The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

# Partial Multi-Label Learning by Low-Rank and Sparse Decomposition

**Lijuan Sun, Songhe Feng,[*] Tao Wang, Congyan Lang, Yi Jin**
School of Computer and Information Technology, Beijing Jiaotong University
{17112082, shfeng, twang, cylang, yjin}@bjtu.edu.cn

## Abstract

Multi-Label Learning (MLL) aims to learn from the training data where each example is represented by a single instance while associated with a set of candidate labels. Most existing MLL methods are typically designed to handle the problem of missing labels. However, in many real-world scenarios, the labeling information for multi-label data is always redundant , which can not be solved by classical MLL methods, thus a novel *Partial Multi-label Learning* (PML) framework is proposed to cope with such problem, i.e. removing the the noisy labels from the multi-label sets. In this paper, in order to further improve the denoising capability of PML framework, we utilize the low-rank and sparse decomposition scheme and propose a novel *Partial Multi-label Learning by Low-Rank and Sparse decomposition* (PML-LRS) approach. Specifically, we first reformulate the observed label set into a label matrix, and then decompose it into a ground-truth label matrix and an irrelevant label matrix, where the former is constrained to be low rank and the latter is assumed to be sparse. Next, we utilize the feature mapping matrix to explore the label correlations and meanwhile constrain the feature mapping matrix to be low rank to prevent the proposed method from being overfitting. Finally, we obtain the ground-truth labels via minimizing the label loss, where the Augmented Lagrange Multiplier (ALM) algorithm is incorporated to solve the optimization problem. Enormous experimental results demonstrate that PML-LRS can achieve superior or competitive performance against other state-of-the-art methods.

## Introduction

As a popular machine learning framework, Multi-Label Learning (MLL) aims to learn a robust classification model from the training data, where each instance is associated with a set of labels instead of a single label (Zhang and Zhou 2014). In recent years, such framework has been widely used in many real-world scenarios, such as image annotation (Sanden and Zhang 2011), web mining (Tang, Rajan, and Narayanan 2009), information retrieval (Gopal and Yang 2010), etc.

Typically multi-label learning methods usually require complete labeling information for training examples (Zhang and Zhou 2014), i.e., each training instance has been precisely annotated with all of its relevant labels. However, in many real-world applications, precise labeling is too scarce to obtain, which makes it infeasible to learn a robust multi-label classifier. Thus, many state-of-the-art studies are designed to handle the problem that the label matrix has missing entries, including treating missing labels as negative labels directly (Wu et al. 2015) (Sun, Zhang, and Zhou 2010) (Bucak, Jin, and Jain 2011), employing matrix completion technique to fill in missing labels (Goldberg et al. 2010) (Cabral et al. 2011), etc. Recently, as the annotation crowdsourcing increasingly becomes popular, redundant labeling information gradually appears in these multi-label data, i.e. annotators may roughly assign each instance with a set of candidate labels, which include both related labels and unrelated labels. For example, as we observed in Figure 1, the image is partially labeled by noisy annotators. Among the candidate labels, *mountain*, *lake*, *grass*, *sky*, *house*, and *tree* are ground-truth labels while *clouds*, *boat*, and *people* are irrelevant labels.

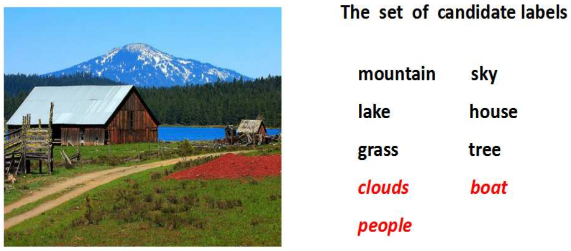


Figure 1: An example of partial multi-label learning.

To overcome the above problem, (Xie and Huang 2018) proposes a novel framework called *Partial Multi-Label Learning* (PML) to learn from the multi-label data with redundant labeling information, where they utilize the label confidence to measure the probability of being the ground-truth label for each candidate label, and obtain the ground-truth labels according to label ranking. However, this approach suffers from some shortcomings, i.e. it only simply utilizes the prior knowledge to obtain the label correlation or directly uses redundant labeling information to ob-

[*] corresponding author

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tain the feature prototype for acquiring the label confidence values, which may reduce the effectiveness of the learning model. Therefore, an intuitive strategy to cope with the PML problem is disambiguation, i.e. how to identify the ground-truth labels from the candidate labels. However, once the irrelevant labels are excessively redundant, such identification work is rather challenging or even impossible. Fortunately, in real-word scenarios, the irrelevant labels are usually sparse among the observed labels, which makes the disambiguation work become possible and easy to implement.

Based on the above consideration, in this paper, we propose **Partial Multi-label Learning by Low-Rank and Sparse decomposition** (PML-LRS) method, which enables simultaneously capturing the ground-truth label matrix from the observed label matrix and learning the prediction model via low-rank and sparse decomposition scheme. Specifically, we firstly introduce $\ell_1$-norm regularization to constrain the redundant label matrix by assuming that the irrelevant labels are sparse. Secondly, a trace norm regularization is introduced to capture the dependence among ground-truth labels. Thirdly, by making full use of the label correlations, the feature mapping matrix is constrained via trace norm regularization. Finally, the desired partial multi-label prediction model is learned by adopting Augmented Lagrange Multiplier (ALM) method. Compared with previous PML algorithm, our method can remove the irrelevant labels and avoid the negative effect of noisy labels, which makes our method become more robust and applicable in real applications. Extensive experimental results on real-world data sets validate the effectiveness of our model against other competitive algorithms.

## Related Work

### Multi-Label Learning with Label Correlation

A significant amount of literatures on multi-label learning has been proposed in recent years, which can be roughly categorized into three groups based on the degree of label correlations (Zhang and Zhou 2014). For the first-order strategy, many approaches tackle the multi-label learning problem in a label-by-label style but they ignore the label correlations (Zhang and Zhou 2007). For the second-order strategy, most algorithms tackle multi-label learning problem by considering pairwise label correlations (Fürnkranz et al. 2008). For the high-order strategy, some methods tackle multi-label learning problem by considering high-order correlations among label subsets or all the classes (Ji et al. 2010) (Tsoumakas, Katakis, and Vlahavas 2011).

### Weakly Supervised Multi-label Learning

In the literature of multi-label learning, most state-of-the-art methods are designed to handle missing labels, which can be roughly divided into the following four categorizes: (1) The first way is to treat the missing labels as negative labels and bring the label bias into the objective function (Chen et al. 2008). However, their performances will greatly decrease when massive ground-truth positive labels are initialized as negative labels. (2) The second way is to transform the missing labels filling as a Matrix Completion (MC) problem (Goldberg et al. 2010) (Cabral et al. 2011), which is often based on the low-rank assumption of the whole label matrix. Recently, (Xu, Tao, and Xu 2016) simultaneously incorporates the sparse constraint and low-rank decomposition into the same framework to solve the multi-label learning problem. (3) The third way is to treat the missing labels as latent variables and embed them into a probabilistic model, such as Bayesian networks (Vasisht et al. 2014) and Conditional Restricted Boltzmann Machines (CR-BM). (4) The last way is to treat missing labels as the three states (Wu et al. 2014), i.e. positive labels +1, negative labels -1 and missing labels 0, to avoid the label bias.

### Partial Label Learning

Partial Label Learning (PLL) deals with the problem where each training example is associated with a set of candidate labels, among which only one is correct. (Cour, Sapp, and Taskar 2011) (Zhang, Yu, and Tang 2017). An intuitive strategy to deal with such problem is disambiguation, i.e., trying to recover the ground-truth label from the candidate label set. One way towards disambiguation is to assume certain parametric model $\mathbf{F}(x, y; \theta)$ where the ground-truth label is first regarded as the latent variable and then refined in an iteration manner (Liu and Dietterich 2012) (Zhang and Yu 2015). Another disambiguation strategy is to assume that each candidate label has equal contribution to the learning model and then it makes prediction for unseen examples by averaging their modeling outputs (Cour, Sapp, and Taskar 2011). Compared to partial label learning, PML problem is much more challenging because the number of correct labels in the candidate set is unknown, which makes disambiguation difficult and inapplicable.

## Proposed Method

In this section, we will first introduce the notations of our method, and then present the details of the proposed framework which combines the low-rank and sparse decomposition scheme for partial multi-label learning. Furthermore, an optimization algorithm will also be described in detail.

### Notations

Our method takes two matrices as input : the instance matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where $d$ is the dimension of the feature vector and $n$ is the number of training instances. And we define the $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k]^\top \in \{0, 1\}^{k \times n}$ to represent the label assignments for the corresponding labeled examples, where $k$ is the number of labels. The values in this matrix are within $\{0,1\}$, i.e., if instance $j$ is annotated with label $\mathbf{y}_i$, $y_{ij} = 1$; otherwise, $y_{ij} = 0$.

### The Regularization Framework

Given the noise-corrupted label matrix, how to identify the ground-truth labels of the instance from the candidate label set and how to train an efficient and robust multi-label classifier for label prediction are two challenging problems in PML. In this paper, we propose a PML-LRS method that acquires the accurate label matrix using the concept of low-rank and sparse decomposition, thereby predicting the labels of unlabeled data more accurately.

Our goal are to use the instance matrix $\mathbf{X}$ and the observed label matrix $\mathbf{Y}$ for training a new PML model and to predict the labels from these redundant labels. For the $i$-th label, the goal is to learn a linear function $f_i$ where $\mathbf{w}_i$ is the model parameter. Here we restrict the prediction function $f_i$ to linear functions for simplicity, i.e., $f_i(\mathbf{X}) = \mathbf{w}_i^\top \mathbf{X}$. Define $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k]^\top \in \mathbb{R}^{k \times d}$ to denote the model parameters for all labels. Following the traditional machine learning discipline, a general multi-label learning model can be learned by solving the following problem:

$$\min_{\mathbf{W}} \frac{\eta}{2} \|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_F^2 + \varPhi(\mathbf{W}) \tag{1}$$

where $\varPhi(\mathbf{W})$ represents a regularization function of $\mathbf{W}$ which is used to control the model complexity. Note that there exists the well-known label correlations among different labels in multi-label learning, so we assume that the feature mapping matrix $\mathbf{W}$ is linearly dependent to effectively capture such label correlations, which leads $\mathbf{W}$ to be low-rank and the optimization problem is defined as:

$$\min_{\mathbf{W}} \frac{\eta}{2} \|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_F^2 + \gamma \mathbf{rank}(\mathbf{W}) \tag{2}$$

The above optimization problem (2) is difficult to solve due to the discrete nature of the rank function. Following recent advances on rank minimization, one popular approach is to replace the rank function by the trace norm (or nuclear norm). Using this relaxation, Eq. (2) is rewritten as follows:

$$\min_{\mathbf{W}} \frac{\eta}{2} \|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_F^2 + \gamma \|\mathbf{W}\|_* \tag{3}$$

where $\| \ \|_*$ denotes the sum of the singular values of the matrix.

However, there is a situation of labels redundancy in real life, i.e., annotators may roughly assign each instance a set of candidate labels, which includes both related labels and some unrelated labels.

To deal with this problem, we assume that the irrelevant labels are sparse among observed candidate labels. The basic idea of our method is to capture the accurate label matrix and irrelevant label matrix from the observed candidate label matrix by utilizing the idea of low-rank and sparse decomposition. Specifically, we assume that the irrelevant labels are sparse and introduce an $\ell_\mathbf{1}$-norm regularization to constrain the redundant label matrix. In order to make full use of the ground-truth label information to compute the prediction model, a trace norm regularization is introduced to capture the dependence among ground-truth labels. So the observed label matrix $\mathbf{Y}$ can be decomposed into a sparse matrix $\mathbf{Q}$ and a low-rank matrix $\mathbf{P}$, which can be represented as follows:

$$\min_{\mathbf{P},\mathbf{Q}} \mathbf{rank}(\mathbf{P}) + \beta \|\mathbf{Q}\|_0, \quad s.t. \, \mathbf{Y} = \mathbf{P} + \mathbf{Q} \tag{4}$$

As aforementioned, Eq. (4) is cumbersome to solve because the rank and cardinality operators are discontinuous and non-convex. Therefore, these operators are respectively relaxed to their convex surrogates: the nuclear norm and the $\ell_\mathbf{1}$-norm. Using this relaxation, Eq. (4) is rewritten as follows:

$$\min_{\mathbf{P},\mathbf{Q}} \|\mathbf{P}\|_* + \beta \|\mathbf{Q}\|_1, \quad s.t. \, \mathbf{Y} = \mathbf{P} + \mathbf{Q} \tag{5}$$

By combining Eq. (5) and Eq. (3) together, the final objective function for the proposed partial multi-label learning model can be formulated as follows:

$$\begin{aligned} \min_{\mathbf{P},\mathbf{Q},\mathbf{W}} \frac{\eta}{2} \|\mathbf{P} - \mathbf{W}\mathbf{X}\|_F^2 + \|\mathbf{P}\|_* + \beta \|\mathbf{Q}\|_1 + \gamma \|\mathbf{W}\|_* \\ s.t. \, \mathbf{Y} = \mathbf{P} + \mathbf{Q} \end{aligned} \tag{6}$$

where $\eta$, $\gamma$ and $\beta$ are trade-off parameters to keep the balance of the model. From the above objective function, we can see that the predictive function is robust to inaccurately labeled instances.

## Optimization

The problem (6) is convex and can be optimized efficiently. We first convert it into the following equivalent problem:

$$\begin{aligned} \min_{\mathbf{P},\mathbf{Q},\mathbf{W},\mathbf{J},\mathbf{T}} \frac{\eta}{2} \|\mathbf{T} - \mathbf{W}\mathbf{X}\|_F^2 + \|\mathbf{P}\|_* + \beta \|\mathbf{Q}\|_1 + \gamma \|\mathbf{J}\|_* \\ s.t. \, \mathbf{Y} = \mathbf{P} + \mathbf{Q}, \quad \mathbf{W} = \mathbf{J}, \quad \mathbf{P} = \mathbf{T} \end{aligned} \tag{7}$$

The optimization problem (7) can be solved with the ALM (Zhang et al. 2017), which minimizes the following augmented Lagrange function:

$$\begin{aligned} \min_{\mathbf{P},\mathbf{Q},\mathbf{W},\mathbf{J},\mathbf{T}} & \frac{\eta}{2} \|\mathbf{T} - \mathbf{W}\mathbf{X}\|_F^2 + \|\mathbf{P}\|_* + \beta \|\mathbf{Q}\|_1 + \gamma \|\mathbf{J}\|_* \\ & + < \mathbf{Y_1}, \mathbf{Y} - \mathbf{P} - \mathbf{Q} > + \frac{\mu_1}{2} \|\mathbf{Y} - \mathbf{P} - \mathbf{Q}\|_F^2 \\ & + < \mathbf{Y_2}, \mathbf{W} - \mathbf{J} > + \frac{\mu_2}{2} \|\mathbf{W} - \mathbf{J}\|_F^2 \\ & + < \mathbf{Y_3}, \mathbf{P} - \mathbf{T} > + \frac{\mu_3}{2} \|\mathbf{P} - \mathbf{T}\|_F^2 \end{aligned} \tag{8}$$

where $\mathbf{Y_1} \in \mathbb{R}^{d \times n}$, $\mathbf{Y_2} \in \mathbb{R}^{d \times n}$ and $\mathbf{Y_3} \in \mathbb{R}^{d \times n}$ are Lagrange multiplier matrices, and $\mu_1$, $\mu_2$ and $\mu_3$ are the penalty parameters. According to the LADMAP method (Lin, Liu, and Su 2011), Eq. (8) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{P},\mathbf{Q},\mathbf{W},\mathbf{J},\mathbf{T}} & \frac{\eta}{2} \|\mathbf{T} - \mathbf{W}\mathbf{X}\|_F^2 + \|\mathbf{P}\|_* + \beta \|\mathbf{Q}\|_1 \\ & + \gamma \|\mathbf{J}\|_* + \frac{\mu_1}{2} \|\mathbf{Y} - \mathbf{P} - \mathbf{Q} + \frac{\mathbf{Y_1}}{\mu_\mathbf{1}}\|_F^2 \\ & + \frac{\mu_2}{2} \|\mathbf{W} - \mathbf{J} + \frac{\mathbf{Y_2}}{\mu_\mathbf{2}}\|_F^2 + \frac{\mu_3}{2} \|\mathbf{P} - \mathbf{T} + \frac{\mathbf{Y_3}}{\mu_\mathbf{3}}\|_F^2 \end{aligned} \tag{9}$$

For each of the five matrices $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{J}$, $\mathbf{T}$ to be solved in particularly Eq. (9), the cost function is convex if the remaining four matrices are kept fixed. Eq. (9) can be solved iteratively via the following subproblems:

1. When keeping $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{J}$, $\mathbf{T}$ fixed, we obtain the following equation for $\mathbf{W}$ by taking the derivative of Eq. (9), denoted by LRS-1,

$$\min_{\mathbf{W}} \frac{\eta}{2} \|\mathbf{T} - \mathbf{W}\mathbf{X}\|_F^2 + \frac{\mu_\mathbf{2}}{2} \|\mathbf{W} - \mathbf{J} + \frac{\mathbf{Y_2}}{\mu_\mathbf{2}}\|_F^2 \tag{10}$$

which is an ordinary least squares regression problem, whose solution is,

$$\mathbf{W} = (\mu_\mathbf{2}\mathbf{J} + \eta\mathbf{T}\mathbf{X}^\mathbf{T} - \mathbf{Y_2})(\eta\mathbf{X}\mathbf{X}^\mathbf{T} + \mu_\mathbf{2}\mathbf{I_d})^{-\mathbf{1}} \tag{11}$$

2. When $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{T}$ are fixed, optimizing Eq. (9) with respect to $\mathbf{J}$ is equivalent to the following problem, denoted by LRS-2,

$$\min_{\mathbf{J}} \gamma\|\mathbf{J}\|_* + \frac{\mu_\mathbf{2}}{\mathbf{2}}\|\mathbf{W}-\mathbf{J}+\frac{\mathbf{Y_2}}{\mu_\mathbf{2}}\|_F^2 \qquad (12)$$

The objective in Eq. (12) can be expressed equivalently as follows:

$$\min_{\mathbf{J}} \frac{\mathbf{1}}{\mathbf{2}}\|\mathbf{J}-(\mathbf{W}+\frac{\mathbf{Y_2}}{\mu_\mathbf{2}})\|_F^2 + \frac{\gamma}{\mu_\mathbf{2}}\|\mathbf{J}\|_* \qquad (13)$$

It turns out that the minimization of the objective in Eq. (13) can be solved by first computing the singular value decomposition (SVD) of $\mathbf{W}+\mathbf{Y_2}/\mu_\mathbf{2}$ and then applying some soft-thresholding on the singular values.

3. Fixing $\mathbf{W}$, $\mathbf{J}$, $\mathbf{T}$, solve (9) for $\mathbf{P}$ and $\mathbf{Q}$ by the following problem, denoted by LRS-3,

$$\begin{aligned}&\min_{\mathbf{P},\mathbf{Q}} \|\mathbf{P}\|_* + \beta\|\mathbf{Q}\|_1\\ &+\frac{\mu_1}{2}\|\mathbf{Y}-\mathbf{P}-\mathbf{Q}+\frac{\mathbf{Y_1}}{\mu_\mathbf{1}}\|_F^2\\ &+\frac{\mu_3}{2}\|\mathbf{P}-\mathbf{T}+\frac{\mathbf{Y_3}}{\mu_\mathbf{3}}\|_F^2\end{aligned} \qquad (14)$$

which is a slight variation of the low-rank representation problem (Liu, Lin, and Yu 2010), and the linear ADM solution is,

$$\mathbf{P^{k+1}} = \mathbf{D_{1/\beta_p}}(\mathbf{P^K} - \mathbf{F_P^K}/\beta_\mathbf{p}) \qquad (15)$$

$$\mathbf{Q^{k+1}} = \mathbf{S}_{\beta/\mu_1}(\mathbf{Y}-\mathbf{P}+\mathbf{Y_1}/\mu_\mathbf{1}) \qquad (16)$$

where $\mathbf{D}$ is the singular value thresholding (Cai and Shen 2010), $\mathbf{S}$ is the shrinkage operator (Zhang et al. 2012), $\beta_P = (\mu_1+\mu_2)\tau_P/2$, $\tau_P > \rho(I^T I)$ is the proximal parameter, $\rho(I^T I)$ denotes the spectral radius of $\rho(I^T I)$, and $\mathbf{F_P^k}$ is the derivative by $\mathbf{P^k}$ for the second and third terms in Eq. (14),

$$\mathbf{F_P^K} = \mu_\mathbf{1}(\mathbf{P}-\mathbf{Y}+\mathbf{Q}) + \mu_\mathbf{3}(\mathbf{P}-\mathbf{T}) + \mathbf{Y_3} - \mathbf{Y_1} \quad (17)$$

4. With $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{J}$ fixed, the computation of $\mathbf{T}$ is independent, we obtain the following optimization problem for $\mathbf{T}$ by taking the derivative of Eq. (9), denoted by LRS-4,

$$\min_{\mathbf{T}} \frac{\eta}{2}\|\mathbf{T}-\mathbf{WX}\|_F^2 + \frac{\mu_3}{2}\|\mathbf{P}-\mathbf{T}+\frac{\mathbf{Y}_3}{\mu_\mathbf{3}}\|_F^2 \qquad (18)$$

which is also an ordinary least squares problem, to which the solution is,

$$\mathbf{T} = (\eta\mathbf{WX} + \mu_\mathbf{3}\mathbf{P} + \mathbf{Y_3})(\eta\mathbf{I_d} + \mu_\mathbf{3}\mathbf{I_d})^{\mathbf{-1}} \qquad (19)$$

Finally, the Lagrange multiplier matrices $\mathbf{Y_1}$, $\mathbf{Y_2}$, $\mathbf{Y_3}$ and regularization terms $\mu_1$, $\mu_2$, $\mu_3$ are updated based on LADM,

$$\begin{aligned}\mathbf{Y_1^{k+1}} &= \mathbf{Y_1^k} + \mu_\mathbf{1}^\mathbf{k+1}(\mathbf{Y}-\mathbf{P}-\mathbf{Q})\\ \mathbf{Y_2^{k+1}} &= \mathbf{Y_2^k} + \mu_\mathbf{2}^\mathbf{k+1}(\mathbf{W}-\mathbf{J})\\ \mathbf{Y_3^{k+1}} &= \mathbf{Y_3^k} + \mu_\mathbf{3}^\mathbf{k+1}(\mathbf{P}-\mathbf{T})\\ \mu_\mathbf{1}^\mathbf{k+1} &= \min(\mu_\mathbf{max}, \rho\mu_\mathbf{1}^\mathbf{k})\\ \mu_\mathbf{2}^\mathbf{k+1} &= \min(\mu_\mathbf{max}, \rho\mu_\mathbf{2}^\mathbf{k})\\ \mu_\mathbf{3}^\mathbf{k+1} &= \min(\mu_\mathbf{max}, \rho\mu_\mathbf{3}^\mathbf{k})\end{aligned} \qquad (20)$$

where $\rho$ is a positive scalar.

The entire optimization procedure will be terminated when $\mathbf{W}$, $\mathbf{P}$ and $\mathbf{Q}$ are all small. Despite the algorithm does not guarantee a global optimum, we found that it performs well in our experiments.

## Computation Complexity Analysis

In the iterations of the proposed method, the computational costs are mainly matrix inversion and SVD. For the sample matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and the label matrix $\mathbf{Y} \in \{0,1\}^{k\times n}$, computation complexity of full SVD is $\mathcal{O}(dk^2)(d > k)$. Each iteration of LRS-3 mainly includes SVD. Then the whole computational complexity for LRS-3 is $\mathcal{O}(t_1 * dk^2)$ and $t_1$ is the iteration number of LRS-3. Similar to LRS-3, the whole computation complexity for LRS-2 is $\mathcal{O}(t_1 * dk^2)$. LRS-1 and LRS-4 contain iterations of matrix inverse, whose complexity is same as $\mathcal{O}(t_1 * d^3)$. Hence, the total complexity for PML-LRS is $\mathcal{O}(T * (t_1 * dk^2 + t_1 * d^3))$.

# Experiments

In this section, we first describe our experimental setup, including the benchmark data sets, comparing algorithms, and evaluation metrics. Then we present three sets of experiments to verify the effectiveness of the proposed PML-LRS approach, where the first experiment reports the detailed experimental results of six comparing algorithms on the data sets respectively, the second experiment use Friedman test (Demšar 2006) as the statistical test to analyze the relative performance among the comparing algorithms, and the third experiment evaluates the parameters sensitivity of the proposed algorithm.

## Data sets

We perform experiments on six data sets. These data sets spanned a broad range of applications: ***corel5k*** for image annotation, ***CAL500*** and ***emotions*** for music classification, ***genbase*** for protein classification, ***medical*** for text categorization and ***delicious*** for web categorization. Ten-fold cross-validation is performed on the benchmark data sets, where the mean metric value, as well as standard deviation, are recorded for each comparing algorithm. Specifically, we illustrate the number of instances, number of classes and domain for each data set in Table 1.

## Experimental Setup

Evaluation Metrics: In this paper, five widely-used multi-label metrics are employed for performance evaluation, including *ranking loss, hamming loss, one error, coverage, and average precision*. These evaluation metrics consider the performance of multi-label predictor from various aspects, whose values all vary between [0,1]. Concrete metric definitions can be found in (Zhang and Zhou 2014). For the *one error, coverage and ranking loss* and *hamming loss*, the smaller the values the better the performance. For the *average precision* metrics, the larger the values, the better the performance.

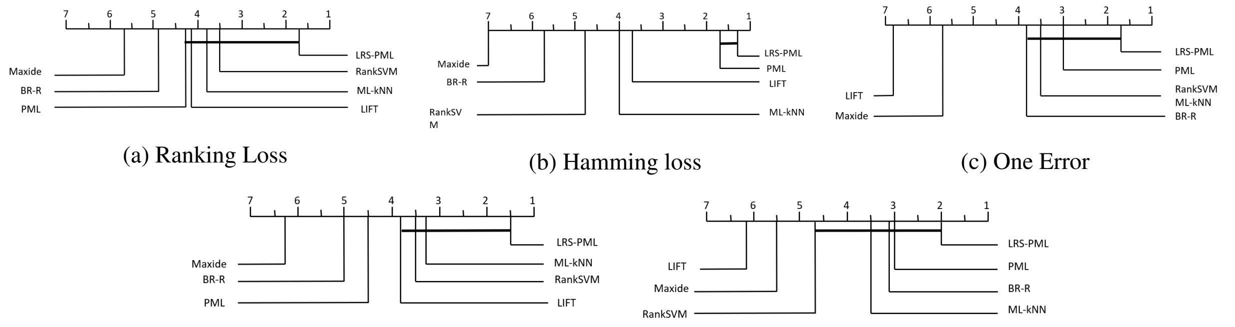
Baselines: To show the advantages of the proposed PML-LRS method, we implemented six state-of-the-art methods

Table 1: Characteristics of the multi-label experimental data sets.

| Data set | Instance | Dim | Class Labels | Domain |
|---|---|---|---|---|
| **emotions (Trohidis et al.2008)** | 593 | 72 | 6 | music |
| **CAL500 (Turnbull et al.2008)** | 500 | 68 | 174 | music |
| **genbase (Diplaris et al.2005)** | 662 | 1186 | 27 | biology |
| **medical(Pestian et al.2007)** | 978 | 1449 | 45 | text |
| **corel5k(Duygulu et al.2002)** | 5000 | 499 | 374 | images |
| **delicious(Tsoumakas et al.2008)** | 14000 | 500 | 983 | text |

Table 2: Comparison of PML-LRS with state-of-the-art multi-label learning approaches on five evaluation criteria. The best performance is bolded.

| Data | BR-R | RankSVM | Maxide | ML-kNN | LIFT | PML-*fp* | PML-LRS |
|---|---|---|---|---|---|---|---|
| Ranking Loss(the smaller, the better) | | | | | | | |
| CAL500 | .266 ± .0193 | .241 ± .020 | .188 ± .033 | .183 ± .007 | .186 ± .011 | .209 ± .006 | **.110 ± .037** |
| Emotions | .176 ± .017 | .138 ± .001 | .375 ± .071 | .157 ± .042 | .262 ± .024 | .170 ± .002 | **.115 ± .024** |
| Medical | .024 ± .008 | **.015 ± .001** | .133 ± .037 | .042 ± .011 | .046 ± .008 | .125 ± .008 | .075 ± .004 |
| Genbase | .002 ± .002 | .005 ± .021 | .184 ± .044 | .007 ± .005 | .037 ± .004 | .002 ± .000 | **.001 ± .001** |
| Corel5k | .233 ± .011 | .130 ± .003 | .098 ± .037 | .134 ± .006 | .125 ± .005 | .065 ± .014 | **.049 ± .029** |
| Delicious | .269 ± .000 | .148 ± .009 | .248 ± .023 | .151 ± .013 | .143 ± .007 | .221 ± .027 | **.088 ± .005** |
| Hamming loss(the smaller, the better) | | | | | | | |
| CAL500 | .449 ± .007 | .437 ± .020 | .461 ± .003 | .345 ± .001 | .342 ± .004 | .170 ± .026 | **.116 ± .009** |
| Emotions | .469 ± .016 | .306 ± .031 | .645 ± .013 | .506 ± .025 | .567 ± .023 | **.233 ± .017** | .255 ± .013 |
| Medical | .659 ± .002 | .591 ± .001 | .699 ± .000 | .584 ± .002 | .531 ± .003 | .523 ± .010 | **.506 ± .088** |
| Genbase | .347 ± .003 | .341 ± .005 | .499 ± .001 | .258 ± .002 | .235 ± .002 | **.034 ± .006** | .035 ± .022 |
| Corel5k | .639 ± .021 | .634 ± .046 | .795 ± .022 | .519 ± .000 | .512 ± .001 | .428 ± .003 | **.409 ± .004** |
| Delicious | .527 ± .000 | .522 ± .012 | .698 ± .035 | .417 ± .022 | .418 ± .009 | .348 ± .018 | **.218 ± .007** |
| One Error(the smaller, the better) | | | | | | | |
| CAL500 | .268 ± .042 | .241 ± .001 | .196 ± .085 | .122 ± .025 | .899 ± .046 | .157 ± .022 | **.102 ± .078** |
| Emotions | .281 ± .051 | .468 ± .036 | .514 ± .065 | .277 ± .079 | .712 ± .067 | .333 ± .016 | **.175 ± .038** |
| Medical | .258 ± .025 | **.163 ± .001** | .722 ± .065 | .245 ± .032 | .895 ± .029 | .471 ± .023 | .420 ± .004 |
| Genbase | .022 ± .012 | .019 ± .024 | .727 ± .049 | .018 ± .014 | .881 ± .037 | **.006 ± .000** | .016 ± .005 |
| Corel5k | .696 ± .031 | .652 ± .000 | .723 ± .024 | .740 ± .023 | .985 ± .007 | .577 ± .013 | **.534 ± .049** |
| Delicious | .357 ± .019 | .337 ± .057 | .635 ± .034 | .395 ± .001 | .319 ± .003 | .341 ± .000 | **.321 ± .007** |
| Coverage(the smaller, the better) | | | | | | | |
| CAL500 | .610 ± .010 | **.541 ± .018** | .815 ± .065 | .754 ± .020 | .769 ± .032 | .641 ± .001 | .574 ± .108 |
| Emotions | .483 ± .027 | .447 ± .030 | .644 ± .059 | .459 ± .042 | .520 ± .032 | .458 ± .008 | **.416 ± .023** |
| Medical | .327 ± .011 | .290 ± .034 | .343 ± .083 | .164 ± .023 | .162 ± .002 | .541 ± .018 | **.112 ± .004** |
| Genbase | .133 ± .005 | .129 ± .018 | .121 ± .025 | .117 ± .032 | .085 ± .007 | .299 ± .017 | **.047 ± .003** |
| Corel5k | .487 ± .020 | .451 ± .045 | .475 ± .173 | .308 ± .015 | **.295 ± .013** | .431 ± .018 | .404 ± .115 |
| Delicious | .752 ± .003 | .739 ± .000 | .808 ± .029 | .597 ± .023 | .774 ± .021 | .521 ± .013 | **.428 ± .028** |
| Average Precision(the greater, the better) | | | | | | | |
| CAL500 | .463 ± .015 | .441 ± .037 | .505 ± .046 | .491 ± .015 | .454 ± .015 | .459 ± .000 | **.638 ± .061** |
| Emotions | .785 ± .023 | .654 ± .001 | .622 ± .056 | .803 ± .044 | .632 ± .030 | .773 ± .063 | **.859 ± .026** |
| Medical | .830 ± .017 | **.889 ± .037** | .425 ± .054 | .811 ± .028 | .518 ± .030 | .636 ± .025 | .663 ± .005 |
| Genbase | .996 ± .002 | .958 ± .000 | .448 ± .047 | .973 ± .013 | .573 ± .021 | **.998 ± .014** | .995 ± .000 |
| Corel5k | .237 ± .014 | .229 ± .003 | .256 ± .022 | .245 ± .010 | .212 ± .005 | .404 ± .029 | **.419 ± .038** |
| Delicious | .344 ± .027 | .340 ± .041 | .278 ± .051 | .333 ± .008 | .328 ± .000 | **.451 ± .000** | .384 ± .007 |



(a) Ranking Loss

(b) Hamming loss

(c) One Error

(d) Coverage

(e) Average Precision

Figure 2: Comparison of PML-LRS (control algorithm) against six comparing algorithms with the Bonferroni-Dunn test. Algorithms not connected with PML-LRS in the CD diagram are considered to have a significantly different performance from the control algorithm (CD=3.2308 at 0.05 significance level).

for comparison. For the comparing algorithms, parameter configurations are adopted by the suggestions in the respective literatures.

- *Binary Reference Model based on RBF Kernel (BR-R)* (Boutell et al. 2004). This is a first-order baseline approach which decomposes the multi-label learning problem into independent binary classification problems, whereas the label correlation is not taken into consideration.
- *Ranking Support Vector Machine (RankSVM)* (Elisseeff and Weston 2002). The basic idea of this algorithm is to adapt maximum margin strategy to deal with multi-label data, where a set of linear classifiers are optimized to minimize the empirical ranking loss and enabled to handle nonlinear cases with kernel tricks.
- *Matrix Completion using Side Information (Maxide)* (Xu, Jin, and Zhou 2013). It is a matrix completion based approach for transductive multi-label learning by exploiting side information matrices.
- *Multi-Label k-Nearest Neighbor (ML-kNN)* (Zhang and Zhou 2007). A nearest neighbor based multi-label classification method. The number of nearest neighbors is chosen by cross-validation. ML-kNN is a very popular baseline in the multi-label learning literature due to its simplicity.
- *Multi-Label Learning with Label Specific Features (LIFT)* (Zhang and Wu 2015). In contrast to existing multi-label learning methods which focus on exploiting label correlations, it tries to exploit label-specific features for multi-label learning.
- *Partial Multi-Label Learning (PML-fp) (Xie and Huang 2018)*. It is a recently proposed partial multi-label learning solution. It introduces confidence value to evaluate the probability of being the ground-truth label for each candidate label, and alternatively optimize the classification model and the confidence values to solve the PML problem. (Xie and Huang 2018) offers two options to further exploit either the local structure of the feature space or the label correlations. Here, we choose PML-*fp* as a comparison algorithm.

Table 3: Friedman statistics $F_F$ in terms of each evaluation metric and the critical value at 0.05 significance level ( # comparing algorithms $k$ = 7, # data sets N = 6).

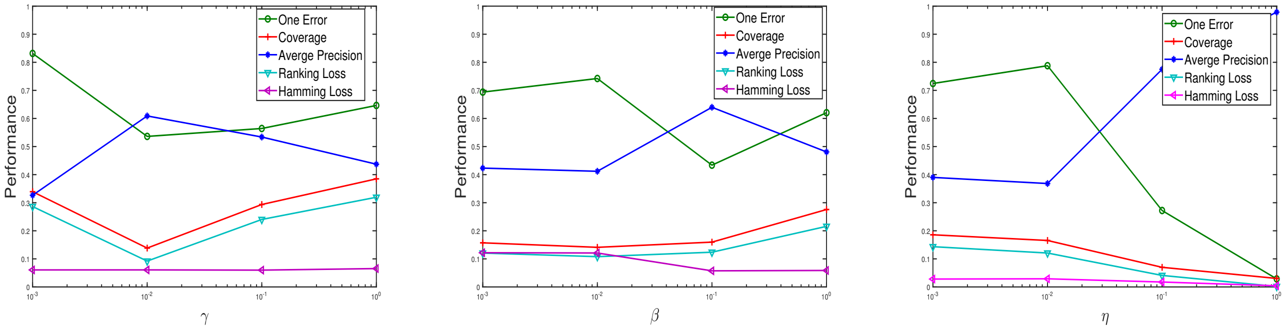
| Evaluation metric | $F_F$ | critical value |
|---|---|---|
| Ranking Loss | 5.6779 | |
| Hamming Loss | 2.5398 | |
| One Error | 8.2020 | 2.4205 |
| Coverage | 4.7622 | |
| Average Precision | 4.5459 | |

To create partial multi-label assignments for the training data, for each sample $\mathbf{x}_i$, we randomly add the irrelevant noisy labels of $\mathbf{x}_i$ with $a\%$ number of ground-truth labels, and we vary the $a\%$ in the range $\{5\%, 10\%, 50\%, 100\%\}$. To examine the performance of the proposed algorithm, we performed experiments with all possible percentages of the noisy labels. PML-LRS is compared with other methods on each data set with respect to each criterion. Statistical significance is examined with the Friedman test at 95% significance level. But considering the page limit, we cannot report all results with every possible percentage of the noisy labels. Instead, we report the detailed results for a more robust noisy label percentage, i.e., the number of irrelevant noisy labels is identical to the number of labeled relevant labels.

## Performance Comparison

Table 2 reports the detailed experimental results of six comparing algorithms, where the best performance among the comparing algorithms is shown in boldface. When compared with other methods, our algorithm shows significant superiority. Among the six compared multi-label approaches, PML-*fp* shows some superiority, and achieves the best performance of 3 criteria on ***genbase***. RankSVM shows the best performance of 3 criteria on ***medical***.

Meanwhile, Friedman test (Demšar 2006) is used as the statistical test to analyze the relative performance among the comparing methods in this paper. Table 3 summarizes the Friedman statistics $F_F$ and the corresponding critical value on each evaluation metric. For each evaluation metric,



(a) performance curve with $\gamma$ changes

(b) performance curve with $\beta$ changes

(c) performance curve with $\eta$ changes

Figure 3: Results of PML-LRS with varying value of trade-off parameters.

the null hypothesis of indistinguishable performance among the comparing algorithms is rejected at the 0.05 significance level.

Therefore, the post-hoc Bonferroni-Dunn test (Demšar 2006) (Zhang, Zhong, and Zhang 2018) is employed to show the relative performance among the comparing algorithms. Here, PML-LRS is treated as the control algorithm whose average rank difference against the comparing algorithm is calibrated with the *critical difference* (CD). Accordingly, PML-LRS is deemed to have a significantly different performance to one comparing algorithm if their average ranks differ by at least one CD (CD=3.2308 in this paper: # comparing algorithms $k$ = 7, # data sets N = 6). Figure 2 illustrates the CD diagrams (Demšar 2006) on each evaluation metric, where the average rank of each comparing algorithm is marked along the axis (lower ranks to the right). In each subfigure, any comparing algorithms whose average rank is within one CD to that of PML-LRS is interconnected to each other with a thick line.

Overall, the following observations can be made based on the above experimental results:

- On six data sets (Table 2) across all evaluation metrics, PML-LRS ranks *1st* in 66.7% cases and ranks *2nd* in 16.7% cases.
- It is noteworthy that PML-LRS achieves optimal (lowest) average rank in terms of all evaluation metrics except average precision. Furthermore, no algorithm significantly outperforms PML-LRS across all evaluation metrics.
- PML-LRS significantly outperforms Maxide and BR-R in terms of all evaluation metrics. PML-*fp* also significantly outperforms other methods in the evaluation metrics of *hamming loss, one error, and average precision* on ***genbase*** and ***delicious***.
- One exception is on the data set ***medical***, where RankSVM outperforms our methods over 3 criteria. This is probably because there are too few training examples to acquire the structure information.
- PML-LRS is comparable to PML-*fp* in terms of *hamming loss, one error, average precision* and significantly outperforms PML-*fp* on all the other cases.

## Parameter Analysis

At last, we study the influences of the three parameters, $\gamma$, $\beta$ and $\eta$ for the proposed method on the medical data set. Our experiment is accomplished by using the grid search method which conducts the parameter analysis by varying three parameters simultaneously. The experimental results are shown in Figure 3 which are measured by the five evaluation metrics. It can be seen that how the performance of our algorithm varies as these parameters change. Therefore we should safely set them in a wide range in practice. From this figure, we can notice that better performances are gained when $\gamma = 0.01$, $\beta = 0.1$, and $\eta = 1$.

## Conclusion

In this paper, we presented a novel approach to address the partial multi-label learning problem in a principled manner. The key to solve the partial multi-label learning problem is identifying the ground-truth labels from the redundant label matrix. The proposed algorithm attempted to utilize the idea of low-rank and sparse decomposition to capture the ground-truth label matrix and irrelevant label matrix from the observed candidate label matrix while training the prediction model simultaneously. Extensive experimental results demonstrated that our approach is effective and outperforms other baseline methods on several data sets.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 61872032, 61673048, 61472028, 61502026, 61872035), in part by the Beijing Natural Science Foundation (No. 4162048), in part by the Fundamental Research Funds for the Central universities (2017JBZ108), and the Joint Research Fund for the Ministry of Education of China and China Mobile (MCM20170201).

## References

Boutell, M.; Luo, J.; Shen, X.; and Brown, C. 2004. Learning multi-label scene classification. *Pattern Recognition* 37(9):1757–1771.

Bucak, S.; Jin, R.; and Jain, A. 2011. Multi-label learning with incomplete class assignments. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2801–2808.

Cabral, R. S.; Torre, F.; Costeira, J.; and Bernardino, A. 2011. Matrix completion for multi-label image classification. In *Advances in Neural Information Processing Systems*, 190–198.

Cai, J.and Candès, E., and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Chen, G.; Song, Y.; Wang, F.; and Zhang, C. 2008. Semi-supervised multi-label learning by solving a sylvester equation. In *SIAM International Conference on Data Mining,*, 410–419.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *Journal of Machine Learning Research* 12(5):1501–1536.

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

Elisseeff, A., and Weston, J. 2002. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems*, 681–687.

Fürnkranz, J.; Hüllermeier, E.; Mencía, E.; and Brinker, K. 2008. Multilabel classification via calibrated label ranking. *Machine learning* 73(2):133–153.

Goldberg, A.; Recht, B.; Xu, J.; Nowak, R.; and Zhu, X. 2010. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems*, 757–765.

Gopal, S., and Yang, Y. 2010. Multilabel classification with meta-level features. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 315–322.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2010. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery from Data* 4(2):1–29.

Lin, Z.; Liu, R.; and Su, Z. 2011. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in Neural Information Processing Systems*, 612–620.

Liu, L., and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems*, 548–556.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning*, 663–670.

Sanden, C., and Zhang, J. 2011. Enhancing multi-label music genre classification through ensemble techniques. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 705–714.

Sun, Y.; Zhang, Y.; and Zhou, Z. 2010. Multi-label learning with weak label. In *AAAI Conference on Artificial Intelligence*, 593–598.

Tang, L.; Rajan, S.; and Narayanan, V. 2009. Large scale multi-label classification via metalabeler. In *International Conference on World Wide Web*, 211–220.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2011. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering* 23(7):1079–1089.

Vasisht, D.; Damianou, A.; Varma, M.; and Kapoor, A. 2014. Active learning for sparse bayesian multilabel classification. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 472–481.

Wu, B.; Liu, Z.; Wang, S.; Hu, B.; and Ji, Q. 2014. Multi-label learning with missing labels. In *International Conference on Pattern Recognition*, 1964–1968.

Wu, F.; Wang, Z.; Zhang, Z.; Yang, Y.; Luo, J.; Zhu, W.; and Zhuang, Y. 2015. Weakly semi-supervised deep learning for multi-label image annotation. *IEEE Transactions on Big Data* 1(3):109–122.

Xie, M., and Huang, S. 2018. Partial multi-label learning. In *AAAI Conference on Artificial Intelligence*, 4303–4309.

Xu, M.; Jin, R.; and Zhou, Z. 2013. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems*, 2301–2309.

Xu, C.; Tao, D.; and Xu, . 2016. Robust extreme multi-label learning. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1275–1284.

Zhang, M., and Wu, L. 2015. Lift: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(1):107–120.

Zhang, M., and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *International Joint Conference on Artificial Intelligence*, 4048–4054.

Zhang, M., and Zhou, Z. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.

Zhang, M., and Zhou, Z. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.

Zhang, X.; Ma, Y.; Lin, Z.; Gao, H.; Zhuang, L.; and Yu, N. 2012. Non-negative low rank and sparse graph for semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2328–2335.

Zhang, Y.; Shi, D.; Gao, J.; and Cheng, D. 2017. Low-rank-sparse subspace representation for robust regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7445–7454.

Zhang, M.; Yu, F.; and Tang, C. 2017. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2155–2167.

Zhang, Q.; Zhong, Y.; and Zhang, M. 2018. Feature-induced labeling information enrichment for multi-label learning. In *AAAI Conference on Artificial Intelligence*, 4446–4453.