Self-Paced Multi-Label Learning with Diversity

Seyed Amjad Seyedi AMJADSEYEDI@ENG.UOK.AC.IR
S. Siamak Ghodsi* S.GHODSI@ENG.UOK.AC.IR
Fardin Akhlaghian F.AKHLAGHIAN@UOK.AC.IR
Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

Mahdi Jalili MAHDI.JALILI@RMIT.EDU.AU
School of Engineering, RMIT University, Melbourne, Australia

Parham Moradi† P.MORADI@UOK.AC.IR
Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

The major challenge of learning from multi-label data has arisen from the overwhelming size of label space which makes this problem NP-hard. This problem can be alleviated by gradually involving easy to hard tags into the learning process. Besides, the utilization of a diversity maintenance approach avoids overfitting on a subset of easy labels. In this paper, we propose a self-paced multi-label learning with diversity (SPMLD) which aims to cover diverse labels with respect to its learning pace. In addition, the proposed framework is applied to an efficient correlation-based multi-label method. The non-convex objective function is optimized by an extension of the block coordinate descent algorithm. Empirical evaluations on real-world datasets with different dimensions of features and labels imply the effectiveness of the proposed predictive model.

Keywords: Self-Paced Learning, Multi-Label Learning, Block Coordinate Descent, Manifold Optimization.

1. Introduction

The paradigm of multi-label learning has become a popular topic in recent years. In many real-world applications, instances are in semantic association with more than one class label Huang et al. (2018). Therefore, it sounds more rational to map each instance into a vector of labels rather than a single one. An effective stage to handle a multi-label problem is to learn dependencies among the labels Zhang and Zhou (2014). Accordingly, numerous studies have been conducted to accomplish this goal. Zhang and Zhou (2007) proposed a lazy learning method, derived from the conventional kNN classifier which classifies instances using a statistical model based on maximum a posteriori principle. However, this algorithm implicitly covers a local definition of correlation; label dependency has gone further. Huang and Zhou (2012) investigated an explicit view of local correlation by encoding their influences into a local code (LOC).

* S. A. Seyedi and S. S. Ghodsi—Equal Contribution.
† Corresponding Author

Another important stage that has drawn much attention is a low-rank representation of label space. Many algorithms with this property are proposed. Yu et al. (2014) presented a large-scale low-rank structure applicable to scaled label spaces meanwhile handling data with missing-labels. The framework of Xu et al. (2014) aims to capture global label correlations by utilizing a low-rank structure on the label correlation matrix and copes with the missing-label challenge by introducing a supplementary label matrix. Xu et al. (2013) proposed a fast matrix completion algorithm with a low-rank representation which exploits side information explicitly to optimize complexity in a transductive manner. Zhu et al. (2018) investigated the concept of label correlation in a new manner. Contrary to previous approaches that only rely on a single definition of correlation i.e. global or local, the GLOCAL framework analyzes both GLObal and loCAL correlations of labels simultaneously in a latent label representation. This method takes advantage of manifold optimization and is capable of dealing with both missing and full label scenarios.

Algorithms mentioned so far have many pros and cons. Many methods in this field have studied the multi-label problem from different perspectives and have made valuable efforts. However, there is a common shortcoming; they lack a mechanism to give a clear order to training instances. Some instances are easy for a specific label. It is beneficial to learn that label with those instances first and then gradually learn harder ones. For example, learning label ”rabbit” with a black rabbit running on grass in a picture is easier than a white rabbit running on snow.

Curriculum and self-paced learning (SPL) are recently proposed regimes with the aim of learning from easier to more complex concepts Bengio et al. (2009); Kumar et al. (2010); Meng et al. (2017). These learning frameworks are inspired by human education system where the major difference arises in identifying the complexity level. Curriculum learning needs a teacher (extra knowledge) to distinguish easy concepts from the complex ones, whereas self-paced learning is like a student who starts to learn a curriculum based on self-abilities.

The SPL framework is widely applied to various fields. Zhao et al. (2015) incorporated SPL with conventional matrix factorization and introduced a new Matrix factorization framework with a generalization of SPL to produce soft weight values along with the original binary weights. Li et al. (2017) proposed a multi-task algorithm with an self-paced regularization, and optimized this learner with efficient development of block coordinate descent. Self-Paced learning is claimed to be a general framework applicable to any learning framework having an objective function with an empirical loss function. It has been successfully applied in various learning fields such as classification Li and Gong (2017), boost learning Pi et al. (2016), object detection Sangineto et al. (2019), Co-saliency detection Zhang et al. (2017), face identification Lin et al. (2018), Multi-view Clustering Xu et al. (2015), and multi-task learning Murugesan et al. (2017).

Li et al. (2018) introduced a self-paced regularization framework for multi-label learning. It is one of the first attempts to tackle the multi-label problem by considering the complexity of instances for labels. However, without a diversity maintenance approach, a self-paced regularizer may cause the learning model to be biased on a subset of labels that are easy to learn Jiang et al. (2014).

In this paper, a self-paced multi-label with diversity (SPMLD) framework is proposed. The diversity regularization term drives the model to be inclined to learn different labels that are easier first and somehow overcomes the problem of being biased on a limited number of easy labels. Besides, this gradual learning scheme can exploit more reliable label dependencies. Finally, to present a comparable example for realizing desired self-paced multi-label learning, SPMLD is applied to a host algorithm, a recent multi-label method Zhu et al. (2018) which has acceptable performance. Empirical results supported by statistical significance tests demonstrate the effectiveness of our method against several well-known algorithms including the host algorithm.

2. Background

Self-paced learning provides a way for simultaneously choosing the easier patterns and re-estimating the learning parameters $\mathbf{w}$ in the form of an iterative process Kumar et al. (2010). We presume a linear function $f(\boldsymbol{x}_i, \mathbf{w})$ with unknown parameter $\mathbf{w}$. SPL is then given by the following objective function to be solved:

$$\min_{\mathbf{w},\mathbf{p}\in\Omega} \sum_{i=1}^{n} p_i \ell_i(y_i, f(\boldsymbol{x}_i, \mathbf{w})) + \Psi(\lambda, \mathbf{p}) \tag{1}$$

where $\Psi(\lambda, \mathbf{p}) = \lambda \sum_{i=1}^{n} p_i$ is the regularization term, Ω is the domain space of $\mathbf{p}$ and $\Psi(\lambda, \mathbf{p})$ is the regularization parameter which determines the complexity of patterns. Equation (1) has two unknowns including $\mathbf{w}$ as learning parameter and λ the parameter of pace control (restricted to the specified domain). Equation (1) then becomes a bi-convex optimization problem over the parameters $\mathbf{w}$ and $\mathbf{p}$, which can be efficiently solved by alternating minimization. The optimal solution of $\mathbf{w}$ with fixed $\mathbf{p}$ can be achieved by any off-the-shelf solver and, the optimal solution of $\mathbf{p}$ with a fixed $\mathbf{w}$ can be obtained by:

$$p_i^* = \begin{cases} 1, & \ell_i(y_i, f(\boldsymbol{x}_i, \mathbf{w})) < \lambda, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

According to (2) easy samples have losses smaller than a determined threshold because they have less prediction errors, when updating $\mathbf{p}$ given a fixed $\mathbf{w}$, so they are chosen for training ($p_i^* = 1$) or otherwise they aren't chosen ($p_i^* = 0$). for updating $\mathbf{w}$ given a fixed $\mathbf{p}$, the training process of learning model only performs on the ”easy” samples selected before. Small values of λ, only pass ”easy” samples with small losses. With gradually increasing λ, larger loss values for ”complex” samples are accepted.

3. Self-Paced Multi-Label Learning

Suppose we are given an input matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ with d-dimensional feature space and $\mathbf{Y} \in \{-1, +1\}^{l\times n}$ be the finite set of l existing labels. Let $\mathcal{D} = (\mathbf{x}_j, \mathbf{y}_j)_{j=1}^{n}$ be a multi-label data, where $\mathbf{x}_j = [x_{1j}, ..., x_{dj}]$ is an arbitrary vector of features with d-dimensions of the jth sample and $\mathbf{y}_j = [y_{1j}, ..., y_{lj}]$ is the interpretation of labels for $\mathbf{x}_j$ and y_{ij} is $+1$ if the ith label is assigned to $\mathbf{x}_j$ and -1 otherwise. Multi-label learning intends to train a predictor $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$ from the training data $\mathcal{D}$, so relevant and irrelevant labels of out-of-sample data are predicted. The general objective function for multi-label learning is:

$$\min_{\mathbf{W}} \sum_{i=1}^{l} \sum_{j=1}^{n} \mathcal{L}(y_{ij}, \mathbf{w}_i^\top \mathbf{x}_j) + \Phi(\mathbf{W}, \mathbf{C}) \tag{3}$$

where $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_l\} \in \mathcal{R}^{d\times l}$ is the learned matrix with each column indicating the weight vector for each independent label. $\mathcal{L}(y_{ij}, \mathbf{w}_i^\top \mathbf{x}_j)$ is the loss of jth sample for the ith label. Correlation matrix $\mathbf{C}$ demonstrates dependency degree between each pair of labels and $\Phi(\mathbf{W}, \mathbf{C})$ is a correlation regularizer with characteristic to let labels with positive dependency degrees encourage their corresponding outputs to be closer, and vice versa.

The aforementioned objective function treats all labels identically and does similar for samples per label. Although in many real-world cases, different labels don't necessarily have identical complexities, and also samples have different complexity levels for labels. Generally, the non-convex objective function of multi-label learning is the potential to get stuck in local optima Zhao et al. (2015), especially with the presence of bad initialization or noisy and corrupted labels. To address these defects, by defining a self-paced regularization, the model can learn a sequence of instances with respect to their degree of complexity.

$$\begin{aligned} \min_{\mathbf{W},\ \mathbf{P}} & \sum_{i=1}^{l} \sum_{j=1}^{n} p_{ij} \mathcal{L}(y_{ij}, \mathbf{w}_i^\top \mathbf{x}_j) + \Phi(\mathbf{W}, \mathbf{C}) + \Psi(\lambda, \gamma, \mathbf{P}) \\ \text{s.t.} \quad & \mathbf{P} \in [0,1]^{l\times n}, \end{aligned} \tag{4}$$

Furthermore, our desirable multi-label learning is expected to learn not only easy but also diverse labels that are sufficiently disparate from the current learning pace. To this end, instance-label weights are introduced. In order to accomplish the easy-to-hard strategy on diverse labels simultaneously, we propose a new self-paced regularizer in (5):

$$\Psi(\lambda, \gamma, \mathbf{P}) = -\lambda \sum_{i=1}^{l} \sum_{j=1}^{n} p_{ij} + \gamma \sum_{i=1}^{l} \left\| \mathbf{p}^{(i)} \right\|_2 \tag{5}$$

Equation (5) consists of two terms, a negative l_1-norm and an adaptive $l_{2,1}$-norm of a matrix. The first term induces a preference to select the easy instances rather than the hard ones per label. Combining this term with (4), implies that small empirical loss $\mathcal{L}$ on the training data point $(\mathbf{x}_j, y_{ij})$ drives the weight p_{ij} to be high. Hence, this optimization process well corresponds with the intuitive notion of starting with the easiest instances (the ones that have the lowest empirical errors). By progressively increasing λ while the learning proceeds, the self-paced weights will increasingly grow higher in consequent. This leads to gradual involvement of more complex samples into training. The $l_{2,1}$-matrix norm term leads to label-wise sparsity. It favors selecting from different categories of labels in the initial steps with higher diversities. As the training proceeds, with gradually decreasing γ, the impact of diversity decreases. By plugging (5) into (4), we obtain the final objective function:

$$\begin{aligned} \min_{\mathbf{W},\ \mathbf{P}} & \sum_{i=1}^{l} \sum_{j=1}^{n} p_{ij} \mathcal{L}(y_{ij}, \mathbf{w}_i^\top \mathbf{x}_j) + \Phi(\mathbf{W}, \mathbf{C}) - \lambda \sum_{i=1}^{l} \sum_{j=1}^{n} p_{ij} + \gamma \sum_{i=1}^{l} \left\| \mathbf{p}^{(i)} \right\|_2 \\ \text{s.t.}\ & \mathbf{p}^{(i)} \in [0,1]^n \end{aligned} \tag{6}$$

3.1. SPMLD with Local and Global Correlation

We give an example to motivate our self-paced learning procedure. We briefly discuss how the proposed algorithm develops the learning pace of the original problem. Host method (GLOCAL) Zhu et al. (2018) simultaneously recovers the missing-labels, trains the learner and exploits both global and local correlations among labels without needing any further prior knowledge, through learning a latent label representation. The objective function of GLOCAL is as follows:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\mathbf{W},\mathbf{Z}} & \left\|\mathbf{J}\circ(\mathbf{Y}-\mathbf{U}\mathbf{V})\right\|_F^2+\alpha\left\|(\mathbf{V}-\mathbf{W}^\top\mathbf{X})\right\|_F^2 \\ & +\sum_{b=1}^{g}[\frac{\beta_1 n_b}{n}\mathrm{tr}(\mathbf{F}^\top\mathbf{Z}_b\mathbf{Z}_b^\top\mathbf{F})+\beta_2\mathrm{tr}(\mathbf{F}_b^\top\mathbf{Z}_b\mathbf{Z}_b^\top\mathbf{F}_b)] \\ \text{s.t.} \quad & \mathrm{diag}(\mathbf{Z}_b\mathbf{Z}_b^\top)=\mathbf{1},\ b=1,2,...,g.\end{aligned} \tag{7}$$

where $\mathbf{V}$ stands for the matrix of the latent labels capturing concepts in a higher level which are more compact and semantically abstract than the original labels; while $\mathbf{U}$ represents the matrix containing the interactions between the original labels and the latent labels. In general, labels may only be partially observed. Low-rank representation is one of the key techniques in matrix completion, and the low-rank decomposition of the observed labels yields a natural solution to recover missing-labels (Let $\mathbf{J}$ be the indicator matrix of the observed labels in $\mathbf{Y}$).

Label correlations may not have the same values in different categories, so we define the local manifold regularization. Assume that the dataset $\mathbf{X}$ is partitioned into b groups $\{\mathbf{X}_1,\dots,\mathbf{X}_b\}$, where $\mathbf{X}_b\in\mathbb{R}^{d\times n_b}$ has n_b instances. This partitioning can be obtained by clustering. If $\mathbf{Y}_b$ is the label submatrix in $\mathbf{Y}$ corresponding to $\mathbf{X}_b$, then $\mathbf{C}_b\in\mathbb{R}^{l\times l}$ are the local correlation of group b. Similar to global label correlations, we force the output to be similar or dissimilar on the relevant or irrelevant correlated labels, and optimize $\mathrm{tr}(\mathbf{F}_b^\top\ \mathbf{L}_b\mathbf{F}_b)$, where $\mathbf{L}_b=\mathbf{Z}_b\mathbf{Z}_b^\top$ is the Laplacian of $\mathbf{C}_b$ that stores our knowledge about the relationship of our labels and is defined as $\mathbf{L}_b=\mathbf{D}_b-\mathbf{C}_b$ where $\mathbf{D}_b$ is a diagonal matrix with $d_{i,i}=\sum_{j=1}^{n}c_{i,j}$ and $\mathbf{F}_b=\mathbf{U}\mathbf{W}^\top\mathbf{X}_b$ is the predicted submatrix for group b.
Finally, We propose the following objective function by considering the diversity of labels in a unified setting:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V},\mathbf{W},\mathbf{Z},\mathbf{P}} & \left\|\mathbf{J}\circ(\mathbf{Y}-\mathbf{U}\mathbf{V})\right\|_F^2+\alpha\left\|\sqrt{\mathbf{P}}\circ(\mathbf{V}-\mathbf{W}^\top\mathbf{X})\right\|_F^2 \\ & +\sum_{b=1}^{g}[\frac{\beta_1 n_b}{n}\mathrm{tr}(\mathbf{F}^\top\mathbf{Z}_b\mathbf{Z}_b^\top\mathbf{F})+\beta_2\mathrm{tr}(\mathbf{F}_b^\top\mathbf{Z}_b\mathbf{Z}_b^\top\mathbf{F}_b)] \\ & -\lambda\sum_{i=1}^{l}\left\|\mathbf{P}^{(i)}\right\|_1+\gamma\left\|\mathbf{P}^\top\right\|_{2,1}+\tau\mathcal{R}(\mathbf{U},\mathbf{V},\mathbf{W}) \\ \text{s.t.} \quad & \mathbf{P}\in[0,1]^{l\times n},\mathrm{diag}(\mathbf{Z}_b\mathbf{Z}_b^\top)=\mathbf{1},\ b=1,2,...,g.\end{aligned} \tag{8}$$

where $\sqrt{\mathbf{P}}$ denotes the element-wise square root of $\mathbf{P}$, the $\circ$ (element-wise product) of matrices and $\mathcal{R}(\mathbf{U},\mathbf{V},\mathbf{W})=\left\|\mathbf{U}\right\|_F^2+\left\|\mathbf{V}\right\|_F^2+\left\|\mathbf{W}\right\|_F^2$ is the regularization term to guarantee generalization ability and numerical stability.

Algorithm 1 SPMLD on GLOCAL

Input: data matrix $\mathbf{X}$, label matrix $\mathbf{Y}$, Observation indicator matrix $\mathbf{J}$, and the group partition
Parameter: $\alpha, \beta_1, \beta_2, \tau, \lambda, \gamma$
Output: $\mathbf{U}$ and $\mathbf{W}$

```
Initialize $\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{Z}$;
while convergence not reached do
    for $b = 1, 2, ..., g.$ do
        fix $\mathbf{P}, \mathbf{U}, \mathbf{V}, \mathbf{W}$, update $\mathbf{Z}_b$ according to (11);
    end for
    fix $\mathbf{P}, \mathbf{U}, \mathbf{W}, \mathbf{Z}$, update $\mathbf{V}$ according to (12);
    fix $\mathbf{P}, \mathbf{V}, \mathbf{W}, \mathbf{Z}$, update $\mathbf{U}$ according to (13);
    fix $\mathbf{P}, \mathbf{U}, \mathbf{V}, \mathbf{Z}$, update $\mathbf{W}$ according to (14);
    fix $\mathbf{U}, \mathbf{V}, \mathbf{W}, \mathbf{Z}$, update $\mathbf{P}$ according to (10);
    $\lambda = \lambda\mu_1; \gamma = \gamma\mu_2$;
end while;
return $\mathbf{U}$ and $\mathbf{W}$.
```

The details of self-paced multi-label learning with diversity (SPMLD) is summarized in Algorithm 1. Implementation is available on GitHub repository[1].

3.2. Optimization

In this section, we discuss how to solve (8) by alternating minimization which gives us capability to tune the variables iteratively and find an optimal solution. It is difficult to find a global optimal answer for this non-convex objective function. To achieve the reliable self-paced weights, we extend a block coordinate descent optimizer. For solving block $\mathbf{p}_{t+1}$ with fixed blocks $\mathbf{U}_t$,$\mathbf{V}_t$,$\mathbf{W}_t$ and $\mathbf{Z}_t$, the optimization problem can be decomposed to k sub-problems for k latent labels, respectively. Thus, objective function of the i-th label, $\mathbf{y}_i$ is given by:

$$\min_{\mathbf{p}^{(i)}} \mathbf{p}^{(i)}\mathcal{L}_t^{(i)} - \lambda \left\|\mathbf{p}^{(i)}\right\|_1 + \gamma \left\|\mathbf{p}^{(i)}\right\|_2, \text{ s.t. } \mathbf{p}^{(i)} \in [0,1]^n, \tag{9}$$

In order to solve (9), we first assume $\mathcal{L}_{1,t}^{(i)} \leq \mathcal{L}_{2,t}^{(i)} \leq ... \leq \mathcal{L}_{n,t}^{(i)}$. Let $r_t^{(i)} = \sum_{\theta_1 < j < \theta_2} (\lambda - \mathcal{L}_{j,t}^{(i)})^2$ and $s_t^{(i)} = \sum_{\theta_1 < j < \theta_2} (\lambda - \mathcal{L}_{j,t}^{(i)})$. For each i and arbitrary $\theta_2 > \theta_1$we define $c_t^*(\theta_1, \theta_2)$, $L_t(\theta_1, \theta_2)$, $G_{i,t}^*$ and $H_{i,t}^*$ for later computation:

1.

$$c_t^*(\theta_1, \theta_2) = \begin{cases} \sqrt{\theta_1/(\gamma^2 - r_t^{(i)})} & , \gamma^2 \neq r_t^{(i)} \\ (\lambda - \mathcal{L}_{\theta_1+1}^{(i)}) & , \gamma^2 = r_t^{(i)}, \gamma^2 < s_t^{(i)} \\ 0 & , \gamma^2 = r_t^{(i)}, \gamma^2 \geq s_t^{(i)}. \end{cases}$$

[1] http://github.com/amjadseyedi/SPMLD

2.

$$L_t(\theta_1,\theta_2) = \sum_{j=1}^{\theta_1} \mathcal{L}_{j,t}^{(i)} - \lambda(\theta_1 + c_t^*(\theta_1,\theta_2)s_t^{(i)}) + \gamma\sqrt{\theta_1 + c_t^*(\theta_1,\theta_2)^2 r_t^{(i)}}.$$

3. $G_{i,t}^*$ be the smallest j such that $\mathcal{L}_{j,t}^{(i)} \geq \lambda$.

4. $H_{i,t}^*$ be the largest j such that $\mathcal{L}_{j,t}^{(i)} \leq \lambda - \gamma$.

Let $\theta_2 = G_{i,t}^*$, and θ_1 be obtained by optimizing the following objective function:

$$\theta_1 = \arg \min_{H_{i,t}^* \leq \theta_1 < \theta_1} L_t(\theta_1,\theta_2)$$

Then, the optimal $\mathbf{p}_{t+1}^{(i)}$ is given by

$$p_{j,t+1}^{(i)} = \begin{cases} 1, & j \leq \theta_1 \\ 0, & j \geq \theta_2 \\ c_t^*(\theta_1,\theta_2)(\lambda - \mathcal{L}_{j,t}^{(i)}), & \theta_1 < j < \theta_2 \end{cases} \tag{10}$$

Then, we discuss update procedures for $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{Z}$. To optimize these variables with the gradient descent method, we utilize the MANOPT toolbox[2] Boumal et al. (2014) with line search on the Euclidian and manifold spaces.

Updating $\mathbf{Z}_b$'s: With $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{p}_k$'s fixed: for each $b \in \{1,...,g\}$. Due to the constraint $\mathrm{diag}(\mathbf{Z}_b\mathbf{Z}_b^\top) = \mathbf{1}$, it has no closed-form solution, and we solve it with projected gradient descent. The gradient of the objective w.r.t. $\mathbf{Z}_b$ is

$$\nabla_{\mathbf{Z}_b} = \frac{\beta_1 n_b}{n}\mathbf{U}\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\mathbf{U}^\top\mathbf{Z}_b + \beta_2\mathbf{U}\mathbf{W}^\top\mathbf{X}_b\mathbf{X}_b^\top\mathbf{W}\mathbf{U}^\top\mathbf{Z}_b \tag{11}$$

To satisfy the constraint $\mathrm{diag}(\mathbf{Z}_b\mathbf{Z}_b^\top) = \mathbf{1}$, we project each row of $\mathbf{Z}_b$ onto the unit norm ball $\mathbf{z}_{b,j,:} \leftarrow \mathbf{z}_{b,j,:}/\|\mathbf{z}_{b,j,:}\|_2$ after each update. where $\mathbf{z}_{b,j,:}$ is the jth row of $\mathbf{Z}_b$.

Updating V: With $\mathbf{U}$, $\mathbf{W}$, $\mathbf{Z}_b$'s and $\mathbf{p}_k$'s fixed: The gradient of the objective in (8) w.r.t. $\mathbf{V}$ is

$$\nabla_{\mathbf{V}} = \mathbf{U}^\top(\mathbf{J}\circ(\mathbf{U}\mathbf{V} - \mathbf{Y})) + (\mathbf{P}\circ(\mathbf{V} - \mathbf{W}^\top\mathbf{X})) + \tau\mathbf{V} \tag{12}$$

Updating U: With $\mathbf{V}$, $\mathbf{W}$, $\mathbf{z}_b$'s and $\mathbf{p}_k$'s fixed: Again, we use gradient descent and the gradient w.r.t. $\mathbf{U}$ is:

$$\nabla_{\mathbf{U}} = (\mathbf{J}\circ(\mathbf{U}\mathbf{V} - \mathbf{Y}))\mathbf{V}^\top + \tau\mathbf{U} + \sum_{b=1}^{g}\mathbf{Z}_b\mathbf{Z}_b^\top\mathbf{U}[\frac{\beta_1 n_b}{n}\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W} + \beta_2\mathbf{W}^\top\mathbf{X}_b\mathbf{X}_b^\top\mathbf{W}] \tag{13}$$

Updating W: With $\mathbf{U}$, $\mathbf{V}$, $\mathbf{z}_b$'s and $\mathbf{p}_k$'s fixed: The gradient w.r.t. $\mathbf{W}$ is:

$$\nabla_{\mathbf{W}} = \alpha\mathbf{X}[(\mathbf{X}^\top\mathbf{W} - \mathbf{V}^\top)\circ\mathbf{P}^\top] + \tau\mathbf{W} + \sum_{b=1}^{g}(\frac{\beta_1 n_b}{n}\mathbf{X}\mathbf{X}^\top + \beta_2\mathbf{X}_b\mathbf{X}_b^\top)\mathbf{W}\mathbf{U}^\top\mathbf{Z}_b\mathbf{Z}_b^\top\mathbf{U} \tag{14}$$

[2] http://www.manopt.org

Table 1: Statistical characteristics of the real-world multi-label datasets.

dataset	#instance	#dimension	#label	#label/instance
Business Ueda and Saito (2003)	5,000	438	30	1.59
Computers Ueda and Saito (2003)	5,000	681	33	1.5
Education Ueda and Saito (2003)	5,000	550	33	1.46
Health Ueda and Saito (2003)	5,000	612	32	1.66
Science Ueda and Saito (2003)	5,000	743	40	1.45
Social Ueda and Saito (2003)	5,000	1,047	39	1.28
Corel5K Duygulu et al. (2002)	5,000	499	374	3.52

4. Experiments

In this section, empirical experiments are conducted to test the validation of our method. In these experiments, seven real-world multi-label datasets including Yahoo text datasets (*Business, Computers, Education, Health, Science and Social*) Ueda and Saito (2003) along with an image classification data (Corel5K) Duygulu et al. (2002) are used[3].

4.1. Experimental Setting

Table 1 lists detailed characteristics of the employed datasets. Each column sequentially represents the number of features, number of instances, number of labels and label per instance ratio for each dataset.

Since prediction in the presence of missing-labels is a more challenging task, we have performed the experiments on missing-label data. We randomly sample $\rho\%$ of the elements in the label matrix as observed, and the rest as missing. ρ is set to 30% and 70% revealed entries, respectively. Coverage, Ranking loss, Average AUC, Instance AUC, MacroF1, MicroF1, and InstanceF1 evaluation metrics are used to measure the performance of the proposed predictive model against baselines. The above-mentioned metrics analyze different aspects of multi-label learning algorithms. The first two metrics are to be minimized and the other metrics are to be maximized according to Wu and Zhou (2017).

For examining the effectiveness of our framework, it is compared with three state-of-the-art multi-label learning algorithms namely LEML Yu et al. (2014), ML-LRC Xu et al. (2014), GLOCAL Zhu et al. (2018). These Baseline methods along with the SPMLD have two common traits, which are the main reasons that we compare our framework with them. All the above-mentioned methods somehow learn in a latent subspace and furthermore, they all handle the missing-label challenge.

- Low-rank empirical risk minimization for multi-label learning (LEML) has a linear low-rank structure to train a model for mapping from instance to label space and utilizes an implicit concept of global label dependency.
- Learning low-rank label correlations for multi-label classification (ML-LRC) learns and exploits low-rank global label correlations for multi-label classification.

[3] http://mulan.sourceforge.net/datasets-mlc.html

- Multi-label learning with global and local label correlation (GLOCAL) learns label correlations in a new manner. It considers both local and global label correlations in latent label representation.

What's more, to statistically measure the significance of performance difference, pairwise t-tests at 95% significance level are conducted between SPMLD and each of the baseline algorithm. Therefore, in the statistical peer test, for each test, the performance of an algorithm is bold-faced denoting that it statistically outperforms the other one. Furthermore, when there is no significant difference between the performance of SPMLD compared to one or more of the baselines on a dataset regarding a specific evaluation metric, their results are shown in italic. Furthermore, self-paced regularization parameter λ controls the complexity of instances and labels. It first considers the least losses corresponding to the easiest samples, then gradually involves harder ones through iterations. λ and its increase ratio μ_1 are searched from $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and $\{1.1, 1.2, 1.3, 1.4, 1.5\}$, respectively. Diversity regularization parameter γ controls to select less similar instances and labels specified as easy. γ needs to be higher in the initial iterations. γ and its decrease ratio μ_2 are tuned using a grid search in ranges [1-10] and $\{0.95, 0.9, 0.85, 0.8, 0.75, 0.7\}$, respectively. Parameters of the host algorithm (GLOCAL) and parameters of the competing methods are all set the same as recommended in the corresponding literature.

4.2. Results on Real-World Datasets

All the results reported in tables and charts are averaged over 10 independent runs. Tables (2-7) exhibit the obtained label prediction results of the multi-label algorithms on six of the mentioned datasets, respectively. Overall, results of all algorithms have improved on 70% observation of samples. In each table three metrics are reported, regarding two different settings of ρ there are six probable cases. On five datasets; *Business*, *Education*, *Health*, *Science* and *Social*, SPMLD has significantly better performance regarding all the measures (and for both observation settings) reported in Tables (2, 4-6), respectively. On *Computers* dataset whose results are shown in Table 3. SPMLD reports significantly better results regarding Rkl measure for both ρ settings. Similarly, it shows statistically better AUC and COV values for 30% entries revealed, while in the case of 70% LEML, ML-LRC and SPMLD show no significant different AUC values compared to each other but they have jointly better AUC values than the GLOCAL. Again, in the case of COV values for 70% entries, the proposed method shows no significant difference compared with ML-LRC but, it gains statistically better results than the two other models. Equivalently, on *Social* dataset SPMLD shows significantly better performance regarding both percentages of observation for Rkl and AUC metrics and also COV with ρ=30% and only for the proportion of 70% it statistically performs equal to GLOCAL while they get jointly better results than the remaining two algorithms and jointly stand at the first place. Subsequently, as the summary of tables, SPMLD ranks first in 91.66% cases according to statistical significance test. Moreover, it shows statistically equal performance in 8.34% cases where it still ranks first but jointly with one ore more of the baselines.

Results obtained on *Corel5K* image dataset are analyzed through a Radar plot which enables one to compare models against multiple metrics. In this straightforward analysis, SPMLD is compared to the baselines in the high-dimensional label space of *Corel5K* in

Table 2: Average performance on *Business* dataset: means and standard deviations over 10 independent runs. The best performance is highlighted in **bold** with respect to paired t-tests at 95% significance level between SPMLD and each baseline

Method	ρ	Rkl ↓	AUC ↑	COV ↓
LEML	30%	0.063 ± 0.0057	0.928 ± 0.0052	3.954 ± 0.2751
	70%	0.058 ± 0.0048	0.942 ± 0.0057	3.303 ± 0.2706
ML-LRC	30%	0.061 ± 0.0024	0.937 ± 0.0055	3.279 ± 0.0666
	70%	0.046 ± 0.0019	0.950 ± 0.0050	2.580 ± 0.0593
GLOCAL	30%	0.054 ± 0.0025	0.937 ± 0.0036	2.863 ± 0.1711
	70%	0.046 ± 0.0022	0.952 ± 0.0031	2.579 ± 0.1658
SPMLD	30%	**0.044 ± 0.0021**	**0.956 ± 0.0032**	**2.361± 0.1688**
	70%	**0.043 ± 0.0018**	**0.958 ± 0.0029**	**2.347± 0.1625**

Table 3: Average performance on *Computers* dataset: means and standard deviations over 10 independent runs. The best performance is highlighted in **bold** with respect to paired t-tests at 95% significance level between SPMLD and each baseline. *Italic* font indicates that SPMLD and the corresponding baselines show no significant difference.

Method	ρ	Rkl ↓	AUC ↑	COV ↓
LEML	30%	0.179 ± 0.0072	0.880 ± 0.0064	7.392 ± 0.2181
	70%	0.141 ± 0.0058	*0.894 ± 0.0069*	6.306 ± 0.2653
ML-LRC	30%	0.152 ± 0.0044	0.873 ± 0.0057	6.052 ± 0.1426
	70%	0.115 ± 0.0026	*0.895 ± 0.0058*	*5.000 ± 0.6228*
GLOCAL	30%	0.132 ± 0.0037	0.876 ± 0.0034	5.647 ± 0.7823
	70%	0.123 ± 0.0028	0.884 ± 0.0062	5.440 ± 0.5340
SPMLD	30%	**0.104 ± 0.0030**	**0.903 ± 0.0047**	**4.600 ± 0.5359**
	70%	**0.113 ± 0.0015**	*0.894 ± 0.0053*	*5.018 ± 0.5712*

Table 4: Average performance on *Education* dataset: means and standard deviations over 10 independent runs. The best performance is highlighted in **bold** with respect to paired t-tests at 95% significance level between SPMLD and each baseline.

Method	ρ	Rkl ↓	AUC ↑	COV ↓
LEML	30%	0.176 ± 0.0084	0.817 ± 0.0075	9.672 ± 0.4461
	70%	0.151 ± 0.0077	0.842 ± 0.0082	7.595 ± 0.5138
ML-LRC	30%	0.144 ± 0.0034	0.845 ± 0.0068	6.350 ± 0.2042
	70%	0.113 ± 0.0028	0.860 ± 0.0063	5.075 ± 0.1866
GLOCAL	30%	0.125 ± 0.0026	0.875 ± 0.0057	5.741 ± 0.2312
	70%	0.122 ± 0.0035	0.878 ± 0.0064	5.784 ± 0.2450
SPMLD	30%	**0.096 ± 0.0019**	**0.904 ± 0.0048**	**4.246 ± 0.1372**
	70%	**0.093 ± 0.0021**	**0.907 ± 0.0052**	**4.162 ± 0.1524**

Table 5: Average performance on *Health* dataset: means and standard deviations over 10 independent runs. The best performance is highlighted in **bold** with respect to paired t-tests at 95% significance level between SPMLD and each baseline.

Method	ρ	Rkl ↓	AUC ↑	COV ↓
LEML	30%	0.095 ± 0.0029	0.896 ± 0.0038	6.248 ± 0.1640
	70%	0.074 ± 0.0033	0.920 ± 0.0045	5.167 ± 0.1827
ML-LRC	30%	0.085 ± 0.0041	0.907 ± 0.0082	4.924 ± 0.1351
	70%	0.071 ± 0.0036	0.920 ± 0.0093	3.960 ± 0.1825
GLOCAL	30%	0.0828 ± 0.0018	0.919 ± 0.0057	4.438 ± 0.1357
	70%	0.0795 ± 0.0023	0.923 ± 0.0078	4.472 ± 0.1414
SPMLD	30%	**0.064 ± 0.0009**	**0.938 ± 0.0063**	**3.355 ± 0.1182**
	70%	**0.059 ± 0.0011**	**0.943 ± 0.0068**	**3.253 ± 0.1200**

Table 6: Average performance on *Science* dataset: means and standard deviations over 10 independent runs. The best performance is highlighted in **bold** with respect to paired t-tests at 95% significance level between SPMLD and each baseline.

Method	ρ	Rkl ↓	AUC ↑	COV ↓
LEML	30%	0.203 ± 0.0052	0.827 ± 0.0053	10.587 ± 0.2011
	70%	0.174 ± 0.0056	0.849 ± 0.0064	9.501 ± 0.2548
ML-LRC	30%	0.169 ± 0.0027	0.830 ± 0.0039	8.794 ± 0.1254
	70%	0.134 ± 0.0024	0.850 ± 0.0033	6.900 ± 0.1273
GLOCAL	30%	0.154 ± 0.0029	0.840 ± 0.0108	7.949 ± 0.1371
	70%	0.134 ± 0.0034	0.866 ± 0.0113	7.106 ± 0.1365
SPMLD	30%	**0.129 ± 0.0024**	**0.871 ± 0.0095**	**6.640 ± 0.1156**
	70%	**0.124 ± 0.0028**	**0.876 ± 0.0087**	**6.432 ± 0.1283**

Table 7: Average performance on *Social* dataset: means and standard deviations over 10 independent runs. The best performance is highlighted in **bold** with respect to paired t-tests at 95% significance level between SPMLD and each baseline. *Italic* font indicates that SPMLD and the corresponding baselines show no significant difference.

Method	ρ	Rkl ↓	AUC ↑	COV ↓
LEML	30%	0.128 ± 0.0067	0.872 ± 0.0065	5.459 ± 0.3084
	70%	0.081 ± 0.0061	0.919 ± 0.0069	3.824 ± 0.3011
ML-LRC	30%	0.123 ± 0.0059	0.877 ± 0.0057	5.167 ± 0.1034
	70%	0.073 ± 0.0052	0.928 ± 0.0046	3.608 ± 0.0975
GLOCAL	30%	0.102 ± 0.0054	0.898 ± 0.0058	4.496 ± 0.2627
	70%	0.073 ± 0.0049	0.929 ± 0.0052	*3.442 ± 0.2583*
SPMLD	30%	**0.068 ± 0.0055**	**0.931 ± 0.0052**	**3.563 ± 0.2554**
	70%	**0.065 ± 0.0048**	**0.934 ± 0.0044**	*3.456 ± 0.2505*

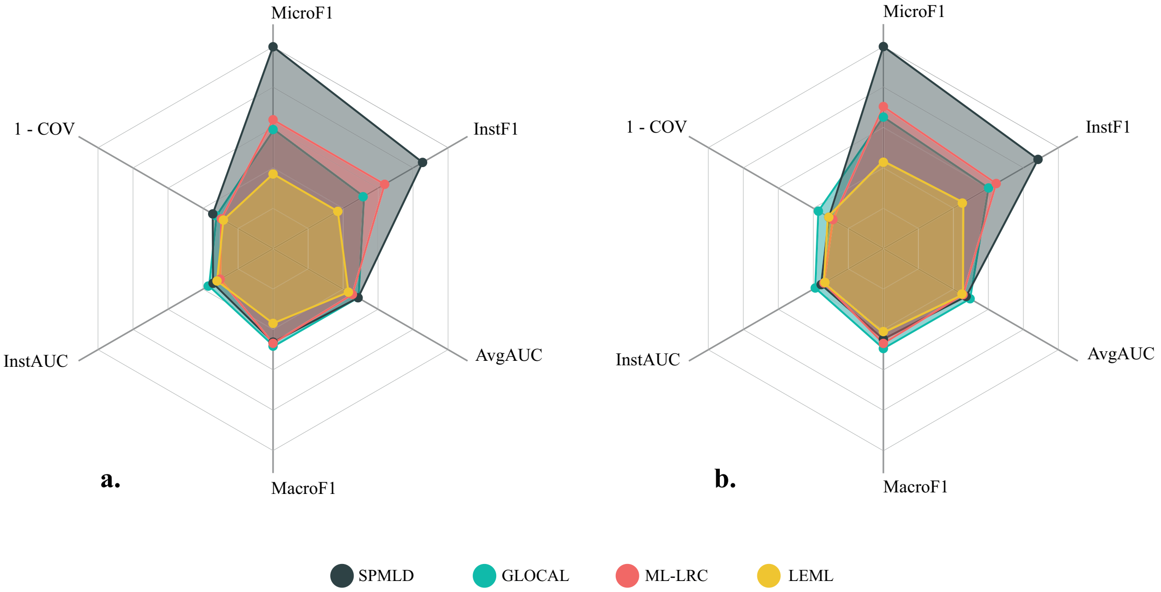

Figure 1: The comparison on Corel5K dataset with respect to several metrics (the coverage measure is normalized). a) 30% label rate , b) 70% label rate.

terms of six evaluation metrics for 30% and 70% revealed data, respectively and the results are shown in Figure 1. Note that, the proposed method endeavors to cover all labels fairly in its learning process. Thus, it is able to distinguish positive and negative labels of an instance by simultaneously making a larger label-wise margin and preserving instance-wise margin. Hence, it can be obviously seen that the SPMLD is far better on label-wise metrics (MicroF1 and InstanceF1) and it obtains comparable results on the other instance-wise metrics such as MacroF1.

4.3. Parameter Analysis

In this subsection, the influence of parameters on the proposed model is analyzed. According to (8) SPMLD has two parameters namely λ and γ which correspond to the self-paced and the diversity regularization terms, respectively. It must be mentioned that parameters of other regularization terms which belong to the base algorithm are analyzed in Zhu et al. (2018). Thus, to make a thorough study on the two mentioned parameters we evaluated them on all datasets through a grid-search strategy and reported the results based on three measures. In addition, each graph consists of 50 evaluations referring to 5 different λs and 10 γs. According to Figure 2, there is a bar next to each graph that indicates the highest and the lowest values obtained on the corresponding dataset regarding each measure which is shown using an spectrum of colors. Subsequently, Light colors (e.g. "orange" to "yellow") represent high amounts and dark colors (e.g. "blue" to "dark-blue") represent low amounts for the measures.

It can be seen that for each value of λ changing the values of γ from 1 to 10 makes a

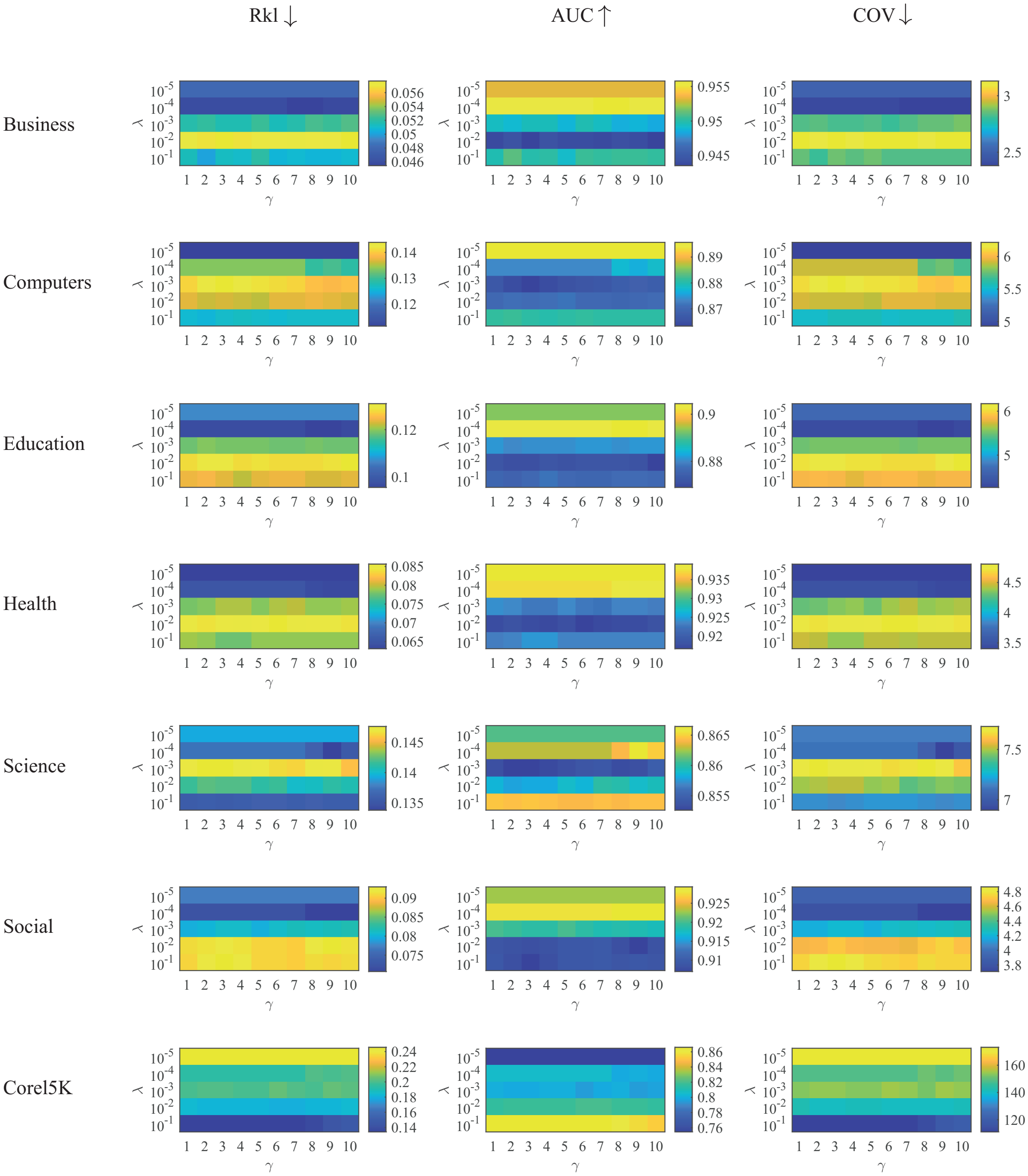

Figure 2: Analysis of influence of λ and γ on SPMLD for 30% data revealed.

significant difference except for $\lambda=10^{-5}$ which lies on the top row of graphs and stays unchanged with increasing γ.

5. Conclusion

In this paper, we propose a novel Self-Paced framework for Multi-Label learning. This framework incorporates the complexity of both instances and labels, and trains its predictive model with gradual involvement of harder samples. It also utilizes an efficient Diversity maintenance mechanism to avoid biasing over a limited subset of labels. The diverse easy-to-hard learning strategy has also an implicit positive effect on correlations exploited. SPMLD is applied to correlation-based multi-label learning as a host algorithm. Experiments on real-world datasets verify the effectiveness of SPMLD compared to the host algorithm itself and two other state-of-the-art methods. For future studies, it is desirable to investigate the direct impact of self-paced regularization on correlation exploitation and we intend to study and analyze the effect of diversity on local dependencies.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, 2009.

Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15 (1):1455–1459, 2014.

Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision*, pages 97–112, 2002.

Sheng-Jun Huang and Zhi-Hua Zhou. Multi-label learning by exploiting label correlations locally. In *Twenty-sixth AAAI conference on artificial intelligence*, pages 949–955, 2012.

Sheng-Jun Huang, Wei Gao, and Zhi-Hua Zhou. Fast multi-instance multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems*, pages 2078–2086, 2014.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

Changsheng Li, Junchi Yan, Fan Wei, Weishan Dong, Qingshan Liu, and Hongyuan Zha. Self-paced multi-task learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 2175–2181, 2017.

Changsheng Li, Fan Wei, Junchi Yan, Xiaoyu Zhang, Qingshan Liu, and Hongyuan Zha. A self-paced regularization framework for multilabel learning. *IEEE transactions on neural networks and learning systems*, 29(6):2660–2666, 2018.

Hao Li and Maoguo Gong. Self-paced convolutional neural networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2110–2116, 2017.

Liang Lin, Keze Wang, Deyu Meng, Wangmeng Zuo, and Lei Zhang. Active self-paced learning for cost-effective and progressive face identification. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):7–19, 2018.

Deyu Meng, Qian Zhao, and Lu Jiang. A theoretical understanding of self-paced learning. *Information Sciences*, 414:319–328, 2017.

Keerthiram Murugesan et al. Self-paced multitask learning with shared knowledge. In *Proceedings of the Twenty-sixth International Joint Conference on Artificial Intelligence*, pages 2522–2528, 2017.

Te Pi, Xi Li, Zhongfei Zhang, Deyu Meng, Fei Wu, Jun Xiao, and Yueting Zhuang. Self-paced boost learning for classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1932–1938, 2016.

Enver Sangineto, Moin Nabi, Dubravko Culibrk, and Nicu Sebe. Self paced deep learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):712–725, 2019.

Naonori Ueda and Kazumi Saito. Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems*, pages 737–744, 2003.

Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3780–3788, 2017.

Chang Xu, Dacheng Tao, and Chao Xu. Multi-view self-paced learning for clustering. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 3974–3980, 2015.

Linli Xu, Zhen Wang, Zefan Shen, Yubo Wang, and Enhong Chen. Learning low-rank label correlations for multi-label classification with missing labels. In *IEEE International Conference on Data Mining*, pages 1067–1072, 2014.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in neural information processing systems*, pages 2301–2309, 2013.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31st International Conference on Machine Learning*, pages 593–601, 2014.

Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):865–878, 2017.

Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

Qian Zhao, Deyu Meng, Lu Jiang, Qi Xie, Zongben Xu, and Alexander G Hauptmann. Self-paced learning for matrix factorization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3196–3202, 2015.

Yue Zhu, James T. Kwok, and Zhi-Hua Zhou. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2018.