Multi-View Multi-Instance Multi-Label Learning based on Collaborative Matrix Factorization*

Yuying Xing[1], Guoxian Yu[1,2,*], Carlotta Domeniconi[3], Jun Wang[1], Zili Zhang[1,4] and Maozu Guo[5]

[1]College of Computer and Information Science, Southwest University, Chongqing, China
[2]Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan, China
[3]Department of Computer Science, George Mason University, Fairfax, USA
[4]School of Information Technology, Deakin University, Geelong, Australia
[5]School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China
{yyxing4148, gxyu, kingjun, zhangzl}@swu.edu.cn, carlotta@cs.gmu.edu, guomaozu@bucea.edu.cn

Abstract

Multi-view Multi-instance Multi-label Learning(M3L) deals with complex objects encompassing diverse instances, represented with different feature views, and annotated with multiple labels. Existing M3L solutions only partially explore the inter or intra relations between objects (or bags), instances, and labels, which can convey important contextual information for M3L. As such, they may have a compromised performance.

In this paper, we propose a collaborative matrix factorization based solution called M3Lcmf. M3Lcmf first uses a heterogeneous network composed of nodes of bags, instances, and labels, to encode different types of relations via multiple relational data matrices. To preserve the intrinsic structure of the data matrices, M3Lcmf collaboratively factorizes them into low-rank matrices, explores the latent relationships between bags, instances, and labels, and selectively merges the data matrices. An aggregation scheme is further introduced to aggregate the instance-level labels into bag-level and to guide the factorization. An empirical study on benchmark datasets show that M3Lcmf outperforms other related competitive solutions both in the instance-level and bag-level prediction.

Introduction

Multi-Instance Multi-Label learning (MIML) is a framework for modeling complex objects, in which each object (or bag) contains one or more instances and is annotated by several semantic labels (Zhou et al. 2012). Let's consider n bags $\mathcal{B}_i = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \cdots, \mathbf{x}_{i_k}\}$ $(i = 1, \ldots, n)$, where each bag encompasses $n_i \geq 1$ instances, and $\mathbf{x}_{i_j} \in \mathbb{R}^d$ is the feature vector of the j-th instance of the i-th bag. The n bags and the $m = \sum_{i=1}^{n} n_i$ instances are annotated with q distinct labels. $\mathbf{Y}_i \in \mathbb{R}^{1\times q}$ is the q-dimensional label vector for the i-th bag. Given a training dataset $\mathcal{D} = \{(\mathcal{B}_i, \mathbf{Y}_i)\}_{i=1}^{n}$, MIML aims at learning an instance-level $f(\mathbf{x}) \in \mathbb{R}^q$ (or bag-level) predictor, which maps the input features of instances (or bags) onto the label space.

Most MIML algorithms focus on single view data, where instances of bags are represented by one set of features. However, in real-world applications, a multi-instance multi-label object can often be represented via different views (Nguyen, Zhan, and Zhou 2013; Shao et al. 2016). For example, as shown in Figure 1, three exemplar bags encompassing diverse instances are represented with V heterogenous feature views. Since there are multi-type relations between bags and between instances, learning from multi-view bags is more difficult and challenging than the recently heavily studied MIML task (Feng and Zhou 2017; Zhu, Ting, and Zhou 2017).

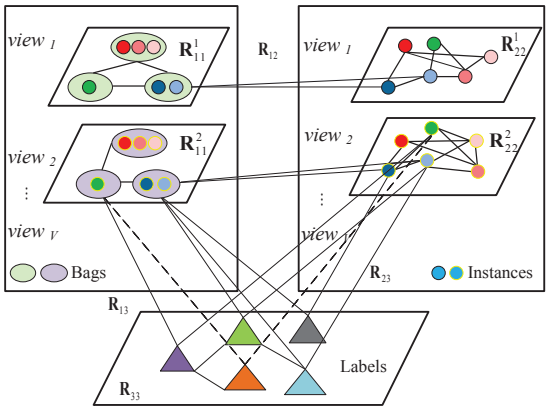


Figure 1: An illustrative example of multi-view multi-instance multi-label objects. $\{\mathbf{R}_{ij}^{v}\}_{v=1}^{V}$ are the multi-type relational data matrices between bags (objects), instances, and labels across V heterogeneous feature views.

Several Multi-view Multi-instance Multi-label Learning (M3L) approaches have been proposed to tackle this challenge (Nguyen, Zhan, and Zhou 2013; Nguyen et al. 2014; Li et al. 2017; Yang et al. 2018). Nguyen, Zhan, and Zhou (2013) pioneered an approach called M3LDA, which employs Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) to explore the visual-label topics from the visual view and the text-label topics from the text view, and then enforces the predicted labels from the two respective views to be consistent. Nguyen et al. (2014) introduced another M3L approach, called MIMLmix, to leverage multiple views using a hierarchical Bayesian network and variational inference. MIMLmix can handle samples which are absent in some views. Li et al. (2017) developed a multi-view multi-instance learning algorithm (M^2IL), which generates different graphs with different parameters to represent various contextual relations between instances of a bag. It then integrates these graphs in-

*Corresponding author, gxyu@swu.edu.cn (Guoxian Yu).

Table 1: Relations exploited by representative M3L and MIML methods.

	Relations					
	bag-bag	instance-instance	label-label	bag-instance	bag-label	instance-label
M3LDA(Nguyen, Zhan, and Zhou 2013)			✓	✓	✓	✓
MIMLmix(Nguyen et al. 2014)			✓	✓	✓	✓
M3DN(Yang et al. 2018)			✓	✓	✓	✓
M^2IL(Li et al. 2017)		✓		✓	✓	
MIMLSVM(Zhou et al. 2008)	✓			✓	✓	
MIMLfast(Huang, Gao, and Zhou 2018)			✓	✓	✓	
MIMLRBF(Zhang and Wang 2009)				✓	✓	✓
Proposed M3Lcmf	✓	✓	✓	✓	✓	✓

to a unified framework for bag classification based on sparse representation and multi-view dictionary learning. Yang et al. (2018) introduced a deep neural network based approach called M3DN. M3DN separately applies a deep network for each view, and requires the bag-based predictions from different views to be consistent within the same bag. In addition, M3DN adopts the Optimal Transport theory (Villani 2008) to capture the geometric information of the underlying label space and to quantify the quality of predictions.

However, these M3L approaches, like MIML solutions, only consider *limited* types of relations between bags or between instances, as summarized in Table 1. M3L approaches generally capture the relations between bags and instances, and the associations between bags and labels. Some approaches additionally exploit the relations between bags (Zhou et al. 2008), between instances (Li et al. 2017), and the correlations between labels (Huang, Gao, and Zhou 2018; Yang et al. 2018). Furthermore, other approaches use the associations between instances and labels (Zhang and Wang 2009; Nguyen et al. 2014) to learn labels of bags at the instance level. All these types of relations simultaneously exist in M3L, however, *none* of the existing solutions explicitly accounts for all these relations.

To take advantage of multiple feature views of instances (or bags), an intuitive solution is to concatenate features from different views into a long vector, and then to apply MIML algorithms on the concatenated vector. However, this concatenation causes over-fitting on a small number of training samples, and ignores the specific statistical property of each view (Xu, Tao, and Xu 2013). Ensemble learning can also work on multi-view data and MIML classifiers are readily available for each view. But the base classifiers are separately trained on individual views; as such, they may have a low performance given the insufficient information of each view and the neglect of complementary information across views. Subspace learning-based approaches (He et al. 2016; Tan et al. 2018) aim at obtaining a latent subspace shared by multiple views under the assumption that the input views are generated from a latent subspace. Latent subspace-based solutions may alleviate the issue of the "curse of dimensionality", but may neglect the intrinsic structure of individual views. For multi-view data, the intrinsic structures of bags and instances may be different across views. Therefore, a competent M3L approach should account for *multiple types of relations* between bags, instances and labels, and the intrinsic structures of different feature views.

In this paper, we introduce an approach called M3Lcmf. M3Lcmf first constructs a heterogeneous network composed of nodes of bags, instances, and labels, to capture the *intra-relations* between nodes of the same type, *inter-relations* between bags and instances, between bags and labels, and between instances and labels. To respect and employ the intrinsic structure of the subnetworks of the intra and inter-relations, it collaboratively factorizes the association matrices of the subnetworks into low-rank matrices to pursue the low-rank representation of the nodes and the latent relationships among them, and also to selectively integrate multiple feature views of bags and instances. M3Lcmf additionally introduces an aggregation term into the factorization objective, which not only can aggregate the instance-label associations into bag-level, but also can reversely guide the prediction of these associations. The main contributions of this work are summarized as follows:

(i) Unlike existing solutions that can only account for several types of relations between bags and instances, M3Lcmf can simultaneously take into account multiple types of relations between bags, instances, and labels.

(ii) Our proposed M3Lcmf can selectively combine multiple feature views of bags and instances, preserve multiple intrinsic intra- and inter-relations without mapping inter-relations into the homologous network of bags or instances. It can make predictions at the instance-level and automatically aggregate the predictions to the bag-level.

(iii) Experimental results on benchmark datasets show that M3Lcmf performs favorably against the recently proposed M3L approaches MIMLmix (Nguyen et al. 2014) and M^2IL (Li et al. 2017), and other representative MIML methods (including MIMLSVM (Zhou et al. 2008), MIMLNN (Zhou et al. 2012), MIMLRBF (Zhang and Wang 2009) and MIMLfast (Huang, Gao, and Zhou 2018)). M3Lcmf is also robust to a wide range of input parameters.

The Proposed Method

Problem Formulation

Without loss of generality, we assume instances (or bags) have V feature views, $\mathcal{B}_i^v = \{\mathbf{x}_{i_1}^v, \mathbf{x}_{i_2}^v, \cdots, \mathbf{x}_{i_k}^v\}$, 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. $\mathbf{Y}_i \in \mathbb{R}^{1\times q}$ is the q-dimensional label space

for the i-th bag across all the views. The task of M3L is to learn a predictive function $f(\{\mathcal{B}^v\}_{v=1}^V, \mathbf{Y}) \in \mathbb{R}^q$, which maps multiple input feature views onto the label space.

To address this task, we first construct a heterogeneous network to encode multiple types of relations between bags, instances, and labels. Next, we collaboratively factorize the relational data matrices of the heterogeneous network into low-rank matrices, and predict the instance-label association based on the respective low-rank matrices; we then aggregate the instance-level predictions onto bag-level. The following two subsections elaborate on the network construction and collaborative matrix factorization.

Heterogeneous Network Construction

As shown in Figure 1, there are three types of nodes in the heterogeneous network: bags, instances, and labels. Each type of nodes has a different intrinsic structure. Bags and instances can have multiple heterogeneous feature views, which often provide complementary information. We first construct a heterogeneous network to represent intrinsic structures between nodes of multiple information sources.

It is recognized that relations among instances in a bag convey important contextual information in multi-instance learning, and they influence the overall performance (Li et al. 2017). To explore the intrinsic structure of instances, we construct a subnetwork of instances for each feature view. For simplicity, we measure the relation between $\mathbf{x}_i^v$ and $\mathbf{x}_j^v$ in the v-th view using the Gaussian heat kernel $\mathbf{R}_{11}^v(i,j) = exp(-\frac{||\mathbf{x}_i^v - \mathbf{x}_j^v||_F^2}{\sigma^2})$, where σ is the average Euclidean distance between all the m instances of the v-th view.

In M3L, a bag contains one or more instances and has its own characteristics, which are different from those of instances. Here, we construct a bag subnetwork to capture the contextual information of bags based on a composite Hausdorff distance for each view as follows:

$$\mathbf{H}(i,j) = \frac{1}{3}\sum_{p\in\eta}\begin{cases} \frac{\sum_{a\in\mathcal{B}_i^v} min_{b\in\mathcal{B}_j^v} d(a,b) + \sum_{b\in\mathcal{B}_j^v} min_{a\in\mathcal{B}_i^v} d(a,b)}{|\mathcal{B}_i^v| + |\mathcal{B}_j^v|} \\ (p = avg), \\ \\ max\{max_{a\in\mathcal{B}_i^v} min_{b\in\mathcal{B}_j^v} d(a,b), \\ max_{b\in\mathcal{B}_j^v} min_{a\in\mathcal{B}_i^v} d(a,b)\}(p = max), \\ \\ min_{a\in\mathcal{B}_i^v} min_{b\in\mathcal{B}_j^v} d(a,b)(p = min) \end{cases}$$

where $p \in \eta = \{`avg', `max', `min'\}$, $d(a,b)$ is the Euclidean distance between two instances (a and b). Then, we define $\mathbf{R}_{22}^v(i,j) = exp(-\frac{\mathbf{H}(i,j)}{\sigma_H^2})$ as the similarity between the i-th bag and j-th bag in the v-th view, and σ_H is set to the average composite Hausdorff distance between all the bags of this view. These three types of Hausdorff distances are widely used in MIML(Zhou et al. 2012). Different Hausdorff distances have different focuses. The minimal Hausdorff distance indicates the minimal distance between all instances of one bag and those of another bag; the maximal Hausdorff distance computes the maximum distance between instances of a bag and the nearest instances of another bag; while the average Hausdorff distance takes into account more geometric relations between instances of two bags (Zhang and Zhou 2009). This composite similarity can integrate the merits of the Hausdorff distance metrics.

In M3L, each bag is simultaneously annotated with several semantic labels, and the labels are not mutually exclusive. Different pairs of labels may have different degrees of correlation. Label correlation can be leveraged to boost the performance multi-label learning (Zhang and Zhou 2014). To quantify label correlations, we adopt the widely used cosine similarity to construct a subnetwork of labels. Since instances and bags share the same label space, only one label subnetwork is constructed. Let $\mathbf{Y}(\cdot, c) \in \mathbb{R}^n$ store the distribution of label c across all the bags. The correlation between two labels c_1 and c_2 can be empirically estimated as follows:

$$\mathbf{R}_{33}(c_1, c_2) = \frac{\mathbf{Y}(\cdot, c_1)^T \mathbf{Y}(\cdot, c_2)}{\| \mathbf{Y}(\cdot, c_1) \| \| \mathbf{Y}(\cdot, c_2) \|} \tag{1}$$

The specific distance metrics used to construct the three types of intra-relations in the subnetworks have been chosen for their simplicity and wide applicability. Other distance metrics can be used as well.

There are three types of inter-relations between bags, instances, and labels. The bag-instance inter-relational data matrix $\mathbf{R}_{12} \in \mathbf{R}^{n\times m}$ can be specified based on the known bag-instance associations, which are readily available in multi-instance data. The bag-label relational matrix $\mathbf{R}_{13} \in \mathbf{R}^{n\times q}$ can be directly specified based on the known labels of bags. For the instance-label relational data matrix $\mathbf{R}_{23} \in \mathbf{R}^{m\times q}$, since the initial labels of instances are generally unknown in multi-instance learning, we initially set $\mathbf{R}_{23} = \mathbf{0}$. If the labels of instances are partially known, we can also specify $\mathbf{R}_{23}$ based on the known labels of instances.

By referring to Table 1, we can say that the heterogeneous network can account for all types of relations between bags, instances, and labels.

Collaborative Matrix Factorization

To combine multiple intra-relational data matrices $\mathbf{R}_{11}^v$ and $\mathbf{R}_{22}^v$, we can project all the data matrices onto a composite instance-instance intra-relational data matrix, or onto a composite bag-bag intra-relational data matrix, and then make prediction on the composite relational data matrix. This projection idea has been used to integrate multiple interconnected subnetworks (Gligorijević and Pržulj 2015). However, this projection may enshroud the intrinsic structures of different relational data matrices and compromise the performance. Zitnik and Zupan (2015) recently introduced a data fusion framework (DFMF) based on matrix factorization. This framework does not need to map a heterogeneous network into a small homologous network, and it can leverage and preserve the intrinsic structures of multiple relational data matrices. The objective function of this framework is as follows:

$$\begin{aligned} min_{\mathbf{G}\geq 0} \mathbf{Z}(\mathbf{G}, \mathbf{S}) = & \sum_{\mathbf{R}_{ij}\in\mathcal{R}} ||\mathbf{R}_{ij} - \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T||_F^2 \\ & + \sum_{t=1}^{max_i t_i} tr(\mathbf{G}^\mathbf{T} \Theta^{(t)} \mathbf{G}) \end{aligned} \tag{2}$$

where $||\cdot||_F^2$ is the Frobenius norm. $\mathbf{R}_{ij} \in \mathbb{R}^{n_i \times n_j}, i, j \in \{1, 2, \cdots, N\}$ stores the inter-relation between the i-th object and the j-th object. $\mathbf{G}_i \in \mathbb{R}^{n_i \times d_i}$, $\mathbf{G}_j \in \mathbb{R}^{n_j \times d_j}$, $\mathbf{S}_{ij} \in \mathbb{R}^{d_i \times d_j} (d_i \ll n_i, d_j \ll n_j)$, $\mathbf{G} = diag(\mathbf{G}_1, ..., \mathbf{G}_N)$ where $\mathbf{G}_i$ is the low rank representation of the i-th object type, and N is the number of object types. Suppose the i-th type of objects has t_i data sources, represented by t_i constraint matrices $\{\Theta_i^t \in \mathbb{R}^{n_i \times n_i}\}_{t=1}^{t_i} (t \in \{1, ...max_i t_i\})$. $\mathbf{\Theta}^{(t)} = diag(\mathbf{\Theta}_1^{(t)}, ..., \Theta_N^{(t)})$, which collectively stores all the block diagonal matrices.

Based on the constructed heterogeneous network, and for the non-negativity of the inter and intra-relational data matrices, we extend Eq. (2) and define the objective function of M3Lcmf as follows:

$$\begin{aligned} min_{\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3 \geq 0} \mathbf{Z}(\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3) &= ||\mathbf{R}_{12} - \mathbf{G}_1 \mathbf{G}_2^T||_F^2 \\ &+ ||\mathbf{R}_{13} - \mathbf{G}_1 \mathbf{G}_3^T||_F^2 \\ &+ ||\mathbf{R}_{13} - \mathbf{\Lambda} \mathbf{R}_{12} \mathbf{G}_2 \mathbf{G}_3^T||_F^2 \\ &+ MR(\mathbf{G}) \end{aligned} \tag{3}$$

where $\mathbf{G}_1 \in \mathbb{R}^{n \times d}$, $\mathbf{G}_2 \in \mathbb{R}^{m \times d}$, and $\mathbf{G}_3 \in \mathbb{R}^{q \times d}$ are the low rank representations of multiple bags, instances, and labels, respectively. M3Lcmf has two prediction objectives. The first one is to predict instance-label associations $\mathbf{R}_{23}$ by approximating it to $\mathbf{G}_2 \mathbf{G}_3^T$. The other objective is to predict labels of bags by approximating $\mathbf{R}_{13}$ to $\mathbf{G}_1 \mathbf{G}_3^T$. Instead of approximating $\mathbf{R}_{13}$ by $\mathbf{G}_1 \mathbf{G}_3^T$, we add an aggregation term $||\mathbf{R}_{13} - \mathbf{\Lambda} \mathbf{R}_{12} \mathbf{G}_2 \mathbf{G}_3^T||_F^2$ into Eq. (3) to aggregate label information of instances to their originating bags. $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ is a diagonal matrix, and $\mathbf{\Lambda}(i, i) = 1/n_i$. This aggregation term is also driven by the multi-instance learning principle that the labels of a bag depend on the labels of its instances. Note, this aggregation term can reversely guide the pursue of $\mathbf{G}_2$ and $\mathbf{G}_3$. As such, the labels of instance can also be learnt from those of bags. The last term $MR(\mathbf{G})$ is the manifold regularization (Belkin, Niyogi, and Sindhwani 2006) on $\mathbf{G}$.

The intra-relations between bags, instances, and labels carry important contextual information, whose usage can improve the overall performance. Since $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$ can be viewed as the latent low-dimensional representation of bags, instances, and labels, we follow the idea of manifold regularization to enforce two data points with a high intra-association value being nearby in the low-dimensional space, and formulate the last term in Eq. (3) as below to use three types of intra-associations:

$$\begin{aligned} MR(\mathbf{G}) &= \sum_{v=1}^{V} \boldsymbol{\alpha}_v tr(\mathbf{G}_1^T (\mathbf{D}_{11}^v - \mathbf{R}_{11}^v) \mathbf{G}_1) \\ &+ \sum_{v=1}^{V} \boldsymbol{\beta}_v tr(\mathbf{G}_2^T (\mathbf{D}_{22}^v - \mathbf{R}_{22}^v) \mathbf{G}_2) \\ &+ tr(\mathbf{G}_3^T (\mathbf{D}_{33} - \mathbf{R}_{33}) \mathbf{G}_3) + \lambda_1 ||\boldsymbol{\alpha}||_F^2 + \lambda_2 ||\boldsymbol{\beta}||_F^2 \\ s.t. &\sum_{v=1}^{V} \boldsymbol{\alpha}_v = 1, \sum_{v=1}^{V} \boldsymbol{\beta}_v = 1. \end{aligned} \tag{4}$$

where $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ are two parameters to balance the importance of the v-th bag view and v-th instance view, respectively. $\mathbf{D}_{11}^v$ and $\mathbf{D}_{22}^v$ are two series of diagonal matrices, with each diagonal entry equal to the row sum of $\mathbf{R}_{11}^v$ and $\mathbf{R}_{22}^v$, respectively; $\mathbf{D}_{33}$ follows a similar definition. $tr(\mathbf{G}_1^T (\mathbf{D}_{11}^v - \mathbf{R}_{11}^v) \mathbf{G}_1)$ can be viewed as the smoothness loss on the v-th bag view. $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are introduced to avoid selecting single view alone. If these two parameters are excluded, only $\mathbf{R}_{11}^v$ and $\mathbf{R}_{22}^v$ with the smallest loss will be selected. Our empirical study shows that $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ can indeed selectively integrate different views and reduce the impact of noisy views by assigning smaller or zero weights to them. We can see that DFMF equally treats all the relational matrices $\{\mathbf{R}_{ij}^v\}_{i,j=1}^3$, it does not differentiate the different degrees of relevance of $\{\mathbf{R}_{11}^v\}_{v=1}^V$ and $\{\mathbf{R}_{22}^v\}_{v=1}^V$ toward the prediction task. Unlike DFMF, which simply reverses the sign of $\{\mathbf{R}_{i,i}^v\}_{v=1}^V (i \in \{1, 2, 3\})$ to fulfil $\mathbf{\Theta}^{(t)}$ in Eq. (2), M3Lcmf uses the graph Laplacian matrix to guide the approximation, and has a good geometric explanation.

From the above analysis, we can conclude that M3Lcmf can predict labels for complicated objects both at instance-level and bag-level, and can simultaneously preserve multi-type relations between bags and instances. Besides the aggregation term, another distinction between M3Lcmf and DFMF is that the former can selectively combine multiple intra-relational data matrices, whereas the latter equally treats all the relational data matrices. As such, M3Lcmf can reduce the impact of noisy (or irrelevant) intra-relational data matrices for the target prediction task.

Following the idea of standard nonnegative matrix factorization (Lee and Seung 2001) and Alternating Direction Method of Multipliers (ADMM), we alternatively optimizes one variable of $\mathbf{G}_1$, $\mathbf{G}_2$, $\mathbf{G}_3$, $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ one time with other variables fixed. Due to page limit, the optimization procedures of these variables are provided in the Supplementary file.

We then use the optimized $\mathbf{G}_2$ and $\mathbf{G}_3$ to approximate $\mathbf{R}_{23}^*$ (instance-label association matrix) as follows:

$$\mathbf{R}_{23}^* = \mathbf{G}_2 \mathbf{G}_3^T \tag{5}$$

To further map the labels of instances onto the corresponding bag, we approximate the bag-label association matrix $\mathbf{R}_{13}^* \in \mathbb{R}^{l \times q}$ as follows:

$$\mathbf{R}_{13}^* = \mathbf{\Lambda} \mathbf{R}_{12} \mathbf{R}_{23}^* \tag{6}$$

As such, M3Lcmf can make label prediction both at the instance and bag levels.

Experiments

Experimental Setup

We perform three experiments to investigate the performance of the proposed M3Lcmf. In the first experiment, six representative and related approaches, including four MIML methods (MIMLSVM (Zhou et al. 2008), MIMLRBF (Zhang and Wang 2009), MIMLNN (Zhou et al. 2012), and MIMLfast (Huang, Gao, and Zhou 2018)) and two M3L methods (MIMLmix (Nguyen et al. 2014) and M^{2}IL(Li et al. 2017)) are compared against M3Lcmf on both the bag-level and

instance-level prediction. In the second experiment, four variants of M3Lcmf are designed to quantify the contribution of different types of relations. The third experiment studies the parameter sensitivity of M3Lcmf.

Nine publicly available multi-instance multi-label datasets from different domains are used for the experiments. The details of the datasets are given in Table 2. The first five datasets are collected from `http://lamda.nju.edu.cn/CH.Data.ashx` and `http://github.com/hsoleimani/MLTM/tree/master/Data`. They only have the bag-level labels and are used for evaluating the bag-level predictions. The original Delicious dataset includes 12234 bags with 223285 instances; to avoid an excessively heavy computational load, we randomly selected 1000 bags with 17613 instances from Delicious for the experiments. The last four datasets have instance-level labels (Winn, Criminisi, and Minka 2005; Briggs, Fern, and Raich 2012), they are used for instance-level prediction and evaluation (Huang, Gao, and Chen 2017; Chen et al. 2018).

Table 2: Statistics of night datasets used for the experiments. bag, $instance$, and $label$ are the number of bags, instances, and labels, respectively. avgBI is the average number of instances per bag, and avgBL is the average number of labels per bag.

Dataset	bag	instance	label	avgBI	avgBL
Haloarcula_marismortui	304	951	234	3.1	3.2
Geobacter_sulfurreducens	379	1214	320	3.2	3.1
Azotobacter_vinelandii	407	1251	340	3.1	4.0
Pyrococcus_furiosus	425	1321	321	3.1	4.5
Delicious	1000	17613	20	17.6	2.8
Letter Frost	144	565	26	3.9	3.6
Letter Carroll	166	717	26	4.3	3.9
MSRC v2	591	1758	23	1.0	2.5
Birds	548	10232	13	18.7	2.1

To evaluate the effectiveness of M3Lcmf, four widely-used multi-label evaluation metrics are adopted, including Ranking Loss (*RankLoss*), macro AUC (Area Under receiver operating Curve) (*macroAUC*), Average Recall (*AvgRecall*), and Average F1-score (*AvgF1*). Due to space limitation, the formal definition of these metrics is omitted here but can be found in (Zhang and Zhou 2014; Gibaja and Ventura 2015). The smaller the values of *RankLoss*, the better the performance is. As such, to be consistent with the other evaluation metrics, we report *1-RankLoss* instead. For the latter metrics, larger values are an indication of a better performance.

Prediction Results at the Bag-Level

We randomly partition the samples of each dataset into a training set (70%) and a testing set (30%), and independently run each algorithm in each partition. We report the average results (10 random partitions) and standard deviations in Table 3. Since there are no off-the-shelf multi-view datasets for multi-instance multi-label learning, for MIMLmix (Nguyen et al. 2014), M^{2}IL(Li et al. 2017) and the proposed M3Lcmf, we divide the original features of each bag into two views by randomly selecting half features for one view, and the remaining features for the other view. We initialize $\mathbf{R}_{12}(i,k) = 1$ when the i-th bag encompasses the k-th instance; $\mathbf{R}_{12}(i,k) = 0$ otherwise. We set $\mathbf{R}_{13}(i,c) = 1$ when the i-th bag is annotated with the c-th label; $\mathbf{R}_{13}(i,c) = 0$ otherwise. Both λ_1 and λ_2 are fixed to 1000, and the low-rank size of $\mathbf{G}_i$ ($i \in \{1,2,3\}$) is fixed to 140. The input parameters of these comparing methods are specified (or optimized) as suggested by the authors in their code or papers, and the setting of the parameters for M3Lcmf will be investigated later.

M3Lcmf generally outperforms these comparing methods across different datasets and the used metrics. We further used the signed-rank test (Demšar 2006) to check the significance between M3Lcmf and these methods (except MIMLRBF). All the p-values are small than 0.02, and the p-value between M3Lcmf and MIMLRBF is 0.13. MIMLmix did not complete the computation on the Delicious dataset over the period of two weeks. As a result, we could not report the results of MIMLmix on this dataset. M3Lcmf, MIMLmix, and M^{2}IL are M3L methods, and M3Lcmf frequently outperforms the latter two, which only use limited types of relations between objects. This fact shows the importance of accounting for multi-type relations in M3L. M3Lcmf has a lower *1-RankLoss* but a higher *AvgRecall* and *AvgF1* than MIMLmix, the possible reason is that MIMLmix captures label correlations by assuming the labels being sampled from Multinomial distribution and it samples a label indicator for each instance, whereas M3Lcmf simply uses the cosine similarity to measure the correlation. M3Lcmf outperforms three MIML solutions (MIMLNN, MIMLfast and MIMLSVM), which utilize much fewer relations between bags, instances and labels than M3Lcmf does. This comparison again corroborates the advantage of leveraging multiple types of relations in M3L, and also suggests the importance of integrating multiple data views. Although MIMLRBF considers limited types of relations between bags and instances, it still obtains a comparable performance with M3Lcmf. The possible cause is that MIMLRBF additionally uses the RBF neural network to learn an enhanced feature representation and a nonlinear classifier.

Prediction Results at the Instance-Level

To investigate the performance of M3Lcmf at the instance-level, we conduct experiments on the last four datasets with instance-level labels in Table 4. MIMLfast, MIMLmix and the proposed M3Lcmf are tested on these datasets under the same experimental protocol at the bag-level. The result values of *1-RankLoss* and *AvgF1* are reported in Table 4.

M3Lcmf outperforms these comparing methods on different datasets in most cases, and it loses to MIMLmix on the Birds dataset. Among these three comparing methods, MIMLmix often ranks the 2nd place and MIMLfast the 3rd place. MIMLmix does not make use of bag-bag relation and instance-instance relation as summarized in Table 1. MIMLfast additionally does not make use of instance-label relation, so it loses to MIMLmix, and say nothing of M3Lcmf, which utilizes all six types of relations. These comparisons again prove the effectiveness of leveraging multi-type relations in M3L. In summary, M3Lcmf can not only accurately predict labels of bags, but also labels of instances.

Table 3: Results of bag-level prediction on different datasets. •/○ indicates whether M3Lcmf is statistically (according to pairwise t-test at 95% significance level) superior/inferior to the other method.

Metric	MIMLNN	MIMLRBF	MIMLSVM	MIMLfast	MIMLmix	M^{2}IL	M3Lcmf
	Haloarcula_marismortui						
1-RankLoss	0.713 ± 0.029•	0.761 ± 0.021○	0.689 ± 0.027•	0.553 ± 0.022•	0.782 ± 0.000○	0.828 ± 0.000○	0.728 ± 0.026
macroAUC	0.624 ± 0.029○	0.658 ± 0.034○	0.603 ± 0.022○	0.717 ± 0.029○	0.547 ± 0.000•	0.442 ± 0.000•	0.582 ± 0.022
AvgRecall	0.079 ± 0.015•	0.184 ± 0.028•	0.175 ± 0.022•	0.007 ± 0.023•	0.002 ± 0.000•	0.016 ± 0.000•	0.299 ± 0.041
AvgF1	0.128 ± 0.019•	0.257 ± 0.027•	0.218 ± 0.022•	0.092 ± 0.022•	0.033 ± 0.000•	0.019 ± 0.000•	0.301 ± 0.022
	Azotobacter_vinelandii						
1-RankLoss	0.656 ± 0.021•	0.693 ± 0.032○	0.681 ± 0.016○	0.537 ± 0.021•	0.813 ± 0.000○	0.805 ± 0.000○	0.663 ± 0.019
macroAUC	0.564 ± 0.048•	0.638 ± 0.040○	0.565 ± 0.028•	0.666 ± 0.021○	0.621 ± 0.000○	0.509 ± 0.000•	0.617 ± 0.045
AvgRecall	0.069 ± 0.024•	0.105 ± 0.024•	0.116 ± 0.021•	0.054 ± 0.018•	0.019 ± 0.000•	0.004 ± 0.000•	0.178 ± 0.022
AvgF1	0.109 ± 0.033•	0.157 ± 0.029•	0.148 ± 0.023•	0.069 ± 0.017•	0.072 ± 0.000•	0.007 ± 0.000•	0.199 ± 0.013
	Geobacter_sulfurreducens						
1-RankLoss	0.656 ± 0.018•	0.688 ± 0.024○	0.694 ± 0.020○	0.552 ± 0.019•	0.798 ± 0.000○	0.821 ± 0.000○	0.684 ± 0.000
macroAUC	0.564 ± 0.027•	0.608 ± 0.033○	0.567 ± 0.015○	0.691 ± 0.022○	0.375 ± 0.000•	0.499 ± 0.000•	0.567 ± 0.000
AvgRecall	0.077 ± 0.016•	0.129 ± 0.021•	0.137 ± 0.018•	0.042 ± 0.009•	0.032 ± 0.000•	0.012 ± 0.000•	0.296 ± 0.000
AvgF1	0.120 ± 0.021•	0.186 ± 0.026•	0.173 ± 0.022•	0.058 ± 0.009•	0.040 ± 0.000•	0.014 ± 0.000•	0.277 ± 0.000
	Pyrococcus_furiosus						
1-RankLoss	0.722 ± 0.014•	0.732 ± 0.000•	0.727 ± 0.027•	0.469 ± 0.035•	0.760 ± 0.000○	0.809 ± 0.000○	0.733 ± 0.015
macroAUC	0.593 ± 0.029○	0.520 ± 0.000•	0.613 ± 0.043○	0.469 ± 0.030•	0.488 ± 0.000•	0.485 ± 0.000•	0.543 ± 0.011
AvgRecall	0.069 ± 0.017•	0.105 ± 0.000•	0.134 ± 0.029•	0.119 ± 0.038•	0.004 ± 0.000•	0.006 ± 0.000•	0.341 ± 0.038
AvgF1	0.086 ± 0.015•	0.116 ± 0.000•	0.174 ± 0.034•	0.115 ± 0.021•	0.056 ± 0.000•	0.008 ± 0.000•	0.307 ± 0.025
	Delicious						
1-RankLoss	0.685 ± 0.012○	0.735 ± 0.008○	0.580 ± 0.053•	0.466 ± 0.023•	−−	0.439 ± 0.000•	0.636 ± 0.000
macroAUC	0.627 ± 0.010○	0.670 ± 0.012○	0.583 ± 0.009○	0.466 ± 0.024•	−−	0.549 ± 0.000○	0.480 ± 0.000
AvgRecall	0.112 ± 0.014•	0.029 ± 0.019•	0.142 ± 0.030•	0.619 ± 0.045○	−−	0.097 ± 0.000•	0.178 ± 0.000
AvgF1	0.180 ± 0.018•	0.054 ± 0.033•	0.201 ± 0.032•	0.264 ± 0.013○	−−	0.136 ± 0.000•	0.252 ± 0.000

Table 4: Results on different multi-instance datasets. •/○ indicates whether M3Lcmf is statistically (according to pairwise t-test at 95% significance level) superior/inferior to the other methods.

Metric	MIMLfast	MIMLmix	M3Lcmf
	Letter Frost		
1-RankLoss	0.426 ± 0.049•	0.667 ± 0.000•	0.734 ± 0.050
AvgF1	0.094 ± 0.015•	0.150 ± 0.000•	0.352 ± 0.107
	Letter Carroll		
1-RankLoss	0.458 ± 0.065•	0.410 ± 0.000•	0.692 ± 0.012
AvgF1	0.096 ± 0.023•	0.086 ± 0.000•	0.104 ± 0.012
	MSRC v2		
1-RankLoss	0.419 ± 0.030•	0.579 ± 0.000•	0.652 ± 0.005
AvgF1	0.111 ± 0.005•	0.333 ± 0.000○	0.208 ± 0.074
	Birds		
1-RankLoss	0.524 ± 0.184•	0.937 ± 0.000○	0.666 ± 0.000
AvgF1	0.061 ± 0.070•	0.503 ± 0.000○	0.286 ± 0.000

Contribution of Different Types of Relations

To further analyze the contribution of different relations used by M3Lcmf, we introduce four variants. (i) M3Lcmf (nR11) does not consider the relation between bags, i.e., $\mathbf{R}_{11}^v = 0$; (ii) M3Lcmf (nR22) does not consider the relation between instances, i.e., $\mathbf{R}_{22}^v = 0$; (iii) M3Lcmf (nR33) does not consider the relation between labels, i.e., $\mathbf{R}_{33} = 0$; (iv) M3Lcmf (nR23) does not consider the relation between instances and labels, i.e., $\mathbf{R}_{13}^* = \mathbf{G}_1\mathbf{G}_3^T$, instead of $\mathbf{R}_{13}^* = \mathbf{\Lambda}\mathbf{R}_{12}\mathbf{G}_2\mathbf{G}_3^T$. We follow the experimental protocol at the bag-level prediction, and report the results of *1-RankLoss* obtained by M3Lcmf and its variants in Fig. 2.

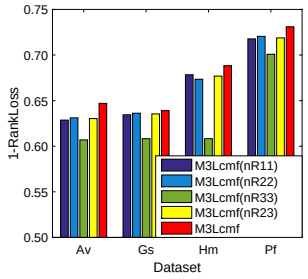


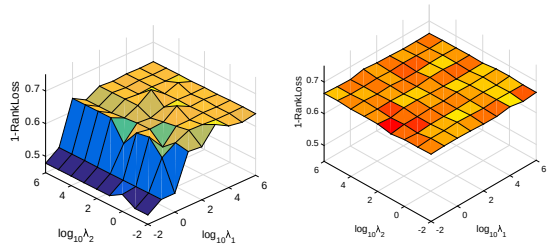
Figure 2: *1-RankLoss* of M3Lcmf and its variants on different datasets. Av: Azotobacter_vinelandii, Gs: Geobacter_sulfurreducens, Hm: Haloarcula_marismortui, Pf: pyrococcus_furiosus.

M3Lcmf significantly outperforms its variants, which separately disregard one type of relations. M3Lcmf often outperforms M3Lcmf (nR11) and M3Lcmf (nR22). This observation suggests the relation between bags and that between instances have an important effect on M3Lcmf. Besides, M3Lcmf(nR33) is outperformed by all the other variants, which shows the importance of considering the label correlation. In addition, we can observe that M3Lcmf (nR23) is outperformed by M3Lcmf. This observation not only proves the effectiveness of the introduced aggregation term, but also shows the importance of instance-label relations in boosting the prediction performance.

From these results, we can conclude that multiple types of relations between bags, instances, and labels should be

simultaneously considered in M3L.

Parameter Sensitivity Three parameters (λ_1, λ_2, and the low-rank size d of $\mathbf{G}$) may affect the performance of M3Lcmf. We conduct additional experiments to investigate the sensitivity of these parameters. For brevity, we only report the results on Azotobacter vinelandii and MSRC v2, and the results on the other datasets lead to similar conclusions.

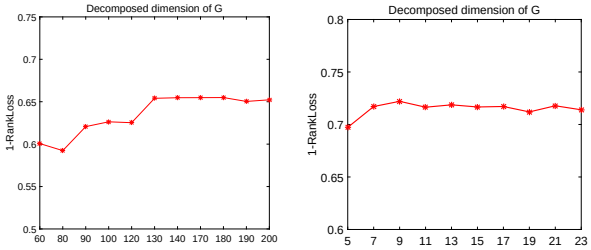


(a) *Azotobacter vinelandii* (b) *MSRC v2*

Figure 3: *1-RankLoss* of M3Lcmf under different combinations of λ_1 and λ_2 on Azotobacter vinelandii and MSRC v2.

From the explicit solution for $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ in the supplementary file, it is clear that once the values λ_1 and λ_2 are specified, the weights assigned to $\mathbf{R}_{11}^v$ and $\mathbf{R}_{22}^v$ can be computed based on the reconstruction loss of those matrices. To investigate the sensitivity of these two parameters, we vary λ_1 and λ_2 in the range $\{10^{-2}, 10^{-1}, \cdots, 10^6\}$, and report the average *1-RankLoss* of M3Lcmf under different combinations of them in Fig. 3. We can see that M3Lcmf achieves a stable performance under a wide range of combinations of values for λ_1 and λ_2. For Azotobacter vinelandii, M3Lcmf achieves a good performance with λ_1 and λ_2 in $[10^2, 10^6]$, and it shows a significantly reduced *1-RankLoss* when either λ_1 or λ_2 are set to a too small value. This is because the predictions are made and evaluated at the bag-level and the bag-level intra-relation plays a more important role, but only one bag-level intra relational data matrix is selected under this setting. Unlike the pattern on Azotobacter vinelandii, M3Lcmf holds a relatively stable performance on MSRC v2 under different combinations of values for λ_1 and λ_2. This is because Azotobacter vinelandii provides more structural information and feature information for the intra-relational data matrices of bags (or instances) than MSRC v2. Particularly, the former has more instances per bag than the latter, and the bag in MSRC v2 generally has one instance. Besides, the feature dimensionality of instances in Azotobacter vinelandii is much larger than that of MSRC v2. This investigation suggests the importance of structural information of bags (or instances) in M3L. From these results, we can conclude that an effective combination of λ_1 and λ_2 can be easily found.

The low-rank size d of $\mathbf{G}$ is an essential parameter for M3Lcmf. Fig. 4 shows the results of M3Lcmf under different input values of d on Azotobacter vinelandii and MSRC v2 with $\lambda_1 = 10^3$ and $\lambda_2 = 10^3$. We observe an increasing trend of *1-RankLoss*, and an overall good performance when $d \geq 140$ or $d \geq 11$. M3Lcmf does not show a high *1-RankLoss* when a small d is adopted, that is because a too small d can not sufficiently encode the latent feature information of



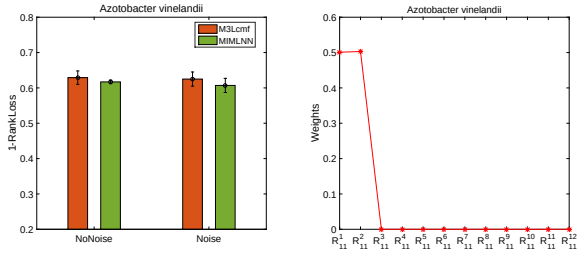
(a) *Azotobacter vinelandii* (b) *MSRC v2*

Figure 4: *1-RankLoss* vs. d (low-rank size) on Azotobacter vinelandii and MSRC v2.

bags, instances, and labels. However, we can still find that an effective input value d can be easily selected.

Contributions of Weighting Intra-Relational Data

To investigate the contribution of weighting intra-relational data and the capability of M3Lcmf on discarding noisy intra-relational data matrices, we added 10 synthetic *noisy* intra-relational data matrices of bags on the Azotobacter vinelandii dataset. Particularly, the 10 noisy data matrices are obtained by randomly shuffling the nonzero entries of each row of two valid matrices, which are constructed in the same way as in the first type of experiments. For reference, we also applied MIMLNN on the same dataset with the same 10 noisy data matrices, and reported the results in Fig. 5(a).



(a) *1-RankLoss* (b) *Weights*

Figure 5: (a) Prediction results with and without the noisy data matrices, (b) Weights assigned by M3Lcmf to 12 intra-relational data matrices of bags. The first 2 are valid data matrices, and the last 10 are noisy ones.

Even with 10 noisy data matrices, M3Lcmf does not show a decreased performance, but MIMLNN shows a clearly reduced performance (by 2%). That is because M3Lcmf explicitly considers the different relevances of intra-relational data matrices, and it can selectively integrate these matrices. In contrast, MIMLNN does not account for the different relevances of these matrices. As a result, it is more impacted by these noisy matrices.

To further investigate the underlying reason for the robust performance of M3Lcmf, we plot weights assigned to these 12 (2 valid and 10 noisy) intra-relational data matrices of bags in Fig. 5(b). We can see that these 10 noisy data matrices are assigned with zero weights. Namely, M3Lcmf discards these noisy data matrices during the collaborative matrix factorization process. This investigation justifies our motivation

to account for different relevances of multiple intra-relation data matrices.

Conclusion

In this paper, we proposed a collaborative matrix factorization based multi-view multi-instance multi-label learning approach called M3Lcmf. M3Lcmf utilizes a heterogeneous network to capture different types of relations in M3L, and collaboratively factorizes the relational data matrices of the network to explore the intrinsic relations between bags, instances, and labels. Extensive experimental results on different datasets corroborate our hypothesis that multiple types of relations can boost the performance of M3L, and their joint usage contributes to a significantly improved performance of M3Lcmf against competitive approaches. The Supplementary file and code of M3Lcmf are available at `http://mlda.swu.edu.cn/codes.php?name=M3Lcmf`.

Acknowledgments

The authors appreciate the reviewers for their helpful comments on improving our work. This work is supported by NSFC (61872300, 61741217, 61873214 and 61871020), NSF of CQ CSTC (cstc2018jcyjAX0228, cstc2016jcyjA0351 and CSTC2016SHMSZX0824), the Open Research Project of Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2017A05), and the National Science and Technology Support Program (2015BAK41B03 and 2015BAK41B04).

References

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR* 7(11):2399–2434.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3(1):993–1022.

Briggs, F.; Fern, X. Z.; and Raich, R. 2012. Rank-loss support instance machines for miml instance annotation. In *KDD*, 534–542.

Chen, X.; Yu, G.; Domeniconi, C.; Wang, J.; Li, Z.; and Zhang, Z. 2018. Cost effective multi-label active learning via querying subexamples. In *ICDM*, 905–910.

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

Feng, J., and Zhou, Z.-H. 2017. Deep miml network. In *AAAI*, 1884–1890.

Gibaja, E., and Ventura, S. 2015. A tutorial on multilabel learning. *ACM Computing Surveys* 47(3):52.

Gligorijević, V., and Pržulj, N. 2015. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface* 12(112):20150571.

He, J.; Du, C.; Zhuang, F.; Yin, X.; He, Q.; and Long, G. 2016. Online bayesian max-margin subspace multi-view learning. In *IJCAI*, 1555–1561.

Huang, S.-J.; Gao, N.; and Chen, S. 2017. Multi-instance multi-label active learning. In *IJCAI*, 1886–1892.

Huang, S.-J.; Gao, W.; and Zhou, Z.-H. 2018. Fast multi-instance multi-label learning. *TPAMI* 99(1):1–14.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *NIPS*, 556–562.

Li, B.; Yuan, C.; Xiong, W.; Hu, W.; Peng, H.; Ding, X.; and Maybank, S. 2017. Multi-view multi-instance learning based on joint sparse representation and multi-view dictionary learning. *TPAMI* 39(12):2554–2560.

Nguyen, C. T.; Wang, X.; Liu, J.; and Zhou, Z. H. 2014. Labeling complicated objects: multi-view multi-instance multi-label learning. In *AAAI*, 2013–2019.

Nguyen, C. T.; Zhan, D. C.; and Zhou, Z. H. 2013. Multi-modal image annotation with multi-instance multi-label lda. In *IJCAI*, 1558–1564.

Shao, W.; Zhang, J.; He, L.; and Philip, S. Y. 2016. Multi-source multi-view clustering via discrepancy penalty. In *IJCNN*, 2714–2721.

Tan, Q.; Yu, G.; Domeniconi, C.; Wang, J.; and Zhang, Z. 2018. Incomplete multi-view weak-label learning. In *IJCAI*, 2703–2709.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Winn, J.; Criminisi, A.; and Minka, T. 2005. Object categorization by learned universal visual dictionary. In *ICCV*, 1800–1807.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

Yang, Y.; Wu, Y.-F.; Zhan, D.-C.; Liu, Z.-B.; and Jiang, Y. 2018. Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport. In *KDD*, 2594–2603.

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

Zhang, M.-L., and Zhou, Z.-H. 2009. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence* 31(1):47–68.

Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *TKDE* 26(8):1819–1837.

Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; and Li, Y.-F. 2008. Miml: A framework for learning with ambiguous objects. *arXiv: 0808.3231*.

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.

Zhu, Y.; Ting, K. M.; and Zhou, Z.-H. 2017. Discover multiple novel labels in multi-instance multi-label learning. In *AAAI*, 2977–2984.

Zitnik, M., and Zupan, B. 2015. Data fusion by matrix factorization. *TPAMI* 37(1):41–53.