Estimating latent relative labeling importances for multi-label learning

He, Shuo; Feng, Lei; Li, Li

2018

He, S., Feng, L., & Li, L. (2018). Estimating latent relative labeling importances for multi-label learning. Proceedings of 2018 IEEE International Conference on Data Mining (ICDM), 1013-1018. doi:10.1109/ICDM.2018.00127

<https://hdl.handle.net/10356/143866>

<https://doi.org/10.1109/ICDM.2018.00127>

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at:
<https://doi.org/10.1109/ICDM.2018.00127>.

Downloaded on 26 Apr 2023 23:33:11 SGT

Estimating Latent Relative Labeling Importances for Multi-Label Learning

Shuo He¹, Lei Feng², Li Li¹

¹College of Computer and Information Science, Southwest University, Chongqing, China

²School of Computer Science and Engineering, Nanyang Technological University, Singapore

Email: hs8207083890@email.swu.edu.cn, feng0093@e.ntu.edu.sg, lily@swu.edu.cn

Abstract—In multi-label learning, each instance is associated with multiple labels simultaneously. Most of the existing approaches directly treat each label in a crisp manner, i.e. one class label is either relevant or irrelevant to the instance. However, the latent relative importance of each relevant label is regrettably ignored. In this paper, we propose a novel multi-label learning approach that aims to estimate the latent labeling importances while training the inductive model simultaneously. Specifically, we present a biconvex formulation with both instance and label graph regularization, and solve this problem using an alternating way. On the one hand, the inductive model is trained by minimizing the least squares loss of fitting the latent relative labeling importances. On the other hand, the latent relative labeling importances are estimated by the modeling outputs via a specially constrained label propagation procedure. Through the mutual adaption of the inductive model training and the specially constrained label propagation, an effective multi-label learning model is therefore built by optimally estimating the latent relative labeling importances. Extensive experimental results clearly show the effectiveness of the proposed approach.

Index Terms—multi-label learning; relevant labels; latent relative labeling importances

I. INTRODUCTION

In multi-label classification tasks, each example is associated with multiple labels simultaneously [1]. So far, multi-label learning has been applied successfully in various application domains, such as image [2], audio [3], text [4], [5], etc. Formally speaking, suppose the given training set is denoted by $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \mid 1 \leq i \leq m\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of the i -th instance and $\mathbf{y}_i \in \{-1, +1\}^l$ is the corresponding label vector, and each element y_{ij} indicates whether the j -th label is relevant (+1) or irrelevant (-1) to the instance $\mathbf{x}_i$. Let $\mathcal{X} = \mathbb{R}^d$ be the d -dimensional feature space and $\mathcal{Y} = \{-1, 1\}^l$ be the l -dimensional categorical label space, the task of multi-label learning is to learn a predictor function $f: \mathcal{X} \rightarrow \mathcal{Y}$, to effectively predict the label vector of unseen instances.

For existing multi-label datasets, the accessible label space is categorical, which fully depicts the relevance or irrelevance of each label to an instance. Therefore, a large number of traditional approaches directly learn from such label space by taking the common assumption of equal labeling importances, i.e. each relevant label assigned to the instance is equally important. However, the importance of each relevant label is normally different in reality, since each relevant label normally makes different contributions to the instance.

To extract such relative labeling importances, a few recent studies resort to numeric labeling information to enrich implicit semantics, which can induce more effective multi-label learning approaches [6]–[8]. However, there are still some issues that these approaches fail to solve. First, only instance graph regularization is exploited to derive a numeric label space, while label correlations in the original categorical label space are discarded. Second, the process of numeric labeling importance matrix estimating and the model training is decoupled. This may result in degraded performance, due to the lack of the mutual adaption of the two tasks. Third, these approaches normally assume that the relevant labels follow the paradigm of *label distribution* [9], which makes the numeric value of each relevant label less than +1. Accordingly, the total strength of relevant labels is regrettably weakened.

To address the above problems, we propose a novel multi-label learning approach named RLIM, i.e. *estimating latent Relative Labeling Importances for Multi-label learning*. RLIM aims to estimate latent relative labeling importances while training the desired predictive model simultaneously. Specifically, a biconvex formulation with both instance and label graph regularization is presented, and an alternating optimization method is adopted to solve this biconvex problem. On the one hand, the inductive model is trained by minimizing the least squares loss of fitting the latent relative labeling importances. On the other hand, the latent relative labeling importances are estimated by the modeling outputs via a specially constrained label propagation procedure. Through the mutual adaption of the inductive model training and the specially constrained label propagation, an effective multi-label learning model is therefore built by optimally estimating the latent relative labeling importances.

This rest of this paper is organized as follows. Section II briefly reviews related work. Section III introduces the RLIM approach. Section IV reports experimental results of comparative studies. Finally, section V concludes this paper and discusses future research issues.

II. RELATED WORK

Existing multi-label approaches can be roughly categorized into three groups based on the *order of label correlations*. For the first-order strategy, label correlations are not considered and the multi-label problem is tackled in a label-by-label style [10], [11]. For the second-order strategy, pairwise relations

between labels is considered [12], [13]. For the high-order strategy, all other labels' influences imposed on each label are taken into account [14]–[16]. Note that these modeling strategies utilize the categorical labeling information in a crisp manner. In contrast, RLIM models high-order correlations by differentiating relative labeling importances.

Recently, there have been a few multi-label approaches which aim to transform the categorical labels to the numeric labels for deriving more meaningful auxiliary labeling information. In [7], based on the paradigm *label distribution learning* [9], relative labeling importance information is obtained by label propagation through the weighted fully-connected graph over training examples. And [6] characterizes the manifold in feature space by the weighted k -nearest neighbor graph defined over training examples, to limit the numeric label space to the label manifold in the Euclidean space. Similarly, another approach [8] enriches labeling information by exploiting the structure of feature space via sparse reconstruction.

Note that these approaches separate the numeric labeling information estimating from inductive model training while RLIM integrates the both into a unified framework. Besides, these approaches only utilize feature space to obtain numeric labeling information while RLIM additionally makes use of label correlations in the original label space. In addition, rather than using the basic assumption of label distributions [7], [9], RLIM poses more reasonable restrictions on the scale of numerical labels, thus the total strength of relevant labels remains intact.

III. THE RLIM APPROACH

As shown in section I, suppose we have a training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \mid 1 \leq i \leq m\}$, $\mathbf{X} \in \mathbb{R}^{m \times d}$ and $\mathbf{Y} \in \{-1, +1\}^{m \times l}$ are the instance matrix and the corresponding label matrix, respectively. Suppose the labeling importance vector of the i -th instance $\mathbf{x}_i$ is denoted by $\mathbf{p}_i \in \mathbb{R}^l$, the labeling importance matrix can be therefore denoted by $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m]^\top$. Integrating the estimation of $\mathbf{P}$ and model training, we propose a unified framework RLIM:

$$\min L(\mathbf{X}, \mathbf{P}, \mathbf{f}) + \lambda \Omega(\mathbf{f}) + \beta \Psi(\mathbf{P}) \quad (1)$$

where L denotes the specified loss function, Ω controls the complexity of the model $\mathbf{f}$ and Ψ guarantees a precise estimation of $\mathbf{P}$, and λ, β are trade-off parameters.

A. Instance and Label Graph Regularization

As shown in the above framework, the key is to estimate numeric labeling information effectively. To achieve this, we exploit both instance and label graph regularization. In this assumption, similar instances are assigned similar numeric label vectors, and similar labels share similar assignments on training instances. The instantiated formulation is given as follows:

$$\beta \Psi(\mathbf{P}) = \beta_1 \sum_{i,j} s_{ij} \|\mathbf{p}_i - \mathbf{p}_j\|^2 + \beta_2 \sum_{i,j} c_{ij} \|\mathbf{P}_i - \mathbf{P}_j\|^2 \quad (2)$$

where $\mathbf{p}_i$ and $\mathbf{P}_i$ are the i -th row and the i -th column of the relative labeling importance matrix $\mathbf{P}$, respectively. Which means, $\mathbf{p}_i$ denotes the labeling importance vector assigned to the i -th instance, $\mathbf{P}_i$ denotes the labeling importance vector of the i -th label. Besides, s_{ij} represents the similarity of the i -th instance $\mathbf{x}_i$ and the j -th instance $\mathbf{x}_j$, c_{ij} characterizes the correlation of the i -th label and the j -th label. β_1 and β_2 control the relative importance of the instance graph regularization and label graph regularization, respectively. Here, we denote by $\mathbf{S} = [s_{ij}]_{m \times m}$ the instance similarity matrix and $\mathbf{C} = [c_{ij}]_{l \times l}$ the label correlation matrix. The construction of $\mathbf{S}$ and $\mathbf{C}$ is presented as follows.

Let $G = (V, E)$ denote a k -NN graph, where V is the set of vertices with each of them representing an instance, E is the set of corresponding edges. We put a directed edge from $\mathbf{x}_j$ to $\mathbf{x}_i$ if $\mathbf{x}_j$ is a member of the k nearest neighbors of $\mathbf{x}_i$. Following this k -NN graph, the instance similarity matrix $\mathbf{S}$ is constructed as follows:

$$s_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\sigma^2}), & \text{if } \mathbf{x}_j \in \mathcal{N}_i \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $\mathcal{N}_i$ is the set that stores the k -nearest neighbors of $\mathbf{x}_i$ and the parameter σ is empirically set to $\sigma = \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{x}_{i_k}\| / m$ where $\mathbf{x}_{i_k}$ denotes the k -th nearest neighbor of $\mathbf{x}_i$. Finally, we set $\mathbf{S} = \mathbf{S} + \mathbf{S}^\top$ to ensure the symmetry of the instance similarity matrix [17].

To construct the label correlation matrix, we exploit the original categorical label space by using the cosine similarity:

$$c_{ij} = \frac{\mathbf{Y}_i^\top \mathbf{Y}_j}{\|\mathbf{Y}_i\| \|\mathbf{Y}_j\|} \quad (4)$$

where $\mathbf{Y}_i$ is the i -th column of the original categorical label space $\mathbf{Y}$, which corresponds to the label vector of the i -th label. To ensure that c_{ij} is no less than 0, we normalize $\mathbf{C}$ by $\mathbf{C} = (\mathbf{C} + 1)/2$. Note that previous approaches that extract numeric labeling information normally discard the original categorical label space and only rely on instance graph regularization. In contrast, RLIM utilizes both instance and label graph regularization, and label correlations derived from the original categorical label space can be further exploited.

To further optimally estimate the latent relative labeling importance matrix $\mathbf{P}$, we pose some constraints:

$$\begin{aligned} p_{ij} &\geq 0, \forall (i, j) \in \{(i, j) \mid y_{ij} = 1\} \\ \sum_j p_{ij} &= |\mathbf{y}_i|, \forall i \in [m], j \in \{j \mid y_{ij} = 1\} \\ -\mathbf{1}_{m \times l} &\leq \mathbf{P} \leq \mathbf{G} \end{aligned} \quad (5)$$

where $|\mathbf{y}_i|$ denotes the number of relevant labels for the i -th instance $\mathbf{x}_i$ and $\mathbf{G} = [g_{ij}]_{m \times l}$ is a matrix with $g_{ij} = |\mathbf{y}_i|$ if $y_{ij} = 1$, and $g_{ij} = -1$ if $y_{ij} = -1$, thus $\mathbf{G}$ indicates the whole scale of labeling importances. The first constraint guarantees that each relevant label must have positive labeling importance. The second constraint sets the total labeling importances of relevant labels to the total number of relevant labels, which controls the scale of the latent labeling importance matrix $\mathbf{P}$. Note that this setting significantly differs from that of existing

approaches, which traditionally follows the scheme of *label distribution* [7], [9]. By allowing the labeling importances of some relevant labels to be larger than +1, our setting retains the total labeling strength of relevant labels conveyed by the original categorical label space. The third constraint strictly confines that the latent labeling importance of each irrelevant label must be -1. This is crucial, since each irrelevant labels make no contributions to the corresponding instance. In summary, these three constraints strictly guarantees that the labeling importances of relevant labels are in the range from 0 to the total number of relevant labels, while the labeling importances of irrelevant labels are always -1.

B. Predictive Model Induction

Following the above proposed framework, we propose to train the model by minimizing the least squares loss of fitting the labeling importances:

$$L(\mathbf{X}, \mathbf{P}, \mathbf{f}) = \|\mathbf{z}_i \mathbf{W} + \mathbf{b}^\top - \mathbf{p}_i\|_2^2 \quad (6)$$

where $\mathbf{z}_i = \varphi(\mathbf{x}_i)$ is a nonlinear transformation of $\mathbf{x}_i$ to a higher dimensional feature space $\mathbb{R}^{\mathcal{H}}$. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_l] \in \mathbb{R}^{\mathcal{H} \times l}$ and $\mathbf{b} = [b_1, b_2, \dots, b_l]^\top \in \mathbb{R}^l$ are the model parameters. To control the model complexity, we utilize the widely-used squared Frobenius norm of $\mathbf{W}$:

$$\Omega(\mathbf{f}) = \|\mathbf{W}\|_F^2 \quad (7)$$

Finally, in order to easily introduce the kernel extension for nonlinear case, we present the formulation as a constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{b}, \mathbf{P}} \quad & \sum_{i=1}^m \|\mathbf{e}_i\|_2^2 + \lambda \|\mathbf{W}\|_F^2 \\ & + \beta_1 \sum_{i,j} s_{ij} \|\mathbf{p}_i - \mathbf{p}_j\|_2^2 + \beta_2 \sum_{i,j} c_{ij} \|\mathbf{P}_i - \mathbf{P}_j\|_2^2 \\ \text{s.t.} \quad & \mathbf{p}_i = \mathbf{z}_i \mathbf{W} + \mathbf{b}^\top + \mathbf{e}_i, \forall i \in [m], j \in [l] \\ & p_{ij} \geq 0, \forall (i, j) \in \{(i, j) | y_{ij} = 1\} \\ & \sum_j p_{ij} = |\mathbf{y}_i|, \forall i \in [m], j \in \{j | y_{ij} = 1\} \\ & -\mathbf{1}_{m \times l} \leq \mathbf{P} \leq \mathbf{G} \end{aligned} \quad (8)$$

C. Alternating Optimization

Obviously, the above optimization problem (8) is a biconvex problem [18], so that we can solve this problem in an alternating way. In each iteration, we choose to update one and leave another fixed. Firstly, we optimize the objective function w.r.t. $\mathbf{W}$ and $\mathbf{b}$ with $\mathbf{P}$ fixed, and then optimize the objective function w.r.t. $\mathbf{P}$ with $\mathbf{W}$ and $\mathbf{b}$ are both fixed.

1) *Optimizing $\mathbf{W}$ and $\mathbf{b}$* : When $\mathbf{P}$ is fixed, the problem (8) reduces to

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{b}} \quad & \text{tr}(\mathbf{E}^\top \mathbf{E}) + \lambda \text{tr}(\mathbf{W}^\top \mathbf{W}) \\ \text{s.t.} \quad & \mathbf{P} = \mathbf{Z}\mathbf{W} + \mathbf{1}_m \mathbf{b}^\top + \mathbf{E} \end{aligned} \quad (9)$$

where $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m]$, $\text{tr}(\cdot)$ is the trace norm operator and $\mathbf{1}_m$ is a vector with all components set to 1. Then, the Lagrangian of the problem (9) can be expressed as:

$$\begin{aligned} \Gamma(\mathbf{W}, \mathbf{b}, \mathbf{E}, \mathbf{A}) = & \text{tr}(\mathbf{E}^\top \mathbf{E}) + \lambda \text{tr}(\mathbf{W}^\top \mathbf{W}) \\ & - \text{tr}(\mathbf{A}^\top (\mathbf{Z}\mathbf{W} + \mathbf{1}_m \mathbf{b}^\top + \mathbf{E} - \mathbf{P})) \end{aligned} \quad (10)$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m]^\top \in \mathbb{R}^{m \times l}$ saves the Lagrangian multipliers. Thus, the gradients w.r.t. each variable are:

$$\begin{aligned} \nabla \mathbf{W} &= 2\lambda \mathbf{W} - \mathbf{Z}^\top \mathbf{A} \\ \nabla \mathbf{b} &= -\mathbf{A}^\top \mathbf{1}_m \\ \nabla \mathbf{E} &= 2\mathbf{E} - \mathbf{A} \\ \nabla \mathbf{A} &= \mathbf{Z}\mathbf{W} + \mathbf{1}_m \mathbf{b}^\top + \mathbf{E} - \mathbf{P} \end{aligned} \quad (11)$$

Let all of gradients (11) equal to $\mathbf{0}$, we can obtain the following compact equations:

$$\begin{aligned} \mathbf{A}^\top \mathbf{1}_m &= \mathbf{0}_l \\ \frac{1}{2\lambda} \mathbf{Z}\mathbf{Z}^\top \mathbf{A} + \mathbf{1}_m \mathbf{b}^\top + \frac{1}{2} \mathbf{A} &= \mathbf{P} \end{aligned} \quad (12)$$

For mathematical convenience, we define the positive definite matrix $\mathbf{H} = \frac{1}{2\lambda} \mathbf{K} + \frac{1}{2} \mathbf{I}_{m \times m}$, where $\mathbf{I}_{m \times m}$ is an identity matrix of size $m \times m$, and $\mathbf{K} = \mathbf{Z}\mathbf{Z}^\top \in \mathbb{R}^{m \times m}$ is a kernel matrix with its element $k_{ij} = \varphi(\mathbf{x}_i) \varphi(\mathbf{x}_j)^\top = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$. $\mathcal{K}(\cdot)$ is a kernel function, in this paper, Gaussian kernel $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / (2\sigma'^2))$ is employed, where σ' is the average distance of all pairwise instances. Then according to (12), we can obtain the following closed-form solutions:

$$\begin{aligned} \mathbf{b} &= \frac{\mathbf{1}_m^\top \mathbf{H}^{-1} \mathbf{P}}{\mathbf{1}_m^\top \mathbf{H}^{-1} \mathbf{1}_m} \\ \mathbf{A} &= \mathbf{H}^{-1} (\mathbf{P} - \mathbf{1}_m \mathbf{b}^\top) \end{aligned} \quad (13)$$

2) *Optimizing $\mathbf{P}$* : With $\mathbf{W}$ and $\mathbf{b}$ fixed, we can obtain the modeling output matrix $\mathbf{Q} = \mathbf{Z}\mathbf{W} + \mathbf{1}_m \mathbf{b}^\top = \frac{1}{2\lambda} \mathbf{K}\mathbf{A} + \mathbf{1}_m \mathbf{b}^\top$. By eliminating $\mathbf{E} = \mathbf{P} - \mathbf{Q}$, the problem (8) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{P}} \quad & \|\mathbf{P} - \mathbf{Q}\|_F^2 + \beta_1 \sum_{i,j} s_{ij} \|\mathbf{p}_i - \mathbf{p}_j\|_2^2 \\ & + \beta_2 \sum_{i,j} c_{ij} \|\mathbf{P}_i - \mathbf{P}_j\|_2^2 \\ \text{s.t.} \quad & p_{ij} \geq 0, \forall (i, j) \in \{(i, j) | y_{ij} = 1\} \\ & \sum_j p_{ij} = |\mathbf{y}_i|, \forall i \in [m], j \in \{j | y_{ij} = 1\} \\ & -\mathbf{1}_{m \times l} \leq \mathbf{P} \leq \mathbf{G} \end{aligned} \quad (14)$$

Actually, the first two terms coincides with the common label propagation problem [19], [20], where β_1 specifies the relative amount of labeling information from the nearest neighbors and the modeling outputs. The third term additionally incorporates label correlations, which is specially designed for multi-label learning. In addition, the first and the third constraints guarantee that the labeling importance information is consistently

Algorithm 1 The RLIM Algorithm**Inputs:**

$\mathcal{D}$: the multi-label training set $\{(\mathbf{x}_i, \mathbf{y}_i) \mid (1 \leq i \leq m)\}$
 k : the number of nearest neighbors
 $\mathbf{x}$: the unseen instance
 $\lambda, \beta_1, \beta_2$: the regularization parameters

Output:

$\mathbf{y}$: the predicted label vector for $\mathbf{x}$

Process:

- 1: construct the instance similarity matrix according to (3);
- 2: construct the label similarity matrix according to (4);
- 3: calculate the kernel matrix $\mathbf{K} = [\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)]_{m \times m}$;
- 4: initialize $\mathbf{P}$ as $\mathbf{Y}$;
- 5: **repeat**
- 6: update $\mathbf{b}$ and $\mathbf{A}$ according to (13);
- 7: update $\mathbf{Q} = \frac{1}{2\lambda} \mathbf{K} \mathbf{A} + \mathbf{1}_m \mathbf{b}^\top$;
- 8: calculate $\tilde{\mathbf{p}}$ by solving (16);
- 9: update $\mathbf{P}$ by reshaping $\tilde{\mathbf{p}} \in \mathbb{R}^{ml}$ into $\mathbf{P} \in \mathbb{R}^{m \times l}$;
- 10: **until** convergence or the maximum number of iterations.
- 11: return $\mathbf{y}$ according to (15).

propagated among relevant labels under their corresponding scales, while the labeling importances of irrelevant labels are fixed at -1. The second constraint controls the scale of the latent labeling importances. Furthermore, we use the modeling output matrix $\mathbf{Q}$ as the initial label matrix rather than the original categorical label matrix $\mathbf{Y}$ for each iteration.

Note that the optimization problem (14) is actually a quadratic programming problem (the detail derivation process is shown in appendix). In this way, (16) is a standard quadratic programming and can be efficiently solved by any off-the-shelf QP toolbox. Before alternating optimization, we initialize the matrix $\mathbf{P}$ as the original matrix $\mathbf{Y}$. After the optimization process, the predicted label vector $\mathbf{y} = [y_k] \in \{-1, +1\}^l$ for the unseen instance $\mathbf{x}$ is given as follows:

$$y_k = \begin{cases} +1, & \text{if } \mathbf{f}_k \geq 0, k \in [l] \\ -1, & \text{if } \mathbf{f}_k < 0, k \in [l] \end{cases} \quad (15)$$

Where $\mathbf{f}_k = \sum_{i=1}^m a_{ik} \mathcal{K}(\mathbf{x}, \mathbf{x}_i) + b_k$ is the output value of trained model on the k -th label. The pseudo code of RLIM is shown in Algorithm 1.

IV. EXPERIMENTS

A. Experimental Setup

1) *Data Sets*: We collect fifteen real-world data sets for comprehensive performance evaluation.¹ For each data set D , $|D|$, $\dim(D)$, $L(D)$, $F(D)$ and $LCard(D)$ are employed to represent its number of examples, number of features, number of class labels, feature type, label cardinality respectively. Table I summarizes detailed characteristics of the benchmark data sets, which are roughly organized in ascending order

TABLE I: CHARACTERISTICS OF THE BENCHMARK MULTI-LABEL DATA SETS.

Data set	$ D $	$\dim(D)$	$L(D)$	$F(D)$	$LCard(D)$
cal500	502	68	174	numeric	26.044
emotions	593	72	6	numeric	1.868
genbase	662	1185	27	nominal	1.25
medical	978	1449	45	nominal	1.245
enron	1702	1001	53	nominal	3.378
image	2000	294	5	numeric	1.236
scene	2407	294	5	numeric	1.074
yeast	2417	103	14	numeric	4.237
science	5000	743	40	numeric	1.45
arts	5000	462	26	numeric	1.64
rcv1-s1	6000	944	101	nominal	2.880
rcv1-s2	6000	944	101	nominal	2.634
rcv1-s3	6000	944	101	nominal	2.614
rcv1-s4	6000	944	101	nominal	2.484
rcv1-s5	6000	944	101	nominal	2.642

of $|D|$, with eight of them being regular-scale (first part, $|D| < 5,000$) and seven of them being large-scale (second part, $|D| \geq 5,000$).

2) *Comparing Algorithms*: In this paper, we choose to compare the performance of RLIM with five multi-label learning algorithms, including three well-established approaches *binary relevance* (BR) [10], *ensemble of classifier chains* (ECC) [15], *random k-labelsets* (RAkEL) [16] and two approaches based on numeric labeling information ML^2 [6] and MFLE [8].

Specifically, three well-established approaches i.e. BR, ECC, RAkEL, are implemented under the MULAN multi-label learning package [21]. For parameter configurations, ECC: ensemble size 10; RAkEL: ensemble size $2q$ with $k = 3$; ML^2 : balance parameter $\lambda = 1$, cost parameters C_1 and C_2 chosen among $\{1, 2, \dots, 10\}$; MLFE: parameters β_1 , β_2 and β_3 are chosen among $\{1, 2, \dots, 10\}$, $\{1, 10, 15\}$ and $\{1, 10\}$. For RLIM², $\lambda, \beta_1, \beta_2$ in Eqn. (8) are chosen among $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 5, 10\}$ respectively with 5-fold cross-validation on the training set.

Specifically, on each data set, 50% examples are randomly sampled without replacement to form the training set, and the rest 50% examples are used to form the test set. The sample process is repeated for ten times and the mean metric value as well as the standard deviation are recorded for comparative studies.

3) *Evaluation Metrics*: In this paper, we use seven popular metrics in multi-label learning, including five example-based metrics *one-error*, *hamming loss*, *coverage*, *ranking loss*, *average precision*, two label-based metrics *macro-averaging F1*, *micro-averaging F1*. Note that *coverage* is normalized by the number of class labels and values of these evaluation metrics all vary between $[0, 1]$. Concrete metric definitions can be found in [1]. Furthermore, For *one-error*, *coverage*, *hamming loss*, and *ranking loss*, the smaller the values the better the performance. For the other three metrics, the larger the values the better the performance.

¹Publicly available at <http://mulan.sourceforge.net/datasets.html>, <http://meka.sourceforge.net/#datasets>

²Code package for RLIM is available at: <https://github.com/LowKeyH/RLIM.git>

TABLE II: Predictive performance of each comparing algorithm (mean $\pm$ std. deviation) on the regular-scale data sets.

Comparing algorithms	One-error $\downarrow$							
	cal500	emotions	genbase	medical	enron	image	scene	yeast
RLIM	0.125$\pm$0.011	0.313 $\pm$ 0.030	0.001$\pm$0.002	0.134$\pm$0.019	0.218$\pm$0.010	0.260$\pm$0.009	0.186$\pm$0.011	0.238 $\pm$ 0.010
MLFE	0.140 $\pm$ 0.023	0.339 $\pm$ 0.023	0.002 $\pm$ 0.002	0.142 $\pm$ 0.010	0.222 $\pm$ 0.011	0.265 $\pm$ 0.019	0.189 $\pm$ 0.007	0.226$\pm$0.011
ML2	0.147 $\pm$ 0.018	0.343 $\pm$ 0.035	0.005 $\pm$ 0.006	0.139 $\pm$ 0.011	0.221 $\pm$ 0.011	0.279 $\pm$ 0.006	0.198 $\pm$ 0.010	0.247 $\pm$ 0.006
BR	0.759 $\pm$ 0.013	0.417 $\pm$ 0.018	0.006 $\pm$ 0.002	0.209 $\pm$ 0.006	0.402 $\pm$ 0.007	0.367 $\pm$ 0.007	0.325 $\pm$ 0.005	0.383 $\pm$ 0.018
ECC	0.223 $\pm$ 0.007	0.306$\pm$0.014	0.007 $\pm$ 0.003	0.158 $\pm$ 0.007	0.259 $\pm$ 0.007	0.322 $\pm$ 0.006	0.262 $\pm$ 0.005	0.260 $\pm$ 0.004
RAKEL	0.367 $\pm$ 0.011	0.314 $\pm$ 0.010	0.009 $\pm$ 0.001	0.205 $\pm$ 0.008	0.304 $\pm$ 0.008	0.334 $\pm$ 0.004	0.286 $\pm$ 0.003	0.294 $\pm$ 0.005
Comparing algorithms	Hamming loss $\downarrow$							
	cal500	emotions	genbase	medical	enron	image	scene	yeast
RLIM	0.137$\pm$0.001	0.223 $\pm$ 0.013	0.001$\pm$0.001	0.010$\pm$0.001	0.045$\pm$0.001	0.156$\pm$0.003	0.077$\pm$0.001	0.201$\pm$0.004
MLFE	0.140 $\pm$ 0.001	0.239 $\pm$ 0.009	0.001$\pm$0.001	0.011 $\pm$ 0.001	0.043 $\pm$ 0.001	0.159 $\pm$ 0.005	0.085 $\pm$ 0.002	0.205 $\pm$ 0.003
ML2	0.139 $\pm$ 0.002	0.235 $\pm$ 0.012	0.003 $\pm$ 0.001	0.010$\pm$0.001	0.048 $\pm$ 0.001	0.160 $\pm$ 0.003	0.081 $\pm$ 0.003	0.207 $\pm$ 0.004
BR	0.171 $\pm$ 0.002	0.262 $\pm$ 0.002	0.005 $\pm$ 0.001	0.011 $\pm$ 0.002	0.054 $\pm$ 0.003	0.236 $\pm$ 0.002	0.143 $\pm$ 0.001	0.259 $\pm$ 0.004
ECC	0.146 $\pm$ 0.001	0.211$\pm$0.003	0.002 $\pm$ 0.001	0.011 $\pm$ 0.001	0.049 $\pm$ 0.001	0.173 $\pm$ 0.002	0.098 $\pm$ 0.001	0.211 $\pm$ 0.002
RAKEL	0.170 $\pm$ 0.001	0.226 $\pm$ 0.002	0.002 $\pm$ 0.001	0.012 $\pm$ 0.001	0.051 $\pm$ 0.001	0.188 $\pm$ 0.001	0.107 $\pm$ 0.002	0.234 $\pm$ 0.002
Comparing algorithms	Coverage $\downarrow$							
	cal500	emotions	genbase	medical	enron	image	scene	yeast
RLIM	0.773$\pm$0.007	0.323$\pm$0.011	0.020 $\pm$ 0.006	0.038 $\pm$ 0.009	0.276 $\pm$ 0.006	0.176 $\pm$ 0.004	0.070$\pm$0.002	0.468 $\pm$ 0.003
MLFE	0.780 $\pm$ 0.009	0.339 $\pm$ 0.009	0.016$\pm$0.004	0.040 $\pm$ 0.005	0.238$\pm$0.007	0.172 $\pm$ 0.005	0.084 $\pm$ 0.004	0.454$\pm$0.005
ML2	0.802 $\pm$ 0.015	0.341 $\pm$ 0.015	0.019 $\pm$ 0.006	0.036$\pm$0.005	0.243 $\pm$ 0.007	0.171$\pm$0.005	0.078 $\pm$ 0.003	0.460 $\pm$ 0.005
BR	0.976 $\pm$ 0.001	0.441 $\pm$ 0.006	0.021 $\pm$ 0.005	0.096 $\pm$ 0.003	0.455 $\pm$ 0.006	0.214 $\pm$ 0.010	0.106 $\pm$ 0.010	0.677 $\pm$ 0.005
ECC	0.799 $\pm$ 0.000	0.328 $\pm$ 0.004	0.021 $\pm$ 0.006	0.054 $\pm$ 0.003	0.244 $\pm$ 0.005	0.199 $\pm$ 0.002	0.092 $\pm$ 0.002	0.478 $\pm$ 0.003
RAKEL	0.950 $\pm$ 0.004	0.332 $\pm$ 0.006	0.022 $\pm$ 0.005	0.106 $\pm$ 0.002	0.488 $\pm$ 0.008	0.209 $\pm$ 0.002	0.107 $\pm$ 0.001	0.546 $\pm$ 0.002
Comparing algorithms	Ranking loss $\downarrow$							
	cal500	emotions	genbase	medical	enron	image	scene	yeast
RLIM	0.183$\pm$0.002	0.193 $\pm$ 0.012	0.006 $\pm$ 0.004	0.025$\pm$0.008	0.096 $\pm$ 0.003	0.144$\pm$0.005	0.063$\pm$0.004	0.173$\pm$0.004
MLFE	0.192 $\pm$ 0.003	0.212 $\pm$ 0.012	0.003$\pm$0.002	0.026 $\pm$ 0.004	0.081$\pm$0.003	0.147 $\pm$ 0.007	0.067 $\pm$ 0.003	0.169 $\pm$ 0.004
ML2	0.198 $\pm$ 0.009	0.216 $\pm$ 0.017	0.005 $\pm$ 0.003	0.028 $\pm$ 0.005	0.081$\pm$0.003	0.146 $\pm$ 0.006	0.065 $\pm$ 0.004	0.175 $\pm$ 0.005
BR	0.342 $\pm$ 0.004	0.312 $\pm$ 0.006	0.007 $\pm$ 0.004	0.073 $\pm$ 0.002	0.184 $\pm$ 0.004	0.201 $\pm$ 0.002	0.110 $\pm$ 0.002	0.328 $\pm$ 0.004
ECC	0.210 $\pm$ 0.002	0.190$\pm$0.005	0.008 $\pm$ 0.004	0.037 $\pm$ 0.003	0.086 $\pm$ 0.003	0.179 $\pm$ 0.004	0.093 $\pm$ 0.003	0.191 $\pm$ 0.003
RAKEL	0.292 $\pm$ 0.001	0.231 $\pm$ 0.003	0.008 $\pm$ 0.003	0.085 $\pm$ 0.003	0.209 $\pm$ 0.004	0.194 $\pm$ 0.002	0.111 $\pm$ 0.001	0.223 $\pm$ 0.001
Comparing algorithms	Average precision $\uparrow$							
	cal500	emotions	genbase	medical	enron	image	scene	yeast
RLIM	0.508$\pm$0.004	0.772$\pm$0.015	0.996$\pm$0.004	0.895$\pm$0.015	0.692 $\pm$ 0.008	0.822$\pm$0.006	0.888$\pm$0.006	0.755 $\pm$ 0.006
MLFE	0.492 $\pm$ 0.006	0.754 $\pm$ 0.014	0.994 $\pm$ 0.003	0.888 $\pm$ 0.011	0.703 $\pm$ 0.005	0.819 $\pm$ 0.009	0.887 $\pm$ 0.004	0.765$\pm$0.005
ML2	0.495 $\pm$ 0.009	0.750 $\pm$ 0.020	0.992 $\pm$ 0.005	0.894 $\pm$ 0.008	0.708$\pm$0.004	0.820 $\pm$ 0.005	0.884 $\pm$ 0.006	0.755 $\pm$ 0.006
BR	0.325 $\pm$ 0.004	0.684 $\pm$ 0.009	0.986 $\pm$ 0.004	0.818 $\pm$ 0.005	0.575 $\pm$ 0.006	0.762 $\pm$ 0.003	0.808 $\pm$ 0.004	0.612 $\pm$ 0.005
ECC	0.462 $\pm$ 0.002	0.768 $\pm$ 0.005	0.987 $\pm$ 0.005	0.872 $\pm$ 0.008	0.673 $\pm$ 0.003	0.788 $\pm$ 0.004	0.842 $\pm$ 0.003	0.739 $\pm$ 0.003
RAKEL	0.393 $\pm$ 0.002	0.745 $\pm$ 0.003	0.985 $\pm$ 0.003	0.808 $\pm$ 0.008	0.597 $\pm$ 0.002	0.777 $\pm$ 0.002	0.825 $\pm$ 0.002	0.707 $\pm$ 0.002
Comparing algorithms	Macro-averaging F1 $\uparrow$							
	cal500	emotions	genbase	medical	enron	image	scene	yeast
RLIM	0.065 $\pm$ 0.004	0.557 $\pm$ 0.030	0.844$\pm$0.040	0.503$\pm$0.056	0.185$\pm$0.018	0.648$\pm$0.009	0.766$\pm$0.006	0.382 $\pm$ 0.014
MLFE	0.051 $\pm$ 0.004	0.556 $\pm$ 0.021	0.755 $\pm$ 0.039	0.437 $\pm$ 0.038	0.120 $\pm$ 0.016	0.646 $\pm$ 0.010	0.760 $\pm$ 0.005	0.368 $\pm$ 0.011
ML2	0.087 $\pm$ 0.005	0.545 $\pm$ 0.026	0.841 $\pm$ 0.074	0.438 $\pm$ 0.054	0.168 $\pm$ 0.024	0.643 $\pm$ 0.006	0.757 $\pm$ 0.008	0.413$\pm$0.007
BR	0.125 $\pm$ 0.002	0.569 $\pm$ 0.009	0.793 $\pm$ 0.007	0.481 $\pm$ 0.018	0.162 $\pm$ 0.006	0.543 $\pm$ 0.003	0.625 $\pm$ 0.003	0.381 $\pm$ 0.002
ECC	0.093 $\pm$ 0.001	0.612 $\pm$ 0.003	0.785 $\pm$ 0.013	0.450 $\pm$ 0.012	0.157 $\pm$ 0.009	0.589 $\pm$ 0.007	0.690 $\pm$ 0.006	0.367 $\pm$ 0.004
RAKEL	0.137$\pm$0.001	0.614$\pm$0.004	0.793 $\pm$ 0.007	0.479 $\pm$ 0.020	0.162 $\pm$ 0.006	0.616 $\pm$ 0.002	0.688 $\pm$ 0.003	0.390 $\pm$ 0.002
Comparing algorithms	Micro-averaging F1 $\uparrow$							
	cal500	emotions	genbase	medical	enron	image	scene	yeast
RLIM	0.355$\pm$0.009	0.599 $\pm$ 0.021	0.989$\pm$0.005	0.809$\pm$0.017	0.566$\pm$0.010	0.639 $\pm$ 0.007	0.759$\pm$0.006	0.628 $\pm$ 0.010
MLFE	0.248 $\pm$ 0.007	0.580 $\pm$ 0.018	0.987 $\pm$ 0.004	0.788 $\pm$ 0.013	0.525 $\pm$ 0.007	0.650$\pm$0.010	0.755 $\pm$ 0.005	0.628 $\pm$ 0.010
ML2	0.351 $\pm$ 0.009	0.574 $\pm$ 0.026	0.979 $\pm$ 0.011	0.802 $\pm$ 0.010	0.557 $\pm$ 0.007	0.640 $\pm$ 0.006	0.749 $\pm$ 0.009	0.633$\pm$0.006
BR	0.349 $\pm$ 0.004	0.578 $\pm$ 0.009	0.979 $\pm$ 0.002	0.787 $\pm$ 0.009	0.511 $\pm$ 0.005	0.539 $\pm$ 0.003	0.610 $\pm$ 0.003	0.564 $\pm$ 0.008
ECC	0.345 $\pm$ 0.003	0.632$\pm$0.003	0.977 $\pm$ 0.003	0.792 $\pm$ 0.007	0.555 $\pm$ 0.003	0.587 $\pm$ 0.007	0.684 $\pm$ 0.006	0.614 $\pm$ 0.003
RAKEL	0.350 $\pm$ 0.001	0.626 $\pm$ 0.005	0.979 $\pm$ 0.002	0.787 $\pm$ 0.008	0.541 $\pm$ 0.006	0.612 $\pm$ 0.003	0.677 $\pm$ 0.004	0.609 $\pm$ 0.002

B. Experimental Results

Table II reports the detailed experimental results of each comparing algorithm on the regular-scale³. And the best performance among the comparing algorithms is shown in boldface. For each evaluation metric, " $\downarrow$ " indicates the smaller the better while " $\uparrow$ " indicates the larger the better.

From the result table we can see, across all evaluation metrics, RLIM ranks *1st* in 66.1% and 57.1% cases respectively. Comparing with the two recent state-of-the-art approaches MLFE and ML² across all evaluation metrics, RLIM outperforms MFLE in 70.5% cases and ML² in 68.6% cases respectively. Thus the effectiveness of RLIM is validated.

V. CONCLUSION

In this paper, we propose a novel unified multi-label learning framework which estimates latent relative label importances

³Duo to space limit, the result of large-scale data sets is also shown at: <https://github.com/LowKeyH/RLIM.git>

while training the inductive model simultaneously. Specifically, a biconvex formulation is proposed to estimate optimally the relative labels importances for the model training. Extensive experimental results validate the effectiveness of the proposed approach RLIM against state-of-the-art multi-label learning approaches. For future work, we will try another way to optimally estimate the numeric labeling information, under the proposed framework. In addition, since graph regularization plays a key role in this framework and the constructed graphs can be easily affected by high-dimensional data, it is also interesting to exploit more robust graph structure for data representation.

APPENDIX

To solve (14), we first transform the matrix $\mathbf{P} \in \mathbb{R}^{m \times l}$ into a vector $\tilde{\mathbf{p}} = [p_{11}, p_{21}, p_{31}, \dots, p_{12}, p_{22}, p_{32}, \dots, p_{ml}]^T \in \mathbb{R}^{ml}$, $\mathbf{Q} \in \mathbb{R}^{m \times l}$ into a vector $\tilde{\mathbf{q}} = [q_{11}, q_{21}, q_{31}, \dots, q_{12}, q_{22}, q_{32}, \dots, q_{ml}]^T \in \mathbb{R}^{ml}$, $\mathbf{Y}$ into a vector $\tilde{\mathbf{y}} = [y_{11}, y_{21}, y_{31}, \dots, y_{12}, y_{22}, y_{32}, \dots, y_{ml}]^T \in [0, 1]^{ml}$, $\mathbf{G}$ into a vector $\tilde{\mathbf{g}} = [g_{11}, g_{21}, g_{31}, \dots, g_{12}, g_{22}, g_{32}, \dots, g_{ml}]^T \in \mathbb{R}^{ml}$. In order to deal with the equality and inequality constraints using $\tilde{\mathbf{p}}$, we define two indicator matrices $\mathbf{U} = [u_{ij}]^{ml \times ml}$ and $\mathbf{V} = [v_{ij}]^{m \times ml}$, where $\mathbf{U}$ is a diagonal matrix with its diagonal element $u_{ij} = 1$ if $\tilde{y}_i = 1$ and $u_{ij} = 0$ otherwise, and $v_{ij} = 1$ if $\tilde{y}_{[(j-1)*m+i]} = 1$ and $v_{ij} = 0$ otherwise. Using these notations, the optimization problem (14) can be rewritten as:

$$\begin{aligned} \min_{\tilde{\mathbf{p}}} \quad & \frac{1}{2} \tilde{\mathbf{p}}^T \mathbf{H} \tilde{\mathbf{p}} + \tilde{\mathbf{f}}^T \tilde{\mathbf{p}} \\ \text{s.t.} \quad & \mathbf{U} \tilde{\mathbf{p}} \geq \mathbf{0}_{ml} \\ & \mathbf{V} \tilde{\mathbf{p}} = \hat{\mathbf{y}} \\ & -\mathbf{1}_{ml} \leq \tilde{\mathbf{p}} \leq \tilde{\mathbf{g}} \end{aligned} \quad (16)$$

where $\tilde{\mathbf{f}} = -2\tilde{\mathbf{q}}$ and $\hat{\mathbf{y}} \in \mathbb{R}^m$ stores the number of relevant labels of each instance, with its element $\hat{y}_i = |\mathbf{y}_i|$. Let $\mathbf{D}_s$ and $\mathbf{D}_c$ be the diagonal matrices with theirs (i, i) -th elements equaling to the sum of the i -th row of $\mathbf{S}$ and $\mathbf{C}$, $\mathbf{H} \in \mathbb{R}^{ml \times ml}$ is defined as follows:

$$\mathbf{H} = \begin{bmatrix} \mathbf{T}^1 & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1m} \\ \mathbf{C}_{21} & \mathbf{T}^2 & \cdots & \mathbf{C}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{m1} & \mathbf{C}_{m2} & \cdots & \mathbf{T}^i \end{bmatrix} \quad (17)$$

Here, $\mathbf{T}^i = \beta_1(\mathbf{D}_s^i - \mathbf{S}) + \beta_2 \mathbf{D}_c^i \in \mathbb{R}^{m \times m}$, where $\mathbf{D}_s^i$ and $\mathbf{D}_c^i$ are diagonal matrices where all diagonal elements of them are the corresponding i -th diagonal elements of $\mathbf{D}_s$ and $\mathbf{D}_c$. And $\mathbf{C}_{ij} \in \mathbb{R}^{m \times m}$ is the diagonal matrix where its all diagonal elements are the (i, j) -th element of $\mathbf{C}$.

REFERENCES

- [1] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
- [2] H. Wang, H. Huang, and C. Ding, "Image annotation using multi-label correlated green's function," in *Proceedings of the 12th IEEE International Conference on Computer Vision*, pp. 2029–2034, 2009.
- [3] H.-Y. Lo, J.-C. Wang, H.-M. Wang, and S.-D. Lin, "Cost-sensitive multi-label learning for audio tag annotation and retrieval," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 518–529, 2011.
- [4] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning*, vol. 88, no. 1-2, pp. 157–208, 2012.
- [5] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, "Effective multi-label active learning for text classification," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 917–926, 2009.
- [6] P. Hou, X. Geng, and M.-L. Zhang, "Multi-label manifold learning," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 1680–1686, 2016.
- [7] Y.-K. Li, M.-L. Zhang, and X. Geng, "Leveraging implicit relative labeling-importance information for effective multi-label learning," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 251–260, 2015.
- [8] Q.-W. Zhang, Y. Zhong, and M.-L. Zhang, "Feature-induced labeling information enrichment for multi-label learning," in *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, 2018.
- [9] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.
- [10] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.
- [11] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.
- [12] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in Neural Information Processing Systems*, pp. 681–687, 2002.
- [13] J. Fürnkranz, E. Hüllermeier, E. L. Mencia, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, 2008.
- [14] S. Ji, L. Tang, S. Yu, and J. Ye, "A shared-subspace learning framework for multi-label classification," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 2, p. 8, 2010.
- [15] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.
- [16] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.
- [17] W. Liu and S. Chang, "Robust multi-class transductive learning with graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 381–388, 2009.
- [18] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: a survey and extensions," *Mathematical Methods of Operations Research*, vol. 66, no. 3, pp. 373–407, 2007.
- [19] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems*, pp. 321–328, 2004.
- [20] L. Feng and B. An, "Leveraging latent label distributions for partial label learning," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2107–2113, 2018.
- [21] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, "Mulan: A java library for multi-label learning," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2411–2414, 2011.