$, from 10% to 40% with 10% as the interval.

The compared algorithms are all set to the best parameters recommended in their corresponding papers. For MIMLSVM, the number of clusters is set to 20% of the training bags and the Gaussian kernel width is set to 0.2. For MIMLBoost, the number of boosting rounds is set to 25. For RANKLOSS, the regularization parameter is set to the default value. For KISAR, the number of clusters is set to 50% of training bags the parameter γ is set to default value. For WELL, the parameters α and β are set to default values, and parameter γ is tuned from $\{10^0, \ldots, 10^4\}$ based on the best performance on kernel alignment using five-fold cross-validation on training data. For MIMLwel, the parameters η, β and ϵ are simply fixed to 50, 2 and 1, respectively.

We use three popular multi-label learning evaluation criteria, i.e., Hamming Loss (*HL*), Macro-F1 (*maF1*) and Micro-F1 (*miF1*). Given a testing data set $\mathcal{Q} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_q, Y_q)\}$, and let $h(X_i)$ denote the binary label vector which is predicted by the classifier for X_i.

Hamming Loss(*HL*):

$$HL_{\mathcal{Q}}(h) = \frac{1}{q}\sum_{i=1}^{q}\frac{|h(X_i) \ominus Y_i|_1}{L},$$

where $\ominus$ stands for the symmetric difference of two sets. Hamming Loss is one of the most important criteria for multi-label learning. It evaluates how many times on average a bag-label pair is incorrectly predicted. The smaller the value of hamming loss, the better the performance.

Table 2: Experimental results (mean± std) on image data. ↑ (↓) indicates the larger (smaller), the better. • (○) indicates the compared method is significantly worse (better) than MIMLwel (pairwise t-tests at 95% significance level).

	W.L.R.	MIMLwel	MIMLwel-D	WELL+	KISAR	MIMLSVM	MIMLBoost	RANKLOSS
HL↓	10%	.246 ± .002	.246 ± .001	.250 ± .003 •	.247 ± .121 •	.340 ± .003 •	.250 ± .006 •	.250 ± .005 •
	20%	.231 ± .001	.232 ± .002	.248 ± .006 •	.254 ± .001 •	.339 ± .003 •	.250 ± .001 •	.249 ± .003 •
	30%	.225 ± .001	.232 ± .003 •	.247 ± .010 •	.247 ± .136 •	.338 ± .002 •	.249 ± .002 •	.247 ± .005 •
	40%	.220 ± .003	.225 ± .002 •	.242 ± .008 •	.246 ± .112 •	.337 ± .003 •	.248 ± .001 •	.241 ± .003 •
maF1↑	10%	.007 ± .010	.007 ± .006	.005 ± .002 •	.001 ± .002 •	.125 ± .007 ○	.001 ± .001 •	.001 ± .001 •
	20%	.167 ± .007	.123 ± .007 •	.030 ± .003 •	.025 ± .004 •	.128 ± .006 •	.002 ± .003 •	.022 ± .001 •
	30%	.330 ± .005	.244 ± .008 •	.108 ± .007 •	.116 ± .106 •	.128 ± .004 •	.002 ± .005 •	.151 ± .002 •
	40%	.416 ± .008	.343 ± .007 •	.118 ± .008 •	.006 ± .003 •	.126 ± .007 •	.003 ± .001 •	.459 ± .008 ○
miF1↑	10%	.007 ± .011	.014 ± .009 ○	.004 ± .002 •	.002 ± .003 •	.241 ± .007 ○	.001 ± .002 •	.001 ± .001 •
	20%	.170 ± .006	.191 ± .010 ○	.041 ± .004 •	.025 ± .004 •	.243 ± .006 ○	.003 ± .006 •	.022 ± .001 •
	30%	.333 ± .003	.332 ± .010 •	.113 ± .006 •	.114 ± .103 •	.243 ± .005 •	.006 ± .005 •	.181 ± .002 •
	40%	.425 ± .007	.420 ± .006 •	.155 ± .003 •	.006 ± .002 •	.241 ± .007 •	.008 ± .003 •	.466 ± .003 ○

Table 3: Experimental results (mean± std) on msra data. ↑ (↓) indicates the larger (smaller), the better. • (○) indicates the compared method is significantly worse (better) than MIMLwel (pairwise t-tests at 95% significance level).

	W.L.R.	MIMLwel	MIMLwel-D	WELL+	KISAR	MIMLSVM	RANKLOSS
HL↓	10%	.098 ± .003	.099 ± .002 •	.104 ± .004 •	.101 ± .003 •	.128 ± .036 •	.098 ± .003
	20%	.092 ± .003	.092 ± .004	.097 ± .003 •	.100 ± .004 •	.128 ± .046 •	.093 ± .003 •
	30%	.085 ± .002	.087 ± .003 •	.097 ± .003 •	.100 ± .003 •	.127 ± .039 •	.096 ± .002 •
	40%	.082 ± .003	.083 ± .003 •	.096 ± .003 •	.098 ± .003 •	.126 ± .039 •	.091 ± .003 •
maF1↑	10%	.010 ± .013	.005 ± .011 •	.001 ± .001 •	.005 ± .002 •	.029 ± .034 ○	.041 ± .004 ○
	20%	.058 ± .012	.035 ± .009 •	.031 ± .004 •	.020 ± .005 •	.027 ± .022 •	.044 ± .002 •
	30%	.061 ± .011	.056 ± .006 •	.040 ± .006 •	.024 ± .005 •	.033 ± .018 •	.042 ± .005 •
	40%	.098 ± .008	.086 ± .007 •	.071 ± .003 •	.041 ± .008 •	.034 ± .018 •	.068 ± .007 •
miF1↑	10%	.055 ± .004	.036 ± .041 •	.001 ± .009 •	.018 ± .008 •	.001 ± .002 •	.259 ± .016 ○
	20%	.206 ± .022	.200 ± .038 •	.183 ± .016 •	.069 ± .014 •	.001 ± .018 •	.285 ± .005 ○
	30%	.341 ± .020	.328 ± .025 •	.293 ± .012 •	.079 ± .013 •	.072 ± .038 •	.265 ± .022 •
	40%	.455 ± .014	.443 ± .021 •	.382 ± .020 •	.145 ± .020 •	.072 ± .038 •	.419 ± .025 •

Macro-F1 (*maF1*):

$$maF1_{\mathcal{Q}}(h) = \frac{2}{L}\sum_{l=1}^{L}\frac{\sum_{i=1}^{q} y_{i,l}h_l(X_i)}{\sum_{i=1}^{q} y_{i,l} + \sum_{i=1}^{q} h_l(X_i)},$$

where $h_l(\cdot)$ and $y_{i,l}$ denotes the lth element of $h(\cdot)$ and Y_i, respectively. It can be seen that Macro-F1 calculates F1 measure on individual class labels at first, and then averages over all class labels. Macro-F1 is more affected by the performance of the classes containing less examples . The larger the value of Macro-F1, the better the performance.

Micro-F1 (*miF1*):

$$miF1_{\mathcal{Q}}(h) = \frac{2\sum_{i=1}^{q}\langle h(X_i), Y_i\rangle}{\sum_{i=1}^{q}|h(X_i)|_1 + \sum_{i=1}^{q}|Y_i|_1},$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. It can be seen that Micro-F1 globally calculates the F1 measure on the predictions over all bags and all class labels. Micro-F1 is more affected by the performance of the classes containing more examples. The larger the value of Micro-F1, the better the performance.

Text Categorization

The text categorization data[1] is collected from Reuters-21578 collection [Sebastiani, 2002]. The data set we used here contains 2,000 examples and 7 class labels, where the average number of labels for each instance is 1.15 ± 0.37. The dimension of each instance is 243. For MIMLwel, when experts have domain knowledge about label relations, they can assign values to the elements of R. When lack of domain knowledge, the entries of the label relation matrix R can be set with alternative method, such as the estimation of the concurrence of positive bags, i.e., $R_{l,\tilde{l}} = \mathbf{I}(\sum_{i=1}^{m} I(\hat{Y}_{i,l} = 1 \wedge \hat{Y}_{i,\tilde{l}} = 1) > \theta)$, where $\mathbf{I}$ is identity function and θ is set as the average concurrence between all label pairs. Ten times 10-fold cross-validation is conducted and the average performances are reported in Table 1.

It can be seen that MIMLwel achieves highly competitive performance with compared methods. Specifically, pairwise *t*-tests at 95% significance level indicates that MIMLwel achieves significantly better performance than MIMLwel-D as well as other compared methods in most cases, as shown by the overwhelming •'s in Table 1.

Scene Classification

The scene classification data set[2] contains 2,000 nature scene images and 5 class labels: *desert*, *mountain*, *sea*, *sunset* and *tree*. About 22% of images have multiple labels, and the average number of labels per image is 1.24 ± 0.44. By using the

[1] http://lamda.nju.edu.cn/datacode/miml-text-data.htm

[2] http://lamda.nju.edu.cn/datacode/miml-image-data.htm.

training image	GroundTruth	Tagged Labels	Prediction					
			MIMLwel	MIMLwel-D	WELL+	KISAR	RANKLOSS	MIMLSVM
	people clothing cloud sky water sea nature	people clothing sky	people clothing cloud sky water sea nature *leaf*	people clothing cloud sky water sea	people clothing cloud sky	people clothing cloud sky	people clothing cloud sky	people clothing sky
test image	GroundTruth	Tagged Labels	Prediction					
			MIMLwel	MIMLwel-D	WELL+	KISAR	RANKLOSS	MIMLSVM
	building cloud sea sky water	null	building cloud sea sky water *landscape*	building cloud sky *landscape*	cloud sky *leaf*	cloud sky	water sea	water

Figure 1: Two example images, ground-truth labels, tagged labels, and labels predicted by compared methods

SBN bag generator [Maron and Lozano-Pérez, 1998], each image is represented by a bag of nine 15-dimensional instances each corresponding to a patch in the image. The label relation matrix used in MIMLwel is set as the same as that in text categorization. Ten times 10-fold cross-validation is conducted on this data set and the average performances are reported in Table 2.

It can be seen that MIMLwel achieves highly competitive performance with compared methods. Specifically, pairwise *t*-tests at 95% significance level indicates that the performances of MIMLwel is significantly better than MIMLwel-D and other compared methods in most cases.

Image Annotation

A subset of MSRA-MM database [Li *et al.*, 2009a] is used in this experiment. In this data set, 38 class labels are considered, and there are 1,605 examples in total where around 92% of them are with more than one label, and the average number of labels per image is 3.85 ± 1.75. Each image is represented by a bag of 6-dimensional instances via image segmentation [Wang *et al.*, 2001], where each instance corresponds to the cluster center of one segment. The label relation matrix used in MIMLwel is set as same as that in text categorization. For each trial, we randomly selected 1,400 images for training and used the remaining images for testing. Experiments are conducted for ten times and the average performances and standard deviations are reported in Table 3. MIMLBoost is not listed here because it did not return results within a reasonable time period (24 hours in our experiments) in a single trial.

It can be seen that MIMLwel obtains highly competitive performance with compared methods. Specifically, pairwise *t*-tests at 95% significance level confirms that MIMLwel achieves significantly better performance on this real data with a lot of labels.

Figure 1 presents two example results. The first image has seven ground-truth labels: {people, clothing, cloud, sky, water, sea, nature}. During the training process, only three labels {people, clothing, sky} are given. After training with different methods, the trained model is then used to annotate the image, and thus we get the labels predicted by each method. It can be seen that MIMLwel successfully predicts all the ground-truth labels; however, it predicts one more label, i.e., *leaf*, which is not in the ground-truth. Nevertheless, it is easy to see that the label *leaf* is consistent with the scene semantics and is not a real mistake. MIMLwel-D misses the ground-truth label *nature*. This might be caused by the fact that MIMLwel-D does not consider the learning of prototypes C, and thus, it may lead to a suboptimal performance. WELL+ was originally designed for single-instance learning with weak label, and it predicts some untagged ground-truth labels, but still misses a lot. The other methods KISAR, RANKLOSS and MIMLSVM almost predict as same as the tagged labels, because they were not designed to handle the weak label setting. Similar observations can be found for the second example in Figure 1.

4 Conclusion

Previous studies on multi-instance multi-label learning (MIML) typically assumed that the complete label assignment for all labels are known. In many real applications such as image annotation, however, the learning problem often suffers from *weak label*; that is, users usually tag only a subset of positive labels, and the untagged labels are not necessarily negative. In this paper, we propose the MIMLwel approach by assuming that highly relevant labels generally share common instances, and the class means of bags for each label are with a large margin. We formulate the problem in a general framework and provide an efficient block coordinate descent solution. Experiments show that MIMLwel is superior to state-of-the-art methods in handling the weak label setting.

There are many future works. For example, instead of using a label relation matrix R specified by domain knowledge or estimated by label co-occurrence, inspired by [Huang *et al.*, 2012], it will be interesting to develop better approaches that incorporate a process of learning R.

Acknowledgments: We thank Yu-Feng Li for helpful discussions.

References

[Argyriou *et al.*, 2008] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[Briggs *et al.*, 2012] 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, Beijing, China, 2012.

[Bucak *et al.*, 2011] S.S. Bucak, R. Jin, and A.K. Jain. Multi-label learning with incomplete class assignments. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2801–2808, Spring, CO, 2011.

[Feng and Xu, 2010] S. Feng and D. Xu. Transductive multi-instance multi-label learning algorithm with application to automatic image annotation. *Expert Systems with Applications*, 37(1):661–670, 2010.

[Huang *et al.*, 2012] S.-J. Huang, Y. Yu, and Z.-H. Zhou. Multi-label hypothesis reuse. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 525–533, Beijing, China, 2012.

[Jin *et al.*, 2009] R. Jin, S. Wang, and Z.-H. Zhou. Learning a distance metric from multi-instance multi-label data. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 896–902, Miami, FL, 2009.

[Li *et al.*, 2009a] H. Li, M. Wang, and X.-S. Hua. MSRA-MM 2.0: A large-scale web multimedia dataset. In *Proceedings of the 9th International Conference on Data Mining Workshops*, pages 164–169, Miami, FL, 2009.

[Li *et al.*, 2009b] Y.-F. Li, J.T. Kwok, and Z.-H. Zhou. Semi-supervised learning using label mean. In *Proceedings of the 26th International Conference on Machine Learning*, pages 633–640, Montreal, Canada, 2009.

[Li *et al.*, 2009c] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1445–1450, Pasadena, CA, 2009.

[Li *et al.*, 2012] Y.-F. Li, J.-H. Hu, Y. Jiang, and Z.-H. Zhou. Towards discovering what patterns trigger what labels. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 1012–1018, Toronto, Canada, 2012.

[Liu *et al.*, 2003] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 179–188, Melbourne, FL, 2003.

[Maron and Lozano-Pérez, 1998] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 570–576. MIT Press, Cambridge, MA, 1998.

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

[Sebastiani, 2002] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[Shor *et al.*, 1985] N.Z. Shor, K.C. Kiwiel, and A. Ruszczynski. *Minimization Methods for Non-Differentiable Functions*. Springer, 1985.

[Sun *et al.*, 2010] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou. Multi-label learning with weak label. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 593–598, Atlanta, GA, 2010.

[Tseng, 2001] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[Wang *et al.*, 2001] J. Wang, J. Li, and G. Wiederholdy. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[Xu *et al.*, 2012] X.-S. Xu, Y. Jiang, X. Xue, and Z.-H. Zhou. Semi-supervised multi-instance multi-label learning for video annotation task. In *Proceedings of the 20th ACM Multimedia Conference*, pages 737–740, Nara, Japan, 2012.

[Yang *et al.*, 2009] S.-H. Yang, H. Zha, and B.-G. Hu. Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2143–2150. MIT Press, Cambridge, MA, 2009.

[Zha *et al.*, 2008] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AL, 2008.

[Zhang and Wang, 2009] M.-L. Zhang and Z.-J. Wang. MIMLRBF: RBF neural networks for multi-instance multi-label learning. *Neurocomputing*, 72(16-18):3951–3956, 2009.

[Zhou and Zhang, 2006] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1609–1616. MIT Press, Cambridge, MA, 2006.

[Zhou *et al.*, 2012] 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.