Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Semi-Supervised Multi-Label Learning with Incomplete Labels

Feipeng Zhao and **Yuhong Guo**
Department of Computer and Information Sciences
Temple University, Philadelphia, PA 19122, USA
{feipeng.zhao, yuhong}@temple.edu

Abstract

The problem of incomplete labels is frequently encountered in many application domains where the training labels are obtained via crowd-sourcing. The label incompleteness significantly increases the difficulty of acquiring accurate multi-label prediction models. In this paper, we propose a novel semi-supervised multi-label method that integrates low-rank label matrix recovery into the manifold regularized vector-valued prediction framework to address multi-label learning with incomplete labels. The proposed method is formulated as a *convex* but non-smooth joint optimization problem over the latent label matrix and the prediction model parameters. We then develop a fast proximal gradient descent with continuation algorithm to solve it for a global optimal solution. The efficacy of the proposed approach is demonstrated on multiple multi-label datasets, comparing to related methods that handle incomplete labels.

1 Introduction

Multi-label classification is an essential problem in many application domains, including image annotation [Huiskes and Lew, 2008], video classification [Snoek *et al.*, 2006], document categorization [Srivastava and Zane-Ulman, 2002], and gene function prediction [Elisseeff and Weston, 2002]. Different from standard multi-class classification problems, a multi-label prediction function maps an input instance to a vector of interdependent multiple labels. Although multi-label classification can be transformed into a set of independent binary classification problems [Joachims, 1998], this mechanism fails to take the label interdependency into account. Many multi-label learning methods hence have been developed with a central theme of capturing label dependence; e.g., [Guo and Schuurmans, 2011; Elisseeff and Weston, 2002; Minh and Sindhwani, 2011]. A few works have also investigated exploiting unlabeled data to perform semi-supervised multi-label learning [Guo and Schuurmans, 2012; Luo *et al.*, 2013]. These methods nevertheless have all assumed training data with complete label assignments.

Complete training labels however are hard to collect in real world problems. In many applications training labels are obtained via crowd-sourcing, which typically leads to label incompleteness [Heymann *et al.*, 2008]. For example, in image or text tagging, human labelers tend to provide only a few keyword labels that describe the most obvious visual or semantic contents, while contents that are rare or ambiguous can be simply omitted. Moreover, the keyword labels provided by different labelers for even similar objects can be different, which can also lead to incomplete labels in the unified label vocabularies from multiple labelers. Label incompleteness can severely degrade the performance of the learned multi-label classification models, since it will build negative prediction patterns between the input instances and the missing labels and further propagate the mistakes into the prediction phase on the test data. This raises the need for learning multi-label classification models that handle incomplete labels. Although standard multi-label learning has received significant attention, multi-label learning with incomplete labels is still far from being well investigated. A number of previous methods only take label imputation as a preprocessing step [Lin *et al.*, 2013; Wu *et al.*, 2013; Zhu *et al.*, 2010; Liu *et al.*, 2010; Sun *et al.*, 2010], which recover labels that are not optimized for the target prediction models. A few others have attempted to take the process of imputing missing labels as part of the prediction model training [Qi *et al.*, 2011; Bucak *et al.*, 2011; Chen *et al.*, 2013], which however either involve complex local training or are limited to supervised learning over the sparsely labeled data.

In this paper, we propose a novel semi-supervised approach to perform multi-label learning with incomplete labels based on a manifold regularized vector-valued learning framework. The approach conducts automatic label imputation with a low-rank matrix recovery model that encodes the natural properties of label existence, while simultaneously performing vector-valued multi-label learning on the completed label matrix by exploiting label correlations. The unlabeled data is incorporated into the learning process by enforcing a vector-valued Laplacian manifold regularization. We formulate the learning process as a joint convex optimization problem, and develop a fast proximal gradient descent with continuation algorithm to solve the optimization problem for a global solution. Experiments are conducted on a variety types of multi-label datasets, and the proposed method demonstrates superior performance over a number of state-of-the-art multi-label methods with incomplete labels.

2 Related Work

Multi-label learning with incomplete labels is a problem that hinders information retrieval in many application domains, and has been addressed in a number of previous works. One class of methods attempts to complete the missing labels before classifier training as a pre-processing step; for example, by training only on the labels provided [Yu *et al.*, 2014], performing label completion based on visual similarity and label co-occurrence [Wu *et al.*, 2013; Zhu *et al.*, 2010], performing image-specific and tag-specific linear sparse reconstructions [Lin *et al.*, 2013], enriching an incomplete tagging with synonyms and hypernyms [Liu *et al.*, 2010], or employing label graph propagation [Sun *et al.*, 2010]. Unfortunately, these methods do not directly consider the consequences on multi-label prediction accuracy, which limits their effectiveness.

A few other works have attempted to consider prediction accuracy as part of the process of imputing missing labels. For example, Qi *et al.* [2011] developed a statistical generative model for mining partially annotated images. This approach however involves complex EM training for local optimal solutions, and is not discriminatively optimized for the target prediction task. Bucak *et al.* [2011] proposed a ranking based multi-label learning method for image annotation with incomplete labels. Their method exploits the group lasso regularizer to handle incompletely labeled training data when estimating the errors in ranking the assigned labels against unassigned labels. This method however focuses on the prediction model training without exploiting label correlations for label imputation. It does not exploit the unlabeled data either. Chen *et al.* [2013] proposed a fast image tagging method for multi-label learning from incompletely labeled data. It learns two classifiers to predict tag annotations: one attempts to reconstruct the (unknown) complete tag (label) set from the few observed tags; the other learns a linear mapping from image features to the reconstructed tag set. The two classifiers are combined within a joint convex loss function via co-regularization. Their method again is limited to supervised training on the sparse and partially labeled data.

Our proposed method in this paper shares similarity with the work [Chen *et al.*, 2013] in simultaneously learning classifiers from the input features and reconstructing the labels in the label space. But our method provides a principled framework for exploiting key properties of multi-label learning problems, including the natural label sparsity property of multi-label data, and the low-rank property of label matrix induced by the existence of label correlations. Moreover, our method exploits a vector-valued and kernelized multi-label classifier to capture label dependence in the learning process and exploit the unlabeled data via manifold regularization.

3 Approach

In this section, we present a semi-supervised method that simultaneously performs manifold regularized vector-valued multi-label learning and low-rank label matrix recovery on the given training data. By integrating label imputation and multi-label prediction in a mutually beneficial manner, this approach is expected to exploit unlabeled data and label interdependencies to improve multi-label prediction performance. In particular, we consider a set of ℓ labeled instances $\mathcal{D}_\ell = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\ell}$ and a large set of u unlabeled instances $\mathcal{D}_u = \{\mathbf{x}_i\}_{i=\ell+1}^{N=\ell+u}$ for multi-label learning, where $\mathbf{x}_i \in \mathbb{R}^d$ is the input feature vector for the i-th instance, and $\mathbf{y}_i \in \{+1, -1\}^n$ is the label vector for the i-th instance. With incomplete labels, the $+1$ value in the label vector $\mathbf{y}_i$ indicates the i-th instance is assigned into the corresponding class, but -1 value only indicates the unknown status of the corresponding label.

3.1 Vector-valued Multi-label Learning

Multi-label prediction functions map each instance $\mathbf{x}$ into a label vector $\mathbf{y} \in \mathcal{Y}$ in the vector space $\mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}^n$ and n is the number of classes. That is, given labeled training data, we aim to learn a vector-valued function $f : \mathcal{X} \mapsto \mathcal{Y}$. The formalism of vector-valued Reproducing Kernel Hilbert Spaces (RKHS) [Micchelli and Pontil, 2005] can be naturally adopted for multi-label function estimation. With regularized least squares loss function, the multi-label learning in the vector-valued RKHS [Minh and Sindhwani, 2011] can be formulated as

$$\arg\min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^{\ell} \|f(\mathbf{x}_i) - \mathbf{y}_i\|_{\mathcal{Y}}^2 + \gamma_A \|f\|_K^2 \tag{1}$$

for $\gamma_A > 0$, where K denotes an operator-valued positive definite kernel and $\mathcal{H}_K$ is a $\mathcal{Y}$-valued RKHS with reproducing kernel K. We expect the operator-valued kernel K can provide a mechanism to capture the dependencies among the multiple labels. Hence in this work, we consider the following matrix-valued kernel employed in [Minh and Sindhwani, 2011]

$$K(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) Q, \tag{2}$$

$$\text{for} \qquad Q = \gamma_O L_{out}^{\dagger} + (1 - \gamma_O) I_n, \tag{3}$$

where $k(\cdot, \cdot)$ is a scalar-valued kernel function, I_n is an $n \times n$ identity matrix, $L_{out}^{\dagger}$ is the pseudo-inverse of the $n \times n$ Laplacian matrix L_{out} over the output label graph. The output nearest-neighbor label graph can be constructed from the label matrix,

$$Y_\ell = [\mathbf{y}_1, \cdots, \mathbf{y}_\ell]^\top \in \{+1, -1\}^{\ell \times n}, \tag{4}$$

by taking each label, i.e., each column of Y_ℓ, as a data point. The Laplacian matrix L_{out} can then be computed on this label graph. The trade-off parameter γ_O satisfies $\gamma_O \in [0, 1]$, and Q is thus a symmetric and positive semi-definite matrix. The solution of (1) takes the form of $f^* = \sum_{i=1}^{\ell} K_{\mathbf{x}_i} \mathbf{a}_i$ with $\{\mathbf{a}_i, i = 1, \cdots, \ell\} \subseteq \mathcal{Y}$ [Micchelli and Pontil, 2005].

3.2 Laplacian Manifold Regularization

To exploit the geometric structural information from the large number of unlabeled instances for prediction function learning, graph Laplacian based manifold regularization [Belkin *et al.*, 2005; 2006] can be incorporated into the learning process. Let $W \in \{0, 1\}^{N \times N}$ be the adjacency matrix of the k-nearest-neighbor graph constructed from the input data, such that $W_{ij} = W_{ji} = 1$ if $i \in \mathcal{N}_k(j)$ or $j \in \mathcal{N}_k(i)$, and

$W_{ij} = W_{ji} = 0$ otherwise, where $\mathcal{N}_k(i)$ denotes the index set of the k nearest neighbors of the i-th instance. The Laplacian L of the input graph can then be obtained as $L = D - W$ with the diagonal matrix $D_{ii} = \sum_{j=1}^{N} W_{ij}$.

Let $\mathcal{Y}^N$ denote the N-direct product of $\mathcal{Y}$. For $f \in \mathcal{H}_K$, the function prediction over all the data instances is $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)) \in \mathcal{Y}^N$. Then based on the Laplacian matrix, the semi-supervised multi-label learning with vector-valued manifold regularization can be formulated as

$$\arg\min_{f \in \mathcal{H}_K} \frac{1}{\ell} \sum_{i=1}^{\ell} \|f(\mathbf{x}_i) - \mathbf{y}_i\|_{\mathcal{Y}}^2 + \gamma_A \|f\|_K^2 + \gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{Y}^N} \tag{5}$$

where $\gamma_A, \gamma_I > 0$. The matrix M is a symmetric positive operator that satisfies $\langle \mathbf{y}, M\mathbf{y} \rangle_{\mathcal{Y}^N} \geq 0$. Here it will be set as $M = L \otimes I_n$, where $\otimes$ denotes the Kronecker matrix product, I_n is the $n \times n$ identity matrix, L is the $N \times N$ graph Laplacian matrix described above.

Let G be the $N \times N$ symmetric and positive semi-definite Gram matrix produced by the scalar kernel $k(\cdot, \cdot)$ over all the N data points. Following [Minh and Sindhwani, 2011], we have the following proposition.

Proposition 1. *The minimization problem (5) has a unique solution*

$$f = \sum_{i=1}^{N} K_{\mathbf{x}_i} \mathbf{a}_i, \tag{6}$$

for vectors $\mathbf{a}_i \in \mathcal{Y}$ that satisfy the following linear equation

$$(J_\ell^N G + \ell \gamma_I L G) A Q + \ell \gamma_A A = Y, \tag{7}$$

where J_ℓ^N is an $N \times N$ diagonal matrix, whose first ℓ diagonal entries have value 1, and the rest have value 0. The matrices A and Y are defined as

$$A = [\mathbf{a}_1, \ldots, \mathbf{a}_N]^\top, \quad Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^\top, \tag{8}$$

where $\mathbf{y}_i = 0^n$ for $i = \ell + 1, ..., N$.

Proof: The unique solution (6) can be obtained through Representer Theorem [Micchelli and Pontil, 2005; Minh and Sindhwani, 2011]. We can then have

$$f(\mathbf{x}) = \sum_{i=1}^{N} K(\mathbf{x}, \mathbf{x}_i) \mathbf{a}_i. \tag{9}$$

By plugging the instantiations of (9) over all data points back into the objective function (5), and then setting its derivative regarding A to zeros, one can get (7). □

3.3 Semi-supervised Learning with Incomplete Labels

When the label matrix Y_ℓ in (4) is incomplete, that is, $Y_{ij} = -1$ may either indicate that the i-th instance does not have the j-th label or the j-th label is omitted for the i-th instance, the performance of semi-supervised multi-label learning (5) based on Y_ℓ will inevitably degrade. To address this problem, we propose to recover the underlying true label matrix Z from Y_ℓ by augmenting Y_ℓ via $Z = Y_\ell + E$, where E is an $\ell \times n$ nonnegative augmenting matrix that contains the labels missed from Y_ℓ, and perform vector-valued multi-label learning over the latent true label matrix Z. Since labels generally present strong correlation patterns and dependence relationships, the latent true label matrix Z should naturally be low-rank. Moreover, since each instance is normally assigned only a few positive labels, Z should also be sparse. Hence we formulate the learning process with incomplete labels as the following joint optimization problem

$$\begin{aligned} \min_{f \in \mathcal{H}_K, Z, E} \quad & \frac{1}{\ell} \sum_{i=1}^{\ell} \|f(\mathbf{x}_i) - \mathbf{z}_i\|_{\mathcal{Y}}^2 + \gamma_A \|f\|_K^2 \\ & + \gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{Y}^N} + \mu \left(\|Z\|_* + \lambda \|E\|_1 \right) \\ \text{subject to} \quad & Z = Y_\ell + E; \; E \geq 0 \end{aligned} \tag{10}$$

where $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_\ell]^\top$, μ and λ are trade-off parameters, other parameters are same as in (5); $\|\cdot\|_*$ and $\|\cdot\|_1$ denote the trace norm and L1-norm over matrix respectively. Trace norm is a convex envelope of the rank function. By adding $\|Z\|_*$ into the minimization objective function, we encode the low-rank property of the underlying true label matrix Z to capture label correlations. Since Y_ℓ is given, the L1-norm regularization over E will encode the sparsity property of Z. This joint minimization problem simultaneously performs label matrix recovery within a low-rank sparse matrix recovery framework and conducts semi-supervised multi-label learning with the manifold regularized vector-valued model.

Note the minimization over $f \in \mathcal{H}_K$ in (10) will again lead to the solution in (6) or (9). By plugging the instantiations of (9) over the training data back into the objective function in (10), the optimization problem above can be transformed into the following concrete formulation:

$$\begin{aligned} \min_{A, Z, E} \quad & \frac{1}{\ell} \|S_\ell^N G A Q - Z\|_F^2 + \gamma_A \mathrm{tr}(G A Q A^\top) \\ & + \gamma_I \mathrm{tr}(Q^\top A^\top G^\top L G A Q) + \mu \left(\|Z\|_* + \lambda \|E\|_1 \right) \\ \text{subject to} \quad & Z = Y_\ell + E; \; E \geq 0 \end{aligned} \tag{11}$$

where $S_\ell^N = [I_\ell, 0_{\ell,u}]$ is a $\ell \times N$ selection matrix, which has a left identity matrix of size ℓ and zeros in all other entries; $A \in \mathbb{R}^{N \times n}$ is the parameter matrix defined in (8); $\|\cdot\|_F$ denotes the Frobenius norm. Note after solving this optimization problem for A, a new test instance $\mathbf{x}$ can be classified by using the kernelized prediction function in (9).

4 Learning Algorithm

The joint optimization problem in (11) is convex, but non-smooth, due to the existence of the trace norm and L1-norm operators. We develop a fast proximal gradient with continuation algorithm to solve this minimization problem.

For the convenience of optimization, we first consider a relaxed intermediate formulation of (11) by replacing the equality constraint with a penalty regularization term

$$\begin{aligned} \min_{A, Z, E \geq 0} \quad & \frac{1}{\ell} \|S_\ell^N G A Q - Z\|_F^2 + \gamma_A \mathrm{tr}(G A Q A^\top) \\ & + \gamma_I \mathrm{tr}(Q^\top A^\top G^\top L G A Q) \\ & + \mu \left(\|Z\|_* + \lambda \|E\|_1 \right) + \rho \|E + Y_\ell - Z\|_F^2 \end{aligned} \tag{12}$$

where ρ is a parameter that controls the enforcement degree of the equality constraint. With a continuation optimization algorithm below, we can gradually increase ρ to enforce the equality constraint in (11) and solve (12) as an intermediate step for solving the optimization problem (11).

For simplicity of presentation, we use Θ to denote all the parameters such that $\Theta = [A; Z; E]$. Then the objective function of (12) can be expressed as the sum of two functions, such as

$$\min_{\Theta: E \geq 0} \quad F(\Theta) = h(\Theta) + g(\Theta) \tag{13}$$

where

$$h(\Theta) = \frac{1}{\ell}\|S_\ell^N GAQ - Z\|_F^2 + \gamma_A \text{tr}(GAQA^\top) \tag{14}$$
$$+ \gamma_I \text{tr}(Q^\top A^\top G^\top LGAQ) + \rho\|E + Y_\ell - Z\|_F^2$$
$$g(\Theta) = \mu\left(\|Z\|_* + \lambda\|E\|_1\right) \tag{15}$$

The first function $h(\Theta)$ is a convex and smooth function and the second function $g(\Theta)$ is convex but non-smooth.

To develop a proximal gradient optimization method, we consider the second order approximation of the objective function $F(\Theta)$. For $\nu > 0$, at a given point $\Theta^{(t)}$, we define

$$\mathcal{M}_\nu(\Theta, \Theta^{(t)}) = h(\Theta^{(t)}) + \langle \Theta - \Theta^{(t)}, \nabla h(\Theta^{(t)})\rangle + \frac{\nu}{2}\|\Theta - \Theta^{(t)}\|_F^2 + g(\Theta) \tag{16}$$

where $\nabla h(\Theta^{(t)})$ is the gradient function of $h(\cdot)$ with respect to Θ at the point $\Theta^{(t)}$. Let $\ell_c(h)$ denote the Lipschitz constant of $\nabla h(\Theta)$. Then for $\nu \geq \ell_c(h)$, this approximate function becomes an upper bound such that $\mathcal{M}_\nu(\Theta, \Theta^{(t)}) \geq F(\Theta)$ [Beck and Teboulle, 2009]. By minimizing $\mathcal{M}_\nu(\Theta, \Theta^{(t)})$, we can reach the next point $\Theta^{(t+1)}$

$$p_\nu(\Theta^{(t)}) = \arg\min_{\Theta: E\geq 0} \mathcal{M}_\nu(\Theta, \Theta^{(t)}) = \arg\min_{\Theta: E\geq 0} \left\{ g(\Theta) + \frac{\nu}{2}\left\|\Theta - \widehat{\Theta}^{(t)}\right\|_F^2 \right\}, \tag{17}$$

where $\widehat{\Theta}^{(t)} = \Theta^{(t)} - \frac{1}{\nu}\nabla h(\Theta^{(t)})$. It is equivalent to the following sub-problems

$$p_\nu(A^{(t)}) = \arg\min_A \ \frac{\nu}{2}\left\|A - \widehat{A}^{(t)}\right\|_F^2, \tag{18}$$

$$p_\nu(Z^{(t)}) = \arg\min_Z \ \mu\|Z\|_* + \frac{\nu}{2}\left\|Z - \widehat{Z}^{(t)}\right\|_F^2, \tag{19}$$

$$p_\nu(E^{(t)}) = \arg\min_{E\geq 0} \ \mu\lambda\|E\|_1 + \frac{\nu}{2}\left\|E - \widehat{E}^{(t)}\right\|_F^2, \tag{20}$$

which have closed-form solutions. The solution for (18) is simply $p_\nu(A^{(t)}) = \widehat{A}^{(t)}$. For the convex minimization with trace norm in (19), its solution can be computed via singular value decomposition $\widehat{Z}^{(t)} = U\Sigma V^\top$ such that

$$\Sigma_{\mu/\nu} = \max\left(0, \Sigma - \frac{\mu}{\nu}\right), \quad p_\nu(Z^{(t)}) = U\Sigma_{\mu/\nu}V^\top. \tag{21}$$

Algorithm 1 Fast Proximal Gradient with Continuation

Input: $X \in \mathbb{R}^{N\times d}; Y_\ell \in \{-1, +1\}^{\ell\times n}$;
$\gamma_A, \gamma_I, \gamma_O, \mu > 0;\ \lambda = \sqrt{\max(\ell, n)}$;
$0 < \rho_1 < \rho_2 < \ldots < \rho_k;\ \eta > 1$.
Initialize $Z^{(0)} = Y_\ell,\ E^{(0)} = 0^{\ell\times n}, A^{(0)} = 0^{N\times n}$,
$\Theta^{(0)} = [A^{(0)}; Z^{(0)}; E^{(0)}],\ \nu = 1$.
Compute G, L, Q, S_ℓ^N.
for $\rho = [\rho_1, \rho_2, \ldots, \rho_k]$ **do**
  Initialize $\Omega^{(1)} = \Theta^{(0)},\ t = 1,\ q_1 = 1$.
  for iter =1:maxiters **do**
    1.**while** $F(p_\nu(\Omega^{(t)})) > \mathcal{M}_\nu(p_\nu(\Omega^{(t)}), \Omega^{(t)})$ **do** $\{\nu = \eta\nu\}$
    end while
    2. $\Theta^{(t)} = p_\nu(\Omega^{(t)}),\ \ q_{t+1} = \frac{1+\sqrt{1+4q_t^2}}{2}$,
    $\Omega^{(t+1)} = \Theta^{(t)} + \frac{q_t - 1}{q_{t+1}}(\Theta^{(t)} - \Theta^{(t-1)})$
    3. Set $t = t + 1$.
  end for
  Set $\Theta^{(0)} = \Theta^{(t-1)}$.
end for

The closed-form solution of the minimization problem (20) with sparsity inducing norm and nonnegativity constraint can be computed by applying soft-thresholding operator

$$p_\nu(E^{(t)}) = \left(abs(\widehat{E}^{(t)}) - \frac{\mu\lambda}{\nu}\right)_+ \tag{22}$$

where $abs(\cdot)$ is the absolute value function, and the operator $(\cdot)_+ = \max(\cdot, 0)$.

By setting the next point as $\Theta^{(t+1)} = p_\nu(\Theta^{(t)})$, and conducting iterative proximal gradient updates according to (17), a sequence of points, $\{\Theta^{(0)}, \Theta^{(1)}, \ldots\}$, can be obtained to minimize the objective function $F(\Theta)$. To speed up the convergence of the optimization procedure, we further adopt the fast proximal gradient update scheme from [Beck and Teboulle, 2009]. Moreover, to enforce the equality constraint $Z = Y_\ell + E$, the ρ parameter in (12) needs to be set as a large value. To improve the convergence speed, we employ a *batch-mode continuation strategy* which starts with a relatively smaller ρ and then gradually increases the ρ value after a certain number of iterations. The overall algorithm is given in Algorithm 1. In this algorithm, we set λ parameter as $\lambda = 1/\sqrt{\max(\ell, n)}$, where (ℓ, n) is the dimension size of Y_ℓ, according to [Wright *et al.*, 2009], to ensure a robust label matrix recovery.

5 Experiments

We report our experimental setting and results in this section.

5.1 Experimental Setting

Datasets. We conducted experiments with six multi-label datasets: *corel5k, msrc, mirflickr, mediamill, tmc2007* and *yeast*. *Msrc* is a Microsoft Research labeled image dataset with 591 images in 23 object classes. Each image is represented as a vector with 960 GIST [Oliva and Torralba, 2001]

Table 2: The average comparison results and their standard deviations in terms of Micro F1 Score.

Dataset	Proposed	MLR-GL	FastTag	LEML	BR	BR-C
corel5k	**0.260 ± 0.005**	0.212 ± 0.007	0.218 ± 0.004	0.187 ± 0.002	0.168 ± 0.005	0.214 ± 0.003
msrc	**0.571 ± 0.016**	0.477 ± 0.010	0.485 ± 0.010	0.398 ± 0.010	0.438 ± 0.023	0.537 ± 0.014
mirflickr	**0.431 ± 0.005**	0.375 ± 0.011	0.344 ± 0.009	0.318 ± 0.001	0.286 ± 0.003	0.326 ± 0.001
mediamill	0.528 ± 0.003	0.446 ± 0.011	0.545 ± 0.011	**0.556 ± 0.001**	0.423 ± 0.004	0.519 ± 0.001
tmc2007	**0.612 ± 0.004**	0.443 ± 0.016	0.452 ± 0.004	0.388 ± 0.006	0.440 ± 0.005	0.544 ± 0.004
yeast	**0.641 ± 0.003**	0.608 ± 0.011	0.599 ± 0.007	0.608 ± 0.006	0.443 ± 0.005	0.559 ± 0.003

Table 3: The average comparison results and their standard deviations in terms of Macro F1 Score.

Dataset	Proposed	MLR-GL	FastTag	LEML	BR	BR-C
corel5k	**0.182 ± 0.006**	0.160 ± 0.003	0.174 ± 0.002	0.142 ± 0.002	0.107 ± 0.005	0.153 ± 0.005
msrc	0.446 ± 0.026	0.425 ± 0.014	0.398 ± 0.018	0.327 ± 0.010	0.356 ± 0.002	**0.468 ± 0.016**
mirflickr	**0.243 ± 0.002**	0.236 ± 0.003	0.171 ± 0.007	0.153 ± 0.001	0.164 ± 0.001	0.192 ± 0.001
mediamill	0.252 ± 0.005	0.257 ± 0.002	0.206 ± 0.011	0.225 ± 0.004	0.215 ± 0.001	**0.267 ± 0.001**
tmc2007	**0.403 ± 0.007**	0.301 ± 0.012	0.323 ± 0.005	0.265 ± 0.009	0.290 ± 0.009	0.399 ± 0.004
yeast	0.396 ± 0.004	**0.427 ± 0.004**	0.390 ± 0.010	0.369 ± 0.002	0.315 ± 0.003	0.391 ± 0.003

Table 1: Statistical information of the datasets

Dataset	# instances	# features	# labels	label card.
corel5k	4,609	499	30	2.07
msrc	591	960	23	2.51
mirflickr	5,000	512	38	4.77
mediamill	42,023	120	30	4.21
tmc2007	28,596	500	22	2.16
yeast	2,417	103	14	4.24

features. *Corel5k* [Duygulu *et al.*, 2002] is a scene classification dataset. Some labels in this dataset rarely appear, and we selected its top 30 labels to use. This leads to a subset with 4609 images, each of which is expressed as a vector with 499 features. *Mirflickr* [Huiskes and Lew, 2008] is a large collection of images. We randomly sampled a subset of 5000 images to use, which has 38 labels, and each image is represented as a vector of 512 GIST features. *Mediamill* [Snoek *et al.*, 2006] is a large video dataset. We used the top 30 labels, which leads to 42,023 instances. Each of its instances is expressed with 120 low-level features. *Tmc2007* dataset [Srivastava and Zane-Ulman, 2002] is a large text dataset with 28,596 instances and 22 labels in total. We used its short version with 500 features. *Yeast* dataset [Elisseeff and Weston, 2002] is a gene function classification dataset with 2417 genes and 14 classes. Each gene is expressed with 103 microarray expression features. The statistical information of the six datasets are summarized in Table 1, where label cardinality (label card.) denotes the average number of labels assigned to each instance.

Approaches. In the experiments, we compared our proposed approach to the following methods: (1) the multi-label ranking method with group lasso regularizer (*MLR-GL*) [Bucak *et al.*, 2011]; (2) the fast tagging method (*FastTag*) [Chen *et al.*, 2013]; (3) the large scale empirical risk minimization method with missing labels (*LEML*) [Yu *et al.*, 2014]; (4) the baseline method, binary relevance (BR); and (5) binary relevance with complete labels (BR-C). The first two methods are state-of-the-art methods for multi-label learning with incomplete labels. *LEML* was used to handle missing labels directly in [Yu *et al.*, 2014]. The *BR* is a baseline method, where we train a binary SVM classifier for each label on the observed incompletely labeled data. The *BR-C* is same as *BR* except that we use complete labels. We used libsvm [Chang and Lin, 2011] as implementation for *BR* and *BR-C*. For the proposed approach, we used RBF kernels as the input kernel $k(\cdot,\cdot)$, and set γ_O=0.5 to compute the matrix-valued kernel. We used 5 nearest neighbor graph to construct the Laplacian matrix on the input data, and set the number of nearest neighbors approximately as 30% of the label dimension size to construct the Laplacian matrix on the output label matrix.

Experiment Setup. For each dataset, we simulate the incomplete label condition by randomly dropping 30% of the observed labels on the labeled training data. For *msrc*, we randomly selected 80% of the data for training (30% labeled and 50% unlabeled) and used the rest 20% for testing. For all the other five datasets, we randomly selected 500 instances as labeled data and 1000 instances as unlabeled data, and used the remaining data for testing. We compared all methods using the same data setting, and repeated each experiment five times with different random partitions.

For all the methods, we conducted parameter selection by performing 5-fold cross-validation on the training set. We further dropped 20% labels on the labeled training data of each cross-validation to simulate the missing label situation. For our proposed approach, we selected the trade-off parameters γ_A and γ_I from $\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$, and selected μ from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Parameter selections via cross validation are also conducted for the other comparison methods accordingly.

5.2 Experiment Results

We measure the classification results in terms of two standard multi-label evaluation criteria: micro-F1 measure and macro-F1 measure, which take both precision and recall into

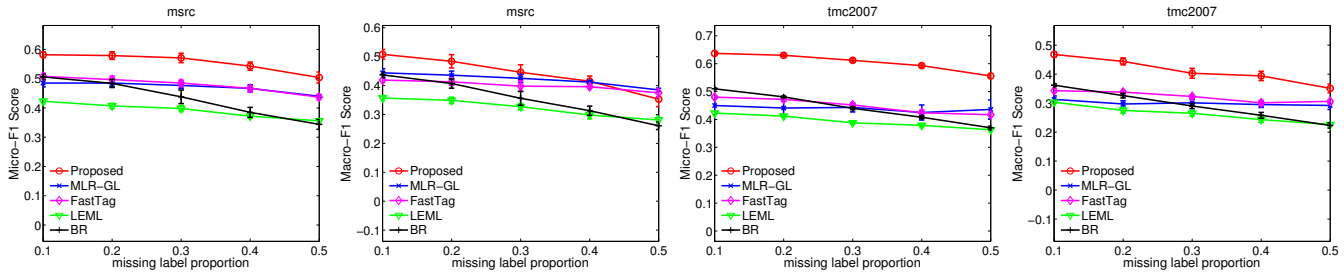

Figure 1: Performance vs. missing label proportion on *msrc* and *tmc2007*.

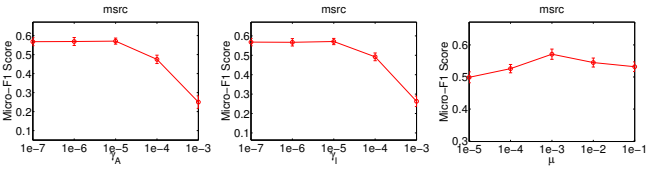

Figure 2: Parameter sensitivity analysis on *msrc*.

accounts. The average classification results and their standard deviations in terms of these two criteria are reported in Table 2 and Table 3 respectively.

From Table 2, we can see the three methods, *Proposed, MLR-GL* and *FastTag*, which handle incomplete labels, consistently outperform the baseline *BR* method across all the datasets. The *LEML* also outperforms *BR* with large margins on four datasets. This suggests that it is important to address the missing label problem when learning multi-label classifiers on data with incomplete labels. Among the advanced learning methods, *MLG-GL, FastTag* and *LEML* demonstrated strengths over each other on different datasets. Our proposed approach on the other hand produced the best results and outperform all the other methods with large margins over five out of the total six datasets. Even on the remaining dataset, *mediamill*, our result is still very good. Moreover, the proposed approach also consistently outperforms *BR-C*, which used the complete labels for training. Similar results are presented in Table 3 as well, where the values are given in terms of the macro-F1 score. Among all the methods that handle incomplete labels, the proposed approach produces the best results on four out of the six datasets, while producing the second best results on the other two datasets. The proposed approach also outperforms *BR-C* on four datasets.

Overall, these results demonstrate the benefit of handling incomplete labels in the learning process. It also clearly shows the advantage of our proposed semi-supervised learning approach which is able to exploit unlabeled data and simultaneously perform label imputation and multi-label learning by exploiting the label correlations.

Impact of missing labels fractions. The results above were conducted with a given label missing fraction. There is one remaining question: How do the comparison methods, especially our proposed approach, perform with different fractions of missing labels? To answer this question, we conducted another set of experiments on the *msrc* and *tmc2007* datasets with a number of different fraction values of missing labels on the labeled training data: $\zeta \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$. For each given missing label fraction value ζ, we randomly dropped ζ portion of the original observed labels from the labeled training instances, and conducted experiments using the same setting as above. The average and standard deviation results in terms of the two F1 measures are reported in Figure 1.

From the results on both *msrc* and *tmc2007*, we can see that the proposed approach greatly outperforms all the other methods in terms of the micro-F1 measure across the whole range of different missing label fractions. In terms of macro-F1, the proposed approach outperforms the other four methods across $\zeta \in \{10\%, 20\%, 30\%, 40\%\}$ on the *msrc* dataset, and outperforms all the other methods on the *tmc2007* dataset. These results again verified the efficacy of the proposed approach on addressing multi-label learning with incomplete labels.

Parameter sensitivity analysis. We have also conducted parameter sensitivity analysis for the proposed approach on the *msrc* dataset over the trade-off parameters, γ_A, γ_I and μ. The sensitivity analysis results in terms of micro-F1 regarding each parameter are reported in Figure 2, given the other two parameters fixed ($\gamma_A = 10^{-5}, \gamma_I = 10^{-5}, \mu = 10^{-3}$). We can see, within the considered range of values, the approach is quite robust in γ_A and γ_I given their values are no bigger than 10^{-5}. The μ parameter that controls the level of low-rank and sparse properties of the label matrix however does not have a very robust value zone and hence needs a good parameter selection procedure such as cross-validation.

6 Conclusion

In this paper, we proposed a semi-supervised method to address multi-label learning with incomplete labels, which integrates two functions, label imputation and multi-label prediction, in a mutually beneficial manner. Specifically, the proposed method conducts automatic label imputation within a low-rank and sparse matrix recovery framework, while simultaneously performing vector-valued multi-label learning and exploiting unlabeled data with vector-valued manifold regularization. With a least squares loss function, we formulated this problem as a joint convex optimization problem over the latent label matrix and the classification model parameters. We then developed a fast proximal gradient descent with continuation algorithm to solve it. We conducted experiments on a variety types of multi-label datasets, and compared our pro-

posed approach with a few related methods. Our experimental results suggest the proposed approach can effectively improve multi-label classification performance on datasets with incomplete labels over the existing state-of-the-art methods.

Acknowledgments

This research was supported in part by NSF grant IIS-1422127.

References

[Beck and Teboulle, 2009] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. on Imaging Sciences*, 2(1):183–202, 2009.

[Belkin *et al.*, 2005] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In *Proc. of AISTATS*, 2005.

[Belkin *et al.*, 2006] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[Bucak *et al.*, 2011] S. Bucak, J. Rong, and A. Jain. Multi-label learning with incomplete class assignments. In *Proc. of CVPR*, 2011.

[Chang and Lin, 2011] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transact. on Intelligent Systems and Technology*, 2, 2011.

[Chen *et al.*, 2013] M. Chen, A. Zhang, and K. Weinberger. Fast image tagging. In *Proc. of ICML*, 2013.

[Duygulu *et al.*, 2002] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV*, 2002.

[Elisseeff and Weston, 2002] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Proc. of NIPS*, 2002.

[Guo and Schuurmans, 2011] Y. Guo and D. Schuurmans. Adaptive large margin training for multilabel classification. In *Proc. of AAAI*, 2011.

[Guo and Schuurmans, 2012] Y. Guo and D. Schuurmans. Semi-supervised multi-label classification: A simultaneous large-margin, subspace learning approach. In *Proc. of ECML-PKDD*, 2012.

[Heymann *et al.*, 2008] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve Web search? In *Proc. of the Intern. Conf. on Web Search and Web Data Mining*, 2008.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The MIR Flickr retrieval evaluation. In *Proc. of the ACM Inter. Conf. on Multimedia Info. Retrieval*, 2008.

[Joachims, 1998] T. Joachims. Text categorization with support vector machines: learn with many relevant features. In *Proc. of ECML*, 1998.

[Lin *et al.*, 2013] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *Proc. of CVPR*, 2013.

[Liu *et al.*, 2010] D. Liu, X. Hua, M. Wang, and H. Zhang. Image retagging. In *Proc. of the ACM Inter. Conf. on Multimedia*, 2010.

[Luo *et al.*, 2013] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transact. on Image Process.*, 22(2):523–536, 2013.

[Micchelli and Pontil, 2005] C. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Comput.*, 17(1):177–204, 2005.

[Minh and Sindhwani, 2011] H. Minh and V. Sindhwani. Vector-valued manifold regularization. In *Proc. of ICML*, 2011.

[Oliva and Torralba, 2001] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[Qi *et al.*, 2011] Z. Qi, M. Yang, Z. Zhang, and Z. Zhang. Mining partially annotated images. In *Proc. of KDD*, 2011.

[Snoek *et al.*, 2006] C. Snoek, M. Worring, J. van Gemert, J. Geusebroek, and A. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. of the ACM Inter. Conf. on Multimedia*, 2006.

[Srivastava and Zane-Ulman, 2002] A.N. Srivastava and B. Zane-Ulman. Discovering recurring anomalies in text reports regarding complex space systems. In *Aerospace Conference*, 2002.

[Sun *et al.*, 2010] Y. Sun, Y. Zhang, and Z. Zhou. Multi-label learning with weak label. In *Proc. of AAAI*, 2010.

[Wright *et al.*, 2009] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. In *Proc. of NIPS*, 2009.

[Wu *et al.*, 2013] L. Wu, R. Jin, and A. Jain. Tag completion for image retrieval. *IEEE TPAMI*, 35(3):716–727, 2013.

[Yu *et al.*, 2014] H. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *Proc. of ICML*, 2014.

[Zhu *et al.*, 2010] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proc. of Inter. Conf. on Multimedia*, 2010.