Incomplete Multi-View Weak-Label Learning*

Qiaoyu Tan[1], Guoxian Yu[1,*], Carlotta Domeniconi[2], Jun Wang[1] and Zili Zhang[1,3]
[1]College of Computer and Information Science, Southwest University, Chongqing 400715, China
[2]Department of Computer Science, George Mason University, Fairfax 22030, USA
[3]School of Information Technology, Deakin University, Geelong, VIC 3220, Australia
{tqy1995119, gxyu, kingjun, zhangzl}@swu.edu.cn, carlotta@cs.gmu.edu

Abstract

Learning from multi-view multi-label data has wide applications. Two main challenges characterize this learning task: incomplete views and missing (weak) labels. The former assumes that views may not include all data objects. The weak label setting implies that only a subset of relevant labels are provided for training objects while other labels are missing. Both incomplete views and weak labels can lead to significant performance degradation. In this paper, we propose a novel model (iMVWL) to jointly address the two challenges. iMVWL learns a shared subspace from incomplete views with weak labels, local label correlations, and a predictor in this subspace, simultaneously. The latter can capture not only cross-view relationships but also weak-label information of training samples. We further develop an alternative solution to optimize our model; this solution can avoid suboptimal results and reinforce their reciprocal effects, and thus further improve the performance. Extensive experimental results on real-world datasets validate the effectiveness of our model against other competitive algorithms.

1 Introduction

In many real-world applications, a sample may have several heterogenous representations, each one giving a different view of the data, and may also have multiple labels. For example, a web image can be tagged with multiple topics given as labels, such as cattle, grass, and tree. At the same time, the image can also be described using heterogenous features, such as texture descriptors, shape descriptors, color descriptors, surrounding texts, and so on. Multi-view multi-label learning, as a natural formulation for this type of data, has attracted a lot of attention in machine learning and in many application domains [Liu *et al.*, 2015; Luo *et al.*, 2015]. Although many multi-view multi-label learning methods have been proposed in recent years, a main challenge remains for this problem: the lack of fully labeled training samples. In practice, it is rather difficult to collect all the relevant labels of a sample, and only a subset may be available. One such example is image annotation. An annotator may only afford to annotate an image with *some* labels, especially when the number of relevant labels is large. Learning from partially labeled samples is termed as the *weak-label learning* problem [Sun *et al.*, 2010; Bucak *et al.*, 2011; Kong *et al.*, 2014]. Several weak-label learning methods have been proposed in single-view [Yu *et al.*, 2014; Cabral *et al.*, 2015] and multi-view scenarios [Zhang *et al.*, 2013].

However, almost all aforementioned methods do not account for another important challenge: incomplete data. Namely, some samples may be missing their representation in one view. This can happen, in practice, for a variety of reasons, e.g., a temporary failure of sensors, or a man-made error. It has been observed that incomplete data are likely to lead to degradation in multi-view learning performance [Xu *et al.*, 2015].

The more challenging case is when both missing labels and incomplete data *co-exist* in a multi-view multi-label learning problem. To the best of our knowledge, few studies exist that handle incomplete data [Xu *et al.*, 2015] or missing labels [Zhang *et al.*, 2013] in multi-view learning, but no previous work simultaneously takes into account both issues. To bridge this gap, we propose a novel unified model, called *i*ncomplete *M*ulti-*V*iew *W*eak-Label *L*earning (iMVWL), to jointly handle incomplete views and missing labels. The basic strategy of iMVWL is to seek a shared subspace across heterogenous incomplete views, and a robust weak-label classifier in this subspace in a unified learning framework, where label correlations and discriminative information can be learned. In summary, our main contributions are as follows:

- The proposed iMVWL can jointly address incomplete views and missing labels. It learns a shared subspace from incomplete views with weak labels, label correlations, and a predictor in this subspace simultaneously.
- We develop a solution to iteratively optimize our model, avoiding suboptimal problems.
- Experiments on five widely used datasets and comparisons with a number of competitive methods [Yuan *et al.*, 2012; Zhang *et al.*, 2013; Xu *et al.*, 2015; Liu *et al.*, 2015] demonstrate the superiority of the proposed work.

2 Related Work

This work is related to two branches of studies, weak-label learning and multi-view learning. Weak-label learning was pioneered by [Sun *et al.*, 2010]. Many weak-label learning algorithms have subsequently been proposed. To name a few,

[*] Guoxian Yu is the corresponding author.

weak-label learning algorithms under a supervised setting [Bucak *et al.*, 2011; Kong *et al.*, 2014], under a semi-supervised setting [Zhao and Guo, 2015; Wu *et al.*, 2015], and under a multi-instance multi-label framework [Yang *et al.*, 2013].

Multi-view learning deals with data represented in different views and has attracted increasing interest in recent years. Previous approaches have considered multi-views in conjunction with semi-supervised learning [Xu *et al.*, 2015; Nie *et al.*, 2017], with multi-label learning [Zhang *et al.*, 2013; Liu *et al.*, 2015] or with active learning [Wang and Zhou, 2010]. Others tried to estimate a latent subspace by assuming that samples (in different views) corresponding to the same object are close to each other when mapped into the latent subspace [Zhang *et al.*, 2013; Liu *et al.*, 2015; Xu *et al.*, 2017].

Almost all previous weak-label learning studies focus on a single-view setting. Likewise, almost all existing multi-view learning studies typically assume completeness of each view (i.e., each sample appears in all views). The only exceptions are LabelMe [Zhang *et al.*, 2013] and MVL-IV [Xu *et al.*, 2015]. LabelMe is a multi-view weak-label learning method, but it assumes complete views of each training sample. As discussed above, this assumption is often violated in practice. MVL-IV is a recently proposed multi-view learning solution that considers incomplete views. It integrates multiple incomplete views by assuming that the views are generated from a common subspace, so that the learned subspace may capture cross-view relationships. Nevertheless, MVL-IV is an unsupervised subspace learning approach, which may be lacking discriminative ability due to missing label information [Xu *et al.*, 2017]. On the other hand, MVL-IV assumes the available labels of training samples are complete, ignoring the widely witnessed weak-label scenarios. Moreover, MVL-IV decouples the subspace learning from the follow-up classification learning tasks, which may result in suboptimal models due to the lack of mutual adaptation of the two steps.

To address these challenges, this paper proposes a novel unified framework (iMVWL) to jointly handle incomplete views and weak labels. iMVWL simultaneously learns the shared subspace from incomplete views, a predictor in this subspace, and local label structure. iMVWL not only achieves a discriminative shared subspace from incomplete views, but also a robust weak-label classifier that can dynamically capture local label correlations. To the best of our knowledge, no previous work has been developed to jointly handle challenges from both incomplete views and weak labels.

3 Proposed Approach

Suppose $\mathcal{X} = \{\mathbf{X}_v\}_{v=1}^{n_v}$ represents a dataset with n samples and n_v views, where $\mathbf{X}_v = [\mathbf{x}_v^1, \mathbf{x}_v^2, ..., \mathbf{x}_v^n] \in \mathbb{R}^{n \times d_v}$ indicates the full feature space in view v. $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T \in \{-1, 1\}^{n \times c}$ is the corresponding weak-label matrix, where $\mathbf{y}_i \in \{-1, 1\}^c$ is the label vector of $\mathbf{x}_i$ and c is the number of distinct labels. $\mathbf{y}_{ic'} = 1$ $(c' = 1, ..., c)$ means the c'-th label is relevant, while $\mathbf{y}_{ic'} = -1$ does not provide any information. In the multi-incomplete view setting, a sample may appear in some views, but not all. That is, the data matrix $\mathbf{X}$ may have a number of missing rows. An easy fix to this problem is to remove any sample missing in at least one view. However, this approach will significantly reduce the number of samples that can be used for training. Our goal is to predict the labels of unlabeled samples based on multiple incomplete feature spaces $\mathcal{X}$ and the weak-label space spanned by $\mathbf{Y}$.

3.1 Problem Formulation

With multi-view multi-label data, how to generate a shared discriminative subspace across views and how to train an efficient and robust multi-label classifier in that subspace for label prediction are two challenging problems. Some subspace learning algorithms have been proposed to seek the shared subspace across views [Zhao *et al.*, 2017], such as multi-view subspace learning methods based on a low rank constraint [Liu *et al.*, 2015], matrix factorization [Žitnik and Zupan, 2015], and nonnegative matrix factorization (NMF) [Wang and Zhang, 2013]. Among them, NMF has been successfully applied in text mining, image annotations, bioinformatics, recommender systems and other domains [Wang and Zhang, 2013], since most data matrices are naturally nonnegative, or can be easily transformed into nonnegative ones. The major difference between NMF and other matrix factorization methods, such as SVD (singular value decomposition), is the nonnegative constraints, which help to obtain a part-based representation as well as to enhance interpretability of the learned subspace. In this paper, we also focus on nonnegative data matrix mining tasks, and adapt NMF to learn a discriminative low-rank representation from incomplete views by using weak-label information. Given a multi-view datasets $\mathcal{X}$, the standard NMF can be adapted to find a shared subspace $\mathbf{V}$ as follows:

$$\min_{\{\mathbf{U}_v, \mathbf{V}\}} \sum_{v=1}^{n_v} ||\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T||_F^2 \;\; \textit{s.t.}\; \mathbf{U} \geq 0, \mathbf{V} \geq 0 \qquad (1)$$

where $\mathbf{U}_v \in \mathbb{R}^{d_v \times k}$, $\mathbf{V} \in \mathbb{R}^{n \times k}$, and k is the desired low-rank size, $||.||_F$ represents the Frobenius norm, $\mathbf{U}_v \geq 0$ and $\mathbf{V} \geq 0$ are the nonnegative constraints for the matrices. The learned subspace $\mathbf{V}$ in Eq. (1) can capture the cross-view relationships since it enables the integration of complementary information across multiple views [Xu *et al.*, 2015]. In many applications, however, Eq. (1) may be unreliable due to the presence of incomplete views. A remedy for Eq. (1) to deal with this problem is to fill the missing samples with average feature values; nonetheless, this approach may introduce errors, especially when the number of missing samples is large, hence not suitable for incomplete view setting. Besides, the above unsupervised subspace learning process is lacking discriminative ability because it ignores label information. To address these drawbacks, we formulate subspace learning from incomplete views as a supervised approach, which considers label information and complementary information across incomplete views as follows:

$$\min_{\{\mathbf{U}_v, \mathbf{V}, \mathbf{W}\}} \sum_{v=1}^{n_v} ||\mathbf{O}^v \odot (\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T)||_F^2 + \alpha||\mathbf{V}\mathbf{W} - \mathbf{Y}||_F^2 \quad (2)$$

where $\odot$ is the Hadamard product (element-wise product). $\mathbf{O}^v \in \mathbb{R}^{n \times d_v}$ is an indicator matrix that denotes the missing entries, where $\mathbf{O}_{i,j}^v = 1$ if (i, j) is an observed entry in $\mathbf{X}_v$; $\mathbf{O}_{i,j}^v = 0$, otherwise. $\mathbf{Y} \in \mathbb{R}^{n \times c}$ denotes the available label matrix of n samples. $\mathbf{W} \in \mathbb{R}^{k\times c}$ is the coefficient matrix, which maps the shared feature subspace into a semantic label space. In Eq. (2), we can achieve two goals. On one hand, the learned subspace can capture the cross-view relationships as discussed before. On the other hand, we can utilize the label information $\mathbf{Y}$ to induce the shared subspace towards a semantic label space via the second term, which not only helps to obtain a discriminative subspace but also may alleviate the widely spread semantic gap [Datta *et al.*, 2008] between the input heterogeneous feature spaces and the semantic label space, since $\mathbf{V}$ can be viewed as a bridge between them.

Eq. (2), however, ignores another important issue in multi-view weak-label learning, i.e., the presence of missing labels. Namely, often in practice, $\mathbf{Y}$ is incomplete and contains many missing entries. As such, we need to avoid the influence of missing labels in $\mathbf{Y}$ and improve the robustness of Eq. (2). Considering that label correlation is very important in weak-label setting and usually can further improve the performance [Dong *et al.*, 2018], we leverage label correlation among weak labels to estimate the predicted likelihood scores and extend Eq. (2) as follows:

$$\min_{\{\mathbf{U}_v,\mathbf{V},\mathbf{W},\mathbf{S}\}} \sum_{v=1}^{n_v} ||\mathbf{O}^v \odot (\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T)||_F^2 + \alpha||\mathbf{M}\odot(\mathbf{VWS}-\mathbf{Y})||_F^2 \tag{3}$$

where $\mathbf{M} \in \mathbb{R}^{n\times c}$ is an indicator matrix for missing labels: $\mathbf{M}_{i,j} = 1$ if (i, j) is an observed entry in $\mathbf{Y}$; $\mathbf{M}_{i,j} = 0$, otherwise. $\mathbf{S} \in \mathcal{R}^{c\times c}$ denotes the label correlation matrix, $\alpha > 0$ is the trade-off parameter. By incorporating the label correlation matrix $\mathbf{S}$, Eq. (3) not only can estimate the predicted likelihood scores, but also can enhance the discriminative ability of the learned subspace by using label correlations among weak labels. However, since the observed relevant label sets of samples are incomplete, we cannot directly compute the label correlation matrix $\mathbf{S}$ from prior knowledge $\mathbf{Y}$; we need to learn it. In addition, we also need to account for the fact that label correlations are naturally *local* [Huang and Zhou, 2012], and can manifest as *direct or indirect* dependencies [Wu *et al.*, 2015]. As such, it is reasonable to assume that label correlations are *locally structured*, that is, there exists a subset of labels, which are closely related to each other through complex correlations, and are independent from the rest. This local structure typically implies a low-rank structure of $\mathbf{S}$, which is common in real-world applications [**?**; Xu *et al.*, 2016]. To capture local label correlations, we add a low-rank constraint on $\mathbf{S}$, and make Eq. (3) more suitable for weak-label problems as follows:

$$\begin{aligned}\min_{\{\mathbf{U}_v,\mathbf{V},\mathbf{W},\mathbf{S}\}} & \sum_{v=1}^{n_v} ||\mathbf{O}^v \odot (\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T)||_F^2 \\ & + \alpha||\mathbf{M}\odot(\mathbf{VWS}-\mathbf{Y})||_F^2 + \beta rank(\mathbf{S})\end{aligned} \tag{4}$$

where β is the trade-off parameter, which balances the relative importance of the low-rank constraint on $\mathbf{S}$. By adding the *rank* term, our model can capture local label correlations among weak labels which is more suitable for real-world applications. It's worth noticing that the work in [**?**] also makes a low-rank assumption among labels, but its usage is different. In particular, the authors in [**?**] multiply the low rank correlation matrix with $\mathbf{Y}$ to replenish missing labels. In contrast, we use the low rank correlation matrix with the predicted likelihood label vectors. As such, since the estimated label correlation values may not be very reliable in practice, our method is less impacted by them. Furthermore, since $\mathbf{V}$ is the low-rank feature representation learned across incomplete views, Eq. (4) is robust to outliers and background noise that may affect the feature space. The rank minimization problem is NP-hard. Here we can relax the rank problem using the nuclear norm $||\cdot||_*$ [Candès and Recht, 2009], and reformulate Eq. (4) as follows:

$$\begin{aligned}\min_{\{\mathbf{U}_v,\mathbf{V},\mathbf{W},\mathbf{S}\}} & \sum_{v=1}^{n_v} ||\mathbf{O}^v \odot (\mathbf{X}_v - \mathbf{V}\mathbf{U}_v^T)||_F^2 \\ & + \alpha||\mathbf{M}\odot(\mathbf{VWS}-\mathbf{Y})||_F^2 + \beta||\mathbf{S}||_*\end{aligned} \tag{5}$$

Eq. (5) considers cross-view relationships and local (low rank) label structure. In addition, it absorbs label information to induce the shared subspace and enhance its discriminative power. Another advantage of our model is that it jointly learns a shared subspace from incomplete views with weak labels, the local label structure, and the predictor in this subspace. This unified model reinforces their reciprocal effects and thus further improves the performance.

4 Optimization

The minimization problem in Eq. (5) is defined with respect to $\{\mathbf{U}_v\}_{v=1}^{v}$, $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{S}$. Since a close-form solution cannot be computed, we develop an alternative optimization method to optimize the objective function.

(I). Keep $\{\mathbf{U}_v\}$, V and S fixed, update W

When $\{\mathbf{U}_v\}$, $\mathbf{V}$ and $\mathbf{S}$ are fixed, we have the following equation for $\mathbf{W}$ by taking the derivative of Eq. (5) w.r.t. $\mathbf{W}$,

$$\mathscr{J}_1(\mathbf{W}) = 2\mathbf{V}^T(\mathbf{M}\odot\mathbf{VWS})\mathbf{S}^T - 2\mathbf{V}^T(\mathbf{M}\odot\mathbf{Y})\mathbf{S}^T \tag{6}$$

We can derive the following fixed-point updating rule for $\mathbf{W}$,

$$\mathbf{W} = \mathbf{W}\odot\frac{\mathbf{V}^T(\mathbf{M}\odot\mathbf{Y})\mathbf{S}^T}{\mathbf{V}^T(\mathbf{M}\odot\mathbf{VWS})\mathbf{S}^T} \tag{7}$$

(II). Keep $\{\mathbf{U}_v\}$, V and W fixed, update S

When $\{\mathbf{U}_v\}$, $\mathbf{V}$ and $\mathbf{W}$ are fixed, optimizing Eq. (5) with respect to $\mathbf{S}$ is equivalent to

$$\mathscr{J}_2(\mathbf{S}) = \alpha||\mathbf{M}\odot(\mathbf{XWS}-\mathbf{Y})||_F^2 + \beta||\mathbf{S}||_* \tag{8}$$

Eq. (8) can be viewed as a matrix completion problem [Candès and Recht, 2009], and many algorithms have been proposed to solve this problem in recent decades. Here we adopt an efficient speedup algorithm, Maxide [Xu *et al.*, 2013], to solve it. Maxide only needs to estimate a $c \times c$ matrix.

(III). Keep $\{\mathbf{U}_v\}$, W and S fixed, update V

When keeping $\{\mathbf{U}_v\}$, $\mathbf{W}$ and $\mathbf{S}$ fixed, we obtain the following equation for $\mathbf{V}$ by taking the derivative of Eq. (5) w.r.t. $\mathbf{V}$:

$$\begin{aligned}\mathscr{J}_3(\mathbf{V}) =& 2\sum_{v=1}^{n_v}(\mathbf{O}^v\odot\mathbf{V}\mathbf{U}_v^T)\mathbf{U}_v + 2\alpha(\mathbf{M}\odot\mathbf{VWS})\mathbf{S}^T\mathbf{W}^T \\ & - 2\sum_{v=1}^{n_v}(\mathbf{O}^v\odot\mathbf{X}_v)\mathbf{U}_v - 2\alpha(\mathbf{M}\odot\mathbf{Y})\mathbf{S}^T\mathbf{W}^T \\ & \textit{s.t. } \mathbf{U}_v \geq 0, \mathbf{V}_v \geq 0,\ v = 1, 2, ..., n_v\end{aligned} \tag{9}$$

Using the Karush-Kuhn-Tucker (KKT) condition [Boyd and Vandenberghe, 2004], we can derive the following updating rule,

$$\mathbf{V}_{i,j} \leftarrow \mathbf{V}_{i,j} \frac{(\sum_{v=1}^{n_v} (\mathbf{O}^v \odot \mathbf{X}_v)\mathbf{U}_v + \alpha(\mathbf{M} \odot \mathbf{Y})\mathbf{S}^T\mathbf{W}^T)_{i,j}}{(\sum_{v=1}^{n_v} (\mathbf{O}^v \odot \mathbf{V}\mathbf{U}_v^T)\mathbf{U}_v + \alpha(\mathbf{M} \odot \mathbf{VWS})\mathbf{S}^T\mathbf{W}^T)_{i,j}} \tag{10}$$

(IV). Keep $\{\mathbf{V}\}$, $\mathbf{W}$ and $\mathbf{S}$ fixed, update $\mathbf{U}_v$

With $\{\mathbf{V}\}$, $\mathbf{W}$ and $\mathbf{S}$ fixed, the computation of $\mathbf{U}_v$ is independent from $\mathbf{U}_{v'}, v' \neq v$. Thus, for each view v, we obtain the following equation for $\mathbf{U}_v$ by taking the derivative of Eq. (5) w.r.t. $\mathbf{U}_v$:

$$\mathcal{J}_4(\mathbf{U}_v) = 2(\mathbf{O}^v \odot \mathbf{U}_v\mathbf{V}^T)\mathbf{V} - 2(\mathbf{O}^v \odot \mathbf{X}_v)^T\mathbf{V} \tag{11}$$

Using the KKT condition, we can derive the following updating rule:

$$(\mathbf{U}_v)_{i,j} \leftarrow (\mathbf{U}_v)_{i,j} \frac{((\mathbf{O}^v \odot \mathbf{X}_v)^T\mathbf{V})_{i,j}}{((\mathbf{O}^v \odot \mathbf{U}_v\mathbf{V}^T)\mathbf{V})_{i,j}} \tag{12}$$

4.1 Complexity Analysis

The time complexity of iMVWL is dominated by matrix multiplication. In each iteration, the time complexities of solving $\mathbf{W}$ and $\mathbf{S}$ in Eq. (7) and Eq. (8) are $O(nck)$ and $O(rc\ln c\ln n)$ respectively, where r is the rank of $\mathbf{S}$; the time complexity of updating $\mathbf{V}$ in Eq. (10) and $\mathbf{U}_v$ in Eq. (12) is less than $O(n_v(nkd_{max} + 2nk^2 + nck))$ and $O(n_v(nkd_{max} + d_{max}k^2))$, respectively. d_{max} represents the largest dimensionality of the views. Since $n \gg k$ and $n \gg c$, the overall time complexity of iMVWL is $O(tn_vnkd_{max})$, where t is the number of iterations to reach convergence. In practice, t does not exceed 60. In our study, some of the views have sparse feature matrices; as such, the actual time cost of the above operations can be further reduced.

5 Experiments

5.1 Experimental Setup

The five multi-view datasets used in the experiments (Core15k, Pascal07, ESPGame, IAPRTC-12, and Mirflicker) are summarized in Table 1. These datasets[1] are obtained from [Guillaumin *et al.*, 2010], and each is represented by six feature views: HUE, SIFT, GIST, HSV, RGB, and LAB. For each dataset, we randomly sample 70% of the data for training, and use the remaining 30% data for testing (unlabeled data). Moreover, to create weak-label scenarios, we follow the protocol given in [Xu *et al.*, 2013]: for each label c' we remove the assignment of c' for $\omega\%$ randomly sampled positive and negative training samples (c' becomes a missing label); to create incomplete-view data scenarios, we randomly remove $\varepsilon\%$ samples from each view, while ensuring each sample appears in at least one view. For each dataset, d_{min} represents the minimum dimensionality of the different views.

Methods: We compare iMVWL against four state-of-the-art methods: LabelMe [Zhang *et al.*, 2013], MVL-IV [Xu *et al.*, 2015], lrMMC [Liu *et al.*, 2015], and iMSF [Yuan *et al.*, 2012]. The first two methods have been introduced in the Related work Section. lrMMC is a matrix completion based multi-view learning method, but it assumes complete views of each training sample and does not explicitly consider missing labels and label correlations. iMSF was initially proposed for single label classification with multiple incomplete sources; we extend it for multi-label classification by training multiple classifiers (one for each label). These comparing methods cannot directly handle incomplete multi-view weak-label settings. For experimental comparisons, we adapt LabelMe and lrMMC by filling missing features with average values, and set the missing labels of MVL-IV and iMSF as negative labels. In addition, we introduce iMVWL-Sp, iMVWL-X, and iMVWL-Nc to investigate the contribution of learning a discriminative shared subspace, separately handling multiple feature views, and capturing local label correlations, respectively. iMVWL-Sp excludes label information during the subspace learning process. iMVWL-X concatenates multi-view features into a single vector. iMVWL-Nc excludes label correlations. Five-fold cross validation on the training set is used to select the optimal parameter values for each competitive method. Optimal parameters for the competitive methods are selected as suggested in the corresponding papers. For our method, we selected the parameters α and β from $\{10^i | i = -5, \cdots, 0\}$. Experimental results show that iMVWL yields relatively stable performance with α around 10^{-2} and β around 10^{-2}, and therefore we use these values. All the experiments are repeated ten times, and both the average and standard deviation are reported. *The source code of iMVWL is publicly available at* http://mlda.swu.edu.cn/codes.php?name=iMVWL.

datasets	n	n_v	c	#avg
Core15k	4999	6	260	3.396
Pascal07	9963	6	20	1.465
ESPGame	20770	6	268	4.686
IAPRTC12	19627	6	291	5.719
Mirflicker	25000	6	38	4.716

Table 1: Statistics of five multi-view datasets: n is the number of samples; n_v is the number of views; c is the number of distinct labels; and #avg is the average number of labels per sample.

Evaluation: Four widely used multi-label evaluation metrics are adopted for performance comparisons, i.e., *Ranking Loss (RL)*, *Average Precision (AP)*, *Hamming Loss (HL)*, and adapted *AUC*. A formal definition of the first three metrics can be found in [Zhang and Zhou, 2014]. The adapted *AUC* is suggested in [Bucak *et al.*, 2011]. To maintain consistency with other evaluation metrics, in our experiments, we report 1-RL instead of *RL*. Thus, as for other metrics, the higher the value of 1-RL, the better the performance is.

5.2 Results On All Datasets

Table 2 and 3 give the results of all methods on five datasets across four evaluation metrics. In the table, •/◦ indicates whether iMVWL is statistically (using a pairwise t-test at 95% significance level) superior/inferior to the corresponding method.

It can be seen that iMVWL outperforms the other methods in most cases. MVL-IV, iMSF, and iMVWL are all designed

[1] Available at http://lear.inrialpes.fr/people/guillaumin/data.php

Dataset	metric	lrMMC	MVL-IV	LabelMe	iMSF	iMVWL
Core15k	1-HL	0.954 ± 0.000●	0.954 ± 0.000●	0.946 ± 0.000●	0.943 ± 0.000●	0.956 ± 0.000
	1-RL	0.762 ± 0.002●	0.756 ± 0.001●	0.638 ± 0.003●	0.709 ± 0.005●	0.822 ± 0.001
	AP	0.240 ± 0.002●	0.240 ± 0.001●	0.204 ± 0.002●	0.189 ± 0.002●	0.313 ± 0.002
	AUC	0.763 ± 0.002●	0.762 ± 0.001●	0.715 ± 0.001●	0.663 ± 0.005●	0.824 ± 0.001
Pascal07	1-HL	0.882 ± 0.000●	0.883 ± 0.000●	0.837 ± 0.000●	0.836 ± 0.000●	0.886 ± 0.000
	1-RL	0.698 ± 0.003●	0.702 ± 0.001●	0.643 ± 0.004●	0.568 ± 0.000●	0.749 ± 0.002
	AP	0.425 ± 0.003●	0.433 ± 0.002●	0.358 ± 0.003●	0.325 ± 0.000●	0.455 ± 0.001
	AUC	0.728 ± 0.002●	0.730 ± 0.001●	0.686 ± 0.005●	0.620 ± 0.001●	0.784 ± 0.001
ESPGame	1-HL	0.970 ± 0.000●	0.970 ± 0.000●	0.967 ± 0.000●	0.964 ± 0.000●	0.971 ± 0.000
	1-RL	0.777 ± 0.001●	0.778 ± 0.000●	0.683 ± 0.002●	0.722 ± 0.002●	0.803 ± 0.001
	AP	0.188 ± 0.000●	0.189 ± 0.000●	0.132 ± 0.000●	0.108 ± 0.000●	0.236 ± 0.001
	AUC	0.783 ± 0.001●	0.784 ± 0.000●	0.734 ± 0.001●	0.674 ± 0.003●	0.808 ± 0.001
IAPRTC12	1-HL	0.967 ± 0.000●	0.967 ± 0.000●	0.963 ± 0.000●	0.960 ± 0.000●	0.969 ± 0.000
	1-RL	0.801 ± 0.000●	0.799 ± 0.001●	0.725 ± 0.001●	0.631 ± 0.000●	0.830 ± 0.001
	AP	0.197 ± 0.000●	0.198 ± 0.000●	0.141 ± 0.000●	0.101 ± 0.000●	0.234 ± 0.002
	AUC	0.805 ± 0.000●	0.804 ± 0.001●	0.746 ± 0.001●	0.665 ± 0.001●	0.832 ± 0.001
Mirflicker	1-HL	0.839 ± 0.000●	0.839 ± 0.000●	0.778 ± 0.000●	0.775 ± 0.000●	0.844 ± 0.001
	1-RL	0.802 ± 0.001●	0.808 ± 0.001●	0.771 ± 0.001●	0.641 ± 0.001●	0.817 ± 0.001
	AP	0.441 ± 0.001●	0.449 ± 0.001●	0.375 ± 0.000●	0.323 ± 0.000●	0.497 ± 0.003
	AUC	0.806 ± 0.001●	0.807 ± 0.000●	0.761 ± 0.000●	0.715 ± 0.001●	0.816 ± 0.001

Table 2: Results on all datasets with $\omega\% = 50\%$, $\varepsilon\% = 50\%$, and $k = 0.5d_{min}$.

Dataset	metric	iMVWL-Sp	iMVWL-Nc	iMVWL-X	iMVWL
Core15k	1-HL	0.955 ± 0.000●	0.955 ± 0.000●	0.955 ± 0.000●	0.956 ± 0.000
	1-RL	0.790 ± 0.001●	0.798 ± 0.002●	0.808 ± 0.000●	0.822 ± 0.001
	AP	0.285 ± 0.003●	0.272 ± 0.003●	0.299 ± 0.000●	0.313 ± 0.002
	AUC	0.791 ± 0.001●	0.798 ± 0.002●	0.811 ± 0.001●	0.824 ± 0.001
Pascal07	1-HL	0.883 ± 0.000●	0.884 ± 0.000●	0.884 ± 0.000●	0.886 ± 0.000
	1-RL	0.721 ± 0.001●	0.728 ± 0.002●	0.726 ± 0.003●	0.749 ± 0.002
	AP	0.436 ± 0.001●	0.440 ± 0.002●	0.446 ± 0.001●	0.455 ± 0.001
	AUC	0.750 ± 0.001●	0.745 ± 0.003●	0.759 ± 0.001●	0.784 ± 0.001
ESPGame	1-HL	0.971 ± 0.000●	0.970 ± 0.000●	0.970 ± 0.000●	0.971 ± 0.000
	1-RL	0.790 ± 0.001●	0.780 ± 0.001●	0.787 ± 0.001●	0.803 ± 0.001
	AP	0.213 ± 0.002●	0.199 ± 0.001●	0.198 ± 0.001●	0.236 ± 0.001
	AUC	0.795 ± 0.001●	0.785 ± 0.001●	0.791 ± 0.000●	0.808 ± 0.001
IAPRTC12	1-HL	0.968 ± 0.000●	0.968 ± 0.000●	0.967 ± 0.000●	0.969 ± 0.000
	1-RL	0.810 ± 0.001●	0.804 ± 0.002●	0.797 ± 0.003●	0.830 ± 0.001
	AP	0.213 ± 0.001●	0.206 ± 0.001●	0.202 ± 0.003●	0.234 ± 0.002
	AUC	0.813 ± 0.001●	0.804 ± 0.002●	0.800 ± 0.002●	0.832 ± 0.001
Mirflicker	1-HL	0.843 ± 0.000	0.839 ± 0.000●	0.841 ± 0.001●	0.844 ± 0.001
	1-RL	0.817 ± 0.001	0.813 ± 0.001●	0.806 ± 0.002●	0.817 ± 0.001
	AP	0.486 ± 0.002●	0.486 ± 0.002●	0.480 ± 0.004●	0.497 ± 0.003
	AUC	0.816 ± 0.001	0.807 ± 0.001●	0.805 ± 0.003●	0.816 ± 0.001

Table 3: Results of variants of iMVWL on all datasets with $\omega\% = 50\%$, $\varepsilon\% = 50\%$, and $k = 0.5d_{min}$.

for incomplete multi-view data, but iMVWL almost always outperforms the other two across the four evaluation metrics. The main reason is that MVL-IV and iMSF assume that the available labels are complete and ignore the widely witnessed weak-label scenarios. Both lrMMC and LabelMe are multi-view learning methods based on subspace learning, and they can handle weak-labels. But LabelMe is outperformed by lrMVL across five datasets. A possible reason is that lrMMC considers the multi-view weak-label learning task as a matrix completion problem, which is more robust to missing values. However, lrMMC is outperformed by iMVWL in many cases. This is mainly because lrMMC assumes the completeness of multiple views. As discussed in the Introduction section, this assumption is often violated in practice.

In Table 3, iMVWL-Sp is a degenerate case of iMVWL, which is obtained by excluding label information during subspace learning, thereby isolating the subspace learning process from the subsequent classification task. iMVWL almost always performs better than iMVWL-Sp on these datasets. This is mainly because in iMVWL-Sp the learned subspace may lack the ability to discriminate between different labels. In addition, when the objectives are treated separately, an optimal subspace can be achieved, but it may not be optimal for the subsequent prediction. These results corroborate our motivation to jointly optimize the two objectives. iMVWL-Nc is obtained from iMVWL by excluding label correlations, and is almost always outperformed by iMVWL. This fact demonstrates the effectiveness of the proposed method in capturing local label correlations. iMVWL-X is another variant of iMVWL; it concatenates all feature view vectors into a single vector, and follows the same process of iMVWL for prediction. It's outperformed by iMVWL in almost every case. These results justify the rationale of handling multiple feature views separately.

An interesting observation is that iMVWL-Sp performs better than (or comparable to) other methods in most cases. This is mainly because iMVWL-Sp addresses both incomplete multi-view and weak-label problems, while the other methods only address one of the two. The performance margin achieved by iMVWL and iMVWL-Sp further justifies our motivation to jointly handle incomplete multi-view data and weak-labels.

5.3 Handling Weak-Labels

We conducted additional experiments on Core15k to investigate the performance of iMVWL and other methods when handling missing labels. We set the dimensionality of the shared subspaces equal to 20%, 50%, and 80% of d_{min}, with $\omega\%$ that varies from 0% to 50% with a step-size of 10%. Since iMSF is not a subspace learning method, its performance is the same for all dimensionality. Since the results on all evaluation metrics are similar, for space limitation, we report only the results of AUC in Figure 1.

We can see that the performance of all the methods decreases when $\omega\%$ increases, and iMVWL outperforms the competitive methods in all the settings. Also, regardless of the dimensionality of the learned subspaces, iMVWL performs consistently better than the other methods under different ratios of missing labels.

5.4 Handling Incomplete Multi-View Data

We also performed experiments to investigate the impact of different percentages ($\varepsilon\%$) of incomplete views on the perfor-

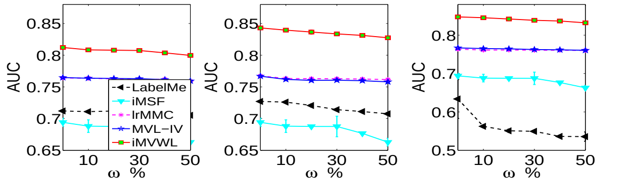

Figure 1: AUC values of compared methods on the Core15k dataset with different missing label proportions $\omega\%$. The dimensionality (k) of the shared subspace is set to $0.2d_{min}$(Left), $0.5d_{min}$(Middle), and $0.8d_{min}$(Right).

	lrMMC	MVL-IV	LabelMe	iMSF	iMVWL
Core15k	50.94	156.77	424.6	5531.75	52.47
Pascal07	184.04	299.06	928.69	2019.98	44.60
ESPGame	410.64	4449.91	2314.93	6887.37	1107.85
IAPRTC-12	405.07	6965.33	1900.96	49540.29	1268.97
Mirflicker	341.03	38404.08	3098.72	729.84	111.24
Total	1392.89	50281.76	8668.69	64778.68	2585.13

Table 4: Runtime comparison (in seconds).

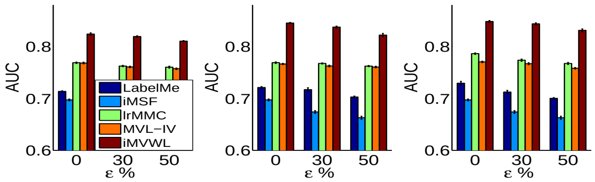

Figure 2: Results on the Core15k dataset with different incomplete view percentages $\varepsilon\%$. The dimensionality (k) of the shared subspace is set to $0.2d_{min}$ (Left), $0.5d_{min}$(Middle), and $0.8d_{min}$(Right).

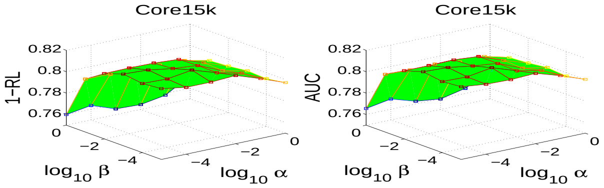

Figure 3: Parameter analysis w.r.t. α and β on Core15k.

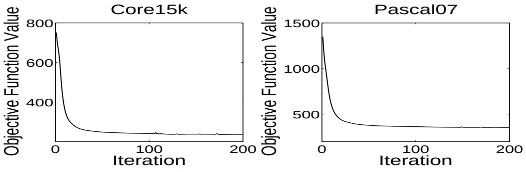

Figure 4: Convergence trend analysis.

mance of various methods. Similarly to previous experimental protocols, we set the dimensionality of the learned subspace equal to 20%, 50%, and 80% of d_{min}, and then increase the percentage of incomplete views $\varepsilon\%$ from 0% to 50% with a step-size of 10%. Due to space limitation, we report only the results for $\varepsilon\%$ equal to 0%, 30%, and 50% in Figure 2. The performance trend for $\varepsilon\%$ equal 10%, 20%, and 40% is similar to those reported in Figure 2.

It can be seen that the performance of all the methods decreases with the increasing of $\varepsilon\%$, and iMVWL gives the best performance in all the cases. In addition, as the dimensionality (k) of the shared subspace increases, the performance of all the methods shows an increasing trend, but iMVWL still performs consistently better than the competitors, under the different percentages of incomplete views.

5.5 Parameter Analysis

In this section, we test the sensitivity of iMVWL w.r.t. α and β. The tested range for α and β is $\{10^i|i = -5, \cdots, 0\}$. For brevity, we only report the 1-RL and AUC results on Core15k in Figure 3; similar results were obtained for the other datasets as well.

From the Figure, we can see that iMVWL achieves relatively stable and good performance when $\alpha \approx 10^{-2}$ and $\beta \approx 10^{-2}$. We also observe that when $\alpha = 10^{-5}$ or $\beta = 10^{-5}$, iMVWL has reduced 1-RL or AUC. This result confirms the contribution of weak-label information and local label correlation in improving the performance of iMVWL. When α or β are close to one, the AUC and 1-RL sharply decrease. This is because large values of α (or β) overweight the effect of weak-label information in subspace learning (or local label correlations), while underweighting the shared subspace $\mathbf{V}$, which encodes cross-view relationships.

5.6 Runtime And Convergence Analysis

We also study the runtime cost of the competing methods on the five datasets, and report the costs in Table 4. The experiments are conducted on CentOS 6.9 with Inter(R) Xeon E5-2678, 64GB RAM and MATLAB 2013a. We can see that iMVWL runs much faster than other comparing methods in most cases. The only exception is lrMMV on ESPGame and IAPRTC-12 datasets. This is mainly because lrMMC needs to estimate a target matrix of size $n \times k$ just once, while iMVWL has to estimate the label correlation matrix of size $c \times c$ in each iteration. As a result, when c is large, iMVWL costs more. These results corroborate the efficiency of the proposed method. The convergence trends on the other datasets are similar.

Figure 4 shows the convergence curve of iMVWL on Core15k and Pascal07 datasets. As we can see, on both datasets, iMVWL tends to converge after 60 iterations. The convergence trends on the other datasets are similar.

6 Conclusion

In this paper, we propose a novel model called iMVWL to learn from data with incomplete views and missing labels. iMVWL learns a discriminative shared subspace from incomplete views with weak labels. At the same time, it learns a robust weak-label classifier in the subspace and the local label structure. An alternative optimization solution is developed to optimize this model, which not only can avoid suboptimal problems, but also reinforces the reciprocal effects of the shared subspace and of the classifier, and further improves the performance. The experimental results show that iMVWL outperforms other competitive methods. How to further improve the efficiency of iMVWL is an interesting future pursue.

Acknowledgments

This work is supported by NSFC (61741217, 61402378 and 61732019), Open Research Project of Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2017A05).

References

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

[Bucak *et al.*, 2011] Serhat Selcuk Bucak, Rong Jin, and Anil K Jain. Multi-label learning with incomplete class assignments. In *CVPR*, pages 2801–2808, 2011.

[Cabral *et al.*, 2015] Ricardo Cabral, Fernando De la Torre, Joao Paulo Costeira, and Alexandre Bernardino. Matrix completion for weakly-supervised multi-label image classification. *TPAMI*, 37(1):121–135, 2015.

[Candès and Recht, 2009] Emmanuel Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

[Datta *et al.*, 2008] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):5, 2008.

[Dong *et al.*, 2018] Haochen Dong, Yufeng Li, and Zhihua Zhou. Learning from semi-supervised weak-label data. In *AAAI, in press*, 2018.

[Guillaumin *et al.*, 2010] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, pages 902–909, 2010.

[Huang and Zhou, 2012] Sheng Jun Huang and Zhi Hua Zhou. Multi-label learning by exploiting label correlations locally. In *AAAI*, pages 949–955, 2012.

[Kong *et al.*, 2014] Xiangnan Kong, Zhaoming Wu, Lijia Li, Ruofei Zhang, Hang Wu, and Wei Fan. Large-scale multi-label learning with incomplete label assignments. In *SDM*, pages 920–928, 2014.

[Liu *et al.*, 2015] Meng Liu, Yong Luo, Dacheng Tao, Chao Xu, and Yonggang Wen. Low-rank multi-view learning in matrix completion for multi-label image classification. In *AAAI*, pages 2778–2784, 2015.

[Luo *et al.*, 2015] Yong Luo, Tongliang Liu, Dacheng Tao, and Chao Xu. Multiview matrix completion for multilabel image classification. *TIP*, 24(8):2355–2368, 2015.

[Nie *et al.*, 2017] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, pages 2408–2414, 2017.

[Sun *et al.*, 2010] Yuyin Sun, Yin Zhang, and Zhihua Zhou. Multi-label learning with weak label. In *AAAI*, pages 1862–1868, 2010.

[Wang and Zhang, 2013] Yuxiong Wang and Yujin Zhang. Nonnegative matrix factorization: a comprehensive review. *TKDE*, 25(6):1336–1353, 2013.

[Wang and Zhou, 2010] Wei Wang and Zhihua Zhou. Multi-view active learning in the non-realizable case. In *NIPS*, pages 2388–2396, 2010.

[Wu *et al.*, 2015] Bao-Yuan Wu, Siwei Lyu, and Bernard Ghanem. Ml-mg: Multi-label learning with missing labels using a mixed graph. In *ICCV*, pages 4157–4165, 2015.

[Xu *et al.*, 2013] Miao Xu, Rong Jin, and Zhihua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *NIPS*, pages 2301–2309, 2013.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao X-u. Multi-view learning with incomplete views. *TIP*, 24(12):5812, 2015.

[Xu *et al.*, 2016] Chang Xu, Dacheng Tao, and Chao Xu. Robust extreme multi-label learning. In *KDD*, pages 1275–1284, 2016.

[Xu *et al.*, 2017] Jinglin Xu, Junwei Han, and Feiping Nie. Multi-view feature learning with discriminative regularization. In *IJCAI*, pages 3161–3167, 2017.

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

[Yu *et al.*, 2014] Hsiangfu Yu, Prateek Jain, Purushottam Kar, and Inderjit S Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, pages 593–601, 2014.

[Yuan *et al.*, 2012] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, and Jieping Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *KDD*, pages 1149–1157, 2012.

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

[Zhang *et al.*, 2013] Wei Zhang, Ke Zhang, Pan Gu, and Xiangyang Xue. Multi-view embedding learning for incompletely labeled data. In *IJCAI*, pages 1910–1916, 2013.

[Zhao and Guo, 2015] Feipeng Zhao and Yuhong Guo. Semi-supervised multi-label learning with incomplete labels. In *IJCAI*, pages 4062–4068, 2015.

[Zhao *et al.*, 2017] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[Žitnik and Zupan, 2015] Marinka Žitnik and Blaž Zupan. Data fusion by matrix factorization. *TPAMI*, 37(1):41–53, 2015.