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

Yuying Xing¹, Guoxian Yu¹,²,*, Jun Wang¹, Carlotta Domeniconi³ and Xiangliang Zhang²

¹College of Computer and Information Sciences, Southwest University, Chongqing, China
²CEMSE, King Abdullah University of Science and Technology, Thuwal, SA
³Department of Computer Science, George Mason University, VA, USA

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

1 Introduction

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.

Abstract

Multi-view, Multi-instance, and Multi-label Learning (M3L) can model complex objects (bags), which are represented with different feature views, including diverse instances, and are simultaneously annotated with 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 different 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
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 reverse 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 et al., 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 (MiMILmix) 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 X_v = \{X_v^1, X_v^2, \cdots, X_v^n_v\}. X_v^i = [x_v^i]_{j=1}^{m_v^i}, a matrix, denotes the i-th bag in the v-th view includes m_v^i ≥ 1 instances, where x_v^i ∈ ℜ^{d_v} (v = 1, 2, \cdots, V) is the feature space of instances in the v-th view. Y_v^i ∈ ℜ^q encodes the currently known labels of X_v^i. Y_v^{iq} = 1 if X_v^i is annotated with the q-th label, Y_v^{iq} = 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, Y_v^i ∈ ℜ^q 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(\{X_v^i\}_{v=1}^{V}, \{Y_v^i\}_{v=1}^{V}) → 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 effec-
ative 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:

$$\argmin_{D_1, D_2, E_1, E_2} \| X_1 - D_1 E_1 \|^2_F + \| X_2 - D_2 E_2 \|^2_F + C(E_1 E_2)$$

where $D_1 \in \mathbb{R}^{d_1 \times d}$ and $D_2 \in \mathbb{R}^{d_2 \times d}$ are the dictionaries, and $E_1 \in \mathbb{R}^{d \times n_1}$ and $E_2 \in \mathbb{R}^{d \times n_2}$ are the coding matrices of the two views, respectively. $d$ is the dictionary size, which can be specified by the designer. The constraint term $C(E_1 E_2)$ 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 representation space for bags. In addition, we seek an encoding matrix of bags per view. Since the same bag may have different numbers 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, $X^v$ 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: $(P^v)^T X^v$, where $P^v \in \mathbb{R}^{d \times s}$ is the projection matrix of the $v$-th view with $P^v (P^v)^T = I \in \mathbb{R}^{s \times s}$. We then use $(P^v)^T X^v$ to seek the shared dictionary and the encoding matrix of bags for each view as follows:

$$\min_{D, E^v, P^v} L_1 = \sum_{v=1}^V \| (P^v)^T X^v - DE^v \|^2_F + \sum_{v=1, u \neq v}^V \| (E^v)^T - M^{uw} (E^w)^T \|^2_F$$

$$s.t. \| d_{uv} \|_2 \leq 1 (v' \in \{1, 2, \ldots, s\}), P^v (P^v)^T = I$$

where $D \in \mathbb{R}^{s \times d}$ is the shared dictionary of bags across views, and $d_{uv} \in \mathbb{R}^d$ is the dictionary vector of $D$. $E^u \in \mathbb{R}^{d \times n_u}$ 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 representation space, which is configured by the shared dictionary. $M^{uw} \in \mathbb{R}^{n_u \times n_w}$ records the mapping information between bags of the $v$-th view and $w$-th view. The term $\sum_{v=1, u \neq v}^V \| (E^v)^T - M^{uw} (E^w)^T \|^2_F$ 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 $(X^v_i, X^w_j)$ as follows:

$$m(X^v_i, X^w_j) = 1 - (1 - f_{ea}(E^v_i, E^w_j)(1 - lab(\tilde{Y}^v_i, \tilde{Y}^w_j) + \epsilon))$$

$$f_{ea}(E^v_i, E^w_j) = \frac{(E^v_i)^T E^w_j}{\| E^v_i \| \| E^w_j \|} lab(\tilde{Y}^v_i, \tilde{Y}^w_j) = \frac{(\tilde{Y}^v_i)^T \tilde{Y}^w_j}{\| \tilde{Y}^v_i \| \| \tilde{Y}^w_j \|}$$

where $f_{ea}(E^v_i, E^w_j)$ and $lab(\tilde{Y}^v_i, \tilde{Y}^w_j)$ are the feature-based and label-based similarity between $X^v_i$ and $X^w_j$, respectively. Two bags may be annotated with the same set of labels, which give a large match score $m(X^v_i, X^w_j) = 1$ and result in a large match score $m(X^v_i, X^w_j)$. 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 $m(X^v_i, X^w_j)$, we can specify the matching matrix $M^{vw}$ between bags of the $v$-th and $w$-th views as follows:

$$M^{vw}_{ij} = \begin{cases} 1, m(X^v_i, X^w_j) \text{is the maximum or } p(X^v_i, X^v_j) = 1 \\ 0, \text{otherwise} \end{cases}$$

where $p$ encodes the previously known matched information of bags across views, $p(X^v_i, X^w_j) = 1$ if $X^v_i$ and $X^w_j$ are known paired; $p(X^v_i, X^w_j) = 0$, otherwise. The first condition shows two matched cases of $X^v_i$ and $X^w_j$: (i) $X^v_i$ and $X^w_j$ are known matched in advance (i.e., $p(X^v_i, X^w_j) = 1$); (ii) $X^v_i$ and $X^w_j$ are calculated to have the maximum value of $m(X^v_i, X^w_j)$. If $X^v_i$ and $X^w_j$ meet the first condition, we set $M^{vw}_{ij} = 1$; otherwise, $M^{vw}_{ij} = 0$. 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:

$$\min_{\tilde{Y}^v} L_2 = \sum_{v=1}^V \| A^v Y^v - \tilde{Y}^v \|^2_F$$

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

$$A^v(i, j) = \begin{cases} 1/k, & \text{if } X^v_i \in N_k(X^v_j) \text{ or } X^v_j \in N_k(X^v_i) \\ 0, & \text{otherwise} \end{cases}$$

where $X^v_i \in N_k(X^v_j)$ is one of the $k$ nearest neighbors of $X^v_j$, 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:

$$\min_{\hat{Y}^v, W^v} \mathcal{L}_3 = \sum_{v=1}^{V} \frac{1}{2} \| \hat{Y}^v - A^v M^v W^v Z^v \|^2_F + \sum_{v=1}^{V} \| Z^v - F^v W^v \|^2_F$$

(7)

where \(\hat{Y}^v\) are the replenished label sets of bags in the \(v\)-th view. Unlike Eq. (6), the computation of \(\hat{Y}^v\) is coordinated with the matched bags (via \(M^v \)) from other views. \(\Lambda^v \in \mathbb{R}^{n_w \times n_v}\) is a diagonal matrix with \(\Lambda^v(i,i) = 1/m(b(i)), \) where \(m(b(i))\) counts the number of matched bags of \(X_i\), including itself. \(R^v \in \mathbb{R}^{n_w \times m_w}\) stores the inter-associations between \(n_w\) bags and \(m_w\) instances in the \(v\)-th view. \(R^v(i,j) = 1\) if the \(i\)-th bag includes the \(j\)-th instance; \(R^v(i,j) = 0\), otherwise. \(F^v \in \mathbb{R}^{m_w \times d}\) stores the feature vectors of instances of the \(v\)-th view, and \(W^v \in \mathbb{R}^{d \times q}\) is the projection matrix for the \(v\)-th view. \(Z^v \in \mathbb{R}^{m_w \times q}\) is the predicted label matrix of instances in the \(v\)-th view, which can be obtained by \(F^v\) and the optimized \(W^v\). 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 \(\Omega = \{D, E^v, P^v, \hat{Y}^v, W^v\}\), we formulate a unified objective function as follows:

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

(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 \(\mathcal{L}_2\) share the same match across views. The parameter \(\alpha\) 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 \(D, E^v, P^v, \hat{Y}^v\) and \(W^v\), 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 \(M^v\) is NP-hard, we update it based on the updated \(E^v, \hat{Y}^v\) in each iteration. Suppose \(t\) is the maximum number of iterations, the time complexity of our model is \(O(tVSd n_w + V d(n_u)^2 + d^2 n_v + m_u q + V(n_u)^2 m_v))\). Our preliminary study shows that WSM3L generally converges within 50 iterations on the used datasets. We give the optimization procedure as a supplementary file.

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

$$f(x_h) = \frac{1}{|V(X_h)|} \sum_{v \in V(X_h)} \left( \frac{1}{m_b} \sum_{j=1}^{m_b} x_{h,j}^v W^v + \frac{1}{k} \sum_{k \neq h \in N_3(X_h)} \hat{Y}_{h,j}^v \right)$$

(9)

where \(V(X_h)\) collects the observed views of \(X_h\), \(x_{h,j}^v\) represents the \(j\)-th instance feature vector of \(X_h\), and \(W^v\) 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 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-RL 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

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#bag</th>
<th>#instance</th>
<th>#label</th>
<th>avgBl</th>
<th>avgBL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrococcus furiosus</td>
<td>425</td>
<td>1321</td>
<td>321</td>
<td>3.1</td>
<td>4.5</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>2512</td>
<td>8509</td>
<td>940</td>
<td>3.4</td>
<td>6.1</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>2605</td>
<td>9146</td>
<td>1035</td>
<td>3.5</td>
<td>6.0</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>3509</td>
<td>6533</td>
<td>1566</td>
<td>1.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Isoform</td>
<td>2000</td>
<td>7907</td>
<td>258</td>
<td>4.0</td>
<td>3.9</td>
</tr>
<tr>
<td>Letter Frost</td>
<td>1414</td>
<td>565</td>
<td>26</td>
<td>3.9</td>
<td>3.6</td>
</tr>
<tr>
<td>Letter Carroll</td>
<td>166</td>
<td>717</td>
<td>26</td>
<td>4.3</td>
<td>3.9</td>
</tr>
<tr>
<td>MSRC v2</td>
<td>591</td>
<td>1758</td>
<td>23</td>
<td>3.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Birds</td>
<td>548</td>
<td>10232</td>
<td>13</td>
<td>18.7</td>
<td>2.1</td>
</tr>
</tbody>
</table>

1http://lamda.nju.edu.cn/CH.Data.ashx
We used the signed-rank test [Demšar, 2006] to check 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 I-RL and mAUC, only the results of I-RL 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.

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>MIMLmix</th>
<th>M3Lcmf</th>
<th>WSM3L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-RL</td>
<td>0.658●</td>
<td>0.644○</td>
<td>0.962●</td>
</tr>
<tr>
<td>1-RL</td>
<td>0.406●</td>
<td>0.732●</td>
<td>0.740○</td>
</tr>
<tr>
<td>AP</td>
<td>0.191●</td>
<td>0.261●</td>
<td>0.286○</td>
</tr>
<tr>
<td>mAUC</td>
<td>0.688●</td>
<td>0.513●</td>
<td>0.535○</td>
</tr>
</tbody>
</table>

We present 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.
Table 4: Results of bag-level prediction on weakly paired bags on different datasets. */o 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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>WSM3L(Bag)</th>
<th>WSM3L(Ins)</th>
<th>WSM3L(nFea)</th>
<th>WSM3L(nMat)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pf: Pyrococcus_furiosus</td>
<td>0.504</td>
<td>0.100</td>
<td>0.099</td>
<td>0.099</td>
</tr>
<tr>
<td>Ce: Caenorhabditis_elegans</td>
<td>0.778</td>
<td>0.008</td>
<td>0.488</td>
<td>0.951</td>
</tr>
<tr>
<td>Dm: Drosophila_melanogaster</td>
<td>0.501</td>
<td>0.012</td>
<td>0.379</td>
<td>0.776</td>
</tr>
<tr>
<td>Sc: Saccharomyces_cerevisiae</td>
<td>0.951</td>
<td>0.131</td>
<td>0.488</td>
<td>0.504</td>
</tr>
</tbody>
</table>

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 WSM3L(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 M3 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 α > 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


