Labeling Complicated Objects: Multi-View Multi-Instance Multi-Label Learning*

Cam-Tu Nguyen[1,3], **Xiaoliang Wang**[1], **Jing Liu**[2] and **Zhi-Hua Zhou**[1]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, 210023, China
[2] National Key Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, China
[3] VNU-University of Engineering and Technology, Hanoi, Vietnam
{nguyenct, zhouzh}@lamda.nju.edu.cn, waxili@nju.edu.cn, jliu@nlpr.ia.ac.cn

Abstract

Multi-Instance Multi-Label (MIML) is a learning framework where an example is associated with multiple labels and represented by a set of feature vectors (multiple instances). In the formalization of MIML learning, instances come from a single source (single view). To leverage multiple information sources (multi-view), we develop a multi-view MIML framework based on hierarchical Bayesian Network, and derive an effective learning algorithm based on variational inference. The model can naturally deal with examples in which some views could be absent (partial examples). On multi-view datasets, it is shown that our method is better than other multi-view and single-view approaches particularly in the presence of partial examples. On single-view benchmarks, extensive evaluation shows that our method is highly competitive or better than other MIML approaches on labeling examples and instances. Moreover, our method can effectively handle datasets with a large number of labels.

Introduction

The last decade has witnessed the development of machine learning to address not only bigger but also more complicated data. Multi-instance multi-label learning (MIML) (Zhou and Zhang 2007; Zhou et al. 2012) provides a natural formulation for complicated objects, where each *example* is represented by *a bag of instances*, and associated with *multiple labels* simultaneously. MIML has a nice property that it allows us to discover labels for examples and instances while, during training, we only need labels for examples, not labels for individual instances. This learning setting is prevalent in practice; for example, a document is often represented by a bag of words, an image can be considered as a bag of regions, and a gene sequence can be treated as a bag of sub-sequences. Annotating these types of objects with multiple labels gives rise to many MIML problems such as image classification and annotation (Zha et al. 2008; Nguyen, Zhan, and Zhou 2013), gene pattern annotation (Li et al. 2012b), relation extraction (Surdeanu et al. 2012), etc.

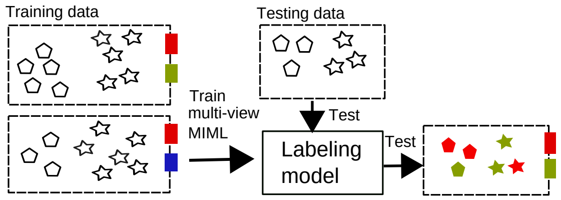


Figure 1: Multi-view MIML learning: examples/bags (e.g. videos) are represented by multi-instances of multi-views (e.g: stars=sound segments/polygons=picture frames); labels (colors) are attached to bags during training.

The original MIML setting only deals with situations where data comes from a single feature set (single-view). For many complicated data, however, it is quite difficult for a single feature set to capture the information required to label a large number of categories. It is thus natural to consider the possibility of leveraging the usefulness of multiple feature sets (multi-view). For example, one can make use of visual and audio signals to label videos; or exploit captions and visual contents to annotate images.

Given the aforementioned scenarios, we formalize the problem of Multi-view Multi-Instance Multi-Label learning as follows. Let $\mathcal{Y}=\{y_l|l=\ldots L\}$ denote a set of L labels, and $\mathcal{D}=\{(X_n,Y_n)|n=1\ldots N\}$ denote a training dataset, where the n-th example X_n is represented by a bag of instances from V views, and $Y_n=\{y_{nl}|l=1\ldots L_n\}\subset\mathcal{Y}$ is the set of L_n (bag) labels of the n-th example. Here, we have $X_n=\{\mathbf{x}_{nvm}|v=1\ldots V, m=1\ldots M_{nv}\}$ where M_{nv} indicates the number of instances in the v-th view of the n-th example, and each instance $\mathbf{x}_{nvm}\in R^{D_v}$ is represented by a D_v dimensional feature vector of view v. The goal of multi-view MIML learning is to predict bag labels $Y_{n'}$ for an unseen example $X_{n'}$ along with labels for its individual instances $\mathbf{x}_{n'vm}$ in V views (see Fig. 1). Additionally, it is desirable for a multi-view approach to work on partial examples, i.e., examples with no instance in some views. This is because multi-view datasets are often corrupted and may have missing values in realistic scenarios. For instance, on sensory datasets, some input signals such as visual, auditory might be missing due to some corruption from environment.

In general, one can simply learn an MIML model in each view separately, and combine the outputs of the single-view

*This research was supported by the 973 Program (2010CB327903) and NSFC (61223003, 61370028).

MIML models, where the learning of single-view MIML models can rely on any previous approach (Zhou et al. 2012). However, by this way the learning in each single-view can not make full use of available information. Moreover, since MIML explores structures among both instances and labels, the effect of utilizing full information could be more significant than that in the traditional single-instance single-label scenario. As a result, it is beneficial to consider the multi-view MIML problem as a whole.

In this paper, we propose a method named Multi-Instance Multi-Label Mixture (MIMLmix) based on hierarchical Bayesian network, where labels are assumed to be sampled from "topics" that capture label relationships; and instances (from multi-views) are sampled from a mixture model where mixture components are representations of labels in multi-views. In continuous feature spaces (continuous views), labels are represented by Gaussian distributions, whereas in discrete feature spaces (discrete views), labels are represented by Multinomial distributions. The usage of Bayesian approach helps handle missing information such as the missing of instance labels, or partial examples.

The remainder of this paper is organized into 4 sections. We first revisit some related works, then present our method (MIMLmix), followed by experiments and conclusions.

Related Work

Zhou et al. (2007; 2012) formulated MIML (Multi-instance Multi-label) framework, proposed several algorithms and applied to image and text applications. Later on, many MIML algorithms have been proposed and many applications have been reported; to name a few, MIML algorithms based on Dirichlet-Bernoulli alignmnent (Yang, Zha, and Hu 2009), based on Conditional Random Fields (Zha et al. 2008), based on single-instance degeneration (Nguyen 2010), based on metric learning (Jin, Wang, and Zhou 2009), etc. MIML techniques have been found well useful in applications such as image retrieval and annotation (Nguyen et al. 2013), video annotation (Xu, Xue, and Zhou 2011), gene pattern annotation (Li et al. 2012b), relation extraction in natural language processing (Surdeanu et al. 2012), etc. A number of MIML methods that discover the relationships between bag labels and instances have been proposed in (Li et al. 2012a; Briggs, Xiaoli, and Raich 2012).

Multi-view learning deals with data in multiple views, i.e., multiple feature sets. The goal is to improve performance or reduce the sample complexity. Multi-view learning has been well studied in learning with unlabeled data. Some studies used multi-views in conjunction with semi-supervised learning (Blum and Mitchell 1998; Wang and Zhou 2010b; Zhou, Zhan, and Yang 2007), or with active-learning (Wang and Zhou 2010a). Others tried to establish a latent subspace by assuming that instances (in different views) belong to the same example are nearby after mapping into the latent subspace (White et al. 2012; Wang, Nie, and Huang 2013). To combine information from multi-views for traditional supervised learning, one can use fusion techniques at feature level, or classifier level (Atrey et al. 2010) .

Almost all previous MIML studies focused on single-view setting and almost all previous multi-view learning studies focused on single-instance and/or single-label learning. To the best of our knowledge, the only exception is (Nguyen, Zhan, and Zhou 2013), where the M3LDA approach was proposed. Our approach, however, is more general and effective than M3LDA, which will be discussed with details in the next section. It is also important to highlight some related Bayesian Network structures such as Dependence-LDA (Rubin et al. 2012), GM-LDA and Corr-LDA (Blei and Jordan 2003). These methods were not designed for MIML, since Dependence-LDA works with textual documents; and GM-LDA, Corr-LDA are for unsupervised learning, i.e. labels have not been exploited.

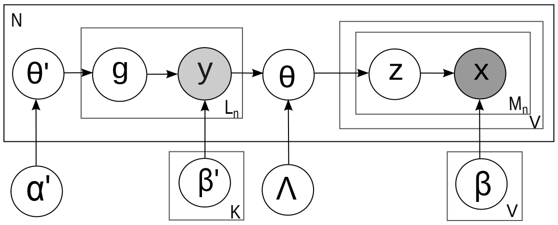


Figure 2: MIMLmix has topic-label part ($\boldsymbol{\theta}'$ to $\mathbf{y}$) and label-instance part ($\boldsymbol{\theta}$ to $\mathbf{x}$). $\Lambda = \{\boldsymbol{\eta}, \xi\}$ is a set of parameters connecting the topic-label part and the label-instance part.

Algorithm 1 Generative Process for MIMLmix

```
for each example $X_n$ do
    ▷ Topic-Label part: LDA model with $K$ topics.
    Sample a topic distribution $\boldsymbol{\theta}'_n \sim Dir(\boldsymbol{\alpha}')$.
    for each label do
        Sample a topic indicator $g \sim Mult(\boldsymbol{\theta}'_n)$.
        Sample a label $y \sim Mult(\boldsymbol{\beta}'_g)$.
    ▷ Label-Instance part: for $V$ views, $L$ labels
    Sample label distribution $\boldsymbol{\theta}_n \sim Dir(\eta \odot \mathbf{y}_n + \boldsymbol{\xi})$.
    for each instance $\mathbf{x}_{nvm}$ in the view $v$ do
        Sample a label indicator $\mathbf{z}_{nvm} \sim Mult(\boldsymbol{\theta}_n)$
        Sample $\mathbf{x}_{nvm} \sim p(\mathbf{x}_{nvm} | \mathbf{z}_{nvm} = y, \boldsymbol{\beta}_v)$.
```

The MIMLmix Model

Inspired by Bayesian Network approaches (Nguyen et al. 2010; 2013; Rubin et al. 2012; Nguyen, Zhan, and Zhou 2013), we propose the MIMLmix model (Multi-instance Multi-label Mixture model) for multi-view MIML (Fig. 2), which consists of two parts: (1) the topic-label part is a LDA topic model of K topics (Blei, Ng, and Jordan 2003), where topics capture label correlations; and (2) the label-instance part where instances are generated from a mixture of Gaussian/Multinomial distributions. The generative process is shown in Alg. 1. and in Fig. 2.

In the label-instance part, for an example X_n, we set the prior for the label distribution $\boldsymbol{\theta}_n \in R^L$ as $\boldsymbol{\alpha}_n = \eta \odot \mathbf{y}_n + \xi$, where $\odot$ is an element-wise product, and $\mathbf{y}_n \in R^L$. During training, $\xi = 0$, $\mathbf{y}_{nl}$ equals to 1 if the l-th label is in Y_n and zero otherwise, thus $\boldsymbol{\alpha}_{nl}$ are zeros for labels not in Y_n. During testing, ξ is set to a nonzero constant and all elements in $\mathbf{y}_n$ are initialized to 1 to trigger all labels for inference, the value of η controls how much the topic distribution affects the label distribution $\boldsymbol{\theta}$. The latent variables $\mathbf{z}$ represent the hidden assignments of bag labels to instances. If $\mathcal{X}_v$ is discrete, we formalize $p(\mathbf{x}_{nvm}|\mathbf{z}_{nvm}=y, \boldsymbol{\beta}_v)=p(\mathbf{x}_{nvm}|\boldsymbol{\beta}_{vy})$ using a Multinomimal distribution with parameter $\boldsymbol{\beta}_{vy} \in R^{D_v}$. Here, we drop indexes n, m, v for simplicity:

$$p(\mathbf{x}|\boldsymbol{\beta}_y) = \binom{||\mathbf{x}||_1}{\mathbf{x}_1 ... \mathbf{x}_D} \prod_{i=1}^{D} (\boldsymbol{\beta}_{yi})^{\mathbf{x}_i} \tag{1}$$

where $||\mathbf{x}||_1 = \sum_i \mathbf{x}_i$. If the feature space $\mathcal{X}_v$ is continuous, we formalize $p(\mathbf{x}_{nvm}|\boldsymbol{\beta}_{vy})$ as a Gaussian distribution where $\boldsymbol{\beta}_{vy} = \{\boldsymbol{\mu}_{vy}, \Sigma_{vy}\}$, $\boldsymbol{\mu}_{vy} \in R^{D_v}, \boldsymbol{\Sigma}_{vy} \in R^{D_v \times D_v}$:

$$p(\mathbf{x}|\boldsymbol{\beta}_y) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}_z^{-1}(\mathbf{x}-\boldsymbol{\mu}_y)\right]}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_y|^{1/2}} \tag{2}$$

where n, m, v have also been removed for simplicity.

As MIMLmix allows instances from discrete and continuous views, it is more general than M3LDA (Nguyen, Zhan, and Zhou 2013), which only works with discrete views. In the sequel, we will derive a training method based on variational inference for MIMLmix that is more effective than Gibbs sampling in M3LDA. Moreover, instead of a "hard" assignment of labels to instances via $\mathbf{z}$ ($\mathbf{z}$ takes value of one out of the bag label set), variational inference introduces a "soft" assignment via $\boldsymbol{\phi}$, ($\sum_y \boldsymbol{\phi}_y = 1$ - the variational variable for $\mathbf{z}$), which allows one instance to be associated with multiple related labels.

Training with MIMLmix

As Y_n are observed during training, the two parts ($p(\boldsymbol{\theta}', \mathbf{g}, Y|\boldsymbol{\beta}', \boldsymbol{\alpha}')$ and $p(\boldsymbol{\theta}, \mathbf{z}, X|\boldsymbol{\beta}, Y, \boldsymbol{\Lambda})$) can be learned independently. In the following, we show a variational inference for the label-instance part. Variational inference places a simpler family of distributions over the latent variables:

$$q(\mathbf{z}, \boldsymbol{\theta}) = \textstyle\prod_n q(\boldsymbol{\theta}_n|\boldsymbol{\gamma}_n) \prod_{v,m} q(\mathbf{z}_{nvm}|\boldsymbol{\phi}_{nvm}) \tag{3}$$

where $\boldsymbol{\theta}_n \sim Dir(\boldsymbol{\gamma}_n)$ and $\mathbf{z}_{nvm} \sim Mult(\boldsymbol{\phi}_{nvm})$ and $\boldsymbol{\gamma}_n \in R^L$ and $\boldsymbol{\phi}_{nvm} \in R^L$ ($\sum_y \boldsymbol{\phi}_{nvm,y} = 1$). We then obtain the *evidence lower bound* (ELBO) $\mathcal{L}$:

$$\mathcal{L} = E_q[\log p(\boldsymbol{\theta}, \mathbf{z}, X|\boldsymbol{\beta}, Y, \boldsymbol{\Lambda})] - E_q[\log q(\mathbf{z}, \boldsymbol{\theta})] \tag{4}$$

The training is performed by maximizing the ELBO using the EM algorithm similar to (Blei, Ng, and Jordan 2003). Here, E-step tries to assign labels in bag labels to instances by alternating the following updates:

$$\boldsymbol{\phi}_{nvmy} \propto \exp\{E_q[\log \boldsymbol{\theta}_{ny}] + \log p(\mathbf{x}_{nvm}|\mathbf{z}_{nvm}=y, \boldsymbol{\beta}_v)\} \tag{5}$$

$$\boldsymbol{\gamma}_{ny} = \boldsymbol{\alpha}_{ny} + \sum_{v,m} \boldsymbol{\phi}_{nvmy} \tag{6}$$

where $E_q[\log \boldsymbol{\theta}_{ny}] = \Psi(\boldsymbol{\gamma}_{ny}) - \Psi(\sum_{y'} \boldsymbol{\gamma}_{ny'})$ (Ψ denotes the digamma function). Note that we do not consider all the labels in $\mathcal{Y}$ for each bag n but *only the labels belonging to* Y_n, thus performing E-step here is efficient.

Given the estimated $\boldsymbol{\phi}_n$ and $\boldsymbol{\gamma}_n$ for all n, M-step updates the global variables that maximizes the ELBO as follows:

- If v is a discrete view, update $\boldsymbol{\beta}_{vy} \in R^{D_v}$:

$$\boldsymbol{\beta}_{vyi} \propto \sum_n \sum_{m=1}^{M_{nv}} \mathbf{x}_{nvm,i} \boldsymbol{\phi}_{nvmy} \tag{7}$$

- If v is a continuous view, update $\boldsymbol{\beta}_{vy} = \{\boldsymbol{\mu}_{vy}, \Sigma_{vy}\}$

$$\boldsymbol{\mu}_{vy} = \frac{\sum_{nm} \boldsymbol{\phi}_{nvmy} \mathbf{x}_{nvm}}{\sum_{nm} \boldsymbol{\phi}_{nvmy}} \tag{8}$$

$$\boldsymbol{\Sigma}_{vy} = \frac{\sum_{nm} \boldsymbol{\phi}_{nvmy} (\mathbf{x}_{nvm})^\top \mathbf{x}_{nvm}}{\sum_{nm} \boldsymbol{\phi}_{nvmy}} - (\boldsymbol{\mu}_{vy})^\top \boldsymbol{\mu}_{vy} \tag{9}$$

The training algorithm is summarized in Alg. 2.

Algorithm 2 Training with MIMLmix

```
▷ Topic-label part
Train a LDA on $Y_{1:N}$ (Blei, Ng, and Jordan 2003).
▷ Label-instance part
Initialize $\boldsymbol{\beta}_{vy}$ for all view $v$ and $y \in \mathcal{Y}$.
while relative improvement in $\mathcal{L} < 10^{-6}$ do
    for n = 1 to $N$ do {E-step}
        Initialize $\boldsymbol{\gamma}_n$.
        repeat
            for each view $v$, ins. $m$ and label $y \in Y_n$ do
                Update $\boldsymbol{\phi}_{nvm,y}$ according to Eq. [5].
            Update $\boldsymbol{\gamma}_{ny}$ according to Eq. [6] for $y \in Y_n$.
        until $1/L \sum_y$[change in$\boldsymbol{\gamma}_{ny}$] $< 10^{-6}$
    for each view $v$, and $y \in \mathcal{Y}$ do {M-step}
        Update $\boldsymbol{\beta}_{vy}$ according to either Eq. [7] or Eqs. [8-9] depending on the view $v$.
```

Algorithm 3 Testing with MIMLmix

```
1: For a test bag $X_{n'}$, initialize $\mathbf{y}_{n'l} = 1, \forall l \in \mathcal{Y}$ and $\boldsymbol{\alpha}_{n'} = \eta \odot \mathbf{y}_{n'} + \xi$.
2: repeat
3:     Perform inference about $\boldsymbol{\phi}_{n'}$, $\boldsymbol{\gamma}_{n'}$ like E-step of Algorithm 2 given the current value of $\boldsymbol{\alpha}_{n'}$.
4:     Sample $Y_{n'}$ according to a Multinomial dist. parameterized by (normalized) $\sum_{vm} \boldsymbol{\phi}_{n'vm}$.
5:     Estimate $\boldsymbol{\theta}'_{n'}$ on $Y_{n'}$ using LDA (Blei et al. (2003)).
6:     Update $\boldsymbol{\alpha}_{n'} = \eta \times \boldsymbol{\beta}'^\top \boldsymbol{\theta}'_{n'} + \xi$.
7: until the change in $\boldsymbol{\theta}'_{n'}$ is smaller than a threshold.
8: Output $\boldsymbol{\gamma}_{n'}$ and $\boldsymbol{\phi}_{n'vm}$ for bag and instance annotation.
```

Testing with MIMLmix

During testing, for a new example $X_{n'}$, we set ξ to a constant larger than 0 ($\xi = 0.1$ by default), initialize $\mathbf{y}_{n'l} = 1, \forall l$; thus we have $\boldsymbol{\alpha}_{n'l} > 0, \forall l$. By doing so, we trigger all the labels for consideration. The algorithm is summarized in Alg. 3. The information is passed from the label-instance part to the topic-label part by exploiting label assignments for instances in multi-views (line 4), and from the topic-label part to label-instance part through $\boldsymbol{\alpha}_{n'}$, where $\mathbf{y}_{n'}$ is implicitly set to $\boldsymbol{\beta}'^\top \boldsymbol{\theta}'_{n'}$ (line 6).

In implementation, in order to reduce the randomness of the sampling step in line 4, we obtain the averaged topic distribution $\bar{\boldsymbol{\theta}}'_{n'}$ over all iterations, then use it to update the prior information for the final bag and instance annotation.

Multi-Views with Unequal Importance

Let l_v be the random variable that represents the number of instances of the v-th view per example, and assume that l_v follows Possion distribution $l_v \sim Po(\lambda_v)$. Suppose we fix the number of instances across views to a constant λ, the conditional distribution $P(l_v|\lambda)$ follows a Multinomial distribution with parameter $\boldsymbol{\rho} = (\rho_1, \dots, \rho_V)$, where $\rho_v = \frac{\lambda_v}{\lambda}$. We then rewrite the joint probability $p(\mathbf{z_n}, \mathbf{X_n}|\boldsymbol{\theta}_\mathbf{n}, \boldsymbol{\beta})$:

$$p(\mathbf{z}_n, X_n|\boldsymbol{\theta}_n, \boldsymbol{\beta}) = \prod\nolimits_{vm} p(\mathbf{x}_{nvm}, \mathbf{z}_{nvm}|\boldsymbol{\theta}_n, \boldsymbol{\beta})^{w_v} \quad (10)$$

where $w_v{=}\lambda\rho_v/\lambda_v{=}1$, the above equation is the same as the original one but with a new insight "*in a bag of one instance of view v repeated λ_v times, instead of repeating the instance as λ_v, we repeat it with $w_v\lambda_v = \lambda\rho_v$ times*". Replacing the constraint $w_v = 1$ with $\sum_v w_v{=}1$, one can change the weights of instances in different views. We modify the variational distributions similarly, and obtain the terms of the modified ELBO related to w_v as follows:

$$\begin{aligned}\mathcal{L}_{[w_v]} = & \sum_i w_v \{E_q[\log p(\mathbf{z}_{vi}|\boldsymbol{\theta}_n) - E_q[\log q(\mathbf{z}_{vi}|\boldsymbol{\phi}_{vi})] \\ & + \; E_q[\log p(\mathbf{x}_{vi}|\mathbf{z}_{vi}, \boldsymbol{\beta}_{1:V})]\} \qquad (11)\end{aligned}$$

where i runs over $\hat{M}_v$ instances in the v-th view from all the training examples. Let $\Delta_v = 1/\hat{M}_v(\partial\mathcal{L}/\partial w_v)$, it is intuitive that Δ_v measures how likely one instance in view v is generated given the bag labels. Maximizing $\sum_v \Delta_v w_v$ with $\sum_v w_v = 1$ yields an extreme result as the view with the largest Δ_v will receive all the weights while the others are zero. We then find a simple solution by setting $w_v \propto \overline{\Delta}_v + \tau_v$ where $\overline{\Delta}_v = \log_2(\Delta_v - \min_i \Delta_i + 2)$ is the scaled values of Δ_v; and τ_v can be interpreted as the prior for the v-th view.

The updates of $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are still the same as in Algorithm 2 as the view weight is eliminated due to normalization or division within each view, the update for γ_{ny} is changed to $\gamma_{ny} = \boldsymbol{\alpha}_{ny} + \sum_{vm} w_v \boldsymbol{\phi}_{nvmy}$. During testing, we sample $Y_{n'}$ in Algorithm 3 by normalizing $\sum_{vm} w_v \boldsymbol{\phi}_{n'vm}$. Note that all the experiments with multi-view datasets in the next session are conducted with this variant of MIMLmix.

Experiments

We perform experiments on 2 multi-view datasets and 3 single-view datasets. The summary of these datasets are given in Table 1. Citeseerx-10k[1] contains scientific papers in two views, i.e, *content* (v1) and *citations* (v2). ImageCLEF (Müller et al. 2010) contains images with two views: *visual* (v1) and *textual* (v2). Here, we use the same subset that has been used in (Nguyen, Zhan, and Zhou 2013). Each example in the visual view is represented by a bag of segmented regions, one region is represented by a frequency vector of 1000 visual words, which are obtained by clustering Opponent SIFTs (Van de Sande, Gevers, and Snoek 2010). Citeseerx-10k has **1072** partial examples, and ImageCLEF has **2114** partial examples; most of partial examples do not have the second view. Among single-view datasets, LetterCarroll, MSRC-v2 were collected by (Briggs, Xiaoli, and Raich 2012); and IAPRTC-12 dataset was selected from (Escalante et al. 2010).

Table 1: Experimental datasets: #ipb is #instances per bag.

Dataset	#bags	#labels	#ipb	#dim
Citeseerx-10K	10,799	500	35.7	2,000
(2 views)			48.3	2,000
ImageCLEF	8,000	78	18.4	1,000
(2 views)			2.6	806
Letter Carroll	166	26	4.3	16
MSRC-v2	591	23	2.97	48
IAPRTC-12	5,000	244	5.09	28

Evaluation: MIML methods are evaluated from three aspects, i.e. example-pivot evaluation using hamming loss (**h.l.**) and average precision (**a.p.**) (Zhou and Zhang 2007); label-pivot evaluation using mean average precision (**m.a.p**) and macro-F1 (**ma-f1**) for labels that appear at least once in training/testing dataset (Rubin et al. 2012); and instance-pivot evaluation in terms of instance accuracy (**ins-acc**) (Briggs, Xiaoli, and Raich 2012). To measure h.l, ma-f1, top $\bar{L}$ labels with highest decision values are selected as the annotation for each example. Here, $\bar{L}$ is chosen based on the average number of labels per example. We conduct 30 times evaluation for ImageCLEF, each time we use 1000 examples for training and 1000 examples for testing; 10-fold cross-validation is conducted for the other datasets. Only single-view datasets have instance labels for ins-acc evaluation.

Compared Methods: On multi-view datasets, the following methods are compared and contrasted: **MIMLmix**; **MIMLmix*** (MIMLmix with $\eta = 0$); **M3LDA** (Nguyen, Zhan, and Zhou 2013); **cs.SVM** which combines decision values of single-view, cost-sensitive SVMs; and MIMLmix with individual views (**MIMLmix.v1** and **MIMLmix.v2**). In order to train single-view SVM, we accumulate multiple instances to obtain a single instance per bag, then use one-vs-all for multi-label learning.

On single-view datasets, we compare **MIMLmix**, **MIMLmix*** with other MIML methods including **RankLossSVM** (Briggs et al. (2012)); **MIMLSVM** (Zhou and Zhang 2007); **cs.MISVM** (Andrews et al. (2002)) which builds a cost sensitive Multi-instance SVM for every label; and **DBA** (Yang et al. (2009)).

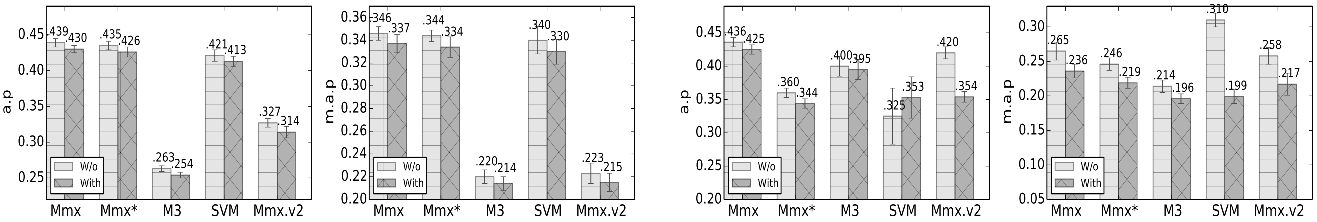
Multi-view Datasets without Partial Examples

We evaluate MIMLmix and the compared methods in the case without partial examples, which are obtained by removing all the partial examples from multi-view datasets. For MIMLmix methods, we set $\alpha' = 0.1$, $K = 200$ as default for both datasets, set $\eta = .3$, $\tau = 5$ for Citeseerx-10k; and $\eta = 10$ and $\tau = 0$ for ImageCLEF. M3LDA is conducted with the same setting as in (Nguyen, Zhan, and Zhou 2013) on ImageCLEF; and with $K = 200$, $\lambda = .5$, and the number of sampling iterations of 300 on Citeseerx-10k. One-vs-all cs.SVM classifiers are trained for every view, every label using LIBSVM (Chang and Lin 2011) with default parameters, except that the weights of positive and negative classes are

[1] collected from http://citeseerx.ist.psu.edu/index

Table 2: Performance on multi-view datasets **without partial examples**. Here, v1/v2 mean *content*/*citations* on Citeseerx-10k; and *visual*/*textual* on ImageCLEF. Here, ● (○) indicates a method is significantly worse (better) than MIMLmix with 95% t-test.

Dataset		MIMLmix	MIMLmix*	M3LDA	cs.SVM	MIMLmix.v1	MIMLmix.v2
CiteSeerx-10K	a.p. ↑	.439 ± .006	.435 ± .006●	.263 ± .004●	.421 ± .008●	.375 ± .007●	.327 ± .006●
	h.l. ↓	.010 ± .000	.010 ± .000	.013 ± .000●	.011 ± .000●	.011 ± .000●	.012 ± .000●
	m.a.p ↑	.346 ± .006	.344 ± .009●	.220 ± .006●	.340 ± .011●	.295 ± .005●	.223 ± .009●
	ma-f1 ↑	.337 ± .006	.336 ± .005	.218 ± .006●	.319 ± .008●	.301 ± .007●	.247 ± .007●
ImageCLEF	a.p. ↑	.436 ± .007	.360 ± .007●	.400 ± .015●	.325 ± .042●	.378 ± .008●	.420 ± .009●
	h.l. ↓	.104 ± .001	.112 ± .002●	.134 ± .003●	.162 ± .011●	.112 ± .001●	.105 ± .002●
	m.a.p ↑	.265 ± .013	.246 ± .009●	.214 ± .009●	.310 ± .010○	.150 ± .004●	.258 ± .012●
	ma-f1 ↑	.237 ± .012	.228 ± .007●	.227 ± .009●	.102 ± .015●	.129 ± .005●	.231 ± .009●



(a) Citeseerx-10k

(b) ImageCLEF

Figure 3: Performances of MIMLmix (Mmx), MIMLmix* (Mmx*), M3LDA (M3), cs.SVM (SVM), MIMLmix.v2 (Mmx.v2) in the presence of partial examples are illustrated with the darker bars. The light-color bars corresponding to "W/o" show the results of these methods in the case without partial examples (Table 2) for references.

set to $\frac{\#pos+\#neg}{\#pos}$ and $\frac{\#pos+\#neg}{\#neg}$, where $\#pos$ and $\#neg$ are the number of positive and negative bags, respectively. We combine decision values of single-view cs.SVM using the rule (.3×v1+.7×v2) on ImageCLEf, and (.6×v1+.4×v2) on Citeseerx, these parameters are selected by trying different values of combination.

The experimental results are represented in Table 2. It can be seen that MIMLmix outperforms other multi-view methods including MIMLmix* in most of the cases. Particularly, the performance of M3LDA is not satisfactory on CiteSeerx dataset, mostly due to the small number of sampling iterations that we set to meet the time constraint. More details about computational comparison will be discussed later in this section. MIMLmix outperforms MIMLmix* by a large gap on ImageCLEF where labels are highly correlated. Compared to single view MIMLmix models, MIMLmix is significantly better on both multiview datasets. This validates the importance of combining multi-views to obtain better performance.

MIMLmix is significantly better than cs.SVM in most of the cases except for m.a.p on ImageCLEF. Interestingly, on ImageCLEF dataset, cs.SVM achieves much worse ma-f1 than MIMLmix although it has higher m.a.p. By examining the combined decision values of cs.SVM, we see that although cs.SVM obtains good ranking of examples with regards to some rare labels, the values for rare labels are not large enough to meet the cut of ma-f1 evaluation. This leads to the fact that a lot of rare labels have zero recalls, consequently low values of ma-f1.

Multi-view Datasets with Partial Examples

This section compares MIMLmix with other methods in the presence of partial examples. Experimental settings are the same as in the previous section, except that we do not remove partial examples from multi-view datasets. The results on Citeseerx-10k and ImageCLEF with partial examples are shown in Fig. 3(a) and Fig. 3(b), respectively. Here, we only show the results of a.p and m.a.p metrics as the results of h.l. (resp. ma-f1) changes in the similar way with a.p (resp. m.a.p) but with smaller magnitudes. Also, we do not show the result of MIMLmix.v1 as most of partial examples have the second view missing.

Fig. 3 shows that most methods degenerate in the presence of partial examples. Nevertheless, MIMLmix outperforms all compared methods on both multi-view datasets. On Citeseerx-10k, the difference in the degeneration magnitude is not so obvious among these methods, probably due to the small rate of partial examples. On ImageCLEF, where there are more partial examples, it is not surprising to see that MIMLmix-v2 suffers more than its multi-view counterparts (MIMLmix, MIMLmix*) on both a.p and m.a.p metrics. M3LDA has comparably low degeneration because it also follows the Bayesian Network approach.

From Fig. 3(b), we can observe some interesting behaviors of cs.SVM on ImageCLEF dataset. It is shown that cs.SVM in the complete case has a.p metric worse than it is in the partial case, where there exist examples without textual view. This is indeed not difficult to understand as textual view tends to be more useful towards rare labels, and a naive

Table 3: Performance on single-view datasets; Here, ●/○ means a method is worse/better than MIMLmix with 95% t-test.

Dataset		MIMLmix	MIMLmix*	DBA	cs.MISVM	MIMLSVM	RankL.SVM
Letter Carroll	a.p. ↑	.770 ± .040	.744 ± .037	.308 ± .025●	.740 ± .048	.443 ± .059●	.672 ± .057●
	h.l. ↓	.126 ± .014	.130 ± .010	.246 ± .016●	.098 ± .014○	.139 ± .011●	.137 ± .017●
	m.a.p ↑	.761 ± .067	.771 ± .072	.388 ± .050●	.627 ± .034●	.397 ± .030●	.632 ± .037●
	ma-f1 ↑	.556 ± .062	.568 ± .046	.218 ± .042●	.372 ± .041●	.227 ± .067●	.442 ± .045●
	ins-acc ↑	.623 ± .053	.604 ± .049●	.122 ± .031●	.571 ± .052●	N/A	.493 ± .048●
MSRCV-v2	a.p. ↑	.688 ± .037	.684 ± .038	.420 ± .028●	.716 ± .056○	.670 ± .054	.692 ± .034
	h.l. ↓	.109 ± .004	.109 ± .005	.174 ± .005●	.101 ± .009○	.111 ± .008	.109 ± .007
	m.a.p. ↑	.600 ± .034	.612 ± .045	.369 ± .034●	.607 ± .043	.588 ± .047	.471 ± .038●
	ma-f1 ↑	.495 ± .032	.486 ± .034	.286 ± .031●	.436 ± .051●	.503 ± .036	.441 ± .038●
	ins-acc ↑	.526 ± .034	.519 ± .031●	.224 ± .029●	.516 ± .054	N/A	.458 ± .043●
IAPRTC-12	a.p. ↑	.529 ± .013	.521 ± .013●	N/A	.559 ± .010○	.255 ± .012●	.407 ± .013●
	h.l. ↓	.023 ± .000	.023 ± .000	N/A	.022 ± .000○	.031 ± .000●	.027 ± .000●
	m.a.p. ↑	.282 ± .021	.292 ± .023○	N/A	.195 ± .006●	.154 ± .014●	.137 ± .006●
	ma-f1 ↑	.231 ± .015	.233 ± .016	N/A	.168 ± .007●	.022 ± .004●	.075 ± .006●
	ins-acc ↑	.400 ± .016	.363 ± .013●	N/A	.411 ± .001	N/A	.260 ± .010●

Table 4: Training time in seconds (here, #examples is the number of training examples in one evaluation).

	#examples	MIMLmix	M3LDA	cs.MISVM	MIMLSVM	RankL.SVM
CiteSeerx (2 views)	9,719	3,448	255,000	N/A	N/A	N/A
ImageCLEF (2 views)	1,000	150	10,000	N/A	N/A	N/A
IAPRTC-12 (1 view)	4,500	915	N/A	3,969	9,545	35,976

combination of the decision values can hurt frequent labels, resulting in the lower value of example-pivot evaluation metrics, which give more credits to frequent labels. On the other hand, we can see that the degeneration of cs.SVM on m.a.p is the most significant one among compared methods. This implies that we may need more investments when applying SVM to multi-view MIML datasets with partial examples.

Single-view Datasets

On single-view datasets, η are chosen from $\{.1, .2, .3\}$, $K{=}50$ is set as default for MIMLmix and MIMLmix*. As single-view datasets are with continuous features, clustering is used to obtain discrete presentation for DBA, which only works with discrete features. We train RankLossSVM with default values, train MIMLSVM and cs-MISVM with RBF kernel with $C{=}2^3$ and $\gamma{=}.5$. cs-MISVM is learned for each label using one-vs-all method where costs are set the same as cs.SVM. The ratio parameter of MIMLSVM is set to 30% for IAPRTC-12 and 20% for the others.

Experimental results are given in Table 3. Due to the quantization error of clustering step, DBA obtains poor performance on two small datasets, and consequently it has not been applied to IAPRTC-12. MIMLmix is better than MIMLmix* on a.p, h.l and inc-acc in most of the cases, but has lower values of label-pivot measures. This shows that by setting $\eta > 0$, we may have to trade off between label-pivot for example-pivot/instance-pivot evaluations on datasets without strong label relationships. Nevertheless, MIMLmix still has better label-pivot evaluation compared to other MIML methods. Particularly on IAPRTC-12, the gap in m.a.p and ma-f1 between MIMLmix and other MIML methods becomes significant. In terms of inc-acc, MIMLmix obtains best results on two datasets and slightly worse than cs.MISVM on IAPRTC-12. MIMLSVM transforms MIML problem into SIML problem, and thus it cannot assign labels to instances, consequently ins-acc is not available.

Time Cost Comparison

Table 4 shows training times of MIML methods on 3 large datasets on the same computer (CPU of 3.3Gz, 4GB memory). MIMLmix is more effective than other MIML methods, particularly in contrast to M3LDA, where sampling method is used for training; and RankLossSVM where all labels in $\mathcal{Y}$ are ranked for every example. Note that RankLossSVM is particularly time consuming when $\mathcal{Y}$ is large.

Conclusion

This paper proposes MIMLmix, a model based on hierarchical Bayesian network for the general problem of multi-view MIML in the presence of partial examples. Extensive evaluation on 5 datasets suggests that (1) MIMLmix can naturally deal with multiple view MIML data with partial examples; (2) our method less suffers from the problem of label-imbalance; and (3) our training method is effective particularly on datasets with a large number of labels.

For the future work, stochastic variational inference (Hoffman et al. (2010)) can be applied to further reduce the computational complexity in training large datasets. It is also interesting to extend MIMLmix to a deep model for a larger representation capacity within each view.

Acknowledgements: We thank anonymous reviewers for suggestions, and Yang Yu for reading our draft.

References

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2002. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 561–568.

Atrey, P. K.; Hossain, M. A.; El-Saddik, A.; and Kankanhalli, M. S. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia System* 16(6):345–379.

Blei, D., and Jordan, M. I. 2003. Modeling annotated data. In *Proceedings of the 26th ACM SIGIR Conference on Information Retrieval*, 127–134.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 92–100.

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

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: a library for SVMs. *ACM Transaction on Intelligent Systems and Technology* 2(3):27:127:27.

Escalante, H. J.; Hernandez, C. A.; Gonzalez, J. A.; Lopez, A.; Montes, M.; Morales, E. F.; Sucar, L. E.; Villasenor, L.; and Grubinger, M. 2010. The segmented and annotated IAPRTC-12 benchmark. *Computer Vision and Image Understanding* 114(4):419 – 428.

Hoffman, M. D.; Blei, D. M.; and Bach, F. R. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, 856–864.

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

Li, Y.-F.; Hu, J.-H.; Jiang, Y.; and Zhou, Z.-H. 2012a. Towards discovering what patterns trigger what labels. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 1012–1018.

Li, Y.-X.; Ji, S.; Kumar, S.; Ye, J.; and Zhou, Z.-H. 2012b. Drosophila gene expression pattern annotation through multi-instance multi-label learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9(1):98–112.

Müller, H.; Clough, P.; Deselaers, T.; and Caputo, B. 2010. *ImageCLEF: Experimental Evaluation of Visual Information Retrieval*. Berlin, German: Springer.

Nguyen, C.-T.; Kaothanthong, N.; Phan, X.-H.; and Tokuyama, T. 2010. A feature-word-topic model for image annotation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1481–1484.

Nguyen, C.-T.; Kaothanthong, N.; Tokuyama, T.; and Phan, X.-H. 2013. A feature-word-topic model for image annotation and retrieval. *ACM Transaction on Web* 7(3):12:1–12:24.

Nguyen, C.-T.; Zhan, D.-C.; and Zhou, Z.-H. 2013. Multimodal image annotation with multi-instance multi-label LDA. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 1558–1564.

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

Rubin, T. N.; Chambers, A.; Smyth, P.; and Steyvers, M. 2012. Statistical topic models for multi-label document classification. *Machine Learning* 88(1-2):157–208.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 455–465.

Van de Sande, K. E. A.; Gevers, T.; and Snoek, C. G. M. 2010. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9):1582–1596.

Wang, W., and Zhou, Z.-H. 2010a. Multi-view active learning in the non-realizable case. In *Advances in Neural Information Processing Systems 23*, 2388–2396.

Wang, W., and Zhou, Z.-H. 2010b. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning*, 1135–1142.

Wang, H.; Nie, F.; and Huang, H. 2013. Multi-view clustering and feature learning via structured sparsity. In *Proceedings of the 28th International Conference on Machine Learning*, 352–360.

White, M.; Zhang, X.; Schuurmans, D.; and Yu, Y.-l. 2012. Convex multi-view subspace learning. In *Advances in Neural Information Processing Systems 25*, 1682–1690.

Xu, X.; Xue, X.; and Zhou, Z. 2011. Ensemble multi-instance multi-label learning approach for video annotation task. In *Proceedings of the 19th ACM international conference on Multimedia*, 1153–1156.

Yang, S.-H.; Zha, H.; and Hu, B.-G. 2009. Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *Advances in Neural Information Processing Systems 22*, 2143–2150.

Zha, Z.-J.; Hua, X.-S.; Mei, T.; Wang, J.; and Wang, Z. 2008. Joint multi-label multi-instance learning for image classification. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

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

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

Zhou, Z.-H.; Zhan, D.-C.; and Yang, Q. 2007. Semi-supervised learning with very few labeled training examples. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, 675–680.