The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Latent Semantic Aware Multi-View Multi-Label Classification

Changqing Zhang,[1] Ziwei Yu,[1] Qinghua Hu,[1] Pengfei Zhu,[1] Xinwang Liu,[2] Xiaobo Wang[3]
[1]School of Computer Science and Technology, Tianjin University, Tianjin, China, 300350
[2]School of Computer National University of Defense Technology Changsha, China, 410073
[3]Center for Biometrics and Security Research & National Laboratory of Pattern Recognition Institute of Automation,
Chinese Academy of Sciences, 100190
{zhangchangqing, yuziwei, huqinghua, zhupengfei}@tju.edu.cn

Abstract

For real-world applications, data are often associated with multiple labels and represented with multiple views. Most existing multi-label learning methods do not sufficiently consider the complementary information among multiple views, leading to unsatisfying performance. To address this issue, we propose a novel approach for multi-view multi-label learning based on matrix factorization to exploit complementarity among different views. Specifically, under the assumption that there exists a common representation across different views, the uncovered latent patterns are enforced to be aligned across different views in kernel spaces. In this way, the latent semantic patterns underlying in data could be well uncovered and this enhances the reasonability of the common representation of multiple views. As a result, the consensus multi-view representation is obtained which encodes the complementarity and consistence of different views in latent semantic space. We provide theoretical guarantee for the strict convexity for our method by properly setting parameters. Empirical evidence shows the clear advantages of our method over the state-of-the-art ones.

Introduction

Multi-label classification, which assigns one example with multiple classes, is of significant interest due to its ubiquity in real-world applications. For example, in computer vision, an image may simultaneously contain more than one type of objects; in web page categorization, a news web page may cover different topics, such as sports, business and entertainment. For this problem, multi-label learning approaches (Boutell et al. 2004; Tsoumakas and Katakis 2006; Zhang and Zhou 2007; Gong et al. 2016) have been proposed over the past decade, such as the early representative methods: binary relevance (BR) (Tsoumakas and Katakis 2006) and label powerset (LP) (Boutell et al. 2004). By directly transforming the multi-label learning task into multiple binary classification tasks, BR neglects the correlation among labels. LP regards each subset of multiple labels as a different class of single-label classification. Although taking the label correlation into consideration, this model lacks of mining the complex label correlation and can not be applied for the task with large label set. Multi-label k-nearest neighbour (MLkNN) (Zhang and Zhou 2007) is one of classic and effective multi-label methods, which builds a Bayesian model by using the k-nearest neighbour method to obtain the prior and likelihood, and then utilizes the max posterior to assign labels for testing example. Some more recent methods focus on other issues in multi-label classification, e.g., label noise (Yu et al. 2014; Yang, Jiang, and Zhou 2013).

Although diverse methods have been proposed in the literature, there still exist the following limitations. On one hand, most existing multi-label learning methods only consider single view data, however, each individual view cannot characterize different labels comprehensively since different views encode different properties of data, which implies the practical necessity of multi-view learning (Xu, Tao, and Xu 2013; Liu et al. 2017; Cao et al. 2015; Zhang et al. 2015). On the other hand, learning with plenty of unlabeled data has shown its power in many real applications. However, most existing multi-label classification models are fully supervised thus they are unable to explore the unlabeled samples. Although a few semi-supervised multi-label learning methods (Liu, Jin, and Yang 2006; Wang, Tu, and Tsotsos 2013) have been developed, these models are not specifically targeted on the multi-view semi-supervised multi-label learning. The most recent and related method in (Liu et al. 2015) also utilizes matrix factorization and common representation. However, it has the following two limitations: firstly, the common representation among multiple views is learned without constraining the bases of different views, which weakens the reasonability of the common representation; secondly, the common representation learning and multi-label learning (label completion) are performed in two separated steps, thus the prediction performance could not be well guaranteed.

In this paper, we propose a new multi-view multi-label learning approach termed as *Latent Semantic Aware Multi-view Multi-label Learning (LSA-MML)*. As shown in Figure 1, given the input data with multiple views, our method simultaneously seeks a predictive common representation of multiple views and the corresponding projection model between the common representation and labels. The bases of V different views, $\{\mathbf{B}^{(v)}\}_{v=1}^{V}$, can be considered as latent semantic components. With the common representation $\mathbf{P}$, the j^{th} bases of different views, i.e, $\{\mathbf{b}_j^v\}_{v=1}^{V}$, encode the same latent semantic, therefore, these bases across different

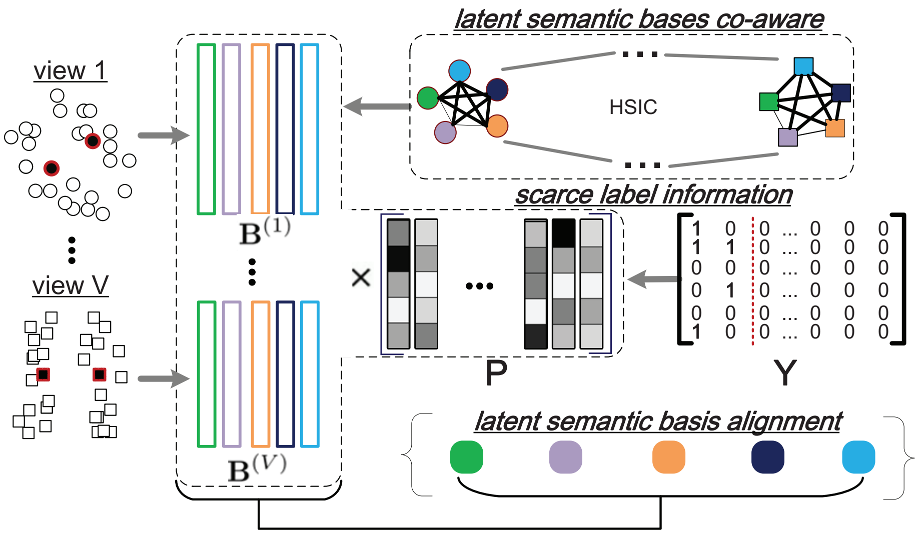


Figure 1: Method overview. The common representation $\mathbf{P}$ is learned by exploring the complementarity of multiple views and scarce labeled samples (solid circles and squares) jointly. The latent semantic basis matrices (i.e., $\mathbf{B}^{(v)}$s) of different views are aligned in kernel spaces, which guarantees the reasonability of the consensus representation $\mathbf{P}$.

views should be consistent with each other. We align these bases of different views with Hilbert-Schmidt Independence Criterion (HSIC) in kernel space, which well addresses the comparability of different views, thus a consensus coefficient matrix (common representation) $\mathbf{P}$ for different views is induced. To solve our problem, we provide the theoretical analysis for the convexity and the instruction for parameter setting to guarantee the strict convexity. Extensive empirical results on benchmark datasets demonstrate that the proposed method outperforms the state-of-the-art methods.

Related Work

From the last decade, multi-label classification has received intensive attention (Boutell et al. 2004; Zhang and Zhou 2007; Yang, Jiang, and Zhou 2013). Generally, existing multi-label methods can be roughly categorized into three lines. The first-order strategy deals with multi-label learning in label-by-label manner, i.e., dividing the multi-label problem into multiple binary classification tasks or its variants (Zhang and Zhou 2007; Clare and King 2001). The second-order methods introduce the pairwise relations between the labels for the multi-label classification, such as the ranking between the relevant label and irrelevant label (Elisseeff and Weston 2002; Fürnkranz et al. 2008; Ghamrawi and McCallum 2005). CLR (Fürnkranz et al. 2008) firstly transforms the multi-label learning problem into label ranking problem by introducing the pairwise comparison, and then constructs binary classifiers to solve the multi-label ranking problem. Rank-SVM (Elisseeff and Weston 2002) conducts multi-label classification by adopting the ranking loss as cost function in SVM. The high-order strategy builds more complex relations among labels for multi-label learning (Read et al. 2011; Tsoumakas and Vlahavas 2007). The representatives include the chain-based method (Read et al. 2011) which transforms the multi-label data to a chain of binary classifiers, and the label-set-based methods (Boutell et al. 2004; Tsoumakas and Vlahavas 2007) that divide the entire label set into multiple overlapping subsets and train one classifier for each subset. Due to the ubiquity of data with multiple views, multi-view learning has been an active research field and shown its effectiveness in a wide range of applications (Xu, Tao, and Xu 2013). Recently, a few multi-view multi-label classification methods (Luo et al. 2013; Liu et al. 2015) were proposed to exploit the complementarity of different types of features for the improved classification performance. The method in (Luo et al. 2013) introduces multi-view vector-valued manifold regularization to integrate multi-view features. The method in (Liu et al. 2015) seeks a common low-dimensional representation under the matrix factorization framework and then conducts classification based on matrix completion. Both the two recent methods (Luo et al. 2013; Liu et al. 2015) perform classification in the transductive semisupervised manner.

LSA-MML: Our Classification Model

Suppose there are L labeled data points $\{\mathbf{x}_l, \mathbf{y}_l\}_{l=1}^{L}$ and U unlabeled data points $\{\mathbf{x}_u\}_{u=L+1}^{N}$, where $N = L+U$. These instance-label pairs are stacked in two matrices, i.e., $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)$ and $\mathbf{Y} = (\mathbf{Y}_l, \mathbf{Y}_u) = (\mathbf{y}_1, ..., \mathbf{y}_N)$. Since we employ the transductive learning manner to simultaneously exploit unsupervised samples, our objective function turns out to be the following general form

$$\min \mathcal{M}(\mathbf{X}; \hat{\mathbf{Y}}) + \lambda \mathcal{S}(\mathbf{Y}_l, \hat{\mathbf{Y}}), \tag{1}$$

where $\hat{\mathbf{Y}}$ is the completed label matrix to be predicted, which is learned with data $\mathbf{X}$ and a few known labels in $\mathbf{Y}_l$. Specifically, to obtain the completed label matrix $\hat{\mathbf{Y}}$, we aim to uncover the underlying structure from data themselves $\mathbf{X}$ by the first term $\mathcal{M}(\mathbf{X};\hat{\mathbf{Y}})$, which is guided by the labeled data $\mathbf{Y}_l$ in the second term $\mathcal{S}(\mathbf{Y}_l,\hat{\mathbf{Y}})$.

Considering the data with multiple views, we generalize the above formulation as

$$\min \mathcal{M}(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(v)};\hat{\mathbf{Y}})+\lambda\mathcal{S}(\mathbf{Y}_l,\hat{\mathbf{Y}}). \tag{2}$$

Under the assumption that different views share the latent common representation, i.e., $\mathbf{X}^{(v)}=\mathbf{B}^{(v)}\mathbf{P}$, we have

$$\begin{aligned}\min \mathcal{M}(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(v)};\mathbf{P})+\lambda_1\Omega(\mathbf{B}^{(1)},\cdots,\mathbf{B}^{(V)})\\ +\lambda_2\mathcal{P}(\mathbf{P},\hat{\mathbf{Y}})+\lambda_3\mathcal{S}(\mathbf{Y}_l,\hat{\mathbf{Y}}),\end{aligned} \tag{3}$$

where $\mathbf{P}\in\mathbb{R}^{K\times N}$ is the consensus multi-view representation which encodes the complementary information from different views. $\mathbf{B}^{(v)}\in\mathbb{R}^{D_v\times K}$ is the basis matrix corresponding to the v^{th} view. Accordingly, the first term searches a comprehensive multi-view representation, and the second term guarantees the reasonability of using a common representation for different views since it aligns the bases of different views in the latent semantic space. The last term delivers the label information on the estimated label matrix and based on which the third term ensures the predictive property of the common multi-view representation. Therefore, the complemented label matrix $\hat{\mathbf{Y}}$ benefits from both multi-view data ($\mathbf{P}$) and supervised label information ($\mathbf{Y}_l$).

Specifically, by using common multi-view representation under matrix factorization framework, we have

$$\mathcal{M}(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(v)};\mathbf{P})=\sum_{v=1}^{V}||\mathbf{X}^{(v)}-\mathbf{B}^{(v)}\mathbf{P}||_F^2, \tag{4}$$

where $||\cdot||_F$ is the well-known Frobenius norm of matrix. For different views, the latent bases should be consistent across different views. To this end, we penalize the independence of bases between different views with

$$\Omega(\mathbf{B}^{(1)},\cdots,\mathbf{B}^{(V)})=\sum_{v\neq w}\text{IND}(\mathbf{B}^{(v)},\mathbf{B}^{(w)}), \tag{5}$$

where the aim of the regularization $\text{IND}(\cdot,\cdot)$ is to enhance the dependence of these bases between different views. Since these bases are in different feature spaces, hence, we introduce HSIC to constrain the consistence across different views in kernel spaces. Specifically, we define $\text{IND}(\mathbf{B}^{(v)},\mathbf{B}^{(w)})=-\text{HSIC}(\mathbf{B}^{(v)},\mathbf{B}^{(w)})$ in our method.

Hilbert-Schmidt Independence Criterion (Gretton et al. 2005). We give the brief description about HSIC as follows. Let us define a mapping $\phi(\mathbf{x})$ from $\mathbf{x}\in\mathcal{X}$ to kernel space $\mathcal{F}$ such that the inner product between vectors in that space is given by a kernel function $k_1(\mathbf{x}_i,\mathbf{x}_j)=\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle$. Let $\mathcal{G}$ be a second kernel space on $\mathcal{Y}$ with kernel function $k_2(\mathbf{y}_i,\mathbf{y}_j)=\langle\varphi(\mathbf{y}_i),\varphi(\mathbf{y}_j)\rangle$. The empirical version of HSIC is induced as:

Definition 1. *Consider a series of N independent observations drawn from $p_{\mathbf{xy}}$, $\mathcal{Z}:=\{(\mathbf{x}_1,\mathbf{y}_1),...,(\mathbf{x}_N,\mathbf{y}_N)\}\subseteq\mathcal{X}\times\mathcal{Y}$, an estimator of HSIC, written as HSIC($\mathcal{Z},\mathcal{F},\mathcal{G}$), is given by:*

$$\mathit{HSIC}(\mathcal{Z},\mathcal{F},\mathcal{G})=(N-1)^{-2}tr(\mathbf{K}_1\mathbf{C}\mathbf{K}_2\mathbf{C}), \tag{6}$$

where $tr(\cdot)$ is the trace of a square matrix. $\mathbf{K}_1$ and $\mathbf{K}_2$ are the Gram matrices with $k_{1,ij}=k_1(\mathbf{x}_i,\mathbf{x}_j)$, $k_{2,ij}=k_2(\mathbf{y}_i,\mathbf{y}_j)$. $c_{ij}=\delta_{ij}-1/N$ centers the Gram matrix to have zero mean in the feature space.

Since HSIC well measures the independence of two variables (Quadrianto, Song, and Smola 2009), we employ it to maximize the dependency between the bases of two views. Note that, according Eq (6), the HSIC in our method can be considered as the penalization for the disagreement of different views in terms of similarity graphs of bases, as shown in Fig .1.

In practice, to ensure the predictive property in terms of labels, the third and fourth terms are integrated as the following formula

$$\Delta=\lambda_2\mathcal{P}(\mathbf{P},\hat{\mathbf{Y}})+\lambda_3\mathcal{S}(\mathbf{Y}_l,\hat{\mathbf{Y}})=||(\mathbf{W}\mathbf{P}-\mathbf{Y})\mathbf{S}||_F^2, \tag{7}$$

where $\hat{\mathbf{Y}}=\mathbf{W}\mathbf{P}$ and $\mathbf{Y}=[\mathbf{Y}_l,\mathbf{Y}_u]$. $\mathbf{W}$ is the prediction model and $\mathbf{S}$ is the filtering matrix used to select the labeled samples with $S_{ii}=1$ if the i^{th} sample is labeled and 0 otherwise. This ensures that the multi-view consensus representation should be predictive corresponding to the known labels. Accordingly, the final form of our objective function turns out to be

$$\begin{aligned}\min_{\mathbf{B}^{(v)},\mathbf{P},\mathbf{W},\alpha_v}&\sum_{v=1}^{V}\alpha_v^r||\mathbf{X}^{(v)}-\mathbf{B}^{(v)}\mathbf{P}||_F^2+\beta||(\mathbf{W}\mathbf{P}-\mathbf{Y})\mathbf{S}||_F^2\\ &+\gamma\sum_{v\neq w}\text{IND}(\mathbf{B}^{(v)},\mathbf{B}^{(w)})\\ &s.t.\ \alpha_v\geq 0,\ \sum_{v=1}^{V}\alpha_v=1;||\mathbf{b}_{.j}^{(v)}||_2\leq 1,\end{aligned} \tag{8}$$

where $\alpha_v>0$ is used to automatically weight different views and $r>1$ is used to avoid a trivial solution that only considers one view and adjusts the complementarity of multiple views (Wang et al. 2007). $\beta>0$ and $\gamma>0$ are tradeoff factors. $\mathbf{B}^{(v)}$ is constrained since without constraint $\mathbf{P}$ can be pushed arbitrarily close to zero only by re-scaling $\mathbf{P}/s$ and $\mathbf{B}^{(v)}s$ ($s>0$) while preserving the same loss.

To summarize, our model has the following merits: 1) our model focuses on seeking the comprehensive common representation of multiple views by enforcing the latent semantic bases of different views to be consistent; 2) our model can be considered as a bi-direction factorization, where the multi-view common representation $\mathbf{P}$ bridges the factorizations between the multi-view input and the label matrix, where the label matrix can be regarded as the description of data in the view of explicit semantic labels; 3) the label correlations are implicitly encoded by the common representation based on the uncovering latent semantic bases and the relations among them.

Optimization

Alternating Optimization Algorithm

We adopt the alternating minimization strategy to solve the optimization problem, which is comprised of four subproblems solved as follows:

Update $\mathbf{P}$ with fixed $\{\alpha_v\}_{v=1}^V$, $\mathbf{W}$ and $\{\mathbf{B}^{(v)}\}_{v=1}^V$. We should minimize the following objective function

$$\mathcal{L}(\mathbf{P}) = \sum_{v=1}^{V} \alpha_v^r ||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2 + \beta||(\mathbf{W}\mathbf{P} - \mathbf{Y})\mathbf{S}||_F^2. \tag{9}$$

By taking the derivative with respect to $\mathbf{P}$ and setting it to zero, then we obtain

$$\begin{aligned}&\sum_{v=1}^{V} \alpha_v^r \big(- (\mathbf{B}^{(v)})^T\mathbf{X}^{(v)} + (\mathbf{B}^{(v)})^T\mathbf{B}^{(v)}\mathbf{P}\big) \\ &\quad + \beta\big(\mathbf{W}^T\mathbf{W}\mathbf{P}\mathbf{S}\mathbf{S}^T - \mathbf{W}^T\mathbf{Y}\mathbf{S}\mathbf{S}^T\big) = 0.\end{aligned} \tag{10}$$

We solve the problem by separating the labeled and unlabeled parts thanks to the diagonal property of $\mathbf{S}$. For the labeled part, we have

$$\begin{aligned}&\sum_{v=1}^{V} \alpha_v^r \big(- (\mathbf{B}_l^{(v)})^T\mathbf{X}_l^{(v)} + (\mathbf{B}_l^{(v)})^T\mathbf{B}_l^{(v)}\mathbf{P}_l\big) \\ &\quad + \beta\big(\mathbf{W}_l^T\mathbf{W}_l\mathbf{P}_l - \mathbf{W}_l^T\mathbf{Y}_l\big) = 0.\end{aligned} \tag{11}$$

where the subscript l and u indicate variables corresponding to labeled and unlabeled data, respectively. Accordingly, we can update $\mathbf{P}_l$ with the following rule

$$\mathbf{P}_l = \frac{\sum_{v=1}^{V} \alpha_v^r (\mathbf{B}_l^{(v)})^T\mathbf{X}_l^{(v)} + \beta\mathbf{W}_l^T\mathbf{Y}_l}{\sum_{v=1}^{V} \alpha_v^r (\mathbf{B}_l^{(v)})^T\mathbf{B}_l^{(v)} + \beta\mathbf{W}_l^T\mathbf{W}_l}. \tag{12}$$

For the unsupervised part, we have

$$\sum_{v=1}^{V} \alpha_v^r (\mathbf{B}_u^{(v)})^T\mathbf{B}_u^{(v)}\mathbf{P}_u = \sum_{v=1}^{V} \alpha_v^r (\mathbf{B}_u^{(v)})^T\mathbf{X}_u^{(v)}. \tag{13}$$

Accordingly, we can update $\mathbf{P}_u$ by the following rule

$$\mathbf{P}_u = \big(\sum_{v=1}^{V} \alpha_v^r (\mathbf{B}_u^{(v)})^T\mathbf{B}_u^{(v)}\big)^{-1}\big(\sum_{v=1}^{V} \alpha_v^r (\mathbf{B}_u^{(v)})^T\mathbf{X}_u^{(v)}\big). \tag{14}$$

After obtaining $\mathbf{P}_l$ and $\mathbf{P}_u$, the common representation corresponding to all data, i.e., $\mathbf{P}$, is obtained as $\mathbf{P} = [\mathbf{P}_l, \mathbf{P}_u]$.

Update $\mathbf{B}^{(v)}$ with fixed α_v,$\mathbf{W}$ and $\mathbf{P}$. We should minimize the following objective function

$$\begin{aligned}&\mathcal{L}(\mathbf{B}^{(v)}) = \alpha_v^r ||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2 - \gamma \sum_{\substack{w=1 \\ w\neq v}}^{V} \text{HSIC}(\mathbf{B}^{(v)}, \mathbf{B}^{(w)}) \\ &\quad s.t. \;\; ||\mathbf{b}_{.j}^{(v)}||_2 \leq 1.\end{aligned} \tag{15}$$

We optimize $\mathbf{B}^{(v)}$-subproblem by following the work in (Gu et al. 2014), which introduces an auxiliary variable $\mathbf{S}^{(v)}$. Then, we have the following objective

$$\begin{aligned}&\mathcal{L}(\mathbf{B}^{(v)}) = \alpha_v^r ||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2 - \gamma \sum_{\substack{w=1 \\ w\neq v}}^{V} \text{HSIC}(\mathbf{B}^{(v)}, \mathbf{B}^{(w)}) \\ &\quad s.t. \;\; \mathbf{B}^{(v)} = \mathbf{S}^{(v)}, \;\; ||\mathbf{s}_{.j}^{(v)}||_2 \leq 1.\end{aligned} \tag{16}$$

We optimize (16) with alternating direction method of multipliers (ADMM). By removing the equality constraint, it turns out to be

$$\begin{aligned}&\mathcal{L}(\mathbf{B}^{(v)}, \mathbf{S}^{(v)}, \mathbf{T}^{(v)}) = \alpha_v^r ||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2 \\ &\quad - \gamma \sum_{\substack{w=1 \\ w\neq v}}^{V} \text{HSIC}(\mathbf{B}^{(v)}, \mathbf{B}^{(w)}) + \mu||\mathbf{B}^{(v)} - \mathbf{S}_r^{(v)} + \mathbf{T}_r^{(v)}||_F^2 \\ &\quad s.t. \;\; ||\mathbf{s}_{.j}^{(v)}||_2 \leq 1,\end{aligned} \tag{17}$$

where $\mu > 0$ is the penalty hyperparameter. The optimal solution of (17) can be obtained with

$$\begin{cases}\mathbf{B}_{r+1}^{(v)} = \arg\min\limits_{\mathbf{B}^{(v)}} \alpha_v^r ||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2 \\ \quad - \gamma \sum\limits_{w\neq v} \text{HSIC}(\mathbf{B}^{(v)}, \mathbf{B}^{(w)}) + \mu||\mathbf{B}^{(v)} - \mathbf{S}_r^{(v)} + \mathbf{T}_r^{(v)}||_F^2 \\ \mathbf{S}_{r+1}^{(v)} = \arg\min\limits_{\mathbf{S}^{(v)}} \mu||\mathbf{B}_{r+1}^{(v)} - \mathbf{S}_r^{(v)} + \mathbf{T}_r^{(v)}||_F^2, \; s.t.||\mathbf{s}_{.j}^{(v)}||_2 \leq 1 \\ \mathbf{T}_{r+1}^{(v)} = \mathbf{T}_r^{(v)} + \mathbf{B}_{r+1}^{(v)} - \mathbf{S}_{r+1}^{(v)}, \; \text{update } \mu \text{ if appropriate,}\end{cases} \tag{18}$$

where $\mathbf{s}_{.j}^{(v)}$ indicates the j^{th} column of $\mathbf{S}$. Note that, Theorem 1 (in subsection 3.2) guarantees the subproblem $\mathcal{L}(\mathbf{B}^{(v)})$ to be convex and thus the optimal solution could be obtained.

Update $\mathbf{W}$ with fixed α_v, $\mathbf{P}$ and $\mathbf{B}^{(v)}$. We minimize the following objective function

$$\mathcal{L}(\mathbf{W}) = \beta||(\mathbf{W}\mathbf{P} - \mathbf{Y})\mathbf{S}||_F^2. \tag{19}$$

Accordingly, we obtain the following updating rule

$$\mathbf{W} = \mathbf{Y}\mathbf{S}\mathbf{S}^T\mathbf{P}^T(\mathbf{P}\mathbf{S}\mathbf{S}^T\mathbf{P}^T)^{-1}. \tag{20}$$

Update $\boldsymbol{\alpha}$ with fixed $\mathbf{B}^{(v)}$ and $\mathbf{P}$. We employ a Lagrange multiplier λ to take the constraint into consideration, obtaining the following Lagrange function

$$Q(\alpha, \lambda) = \sum_{v=1}^{V} \alpha_v^r ||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2 - \lambda(\sum_{v=1}^{V} \alpha_v^r - 1). \tag{21}$$

By setting the derivative of Eq. (21) with respect to $\boldsymbol{\alpha}$ and λ to 0, then we have the following updating rule

$$\alpha_v = \frac{\big(\frac{1}{||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2}\big)^{1/r-1}}{\big(\sum_{v=1}^{V}\big(\frac{1}{||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2}\big)^{1/r-1}\big)^{-1}}. \tag{22}$$

According to the above updating rules, we can alternatively update these variables until convergence condition (i.e., the difference of the objective function value between two consecutive iterations is smaller than 10^{-6}) is reached.

Algorithm Analysis

○ **Convexity analysis.** Note that, due to the HSIC term involved, it is generally not convex due to the negative sign. This leads to a question: is the following function convex?

$$\begin{aligned}&\mathcal{L}(\mathbf{B}^{(v)}) = \alpha_v^r ||\mathbf{X}^{(v)} - \mathbf{B}^{(v)}\mathbf{P}||_F^2 \\ &\quad - \gamma \sum_{\substack{w=1 \\ w\neq v}}^{V} \text{HSIC}(\mathbf{B}^{(v)}, \mathbf{B}^{(w)}) + \mu||\mathbf{B}^{(v)} - \mathbf{S}_r^{(v)} + \mathbf{T}_r^{(v)}||_F^2.\end{aligned} \tag{23}$$

The optimal solution could be obtained if the function $\mathcal{L}(\mathbf{B}^{(v)})$ is strictly convex, which is also a prerequisite for the convergence of the holistic optimization. Therefore, we provide the guarantee for the convexity of $\mathcal{L}(\mathbf{B}^{(v)})$ under proper parameter setting as follows:

Theorem 1. *The subproblem $\mathcal{L}(\mathbf{B}^{(v)})$ is convex given the parameter setting $\mu \geq 4D(V-1)\gamma$, where V and D are number of views and $D = \max_{1\leq v<V}(\{D^{(v)}\}_{v=1}^{V})$.*

Proof 1. *The convexity of $\mathcal{L}(\mathbf{B}^{(v)})$ depends on whether its Hessian matrix $\bigtriangledown^2\mathcal{L}(\mathbf{B}^{(v)})$ is semi-positive definite or not (Boyd and Vandenberghe 2004). Since the first term is convex and we can gives the condition for strict convexity for the last two terms, thus, we only should guarantee the convexity of*

$$\mathcal{L}_c(\mathbf{B}^{(v)}) = -\gamma \sum_{\substack{w=1\\ w\neq v}}^{V} \mathit{HSIC}(\mathbf{B}^{(v)}, \mathbf{B}^{(w)}) + \mu||\mathbf{B}^{(v)} - \mathbf{S}_r^{(v)} + \mathbf{T}_r^{(v)}||_F^2. \tag{24}$$

Fortunately, the Hessian matrix $\bigtriangledown^2\mathcal{L}_c(\mathbf{B}^{(v)})$ can be easily computed as:

$$\nabla^2\mathcal{L}_c(\mathbf{B}^{(v)}) = \mu\mathbf{I} - \gamma \sum_{\substack{w=1\\ w\neq v}}^{V} \mathbf{C}\mathbf{B}^{(w)^T}\mathbf{B}^{(w)}\mathbf{C} = \mathbf{A}. \tag{25}$$

For convenience, we denote $\mathbf{L} = -\sum_{\substack{w=1\\ w\neq v}}^{V} \mathbf{C}\mathbf{B}^{(w)^T}\mathbf{B}^{(w)}\mathbf{C}$. According to the Gerschgorin theorem (Varga 2009), all the eigenvalues η of $\mathbf{A}$ lie in $|\eta - \mu/\gamma - l_{ii}| \leq \sum_{j=1;j\neq i}^{K} |l_{ij}|$, where K is the number of latent components as mentioned above. After transformation, the value of μ/γ has to satisfy the following constraint:

$$\mu/\gamma \geq \max_{1<i<K}(\sum_{j=1;j\neq i}^{K} |l_{ij}| - l_{ii}). \tag{26}$$

It is easy to show that $|l_{ij}| \leq 4(V-1)$, therefore, the lower bound of μ/γ is $4K(V-1)$. Accordingly, we can set $\mu = 4K(V-1)\gamma$ or even larger to ensure the constraint satisfied in practice. □

According to Theorem 1, the convexity of $\mathcal{L}(\mathbf{B}^{(v)})$ and the subsequent optimal solution is ensured.

○ **Complexity and convergence analysis.** There are four main sub-problems in our optimization procedure, i.e., $\mathbf{P}$, $\mathbf{B}^{(v)}$, $\mathbf{W}$ and α_v. For simplicity, we suppose the dimensionality of each view is D. The complexity of these subproblems are $O(K^2D + KDN + K^3)$, $O(L(K^2D + K^3))$, $O(CN^2 + CNK + K^3)$ and $O(DKN)$ respectively, where L is the iteration number in ADMM algorithm for updating $\mathbf{P}$. Since there are closed-form (optimal) solutions for updating $\mathbf{P}$, $\mathbf{B}^{(v)}$, $\mathbf{W}$ and weight vector $\boldsymbol{\alpha}$, the objective is non-decreasing with iterations. Therefore, our algorithm can be guaranteed to converge to a stationary point, which is also empirically validated in experiment as shown in Figure 3.

Experiments

Experiment Settings

Datasets & features. In this section, we evaluate our LSA-MML and compare it with state-of-the-art methods on three benchmark multi-label datasets, i.e., Corel5k (Duygulu et al. 2002), ESP Game (Von Ahn and Dabbish 2005) and PASCAL VOC' 07 (Everingham 2006). The detailed statistics information of these datasets is summarized in Table 1. We employ the standard partitions for training and testing sets [1] as described in Table 1.

Table 1: Statistics of datasets.

dataset	#instance	#training	#testing	#label
Corel5k	4999	4500	499	260
ESP Game	20770	18689	2081	268
PASCAL VOC	9963	5011	4952	20

There are three types of features, i.e., two types of local features: DenseSift (Lowe 2004) and DenseHue (Weijer and Schmid 2006), and one type of global features: Gist (Oliva and Torralba 2001) used in our experiments, where each type of features can be regarded as one view. The dimensionalities of DenseSift, DenseHue and Gist are 1000, 100 and 512, respectively. To comprehensively evaluate the effectiveness of label information, we randomly select a subset of labeled samples from the training sets with ratio $\in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. Due to randomness, we conduct each experiment 10 runs and report the average results with standard deviations.

Compared methods. We compare our method with several state-of-the-art multi-label classification methods. The first line is the traditional single view multi-label methods, while the second line takes advantage of all different views. Specifically, for the traditional single view methods, we report the best performance on the best single view (**BestView**). Moreover, we also conduct experiments for these methods by concatenating all views (**VConcate**). The traditional single view multi-label methods include binary relevance (**BR**) (Tsoumakas and Katakis 2006) and label powerset (**LP**) (Boutell et al. 2004) that act as the baselines. LP utilizes C4.5 as the base classifier and RAkEL is based on LP. There are some advanced comparisons in our experiments: the lazy multi-label methods based on k-nearest neighbor (**ML-kNN**) (Zhang and Zhou 2007) which is a simple but rather effective method, the ensemble methods such as random k-labelsets (**RAkEL**) (Tsoumakas, Katakis, and Vlahavas 2011), ensemble of classifier chains (**ECC**) (Read et al. 2011) and multi-label classification using ensembles of pruned sets (**EPS**) (Read, Pfahringer, and Holmes 2008). We compare our method with two pieces of work highly related with ours, i.e., multiview vector-valued manifold regularization (**MV^{3}MR**) (Luo et al. 2013) and low rank multi-view matrix completion (**LrMMC**) (Liu et al. 2015). The two methods both belong to the category of multi-view multi-label learning in semi-supervised manner.

[1] lear.inrialpes.fr/people/guillaumin/data.php

Table 2: Results (mean±standard deviation) on *Corel5k, ESP Game and PASCAL VOC*.

Dataset		Corel5k		ESP Game		PASCAL VOC	
Method	View	R-Loss↓	Ave-Pre↑	R-Loss↓	Ave-Pre↑	R-Loss↓	Ave-Pre↑
BR	BestView	.266±.002	.268±.004	.264±.000	.229±.003	.407±.007	.372±.009
	VConcate	.360±.022	.214±.006	.360±.002	.198±.004	.396±.008	.393±.003
LP	BestView	..696±.008	.046±.003	.495±.002	.057±.000	.420±.008	.296±.003
	VConcate	.678±.007	.067±.011	.491±.000	.058±.002	.411±.005	.311±.008
RAkEL	BestView	.389±.005	.243±.004	.380±.005	.208±.005	.247±.005	.499±.005
	VConcate	.350±.004	.304±.007	.308±.003	.254±.003	.237±.003	.515±.000
EPS	BestView	.347±.005	.289±.004	.358±.005	.223±.005	.239±.00	.494±.004
	VConcate	.344±.004	.378±.007	.352±.003	.223±.003	.228±.000	.507±.002
ECC	BestView	.150±.006	.372±.005	.182±.001	.294±.001	.203±.001	.253±.001
	VConcate	.150±.000	.382±.010	.179±.001	.309±.001	.195±.002	.552±.001
MLkNN	BestView	.126±.002	.406±.003	.182±.002	.294±.001	.180±.002	.564±.001
	VConcate	.123±.002	.424±.009	.167±.002	.272±.002	.164±.001	.586±.001
MV3MR	MultiView	.135±.005	.425±$\dot{0}$00	.183±.001	.267±.002	.181±$\dot{0}$03	.568±.003
LrMMC-1	MultiView	.112±.000	.382±$\dot{0}$04	**.158±.001**	.258±.003	.217±$\dot{0}$02	.463±.002
LrMMC-2	MultiView	**.101±.000**	.425±$\dot{0}$04	**.153±.001**	.264±.003	.163±$\dot{0}$02	.532±.002
Ours	BestView	.123±.010	.419±.014	.183±.001	.267±.002	.206±.005	.528±.006
	MultiView	.103±.002	**.462±.003**	.161±.001	**.345±.001**	**.149±.001**	**.610±.001**
Rank	——	2	1	3	1	1	1

For LrMMC, we use LrMMC-2 to indicate the method with data preprocessing for label unbalance issue as the authors did in their work while LrMMC-1 indicates using the original training data.

Parameter setting. We conduct parameter tuning on validation sets by following the same settings in (Luo et al. 2013; Liu et al. 2015). In specific, each data set is first partitioned into training and test set. Following the methods (Luo et al. 2013; Liu et al. 2015), 20% samples are then randomly selected from the test set as validation set for parameter tuning, and the rest is used for evaluating the classification performance of each algorithm. We select the value from $\{2, 3, 4, 5\}$ for r and from $\{0.01, 0.1, 1, 10, 100\}$ for β and γ. For optimization convenience, the inner product kernel is employed so there is no kernel parameter. To address the randomness in selecting samples, we have repeated the above procedure 10 times and reported the averaged results.

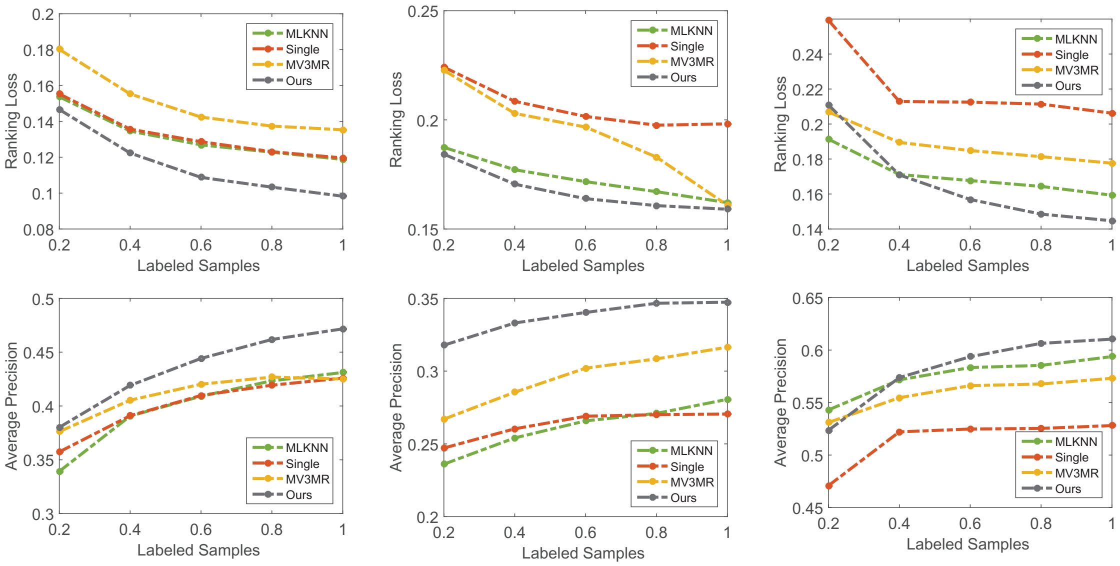
Two evaluation metrics that mostly used for multi-label classification (Zhang and Zhou 2007) are employed. For the *Ranking loss* (R-Loss), smaller value indicates better classification performance, while larger value of *Average precision* (Ave-Pre) means better performance. Limited by space, please refer the work in (Schapire and Singer 2000) for the details of these evaluation metrics.

Experiment Results

Comparison with state-of-the-arts. Table 2 demonstrates the classification comparison of different methods on benchmark datasets with 80% labeled samples used. We report the results under a part of training samples (instead of all training samples), since the performances (with average results and standard deviations) under different random parts of training samples could better characterize the comparison.

Based on the results in these tables, several observations are obtained as follows: 1) In a big picture, our algorithm almost achieves the best performance on all datasets, which clearly demonstrates the advantages of our method in exploring multi-view multi-label data. 2) It is clear that for each traditional single view multi-label method, the view concatenating strategy always obtains better performances than those of the best single view. This validates the effectiveness of multi-view learning over single view learning, since the complementarity among different views is of great importance. 3) We reported both the results of using the best single view and multiple views of our method, and the performance of using multiple views clearly outperforming that of single view validates that the multi-view treatment is essential for the performance. 4) The performance is rather stable with random initiation. Taking the experiment on the dataset Corel5k for example, we run our method 10 times with random initialization, and the standard deviation is 0.02 (same setting as in Table 2). Moreover, we also employed SVD for each single view as initialization (for each $\mathbf{B}^{(v)}$) in our code, and the performance is similar to that of random initialization.

Comparison with different ratios of training samples. To further evaluate the effectiveness of our method in utilizing scarce labeled samples, we provide the comparison for the competitive methods in terms of Ranking Loss and Average Precision with different labeled data ratios (from $20\% \sim 100\%$). Based on the results in Figure 2, the following observations are obtained: 1) With the increment of the ratio of labeled samples, the performances for all the algorithms are getting better, which confirms the valuableness of scarce labeled samples. 2) Compared with other algorithms, our model usually achieves the best result for different supervised ratio on all these datasets. This demonstrates that

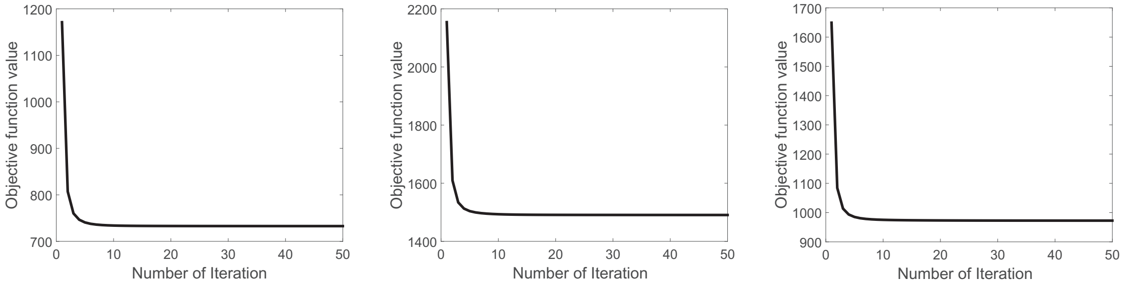


(a) Corel5k

(b) ESP Game

(c) PASCAL

Figure 2: The 1^{st} to 3^{rd} columns correspond to the performances with increase of the ratio of labeled samples. The method named as “Single” denotes LSA-MML with the best single view.



(a) Corel5k

(b) ESP Game

(c) PASCAL

Figure 3: Convergence curves on benchmark datasets.

our model can achieve better performance given the same amount of labeled samples, and empirically validates the effectiveness of our multi-view multi-label model.

Convergence experiments. As shown in Figure 3, we give the convergence experiments on the three datasets. Clearly, the results empirically prove that our algorithm can converge fast within a small number of iterations for all the datasets, and this is generally consistent with the theoretical analysis.

Conclusion

In this paper, we have proposed a new multi-label learning method, i.e., Latent Semantic Aware Multi-view Multi-label Learning, to fully take advantage of multiple views of data. Supervised by the limited label information, our model could well learn the common representations by simultaneously enforcing the consistence of latent semantic bases among different views in kernel spaces. Furthermore, different from the two-step manner (Liu et al. 2015), in our model the common representation learning and label prediction are in a unified framework, where they can improve each other iteratively. We also provided the instruction for parameter setting to guarantee the strict convexity of our algorithm. Experiments on different benchmark datasets clearly validated the superiority of our method over state-of-the-art ones for multi-view multi-label classification. There are several directions for the future work. First, exploring more general correlations between common representations and labels is challenging but of great interest. Second, due to the appealing performance of deep learning, extending our model with deep model will be our future work.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grand No: 61602337, 61732011, 61432011, U1435212 and 61502332), and Science Foundation of Tianjin (Grand No: 17JCZDJC30800).

References

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern Recognition* 1757–1771.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Cao, X.; Zhang, C.; Fu, H.; Liu, S.; and Zhang, H. 2015. Diversity-induced multi-view subspace clustering. In *CVPR*, 586–594.

Clare, A., and King, R. 2001. Knowledge discovery in multi-label phenotype data. *Principles of data mining and knowledge discovery* 42–53.

Duygulu, P.; Freitas, N. D.; Barnard, K.; and Forsyth, D. A. 2002. Object recognition as machine translation. *ECCV*.

Elisseeff, A., and Weston, J. 2002. A kernel method for multi-labelled classification. In *NIPS*, 681–687.

Everingham, M. 2006. The pascal visual object classes challenge 2007 (voc2007) results. *Lecture Notes in Computer Science* 117–176.

Fürnkranz, J.; Hüllermeier, E.; Mencía, E. L.; and Brinker, K. 2008. Multilabel classification via calibrated label ranking. *Machine Learning* 133–153.

Ghamrawi, N., and McCallum, A. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 195–200.

Gong, C.; Tao, D.; Yang, J.; and Liu, W. 2016. Teaching-to-learn and learning-to-teach for multi-label propagation. In *AAAI*, 1610–1616.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, 63–77.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Projective dictionary pair learning for pattern classification. In *NIPS*, 793–801.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-rank multi-view learning in matrix completion for multi-label image classification. In *AAAI*, 2778–2784.

Liu, X.; Li, M.; Wang, L.; Dou, Y.; Yin, J.; and Zhu, E. 2017. Multiple kernel k-means with incomplete kernels. In *AAAI*, 2259–2265.

Liu, Y.; Jin, R.; and Yang, L. 2006. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, 421–426.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 91—110.

Luo, Y.; Tao, D.; Xu, C.; and Xu, C. 2013. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE T-NNLS* 709–722.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *ICCV* 145–175.

Quadrianto, N.; Song, L.; and Smola, A. J. 2009. Kernelized sorting. In *NIPS*, 1289–1296.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine Learning* 333–359.

Read, J.; Pfahringer, B.; and Holmes, G. 2008. Multi-label classification using ensembles of pruned sets. In *ICDM*, 995–1000.

Schapire, R. E., and Singer, Y. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning* 135–168.

Tsoumakas, G., and Katakis, I. 2006. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*.

Tsoumakas, G., and Vlahavas, I. 2007. Random k-labelsets: An ensemble method for multilabel classification. *Machine Learning: ECML 2007* 406–417.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2011. Random k-labelsets for multilabel classification. *IEEE T-KDE* 1079–1089.

Varga, R. S. 2009. *Matrix iterative analysis*.

Von Ahn, L., and Dabbish, L. 2005. Labeling images with a computer game. In *Sigchi Conference on Human Factors in Computing Systems*, 319–326.

Wang, M.; Hua, X.-S.; Yuan, X.; Song, Y.; and Dai, L.-R. 2007. Optimizing multi-graph learning: towards a unified video annotation scheme. In *ACM Multimedia*, 862–871.

Wang, B.; Tu, Z.; and Tsotsos, J. K. 2013. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *ICCV*, 425–432.

Weijer, J. V. D., and Schmid, C. 2006. Coloring local feature extraction. In *ECCV*, 334–348.

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

Yang, S.-J.; Jiang, Y.; and Zhou, Z.-H. 2013. Multi-instance multi-label learning with weak label. In *IJCAI*, 1862–1868.

Yu, H.-F.; Jain, P.; Kar, P.; and Dhillon, I. S. 2014. Large-scale multi-label learning with missing labels. In *ICML*, 593–601.

Zhang, M.-L., and Zhou, Z.-H. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 2038–2048.

Zhang, C.; Fu, H.; Liu, S.; Liu, G.; and Cao, X. 2015. Low-rank tensor constrained multiview subspace clustering. In *ICCV*, 1582–1590.