Weakly-Supervised Multi-view Multi-instance Multi-label Learning

Yuying Xing[1] , **Guoxian Yu**[1,2,*] , **Jun Wang**[1] , **Carlotta Domeniconi**[3] and **Xiangliang Zhang**[2]

[1]College of Computer and Information Sciences, Southwest University, Chongqing, China

[2]CEMSE, King Abdullah University of Science and Technology, Thuwal, SA

[3]Department of Computer Science, George Mason University, VA, USA

{yyxing4148, gxyu, kingjun}@swu.edu.cn, carlotta@cs.gmu.edu, xiangliang.zhang@kaust.edu.sa

Abstract

Multi-view, Multi-instance, and Multi-label Learning (M3L) can model complex objects (bags), which are represented with different feature views, made of diverse instances, and annotated with discrete non-exclusive labels. Existing M3L approaches assume a *complete* correspondence between bags and views, and also assume a *complete* annotation for training. However, in practice, neither the correspondence between bags, nor the bags' annotations are complete. To tackle such a weakly-supervised M3L task, a solution called WSM3L is introduced. WSM3L adapts multimodal dictionary learning to learn a shared dictionary (representational space) across views and individual encoding vectors of bags for each view. The label similarity and feature similarity of encoded bags are jointly used to match bags across views. In addition, it replenishes the annotations of a bag based on the annotations of its neighborhood bags, and introduces a dispatch and aggregation term to dispatch bag-level annotations to instances and to reversely aggregate instance-level annotations to bags. WSM3L unifies these objectives and processes in a joint objective function to predict the instance-level and bag-level annotations in a coordinated fashion, and it further introduces an alternative solution for the objective function optimization. Extensive experimental results show the effectiveness of WSM3L on benchmark datasets.

1 Introduction

Multi-view Multi-instance Multi-label (M3) objects (or bags) are characterized by heterogeneous feature views, including diverse instances, and are simultaneously annotated with non-exclusive labels. For example, in Figure 1, a video is represented by text and image views, where each text (image) bag includes diverse instances (paragraphs or animals) and is annotated with several semantic labels (e.g., seagull, water, and sky). Multi-view Multi-instance Multi-label Learning (M3L) [Nguyen *et al.*, 2013] can simultaneously model bags, instances of bags, and their non-exclusive labels to learn a predictive model to project multiple views of bags (and instances) into the label space, which reflects the semantic meaning of the bags (instances). Due to its capability of modeling complex objects in the real-world, M3L has attracted increasing research interest [Yang *et al.*, 2018; Xing *et al.*, 2019].

Traditional M3L approaches typically assume that the entire data is mapped across views, and the label annotation of objects is complete. Both assumptions are often violated in practical M3L tasks. As an example, in Figure 1, the mapping of a given bag across different views is only partially given. Moreover, the bags have missing annotations, and the number of bags in two views is different. In fact, such weakly-supervised multi-view data are universal in many domains. For example, for medicine development, the relation of a pill and its compounds with the therapy (adverse) effects is typically partially known. However, to the best of our knowledge, none of the existing M3L methods has studied the *partial correspondence* of M3 data. The *incomplete annotation* problem [Xu and Zhou, 2017; Tan *et al.*, 2018; Xing *et al.*, 2018] has also not been investigated. We term these two types of information as *weakly-supervised* information, which restricts the effectiveness and application, or even the adaption, of existing M3L approaches.

To address the weakly-supervised M3 problem, we introduce a weakly-supervised M3L approach (WSM3L) based on multimodal dictionary learning [Mandal and Biswas, 2016; Liu *et al.*, 2018a]. WSM3L introduces a unified objective function to seek the matches between bags across multiple views and to predict labels of bags. It uses the heterogeneous features of bags to learn a multi-view coordinated dictionary (representation space) and individual encoding vectors of bags for each view. Then the feature similarity derived from the encoding vectors and label similarity of bags are leveraged to seek matches between bags across views. Besides, it jointly replenishes the labels of a bag using the labels of its neighborhood bags, distributes the labels of bags to instances, and reversely aggregates the labels of instances to their originating bags. In this way, WSM3L can predict the labels of bags, and also the labels of instances in a coherent fashion. The main contributions of this work are as follows:

(i) WSM3L can handle not only weakly-paired (or even completely-unpaired) bags across views, but also partially

*Corresponding author, guoxian85@gmail.com. This work is supported by NSFC (61872300 and 61873214).

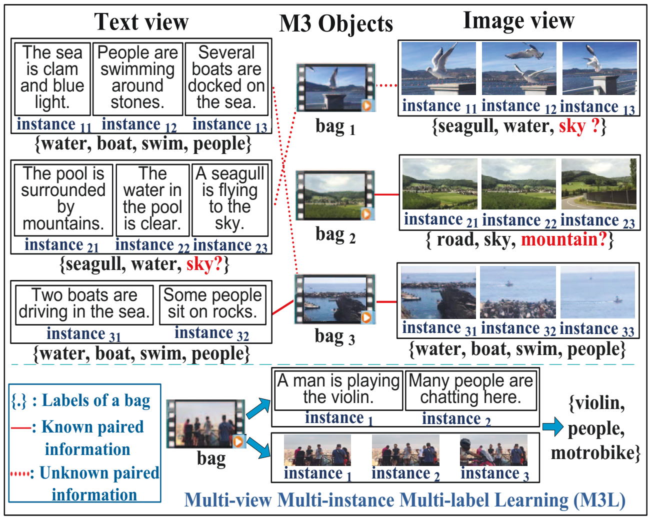


Figure 1: An example of a weakly-supervised multi-view multi-instance multi-label learning scenario. Each bag (video) is represented by an image view and a text view. The red solid (dotted) lines indicate the known (unknown) paired information of bags across views, and the labels highlighted in red with question marks '?' denote the missing annotations of bags. Unpaired bags across views have their own labels, but from the same label space.

annotated training bags. To the best of our knowledge, *none* of the existing M3L approaches can simultaneously make well usage of these two types of weakly-supervising information.
(ii) A matching solution based on labels and features of bags is introduced to discover their correspondence across views. We also introduce a unified objective function to seek the match between bags, to replenish missing labels of bags, to push the bag-level labels to instances, and reversely aggregate the labels of instances to their affiliated bags in a coordinated fashion.
(iii) WSM3L significantly outperforms state-of-the-art M3L approaches [Nguyen *et al.*, 2014; Li *et al.*, 2017; Xing *et al.*, 2019], multi-instance multi-label weak-label learning [Yang *et al.*, 2013], and weakly-paired multi-modal learning [Lampert and Krömer, 2010; Liu *et al.*, 2018a] in different practical settings. In addition, WSM3L can work in open settings (i.e., with different numbers of bags across views and with completely unpaired multi-view bags), in which the competitive methods cannot be applied.

2 Related Work

Multi-instance multi-label learning (M2L) [Zhou *et al.*, 2012; Huang *et al.*, 2019] deals with complex interrelations between bags, instances, and labels. M3L is more difficult, and less well-studied, than M2L, due to the additional heterogeneous feature views and complicated correlations across views. [Nguyen *et al.*, 2013] introduced a Latent Dirichlet Allocation [Blei *et al.*, 2003] based M3L approach, which separately explores the visual-label topics from the visual view and the text-label topics from the text view, and then performs prediction by forcing the label consistency between the two views. [Nguyen *et al.*, 2014] proposed an M3L approach (MIMLmix) that uses a hierarchical Bayesian network and variational inference to leverage multiple feature views. [Li *et al.*, 2017] developed a multi-view multi-instance learning (M2IL) algorithm, which considers different intrinsic structures between instances of a bag across views, and exploits sparse representation [Rubinstein *et al.*, 2010] and multi-view dictionary learning [Wu *et al.*, 2016; Gao *et al.*, 2015] for bag-level label prediction. [Yang *et al.*, 2018] introduced a deep neural network based approach, which separately applies a deep network for each view, and keeps the bag-level predictions across views consistent. Furthermore, a semi-supervised deep M3L approach [Yang *et al.*, 2019] is introduced to leverage label correlation and unlabeled instances for bag-level prediction. The aforementioned M3L approaches only consider limited types of inter-relations and intra-relations between bags, and between instances and labels, which in fact carry important contextual information for M3L to explore. [Xing *et al.*, 2019] recently introduced a collaborative matrix factorization based solution (M3Lcmf), which first constructs multiple inter(intra)-relational data matrices of bags, of instances, and of labels, to capture diverse intrinsic relations among them, and then collaboratively factorizes the matrices into low-rank ones to merge them and to coherently predict the bag(instance)-label associations.

The above M3L solutions optimistically assume that bags are completely paired across heterogeneous views, and are also comprehensively annotated. However, these two assumptions are often violated in practical M3L scenarios. Our study expands the flexibility and capability of M3L by designing a weakly-supervised M3L approach (WSM3L).

3 Proposed Method

Without loss of generality, we assume bags (or instances) have V feature views, and each view has n_v bag sets $\mathcal{X}^v = \{\mathbf{X}_1^v, \mathbf{X}_2^v, \cdots, \mathbf{X}_{n_v}^v\}$. $\mathbf{X}_i^v = [\mathbf{x}_{i,j}^v]_{j=1}^{m_i^v}$, a matrix, denotes the i-th bag in the v-th view includes $m_i^v \geq 1$ instances, where $\mathbf{x}^v \in \mathbb{R}^{d_v} (v = 1, 2, \cdots, V)$ is the feature space of instances in the v-th view. $\mathbf{Y}_i^v \in \mathbb{R}^q$ encodes the currently known labels of $\mathbf{X}_i^v$. $\mathbf{Y}_{iq'}^v = 1$ if $\mathbf{X}_i^v$ is annotated with the q'-th label, $\mathbf{Y}_{iq'}^v = 0$ otherwise. All bags belong to the same label space and paired bags share a same subset of labels. For an M3 dataset with completely paired bags, $\{\mathbf{Y}^v\}_{v=1}^V$ is identical across all the views, but not so for an M3 dataset with weakly-paired (completely-unpaired) bags. The task of M3L is to learn a predictive function $f(\{\mathbf{X}^v\}_{v=1}^V, \{\mathbf{Y}^v\}_{v=1}^V) \rightarrow \mathbb{R}^q$.

The correspondence between bags in M3L is the basis for multi-view data fusion. For weakly-paired M3 data, a simple bypass solution is to exclude unpaired bags and only use the known paired bags across views to train the predictive model. However, these excluded bags (and their member instances) also convey important context information for the task, and disregarding them may distort the underlying data distribution. To make use of as many bags as possible, we first seek matches between bags across views. Different techniques [Zhang *et al.*, 2015; Mandal and Biswas, 2016] can be used to this end, and here we adopt multi-modal dictionary learning [Monaci *et al.*, 2007], which has been successfully adopted to capture and correlate heterogeneous features across modalities [Mandal and Biswas, 2016; Liu *et al.*, 2018b]. The multi-modal dictionary learning technique provides an effective strategy to unify multi-modal data, since each view can be generated from the shared dictionary with individual encoding vectors. As such, the heterogeneous feature vectors are reformulated as comparable encoding vectors. Multi-modal dictionary learning on two feature views [Monaci *et al.*, 2007; Liu *et al.*, 2018a] is formulated as follows:

$$\underset{\mathbf{D}^1,\mathbf{D}^2,\mathbf{E}^1,\mathbf{E}^2}{argmin} \parallel \mathbf{X}^1-\mathbf{D}^1\mathbf{E}^1 \parallel_F^2 + \parallel \mathbf{X}^2-\mathbf{D}^2\mathbf{E}^2 \parallel_F^2 + \mathcal{C}(\mathbf{E}^1\mathbf{E}^2) \tag{1}$$

where $\mathbf{D}^1 \in \mathbb{R}^{d_1\times d}$ and $\mathbf{D}^2 \in \mathbb{R}^{d_2\times d}$ are the dictionaries, and $\mathbf{E}^1 \in \mathbb{R}^{d\times n_1}$ and $\mathbf{E}^2 \in \mathbb{R}^{d\times n_2}$ are the coding matrices of the two views, respectively. d is the dictionary size, which can be specified by the designer. The constraint term $\mathcal{C}(\mathbf{E}^1\mathbf{E}^2)$ has different forms [Mandal and Biswas, 2016], and it can be used to incorporate inter(intra)-modal relations.

3.1 Matching Bags Across Views

To explore the complementary information across views and the matches between bags, we first learn a shared dictionary for bags across views, which also gives a unified representational space for bags. In addition, we seek an encoding matrix of bags per view. Since the same bag may have different number of instances in different views, we first project the instances' features of a bag onto a bag feature vector like [Zhou and Zhang, 2007] for dictionary learning. Thus, $\mathbf{X}^v$ used in the following equations is a matrix storing the projected features of bags in the v-th view. To learn a shared dictionary, we project the feature views onto the same dimensional space: $(\mathbf{P}^v)^T\mathbf{X}^v$, where $\mathbf{P}^v \in \mathbb{R}^{d_v\times s}$ is the projection matrix of the v-th view with $\mathbf{P}^v(\mathbf{P}^v)^T = \mathbf{I} \in \mathbb{R}^{s\times s}$. We then use $(\mathbf{P}^v)^T\mathbf{X}^v$ to seek the shared dictionary and the encoding matrix of bags for each view as follows:

$$\begin{aligned} \min_{\mathbf{D},\mathbf{E}^v,\mathbf{P}^v} \mathcal{L}_1 &= \sum\nolimits_{v=1}^{V} \parallel (\mathbf{P}^v)^T\mathbf{X}^v - \mathbf{D}\mathbf{E}^v \parallel_F^2 \\ &+ \sum\nolimits_{v=1,w\neq v}^{V} \parallel (\mathbf{E}^v)^T - \mathbf{M}^{vw}(\mathbf{E}^w)^T \parallel_F^2 \\ s.t. &\parallel \mathbf{d}_{s'} \parallel_2 \leq 1 (\forall s' \in \{1,2,\cdots,s\}), \mathbf{P}^v(\mathbf{P}^v)^T = \mathbf{I} \end{aligned} \tag{2}$$

where $\mathbf{D} \in \mathbb{R}^{s\times d}$ is the shared dictionary of bags across views, and $\mathbf{d}_{s'} \in \mathbb{R}^d$ is the dictionary vector of $\mathbf{D}$. $\mathbf{E}^v \in \mathbb{R}^{d\times n_v}$ is the coding matrix of bags (and instances therein) of the v-th view. In this way, bags across different views are comparable in the representational space, which is configured by the shared dictionary. $\mathbf{M}^{vw} \in \mathbb{R}^{n_v\times n_w}$ records the mapping information between bags of the v-th view and w-th view. The term $\sum_{v=1,w\neq v}^{V} \parallel (\mathbf{E}^v)^T - \mathbf{M}^{vw}(\mathbf{E}^w)^T \parallel_F^2$ is introduced to force matched bags having similar encoding vectors.

Existing multi-modal learning methods match objects across views solely using the features [Lampert and Krömer, 2010; Mandal and Biswas, 2016], or labels of objects [Liu *et al.*, 2018b]. In contrast, we leverage both label and feature information to improve the matching process. To match bags across views, we leverage the label and feature information of pairwise bags ($\mathbf{X}_i^v$ and $\mathbf{X}_j^w$) as follows:

$$\begin{aligned} &m(\mathbf{X}_i^v,\mathbf{X}_j^w) = 1-(1-fea(\mathbf{E}_i^v,\mathbf{E}_j^w)(1-lab(\widetilde{\mathbf{Y}}_i^v,\widetilde{\mathbf{Y}}_j^w)+\epsilon)) \\ &fea(\mathbf{E}_i^v,\mathbf{E}_j^w) = \frac{(\mathbf{E}_i^v)^T\mathbf{E}_j^w}{\parallel \mathbf{E}_i^v \parallel\parallel \mathbf{E}_j^w \parallel}, lab(\widetilde{\mathbf{Y}}_i^v,\widetilde{\mathbf{Y}}_j^w) = \frac{(\widetilde{\mathbf{Y}}_i^v)^T\widetilde{\mathbf{Y}}_j^w}{\parallel \widetilde{\mathbf{Y}}_i^v \parallel\parallel \widetilde{\mathbf{Y}}_j^w \parallel} \end{aligned} \tag{3}$$

where $fea(\mathbf{E}_i^v,\mathbf{E}_j^w)$ and $lab(\widetilde{\mathbf{Y}}_i^v,\widetilde{\mathbf{Y}}_j^w)$ are the feature-based and label-based similarity between $\mathbf{X}_i^v$ and $\mathbf{X}_j^w$, respectively. Two bags may be annotated with the same set of labels, which give a $lab(\widetilde{\mathbf{Y}}_i^v,\widetilde{\mathbf{Y}}_j^w) = 1$ and result in a large match score $m(\mathbf{X}_i^v,\mathbf{X}_j^w)$. However, these two bags may not be the best match, since they may have a moderate feature similarity. Given that, we add a small constant $\epsilon = 0.01$. The larger the feature-based and label-based similarities, and the more consistent between these two similarities, the more likely these two bags will be matched. To quantify the label and feature similarities between bags, we use the cosine similarity for its simplicity and effectiveness, other similarity metrics can also be used here. Since the feature and label vectors of our used datasets are all nonnegative, thus our cosine similarity actually locates in [0,1].

Based on $m(\mathbf{X}_i^v,\mathbf{X}_j^w)$, we can specify the matching matrix $\mathbf{M}^{vw}$ between bags of the v-th and w-th views as follows:

$$\mathbf{M}_{ij}^{vw} = \begin{cases} 1, m(\mathbf{X}_i^v,\mathbf{X}_j^w) \text{is the maximum or } p(\mathbf{X}_i^v,\mathbf{X}_j^w) = 1 \\ 0, \text{otherwise} \end{cases} \tag{4}$$

where p encodes the previously known matched information of bags across views, $p(\mathbf{X}_i^v,\mathbf{X}_j^w) = 1$ if $\mathbf{X}_i^v$ and $\mathbf{X}_j^w$ are known paired; $p(\mathbf{X}_i^v,\mathbf{X}_j^w) = 0$, otherwise. The first condition shows two matched cases of $\mathbf{X}_i^v$ and $\mathbf{X}_j^w$: (i) $\mathbf{X}_i^v$ and $\mathbf{X}_j^w$ are known matched in advance (i.e., $p(\mathbf{X}_i^v,\mathbf{X}_j^w) = 1$); (ii) $\mathbf{X}_i^v$ and $\mathbf{X}_j^w$ are calculated to have the maximum value of $m(\mathbf{X}_i^v,\mathbf{X}_j^w)$. If $\mathbf{X}_i^v$ and $\mathbf{X}_j^w$ meet the first condition, we set $\mathbf{M}_{ij}^{vw} = 1$; $\mathbf{M}_{ij}^{vw} = 0$, otherwise. As such, WSM3L can not only incorporate the known paired bags to deal with weakly-paired bags, but also deal with completely-unpaired bags, by leveraging feature and label similarities of bags.

3.2 Replenishing Labels of Bags

Most existing M3L approaches typically assume complete label annotations of bags, i.e., no missing labels. However, in practice, the annotation is indeed incomplete. Since each feature view has its distinctiveness and bags across views are only partially paired, we first replenish the missing labels of bags per view. We assume that missing labels of a bag can be replenished based on the labels of its neighborhood bags as follows:

$$\min_{\widetilde{\mathbf{Y}}^v} \mathcal{L}_2 = \sum\nolimits_{v=1}^{V} \parallel \mathbf{A}^v\mathbf{Y}^v - \widetilde{\mathbf{Y}}^v \parallel_F^2 \tag{5}$$

where $\widetilde{\mathbf{Y}}^v \in \mathbb{R}^{n_v\times q}$ represents the replenished label sets of bags in the v-th view. $\mathbf{A}^v \in \mathbb{R}^{n_v\times n_v}$ is the adjacency matrix of the k nearest neighborhood (kNN) graph of bags in the v-th view, and it's specified as follows:

$$\mathbf{A}^v(i,j) = \begin{cases} 1/k, \text{ if } \mathbf{X}_i^v \in \mathcal{N}_k(\mathbf{X}_j^v) \text{ or } \mathbf{X}_j^v \in \mathcal{N}_k(\mathbf{X}_i^v) \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

where $\mathbf{X}_i^v \in \mathcal{N}_k(\mathbf{X}_j^v)$ is one of the k nearest neighbors of $\mathbf{X}_j^v$, and the neighborhood relationship between bags is determined by the cosine similarity.

3.3 Distribution and Aggregation of Labels

In multi-instance learning, a bag includes one or more instances, and its label set depends on the labels of its instances

[Zhou *et al.*, 2012]. Multi-instance learning typically uses the bag-instance relations to predict the labels of bags; some approaches can also identify the labels of instances [Carbonneau *et al.*, 2018]. To perform label prediction for both bags and instances, we introduce a term to distribute the labels of bags to instances, and reversely aggregate instance-level labels for bags, as shown in the following:

$$\min_{\widetilde{\mathbf{Y}}^v, \mathbf{W}^v} \mathcal{L}_3 = \sum_{v=1, w\neq v}^{V} \| \widetilde{\mathbf{Y}}^v - \mathbf{\Lambda}^v \mathbf{M}^{vw} \mathbf{R}^w \mathbf{Z}^w \|_F^2 + \sum_{v=1}^{V} \| \mathbf{Z}^v - \mathbf{F}^v \mathbf{W}^v \|_F^2 \quad (7)$$

where $\widetilde{\mathbf{Y}}^v$ are the replenished label sets of bags in the v-th view. Unlike Eq. (6), the computation of $\widetilde{\mathbf{Y}}^v$ is coordinated with the matched bags (via $\mathbf{M}^{vw}$) from other views. $\mathbf{\Lambda}^v \in \mathbb{R}^{n_v \times n_v}$ is a diagonal matrix with $\mathbf{\Lambda}^v(i,i) = 1/mb(v,i)$, where $mb(v,i)$ counts the number of matched bags of $\mathbf{X}_i^v$, including itself. $\mathbf{R}^w \in \mathbb{R}^{n_w \times m_w}$ stores the inter-associations between n_w bags and m_w instances in the w-th view. $\mathbf{R}^w(i,j) = 1$ if the i-th bag includes the j-th instance; $\mathbf{R}^w(i,j) = 0$, otherwise. $\mathbf{F}^v \in \mathbb{R}^{m_v \times d_v}$ stores the feature vectors of instances of the v-th view, and $\mathbf{W}^v \in \mathbb{R}^{d_v \times q}$ is the projection matrix for the v-th view, $\mathbf{Z}^v \in \mathbb{R}^{m_v \times q}$ is the predicted label matrix of instances in the v-th view, which can be obtained by $\mathbf{F}^v$ and the optimized $\mathbf{W}^v$. As a result, our proposed WSM3L makes predictions for instances, and also aggregates instance labels at the bag-label. Meanwhile, it combines the replenished labels of bags across views.

3.4 The Unified Objective Function

To coordinate the match between bags across views and label replenishment, and to coherently dispatch the bag-level labels to instances and aggregate the instance-level labels onto bags, let $\Omega = \{\mathbf{D}, \mathbf{E}^v, \mathbf{P}^v, \widetilde{\mathbf{Y}}^v, \mathbf{W}^v\}$, we formulate a unified objective function as follows:

$$\min_{\Omega} \mathcal{L}_1 + \alpha(\mathcal{L}_2 + \mathcal{L}_3) \quad (8)$$

where $\mathcal{L}_1$ aims to control the data fidelity and to explore match across bags. $\mathcal{L}_2$ and $\mathcal{L}_3$ target to replenish and predict the labels of bags at bag-level and instance-level, respectively. Notice that $\mathcal{L}_1$ and $\mathcal{L}_3$ share the same match information. The parameter α balances the importance of $\mathcal{L}_1$ and the latter two terms. Eq. (8) makes the potential match between bags across views, label replenishment, bag-level and instance-level label prediction in a coordinated fashion. Thus, both the weakly-paired bags and incomplete labels of weakly-supervised learning on M3 data are jointly accounted for.

To compute $\mathbf{D}$, $\mathbf{E}^v$, $\mathbf{P}^v$, $\widetilde{\mathbf{Y}}^v$ and $\mathbf{W}^v$, we adopt an alternative optimization technique following the idea of the alternating direction method of multipliers (ADMM) [Boyd and Vandenberghe, 2004]. Since directly optimizing the discrete indicator match matrix $\mathbf{M}^{vw}$ is NP-hard, we update it based on the updated $\mathbf{E}^v$, $\widetilde{\mathbf{Y}}^v$ in each iteration. Suppose t is the maximum number of iterations, the time complexity of our model is $O(tV[sdn_v + Vd(n_v)^2 + d^2 n_v + m_v q + V(n_v)^2 m_v])$. Our preliminary study shows that WSM3L generally converges within 50 iterations on the used datasets. We give the optimization procedure as a supplementary file.

Dataset	#bag	#instance	#label	avgBI	avgBL
Pyrococcus_furiosus	425	1321	321	3.1	4.5
Caenorhabditis_elegans	2512	8509	940	3.4	6.1
Drosophila_melanogaster	2605	9146	1035	3.5	6.0
Saccharomyces_cerevisiae	3509	6533	1566	1.9	5.9
Isoform	2000	7907	258	4.0	3.9
Letter Frost	144	565	26	3.9	3.6
Letter Carroll	166	717	26	4.3	3.9
MSRC v2	591	1758	23	3.0	2.5
Birds	548	10232	13	18.7	2.1

Table 1: Statistics of datasets used for experiments. #bag, #instance and #label are the number of bags, instances and labels, respectively. avgBI/avgBL is the average number of instances/labels per bag.

To this end, WSM3L predicts the labels of a new bag $\mathbf{X}_h$ by integrating the aggregated labels from its instances and the known labels of it neighborhood training bags across views (if any) as follows:

$$f(\mathbf{X}_h) = \frac{1}{|\mathcal{V}(\mathbf{X}_h)|} \sum_{v \in \mathcal{V}(\mathbf{X}_h)} \left(\frac{1}{m_h^v} \sum_{j=1}^{m_h^v} \mathbf{x}_{h,j}^v \mathbf{W}^v + \frac{1}{k} \sum_{\mathbf{X}_j^v \in \mathcal{N}_k(\mathbf{X}_h^v)} \widetilde{\mathbf{Y}}_j^v \right) \quad (9)$$

where $\mathcal{V}(\mathbf{X}_h)$ collects the observed views of $\mathbf{X}_h$, $\mathbf{x}_{h,j}^v$ represents the j-th instance feature vector of $\mathbf{X}_h^v$, and $\mathbf{W}^v$ is the optimized coefficient matrix for instance-label prediction in the v-th view. The first term targets to aggregate the prediction from instance-level, and the second term aims to integrate the prediction from neighborhood training bags across views.

4 Experiments

4.1 Experimental Setup

We design experiments to study the performance of WSM3L on completely-paired bags, weakly-paired bags and completely-unpaired bags across views, respectively. We collect eight publicly available multi-instance multi-label datasets and one real M3 dataset from different domains for the experiments. The details of these datasets are listed in Table 1. The first four datasets[1] and Isoform dataset [Yu *et al.*, 2020] are used to evaluate the predicted labels of bags, since bag-level labels are available one. The last four datasets have instance-level labels for evaluation [Briggs *et al.*, 2012].

To evaluate the effectiveness of the proposed WSM3L, four widely-used multi-label evaluation metrics are adopted to evaluate the performance from different perspectives, including Hamming Loss (*HL*), Ranking Loss (*RL*), Average Precision (*AP*), and macro AUC (Area Under receiver operating Curve) (*mAUC*). Due to the page limit, the formal definition of these metrics is omitted here and can be found in [Zhang and Zhou, 2014]. The smaller the values of *HL* and *RL*, the better the performance is. As such, to be consistent with the other evaluation metrics, we report *1-RL* and *1-HL* instead. For the latter metrics, larger values indicate better performance.

4.2 Results on Completely Paired Multi-view Data

We randomly select 70% of the bags of a dataset to train the model, and use the remaining 30% for testing. For the eight multi-instance multi-label datasets, we randomly divide the original features of each bag into two sets of equal size, each providing one view. We then randomly mask 30% of the label

[1] http://lamda.nju.edu.cn/CH.Data.ashx

Metric	MIMLmix	M^2IL	M3Lcmf	MIMLwel	WSM3L	WSM3L(cL)
	Pyrococcus_furiosus					
1-HL	0.904●	0.974●	0.966●	0.630●	0.987	0.987
1-RL	0.527●	0.647●	0.740○	0.649●	0.697	0.718
AP	0.061●	0.148●	0.237	0.269○	0.244	0.281
mAUC	0.503●	0.525●	0.530●	0.563○	0.562	0.584
	Caenorhabditis_elegans					
1-HL	0.914●	0.985○	0.982○	0.631●	0.978	0.981
1-RL	0.641●	0.525●	0.773●	0.783●	0.801	0.819
AP	0.087●	0.089●	0.219●	0.393○	0.270	0.267
mAUC	0.562●	0.518●	0.561●	0.674○	0.669	0.685
	Drosophila_melanogaster					
1-HL	0.917●	0.993○	0.978	0.635●	0.978	0.981
1-RL	0.658●	0.423●	0.779●	0.781●	0.808	0.820
AP	0.089●	0.087●	0.179●	0.375○	0.245	0.253
mAUC	0.510●	0.516●	0.546●	0.689○	0.669	0.698
	Saccharomyces_cerevisiae					
1-HL	0.926●	0.989	0.989	0.650●	0.991	0.993
1-RL	0.666●	0.382●	0.752●	0.662●	0.782	0.783
AP	0.063●	0.063●	0.133●	0.155●	0.173	0.186
mAUC	0.556○	0.505●	0.528●	0.572○	0.552	0.568
	Isoform					
1-HL	0.933●	0.980	0.664●	0.527●	0.981	0.980
1-RL	0.568●	0.450●	0.655●	0.535●	0.676	0.679
AP	0.074●	0.033●	0.100○	0.075●	0.097	0.108
mAUC	0.546○	0.505●	0.505●	0.503●	0.533	0.543

Table 2: Results of bag-level label prediction with **completely paired** bags on different datasets. ●/○ indicates whether WSM3L is statistically (pairwise t-test at 95% significance level) superior/inferior to the other method. Unlike other compared methods, WSM3L(cL) operates on training data with 'complete labels' (or no label missed).

information of each bag in the training set, to study the performance of WSM3L on bags annotated with incomplete labels. For multi-view methods (i.e., MIMLmix [Nguyen *et al.*, 2014], M2IL[Li *et al.*, 2017] and M3Lcmf[Xing *et al.*, 2019]), we use the same datasets as our method, and for MIMLwel[Yang *et al.*, 2013], we directly use the collected datsets. Besides, the input parameters of all comparing methods used in this paper are specified (or optimized) as suggested by the authors in their papers or shared codes. For reference, we also report the results of WSM3L(cL), which does not mask any label but uses *complete labels*. The input parameters of WSM3L are set as follows: $d = 160$, $s = 150$, $k = 30$ and $\alpha = 1$. Tables 2 and 3 report the results of the comparing methods on bag-level and instance-level label prediction, respectively. Only MIMLmix and M3Lcmf can make instance-level prediction, so Table 3 does not report results of other compared methods.

Our proposed WSM3L generally outperforms the comparing methods across different datasets and evaluation metrics, on both the bag-level and instance-level label prediction tasks. We used the signed-rank test [Demšar, 2006] to check the significance of the results between WSM3L and the other methods, and all the p-values are smaller than 0.037. WSM3L frequently outperforms other M3L methods, which shows the effectiveness of WSM3L on completely paired multi-view datasets. Both M^2IL and WSM3L learn a shared dictionary across views, and WSM3L performs much better than M^2IL. This observation shows that WSM3L can learn a more adaptive dictionary. WSM3L outperforms MIMLwel which demonstrates the effectiveness of WSM3L on replenishing the labels of bags. WSM3L obtains a slightly lower performance than WSM3L(cL), which indicates the effectiveness of WSM3L on replenishing labels and also suggests WSM3L is not so sensitive to missing labels of training bags.

The results on instance-level prediction again expresses

Metric	MIMLmix	M3Lcmf	WSM3L
	Letter Frost		
1-HL	0.656●	0.644●	0.962
1-RL	0.406●	0.732	0.740
AP	0.191●	0.261●	0.286
mAUC	0.688○	0.513●	0.535
	Letter Carroll		
1-HL	0.649●	0.648●	0.962
1-RL	0.441●	0.697	0.702
AP	0.237●	0.247●	0.257
mAUC	0.686○	0.516	0.516
	MSRC v2		
1-HL	0.693●	0.768●	0.957
1-RL	0.582●	0.603●	0.704
AP	0.305●	0.368●	0.395
mAUC	0.625●	0.546●	0.727
	Birds		
1-HL	0.539○	0.876○	0.471
1-RL	0.524○	0.530○	0.445
AP	0.271○	0.075●	0.241
mAUC	0.503●	0.506	0.513

Table 3: Results of instance-level prediction on different datasets. ●/○ indicates whether WSM3L is statistically (according to a pairwise t-test at 95% significance level) superior/inferior to the other method.

the effectiveness of the proposed WSM3L in distributing the bag-level labels to instances, which in turn boosts the accuracy of bag-level label prediction. WSM3L sometimes loses to M3Lcmf and MIMLmix on the Birds dataset. The possible reason is that each bag in Birds has a large number of instances, WSM3L does not concretely use the relations between instances or labels as these compared methods, which boost the performance but result in a more complicated model.

4.3 Results on Weakly-paired Multi-view Data

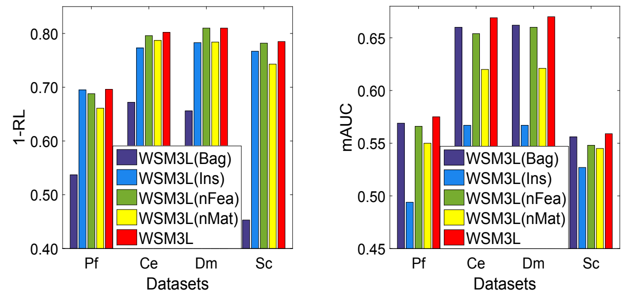
Based on the previous 70-30% split, we simulate three settings for weakly-supervised M3 data. In the first setting, we randomly mask the correspondence between 30% of the training bags, and then randomly remove 30% of the labels of the training bags. The second setting is the same as the previous one, with the additional removal of 30% of the bags in one view, to investigate the flexibility of WSM3L when different numbers of bags are present across views. In the third setting, we completely mask all the mappings between bags across views. For the last two settings, *none* of the comparing methods in Table 2 can be applied. We report the results of WSM3L(dB) and WSM3L(uB) in the last two columns of Table 4. WSM3L(dB) and WSM3L(uB) correspond to the case with *different numbers of bags* across views and to the case with *completely unpaired bags* across views, respectively. For a comprehensive comparison, two multi-view dictionary learning methods for weakly-paired data, WMCA (weakly-paired maximum covariance analysis) [Lampert and Krömer, 2010] and MFCDL (Multimodal Fusion via Common Dictionary Learning) [Liu *et al.*, 2018a] are also included for comparison in the first setting. Since WMCA does not provide the label likelihoods as other comparing methods, which are required by *1-RL* and *mAUC*, only the results of *1-HL* and *AP* are reported in Table 4.

We have the following observations:

(i) In the first setting, all comparing methods use all the training bags, and WSM3L achieves the best performance and holds comparable results with itself on completely paired bags in Table 2. This observation shows the effectiveness of WSM3L on learning from weakly-paired M3 data. Both WSM3L, WMCA and MFCDL can work on weakly-paired multi-view

Metric	MIMLmix	M^2IL	M3Lcmf	WMCA	MFCDL	WSM3L	WSM3L (dB)	WSM3L (uB)
	Pyrococcus_furiosus							
1-HL	0.899●	0.971●	0.969●	0.502●	0.929●	0.987	0.980	0.964
1-R	0.512●	0.661●	0.738○	−−	0.491●	0.698	0.669	0.695
AP	0.074●	0.122●	0.236○	0.093●	0.024●	0.235	0.226	0.232
mAUC	0.500●	0.513●	0.517●	−−	0.509●	0.565	0.551	0.552
	Drosophila_melanogaster							
1-HL	0.923●	0.992○	0.978	0.504●	0.953●	0.978	0.978	0.978
1-RL	0.648●	0.379●	0.776●	−−	0.516●	0.808	0.783	0.808
AP	0.073●	0.115●	0.179●	0.058●	0.012●	0.240	0.259	0.238
mAUC	0.523●	0.508●	0.556●	−−	0.501●	0.669	0.621	0.666
	Saccharomyces_cerevisiae							
1-HL	0.929●	0.993○	0.988	0.509●	0.951●	0.990	0.990	0.990
1-RL	0.684●	0.265●	0.755●	−−	0.488●	0.781	0.744	0.779
AP	0.064●	0.050●	0.131●	0.040●	0.008●	0.172	0.166	0.170
mAUC	0.535●	0.502●	0.561○	−−	0.507●	0.548	0.542	0.548
	Isoform							
1-HL	0.935●	0.979	0.577●	0.523●	0.778●	0.981	0.981	0.981
1-RL	0.539●	0.465●	0.668	−−	0.167●	0.675	0.667	0.674
AP	0.053●	0.036●	0.104○	0.030●	0.033●	0.100	0.099	0.099
mAUC	0.580○	0.502●	0.500●	−−	0.504●	0.536	0.529	0.535

Table 4: Results of bag-level prediction on **weakly paired** bags on different datasets. ●/○ indicates whether WSM3L is statistically (pairwise t-test at 95% significance level) superior/inferior to the other method. WSM3L(dB) and WSM3L(uB) respectively correspond to the results under the setting of *different numbers* of bags across views and the setting of *completely-unpaired* bags across views.



(a) 1-*RL* (b) *mAUC*

Figure 2: *1-RL* and *mAUC* of WSM3L variants on different datasets (Pf: Pyrococcus_furiosus, Ce: Caenorhabditis_elegans, Dm:Drosophila_melanogaster, Sc: Saccharomyces_cerevisiae).

data. WMCA adopts the maximum covariance analysis to match bags. WSM3L achieves a better performance than WMCA. This observation shows the effectiveness of WSM3L on handling multi-view weakly-paired data. Both WSM3L and MFCDL learn a shared dictionary of multiple views, WSM3L outperforms MFCDL, which facts the effectiveness of WSM3L on matching bags across views.
(ii) In the second setting, WSM3L(dB) operates with training data where some bags are missing in one view. WSM3L(dB) obtains a slightly lower performance compared to WSM3L in the first setting. This result shows that WSM3L can also work well in the case of bags with a different number of feature views.
(iii) In the third setting, WSM3L(uB), which operates on completely unpaired bags, achieves a performance comparable to the first setting. This shows that our strategy is reliable in finding the matching between bags. Overall these results prove the effectiveness of WSM3L on M3 data in different open settings.

4.4 Ablation Study

Four variants of WSM3L are designed to further explore the different contribution components of WSM3L with the setting of 70/30% split of training/testing set, and 30% correspondence between training bags randomly masked. The description of these variants is as follows:
(i) **WSM3L(Bag):** only uses neighbourhood information of bags to replenish missing labels of bags.
(ii) **WSM3L(Ins):** only considers the aggregated instance predictions to predict the labels of bags.
(iii) **WSM3L(nFea)):** only uses the label similarity for bag matching.
(iv) **WSM3L(nMat):** does not match bags across view.

From Figure 2, we observe that WSM3L outperforms its variants, which separately disregard different components of WSM3L. WSM3L(Bag) and WSMEL(Ins) are two components for label prediction of bags, they ignore the aggregated label predictions from instances and the predictions from neighbourhood bags, respectively. WSM3L outperforms them both. This observation suggests the significance of integrating these two types of label predictions in WSM3L. WSM3L(nMat) does not seek the matches between bags across views, and WSM3L(nFea) ignores the feature similarity of bags during the bag matching process. Both of them are outperformed by WSM3L, and WSM3L(nMat) achieves the lowest performance. These observations manifest the necessity of matching bags in M3L and the contribution of leveraging feature similarity and label similarity for matching bags. There is a small margin between WSM3L(nFea) and WSM3L in *1-RL* for Drosophila_melanogaster and Saccharomyces_cerevisiae datasets. The reason of such phenomenon is that these two datasets have a relatively large label space, which causes a low distinction of *1-RL*.

From these results, we can safely say that these components of WSM3L indeed deal with the multiplicity of learning on weakly-supervised M3 data.

We further investigated the sensitivity of four input parameters (i.e., α, k, s and d). We run WSM3L with different input values of α, k, combinations of s and d in the range of $[10^{-2}, 10^3]$, $[0, 100]$ and $[50, 300]$, respectively. We summarize the observations here: (i) α maintains a relatively stable and good performance when $\alpha > 0.1$, which suggests the importance of label replenishment; (ii) WSM3L achieves the lowest performance when $k = 0$, it then rises as k increases and has a good performance when k is close to 30; (iii) WSM3L achieves a stable performance under a wide range of combinations of d and s, and it achieves a good performance with d and s in $[150, 250]$. From these results, we can conclude that WSM3L is relatively robust to α, k, s and d. These results are given in the supplementary file(mlda.swu.edu.cn/WSM3L).

5 Conclusions

In this paper, we proposed a weakly-supervised multi-view multi-instance multi-label learning approach (WSM3L), which extends the flexibility of M3L on practical M3 data, whose matches between bags across views are partially (or completely) unknown, and the labels of bags are incomplete. WSM3L outperforms existing M3L algorithms under different practical settings, some of which existing M3L methods cannot handle.

References

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[Briggs *et al.*, 2012] Forrest Briggs, Xiaoli Z Fern, and Raviv Raich. Rank-loss support instance machines for miml instance annotation. In *KDD*, pages 534–542, 2012.

[Carbonneau *et al.*, 2018] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7(1):1–30, 2006.

[Gao *et al.*, 2015] Zan Gao, Hua Zhang, GP Xu, YB Xue, and Alexander G Hauptmann. Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. *Signal Processing*, 112:83–97, 2015.

[Huang *et al.*, 2019] Shengjun Huang, Wei Gao, and Zhihua Zhou. Fast multi-instance multi-label learning. *TPAMI*, 99(1):1–14, 2019.

[Lampert and Krömer, 2010] Christoph H Lampert and Oliver Krömer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *ECCV*, pages 566–579, 2010.

[Li *et al.*, 2017] Bing Li, Chunfeng Yuan, Weihua Xiong, Weiming Hu, Houwen Peng, Xinmiao Ding, and Steve Maybank. Multi-view multi-instance learning based on joint sparse representation and multi-view dictionary learning. *TPAMI*, 39(12):2554–2560, 2017.

[Liu *et al.*, 2018a] Huaping Liu, Fuchun Sun, Bin Fang, and Shan Lu. Multi-modal measurements fusion for surface material categorization. *IEEE Transactions on Instrumentation and Measurement*, 67(2):246–256, 2018.

[Liu *et al.*, 2018b] Huaping Liu, Feng Wang, Xinyu Zhang, and Fuchun Sun. Weakly-paired deep dictionary learning for cross-modal retrieval. *Pattern Recognition Letters*, 99(1):1–8, 2018.

[Mandal and Biswas, 2016] Devraj Mandal and Soma Biswas. Generalized coupled dictionary learning approach with applications to cross-modal matching. *TIP*, 25(8):3826–3837, 2016.

[Monaci *et al.*, 2007] Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhé, Sylvain Lesage, and Rémi Gribonval. Learning multi-modal dictionaries. *TIP*, 16(9):2272–2283, 2007.

[Nguyen *et al.*, 2013] Cam Tu Nguyen, De Chuan Zhan, and Zhi Hua Zhou. Multi-modal image annotation with multi-instance multi-label lda. In *IJCAI*, pages 1558–1564, 2013.

[Nguyen *et al.*, 2014] Cam Tu Nguyen, Xiaoliang Wang, Jing Liu, and Zhihua Zhou. Labeling complicated objects: multi-view multi-instance multi-label learning. In *AAAI*, pages 2013–2019, 2014.

[Rubinstein *et al.*, 2010] Ron Rubinstein, Alfred M Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[Tan *et al.*, 2018] Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *IJCAI*, pages 2703–2709, 2018.

[Wu *et al.*, 2016] Fei Wu, Xiaoyuan Jing, Xinge You, Dong Yue, Ruimin Hu, and Jingyu Yang. Multi-view low-rank dictionary learning for image classification. *Pattern Recognition*, 50:143–154, 2016.

[Xing *et al.*, 2018] Yuying Xing, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zhang Zili. Multi-label co-training. In *IJCAI*, pages 2882–2888, 2018.

[Xing *et al.*, 2019] Yuying Xing, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Zili Zhang, and Maozu Guo. Multi-view multi-instance multi-label learning based on collaborative matrix factorization. In *AAAI*, pages 5508–5515, 2019.

[Xu and Zhou, 2017] Miao Xu and Zhihua Zhou. Incomplete label distribution learning. In *IJCAI*, pages 3175–3181, 2017.

[Yang *et al.*, 2013] Shujun Yang, Yuan Jiang, and Zhihua Zhou. Multi-instance multi-label learning with weak label. In *IJCAI*, pages 1862–1868, 2013.

[Yang *et al.*, 2018] Yang Yang, Yifeng Wu, Dechuan Zhan, Zhibin Liu, and Yuan Jiang. Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport. In *KDD*, pages 2594–2603, 2018.

[Yang *et al.*, 2019] Yang Yang, Zhaoyang Fu, Dechuan Zhan, Zhibin Liu, and Yuan Jiang. Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport. *TKDE*, 2019.

[Yu *et al.*, 2020] Guoxian Yu, Keyao Wang, Carlotta Domeniconi, Maozu Guo, and Jun Wang. Isoform function prediction based on bi-random walks on a heterogeneous network. *Bioinformatics*, 36(1):303–310, 2020.

[Zhang and Zhou, 2014] Minling Zhang and Zhihua Zhou. A review on multi-label learning algorithms. *TKDE*, 26(8):1819–1837, 2014.

[Zhang *et al.*, 2015] Xianchao Zhang, Linlin Zong, Xinyue Liu, and Hong Yu. Constrained nmf-based multi-view clustering on unmapped data. In *AAAI*, pages 3174–3180, 2015.

[Zhou and Zhang, 2007] Zhihua Zhou and Minling Zhang. Solving multi-instance problems with classifier ensemble based on constructive clustering. *KAIS*, 11(2):155–170, 2007.

[Zhou *et al.*, 2012] Zhihua Zhou, Minling Zhang, Shengjun Huang, and Yufeng Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.