Leveraging Implicit Relative Labeling-Importance Information for Effective Multi-Label Learning

Yu-Kun Li[*,†], Min-Ling Zhang[*,†], Xin Geng[*,†]
[*]*School of Computer Science and Engineering, Southeast University, Nanjing 210096, China*
[†]*Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China*
Email: {liyk, zhangml, xgeng}@seu.edu.cn

***Abstract*—In multi-label learning, each training example is represented by a single instance while associated with multiple labels, and the task is to predict a set of relevant labels for the unseen instance. Existing approaches learn from multi-label data by assuming equal labeling-importance, i.e. all the associated labels are regarded to be relevant while their *relative importance* for the training example are not differentiated. Nonetheless, this assumption fails to reflect the fact that the importance degree of each associated label is generally different, though the importance information is not explicitly accessible from the training examples. In this paper, we show that effective multi-label learning can be achieved by leveraging the implicit *relative labeling-importance* (RLI) information. Specifically, RLI degrees are formalized as multinomial distribution over the label space, which are estimated by adapting an iterative label propagation procedure. After that, the multi-label prediction model is learned by fitting the estimated multinomial distribution as regularized with popular multi-label empirical loss. Comprehensive experiments clearly validate the usefulness of leveraging implicit RLI information to learn from multi-label data.**

***Keywords*-multi-label learning; relative labeling-importance; label distribution**

I. Introduction

Multi-label learning deals with training examples each represented by a single instance while associated with multiple labels, and the task is to learn a multi-label predictor which maps from unseen instance to relevant label set [14], [22], [30]. During the past decade, multi-label learning techniques have been widely employed to learn from data with rich semantics, such as text [21], image [3], audio [16], video [25], etc.

Formally speaking, let $\mathcal{X} = \mathbb{R}^d$ be the d-dimensional feature space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_q\}$ be the label space with q possible class labels. Given a multi-label training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq m\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is the d-dimensional instance and $Y_i \subseteq \mathcal{Y}$ is the set of labels associated with $\boldsymbol{x}_i$, the task is to learn a multi-label predictor $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from $\mathcal{D}$ which maps from the space of feature vectors to the space of *label sets*. To learn from multi-label data, existing approaches take the common assumption of equal labeling-importance, i.e. each label associated with the training example is regarded to be relevant while the *relative importance* among them are not differentiated [30].

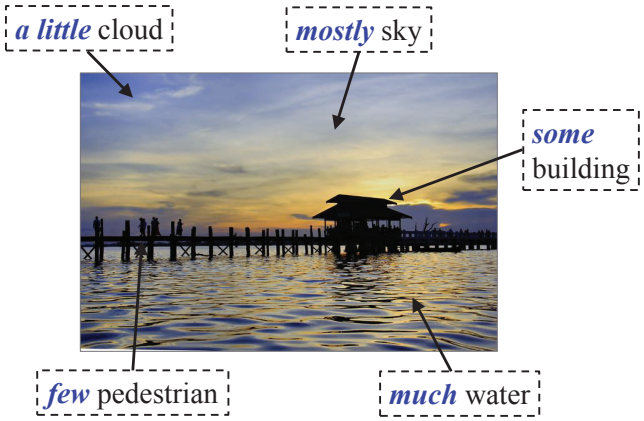

Figure 1. An exemplar natural scene image which has been annotated with multiple labels *sky*, *water*, *building* and *cloud*. The relative importance of each label is illustrated in the figure, which has not been explicitly provided by the annotator. In addition, the label *pedestrian* is not annotated for the image due to its insignificant appearance.

However, for real-world multi-label learning problems the importance degree of each associated label is generally different, though the importance information is not explicitly accessible from the training examples. As shown in Fig. 1, a natural scene image may be annotated with labels *sky*, *water*, *building* and *cloud* simultaneously, while their relative importance for characterizing this image are not explicitly provided by the annotator. Similar situations also hold for other types of multi-label data, e.g. the multiple categories associated with a news document would have different topical importance, the multiple functionalities associated with a gene would have different expression levels, etc.

Based on the above observations, we naturally postulate that effective multi-label learning can be achieved by leveraging the implicit *relative labeling-importance* (RLI) information. Accordingly, a novel multi-label learning approach named RELIAB, i.e. *RElative Labeling-Importance Aware multi-laBel learning*, is proposed in this paper. Firstly, the RLI degrees are formalized as multinomial distribution over the label space, which are estimated by invoking an iterative label propagation procedure over the training examples.

After that, a multi-label predictor is induced by fitting the prediction model with the estimated multinomial distribution along with multi-label empirical loss regularization. Extensive experiments across 17 benchmark multi-label data sets show that RELIAB performs favorably against state-of-the-art multi-label learning approaches.

The rest of this paper is organized as follows. Section II presents technical details of the proposed approach. Section III discusses existing works related to RELIAB. Section IV reports experimental results of comparative studies. Finally, Section V concludes.

II. THE RELIAB APPROACH

As shown in Section I, the task of multi-label learning is to induce a multi-label predictor $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from the training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq m\}$. Given any instance $\boldsymbol{x} = [x_{i1}, x_{i2}, \ldots, x_{id}]^\top \in \mathcal{X}$ and label $y_l \in \mathcal{Y}$, we use $\mu_{\boldsymbol{x}}^{y_l}$ to denote the *implicit RLI degree* of y_l for characterizing $\boldsymbol{x}$. Intuitively speaking, the higher the value of $\mu_{\boldsymbol{x}}^{y_l}$, the more semantics conveyed by y_l in characterizing $\boldsymbol{x}$.

Accordingly, the set of relevant labels Y for $\boldsymbol{x}$ can be determined as: $Y = \{y_l \mid \mu_{\boldsymbol{x}}^{y_l} > t(\boldsymbol{x}), 1 \leq l \leq q\}$, where $t(\boldsymbol{x})$ corresponds to the threshold value which separates relevant labels from irrelevant ones for instance $\boldsymbol{x}$. In this paper, we enlarge the original label space $\mathcal{Y}$ into $\tilde{\mathcal{Y}} = \{y_0\} \bigcup \mathcal{Y}$, where y_0 is the complementary *virtual label* serving as an artificial bipartition point between relevant and irrelevant labels [8], [17], [22], [30]. In this case, $t(\boldsymbol{x})$ can be viewed as the thresholding-importance w.r.t. virtual label y_0, i.e. $\mu_{\boldsymbol{x}}^{y_0}$. Therefore, we have the formal definition on RLI degree as follows:

Definition. *Relative Labeling-Importance (RLI) Degree*

Given any instance $\boldsymbol{x} \in \mathcal{X}$, the RLI degree of label $y_l \in \tilde{\mathcal{Y}}$ for $\boldsymbol{x}$ is denoted as $\mu_{\boldsymbol{x}}^{y_l}$ $(0 \leq l \leq q)$, which satisfies the following constraints:

(i) ***non-negativity***: $\mu_{\boldsymbol{x}}^{y_l} \geq 0$

(ii) ***normalization***: $\sum_{l=0}^{q} \mu_{\boldsymbol{x}}^{y_l} = 1$

Furthermore, the set of relevant labels $Y \subseteq \mathcal{Y}$ for $\boldsymbol{x}$ can be determined as: $Y = \{y_l \mid \mu_{\boldsymbol{x}}^{y_l} > \mu_{\boldsymbol{x}}^{y_0}, 1 \leq l \leq q\}$.

There are three points which need to be noticed for the RLI degree formulated as above. Firstly, the RLI degree is not directly accessible from the multi-label training examples and thus *implicit* to the learning algorithm. Secondly, the RLI degree is instance-dependant which corresponds to the *relative* importance among all labels in characterizing the semantics of one particular instance.[1] Thirdly, the RLI degree for each instance, i.e. $\mu_{\boldsymbol{x}}^{y_l}$, can be viewed as a *label distribution* over the label space $\tilde{\mathcal{Y}}$. For label distribution learning (LDL) [10], [11], [12], the distribution information is assumed to be available while for multi-label learning the RLI information needs to be further inferred.

In this paper, RELIAB learns from multi-label data in two basic stages, i.e. *implicit RLI degree estimation* and *prediction model induction*, which are scrutinized in the following subsections respectively.

A. Implicit RLI Degree Estimation

In the first stage, RELIAB aims to estimate the implicit RLI degree for all training examples, i.e. $\mathcal{U} = \{\mu_{\boldsymbol{x}_i}^{y_l} \mid 1 \leq i \leq m, 0 \leq l \leq q\}$. To fulfill this task, the widely-used iterative label propagation techniques [31], [33] is adapted for the estimation. Let $G = (V, E)$ denote the fully-connected graph constructed over the set of training examples with $V = \{\boldsymbol{x}_i \mid 1 \leq i \leq m\}$. Furthermore, an $m \times m$ symmetric similarity matrix $\mathbf{W} = [w_{ij}]_{m \times m}$ is specified for G as follows:

$$\forall_{i,j=1}^{m}: \quad w_{ij} = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{1}$$

Here, $\sigma > 0$ is the width parameter for similarity calculation, which is fixed to be 1 in this paper.

Correspondingly, a label propagation matrix $\mathbf{P}$ is constructed from the similarity matrix: $\mathbf{P} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$. Here, $\mathbf{D} = \mathrm{diag}[d_1, d_2, \ldots, d_m]$ is a diagonal matrix with its diagonal entry d_i equal to the sum of the i-th row of $\mathbf{W}$: $d_i = \sum_{j=1}^{m} w_{ij}$. Let $\mathbf{F} = [f_{il}]_{m \times (q+1)}$ be an $m \times (q+1)$ matrix with non-negative entries, where $f_{il} \geq 0$ is assumed to be proportional to the labeling-importance $\mu_{\boldsymbol{x}_i}^{y_l}$. Based on the multi-label training set, an initial matrix $\mathbf{F}^{(0)} = \mathbf{\Phi} = [\phi_{il}]_{m \times (q+1)}$ is instantiated as follows:

$$\forall_{i=1}^{m}\ \forall_{l=0}^{q}: \quad \phi_{il} = \begin{cases} \tau, & \text{if } y_l = y_0 \\ 1, & \text{if } y_l \in Y_i \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Here, $\tau \in (0, 1)$ is the initial thresholding-importance parameter for virtual label y_0. As shown in Eq.(2), at the initialization step, all the relevant (irrelevant) labels are assumed to have unit (zero) labeling-importance. At the t-th iteration, $\mathbf{F}$ is updated by propagating labeling-importance information with the label propagation matrix $\mathbf{P}$:

$$\mathbf{F}^{(t)} = \alpha\mathbf{P}\mathbf{F}^{(t-1)} + (1 - \alpha)\mathbf{\Phi} \tag{3}$$

Here, $\alpha \in (0, 1)$ is the balancing parameter which controls the fraction of information inherited from label propagation (i.e. $\mathbf{P}\mathbf{F}^{(t-1)}$) and initial labeling (i.e. $\mathbf{\Phi}$).

By applying Eq.(3) recursively with $\mathbf{F}^{(0)} = \mathbf{\Phi}$, it is not difficult to show that:

$$\mathbf{F}^{(t)} = (\alpha\mathbf{P})^t\mathbf{\Phi} + (1 - \alpha)\sum_{i=0}^{t-1}(\alpha\mathbf{P})^i\mathbf{\Phi} \tag{4}$$

[1] In other words, given two instances $\{\boldsymbol{x}, \boldsymbol{z}\}$ and two labels $\{y_l, y_m\}$, based on RLI degree we are only modeling and interested in the relative magnitude between $\mu_{\boldsymbol{x}}^{y_l}$ and $\mu_{\boldsymbol{x}}^{y_m}$ (or $\mu_{\boldsymbol{z}}^{y_l}$ and $\mu_{\boldsymbol{z}}^{y_m}$), instead of the relative magnitude between $\mu_{\boldsymbol{x}}^{y_l}$ and $\mu_{\boldsymbol{z}}^{y_l}$ (or $\mu_{\boldsymbol{x}}^{y_m}$ and $\mu_{\boldsymbol{z}}^{y_m}$).

As a real symmetric matrix, the label propagation matrix $\mathbf{P}$ can be diagonalized as $\mathbf{P} = \mathbf{C}^\top \mathbf{\Lambda} \mathbf{C}$, where $\mathbf{C}$ is an orthonormal matrix and $\mathbf{\Lambda} = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_m]$ is a diagonal matrix containing eigenvalues of $\mathbf{P}$. Note that $\mathbf{P}$ is similar to $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}} \mathbf{P} \mathbf{D}^{\frac{1}{2}} = \mathbf{D}^{-1}\mathbf{W}$, and therefore $\mathbf{P}$ and $\mathbf{S}$ share identical eigenvalues.

Since $\mathbf{S}$ is a stochastic matrix whose rows consist of non-negative entries and sum to one, the absolute value of each eigenvalue satisfies $|\lambda_i| \leq 1$ $(1 \leq i \leq q)$ as ensured by the Perron-Frobenius theorem [18], [33]. Under the setting of $\alpha \in (0, 1)$, the limit for the first term of Eq.(4) would be:

$$\begin{aligned} \lim_{t\to\infty} (\alpha\mathbf{P})^t \mathbf{\Phi} &= \lim_{t\to\infty} \alpha^t \cdot \left(\mathbf{C}^\top \mathbf{\Lambda} \mathbf{C}\right)^t \mathbf{\Phi} \\ &= \lim_{t\to\infty} \alpha^t \cdot \mathbf{C}^\top \mathbf{\Lambda}^t \mathbf{C}\ \mathbf{\Phi} \\ &= \mathbf{0} \end{aligned} \tag{5}$$

It also holds that $\lim_{t\to\infty} \sum_{i=0}^{t-1} (\alpha\mathbf{P})^i = (\mathbf{I} - \alpha\mathbf{P})^{-1}$ because:

$$\begin{aligned} (\mathbf{I} - \alpha\mathbf{P}) \lim_{t\to\infty} \sum_{i=0}^{t-1} (\alpha\mathbf{P})^i &= \lim_{t\to\infty} (\mathbf{I} - \alpha\mathbf{P}) \sum_{i=0}^{t-1} (\alpha\mathbf{P})^i \\ &= \lim_{t\to\infty} \left(\mathbf{I} - (\alpha\mathbf{P})^t\right) \\ &= \mathbf{I} \end{aligned}$$

Thus, the limit for the second term of Eq.(4) would be:

$$\lim_{t\to\infty} (1-\alpha) \sum_{i=0}^{t-1} (\alpha\mathbf{P})^i \mathbf{\Phi} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1} \mathbf{\Phi} \tag{6}$$

According to Eqs.(5) and (6), $\mathbf{F}^{(t)}$ will converge to $\mathbf{F}^*$ as the number of iterations grow:

$$\mathbf{F}^* = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1} \mathbf{\Phi} \tag{7}$$

Based on $\mathbf{F}^*$, the implicit RLI degree for each training example is estimated as:

$$\forall_{i=1}^m\ \forall_{l=0}^q : \quad \mu_{\boldsymbol{x}_i}^{y_l} = \frac{f_{il}^*}{\sum_{k=0}^q f_{ik}^*} \tag{8}$$

In other words, the set of $q+1$ RLI degrees for each instance $\boldsymbol{x}_i$, i.e. $\{\mu_{\boldsymbol{x}_i}^{y_l} \mid 0 \leq l \leq q\}$, can be regarded as a multinomial distribution over the (enlarged) label space $\tilde{\mathcal{Y}}$, which are obtained by normalizing $\mathbf{F}^*$ on each row.

B. Prediction Model Induction

In the second stage, RELIAB aims to induce the multi-label prediction model by leveraging the implicit RLI information estimated in the first stage, i.e. $\mathcal{U} = \{\mu_{\boldsymbol{x}_i}^{y_l} \mid 1 \leq i \leq m, 0 \leq l \leq q\}$. To facilitate the exploitation of $\mathcal{U}$, we employ the simple maximum entropy model [5], [12] to parametrize the multi-label predictor:

$$\forall_{l=0}^q : \quad f(y_l \mid \boldsymbol{x}, \mathbf{\Theta}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{\theta}_l^\top \boldsymbol{x}\right) \tag{9}$$

Here, $\mathbf{\Theta} = \left[\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q\right]$ represents the set of model parameters and $\boldsymbol{\theta}_l = [\theta_{l1}, \theta_{l2}, \ldots, \theta_{ld}]^\top$ is the d-dimensional weighting parameter vector for the l-th label $y_l \in \tilde{\mathcal{Y}}$. Furthermore, the partition function $Z(\boldsymbol{x}) = \sum_{l=0}^q \exp(\boldsymbol{\theta}_l^\top \boldsymbol{x})$ serves as a normalization term to ensure distributional outputs over $\tilde{\mathcal{Y}}$: $\sum_{l=0}^q f(y_l \mid \boldsymbol{x}, \mathbf{\Theta}) = 1$. In this case, the multi-label predictor h can be derived from f by thresholding the outputs against the virtual label y_0:

$$h(\boldsymbol{x}) = \{y_l \mid f(y_l \mid \boldsymbol{x}, \mathbf{\Theta}) > f(y_0 \mid \boldsymbol{x}, \mathbf{\Theta}), 1 \leq l \leq q\} \tag{10}$$

To induce the parametric model f, RELIAB chooses to optimize the following objective function:

$$V(f, \mathcal{U}, \mathcal{D}) = V_{dis}(f, \mathcal{U}) + \beta \cdot V_{emp}(f, \mathcal{D}) \tag{11}$$

The first term $V_{dis}(f, \mathcal{U})$ considers how the parametric model f fits the estimated RLI degrees $\mathcal{U}$, while the second term $V_{emp}(f, \mathcal{D})$ is used as a regularizer which considers how f classifies the multi-label training examples in $\mathcal{D}$.

On one hand, $V_{dis}(f, \mathcal{U})$ can be measured by the *compatibility* between the importance-based distribution, i.e. $\{\mu_{\boldsymbol{x}}^{y_l} \mid 0 \leq l \leq q\}$, and the model-based distribution, i.e. $\{f(y_l \mid \boldsymbol{x}, \mathbf{\Theta}) \mid 0 \leq l \leq q\}$. Here, the canonical Kullback-Leibler (KL) divergence is employed to instantiate the first term of Eq.(11):

$$\begin{aligned} &V_{dis}(f, \mathcal{U}) \\ &= \sum_{i=1}^m \mathrm{KL}\left(\{\mu_{\boldsymbol{x}_i}^{y_l} \mid 0 \leq l \leq q\}, \{f(y_l \mid \boldsymbol{x}_i, \mathbf{\Theta}) \mid 0 \leq l \leq q\}\right) \\ &= \sum_{i=1}^m \sum_{l=0}^q \left(\mu_{\boldsymbol{x}_i}^{y_l} \ln \frac{\mu_{\boldsymbol{x}_i}^{y_l}}{f(y_l \mid \boldsymbol{x}_i, \mathbf{\Theta})}\right) \end{aligned} \tag{12}$$

On the other hand, $V_{emp}(f, \mathcal{D})$ can be measured by the *empirical loss* of the parametric model f on $\mathcal{D}$. As shown in Eq.(10), by taking the virtual label y_0 as the bipartition point, its modeling output $f(y_0 \mid \boldsymbol{x}_i, \mathbf{\Theta})$ should be less than those of relevant labels in Y_i while larger than those of irrelevant labels in $\overline{Y}_i$ (i.e. $\mathcal{Y} \setminus Y_i$). Accordingly, the second term of Eq.(11) is instantiated as:

$$\begin{aligned} &V_{emp}(f, \mathcal{D}) \\ &= -\sum_{i=1}^m \left(\sum_{y_j \in Y_i} \Big(f(y_j \mid \boldsymbol{x}_i, \mathbf{\Theta}) - f(y_0 \mid \boldsymbol{x}_i, \mathbf{\Theta})\Big) \right. \\ &\qquad \left. + r_i \cdot \sum_{y_k \in \overline{Y}_i} \Big(f(y_0 \mid \boldsymbol{x}_i, \mathbf{\Theta}) - f(y_k \mid \boldsymbol{x}_i, \mathbf{\Theta})\Big) \right) \end{aligned} \tag{13}$$

Here, $r_i = |Y_i| / |\overline{Y}_i|$ is used to account for potential imbalance between the number of relevant and irrelevant labels associated with each example [28]. Note that minimizing the loss in Eq.(13) can be viewed as minimizing one of the most popular multi-label metrics, namely the *ranking loss* [2], [9], [22], [30], which considers pairwise ranking between each relevant-irrelevant label pair. Nonetheless, by incorporating the virtual label y_0, the number of pairwise relationships

to be considered can be reduced from $O(q^2)$ for traditional ranking loss to $O(q)$ for the loss in Eq.(13).

By substituting Eqs.(12) and (13) into the objective function and ignoring constant terms, Eq.(11) can then be rewritten as:

$$V(f,\mathcal{U},\mathcal{D}) = -\sum_{i=1}^{m}\sum_{l=0}^{q}\left(\mu_{\boldsymbol{x}_i}^{y_l}\ln f(y_l \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\right) -\beta\cdot\sum_{i=1}^{m}\left(\sum_{y_j\in Y_i}\Big(f(y_j \mid \boldsymbol{x}_i,\boldsymbol{\Theta}) - f(y_0 \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\Big) +r_i\cdot\sum_{y_k\in\overline{Y}_i}\Big(f(y_0 \mid \boldsymbol{x}_i,\boldsymbol{\Theta}) - f(y_k \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\Big)\right) \quad (14)$$

The final prediction model f^* is obtained by minimizing Eq.(14), i.e. $f^* = \arg\min_f V(f,\mathcal{U},\mathcal{D})$. To solve the corresponding unconstrained nonlinear optimization problem, RELIAB employs the *Limited-memory Broyde-Fletcher-Goldfarb-Shanno* (L-BFGS) algorithm which is particularly suited for problems with large number of variables [19]. As a quasi-Newton algorithm, L-BFGS iteratively optimizes the objective function with resort to gradient of the function:

$$\frac{\partial V}{\partial\boldsymbol{\Theta}} = \left[\frac{\partial V}{\partial\boldsymbol{\theta}_0},\cdots,\frac{\partial V}{\partial\boldsymbol{\theta}_l},\cdots,\frac{\partial V}{\partial\boldsymbol{\theta}_q}\right], \qquad \text{where}$$

$$\frac{\partial V}{\partial\boldsymbol{\theta}_l} = -\sum_{i=1}^{m}\Big(\big(\mu_{\boldsymbol{x}_i}^{y_l} - f(y_l \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\big)\cdot\boldsymbol{x}_i\Big) - \beta\cdot\sum_{i=1}^{m} \left(f(y_l \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\Big(\sum_{y_j\in Y_i\setminus\{y_l\}}\Big(f(y_0 \mid \boldsymbol{x}_i,\boldsymbol{\Theta}) - f(y_j \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\Big) +r_i\cdot\sum_{y_k\in\overline{Y}_i\setminus\{y_l\}}\Big(f(y_k \mid \boldsymbol{x}_i,\boldsymbol{\Theta}) - f(y_0 \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\Big) +\zeta(y_l,Y_i)\Big(1 - f(y_l \mid \boldsymbol{x}_i,\boldsymbol{\Theta}) + f(y_0 \mid \boldsymbol{x}_i,\boldsymbol{\Theta})\Big)\Big)\cdot\boldsymbol{x}_i\right) \quad (15)$$

Here, $\zeta(y_l,Y_i)$ returns 0 if y_l corresponds to the virtual label y_0. Otherwise, $\zeta(y_l,Y_i)$ returns $+1$ if $y_l\in Y_i$ and $-r_i$ if $y_l\in\overline{Y}_i$.

Table I summarizes the complete procedure of the proposed RELIAB approach. After incorporating the virtual label y_0 into the original label space (Step 1), a similarity matrix as well as an initial labeling-importance matrix are constructed based on the training examples (Steps 2-3). After that, the implicit degrees of RLI are estimated via a label propagation procedure (Steps 4-5), and then the multi-label prediction model is learned by leveraging the estimated labeling-importance information (Steps 6-14). Finally, the predicted label set for unseen instance is determined by thresholding the model outputs against the virtual label (Step 15).[2]

[2]Code package for RELIAB is publicly available at http://cse.seu.edu.cn/PersonalPage/zhangml/files/RELIAB.zip

Table I
THE PSEUDO-CODE OF RELIAB.

Inputs:

$\mathcal{D}$: the multi-label training set $\{(\boldsymbol{x}_i,Y_i)\mid 1\le i\le m\}$ ($\boldsymbol{x}_i\in\mathcal{X}, Y_i\subseteq\mathcal{Y}, \mathcal{X}=\mathbb{R}^d, \mathcal{Y}=\{y_1,y_2,\ldots,y_q\}$)
τ: the initial thresholding-importance parameter in $(0,1)$
α: the balancing parameter in $(0,1)$
β: the regularization parameter
$\boldsymbol{x}$: the unseen instance ($\boldsymbol{x}\in\mathcal{X}$)

Outputs:

Y: the predicted label set for $\boldsymbol{x}$

Process:

1: Enlarge the original label space by introducing the virtual label y_0: $\widetilde{\mathcal{Y}}=\{y_0\}\bigcup\mathcal{Y}$;
2: Construct the similarity matrix $\mathbf{W}=[w_{ij}]_{m\times m}$ according to Eq.(1);
3: Construct the initial labeling-importance matrix $\boldsymbol{\Phi}=[\phi_{il}]_{m\times(q+1)}$ according to Eq.(2);
4: Conduct label propagation to yield the converged solution $\mathbf{F}^*$ according to Eq.(7);
5: Estimate the implicit RLI degrees $\mathcal{U}=\{\mu_{\boldsymbol{x}_i}^{y_l}\mid 1\le i\le m, 0\le l\le q\}$ according to Eq.(8);
6: Initialize model parameters $\boldsymbol{\Theta}^{(0)}=\frac{1}{d(q+1)}\cdot\mathbf{1}_{d\times(q+1)}$;
7: Set $t=0$;
8: **repeat**
9: Evaluate $f(y_l \mid \boldsymbol{x}_i,\boldsymbol{\Theta}^{(t)})$ $(1\le i\le m, 0\le l\le q)$ according to Eq.(9);
10: Evaluate gradient $\frac{\partial V}{\partial\boldsymbol{\Theta}}|_{\boldsymbol{\Theta}^{(t)}}$ according to Eq.(15);
11: Update $\boldsymbol{\Theta}^{(t+1)}$ by running one L-BFGS iteration [19] with current parameters $\boldsymbol{\Theta}^{(t)}$ and gradient $\frac{\partial V}{\partial\boldsymbol{\Theta}}|_{\boldsymbol{\Theta}^{(t)}}$;
12: $t=t+1$;
13: **until** convergence
14: Set the final prediction model f^* with $\boldsymbol{\Theta}^*=\boldsymbol{\Theta}^{(t)}$;
15: Return $Y=h(\boldsymbol{x})$ according to Eq.(10).

III. RELATED WORK

Existing works related to RELIAB are briefly discussed in this section, while more comprehensive reviews on multi-label learning can be found in [14], [22], [30].

Existing approaches to multi-label learning can be roughly grouped into three categories based on the *order of label correlations* being considered [22], [30], i.e. first-order approaches assuming independence among class labels [1], [29], second-order approaches considering correlations between a pair of class labels [7], [8], and high-order approaches considering correlations among label subsets or all the class labels [15], [20], [23]. For whichever order of correlations, the common modeling strategy is to treat each label in a crisp manner, i.e. being either relevant or irrelevant for an instance without differentiating its relative importance. In contrast, RELIAB models high-order label correlations by differentiating degrees of RLI over the label

space.

There have been some works which learn from multi-label data with auxiliary labeling-importance information. In [4], an *ordinal scale* is assumed to characterize the membership degree and an ordinal grade is assigned for each label of the training example. In [27], a *full ordering* is assumed to be known to rank relevant labels of the training example. In both cases, those auxiliary labeling-importance information are explicitly given and accessible to the learning algorithm. Obviously, RELIAB differs from them without assuming the availability of such explicit information.

The principle of maximum entropy (MaxEnt) has been employed to design multi-label learning algorithms, which works by modeling $p(\boldsymbol{y} \mid \boldsymbol{x})$, i.e. the joint probabilities of all labels $\boldsymbol{y} = (y_1, y_2, \ldots, y_q) \in \{0,1\}^q$ conditioned on the instance $\boldsymbol{x}$ [13], [30], [32]. Due to the combinatorial nature of $\boldsymbol{y}$, existing MaxEnt-based multi-label learning approaches can not scale well to data set with large number of labels. Actually, the data sets employed in the experiments of [13] and [32] only contain up to 10 labels. In contrast, the MaxEnt model employed by RELIAB (Eq.(9)) corresponds to a multinomial distribution instead of a joint distribution over the label space. This property makes RELIAB scalable for data sets with large number of labels, whose experimental results are reported in the next section.

IV. EXPERIMENTS

A. Preliminary Analysis

As shown in Table I, the implicit RLI degrees estimated from the label propagation (LP) procedure (Steps 1-5) will be employed as the basis for subsequent prediction model induction (Steps 6-14). Therefore, quality of the estimated RLI information will have significant influence on the performance of RELIAB.

Due to the lack of multi-label data sets with known RLI information, several two-dimensional synthetic data sets are generated in this subsection to investigate how well the RLI degrees estimated by RELIAB's LP procedure can recover the ground-truth RLI information. Specifically, to generate one multi-label synthetic data set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq m\}$ with q possible class labels, any two-dimensional instance is drawn randomly according to the following Gaussian Mixture Model (GMM): $p(\boldsymbol{x}) = \sum_{l=1}^{q} \pi_l \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$.

For each Gaussian mixture component, the mixture coefficient π_l is set to be $\frac{1}{q}$. In addition, elements of the mean vector $\boldsymbol{\mu}_l$ are chosen randomly from the pool $\{0, 0.5, 1.0, 1.5, 2.0\}$, and diagonal values of the diagonal covariance matrix $\boldsymbol{\Sigma}_l$ are chosen randomly from the pool $\{0.5, 1.0, 1.5, 2.0\}$.

For each instance $\boldsymbol{x}_i$ drawn according to the GMM distribution, the posteriori probability of $\boldsymbol{x}_i$ belonging to the j-th mixture component will be regarded as the *ground-truth* RLI degree of label y_j for $\boldsymbol{x}_i$, i.e.:

$$p(y_j \mid \boldsymbol{x}_i) = \frac{\pi_j \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{q} \pi_l \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \quad (1 \leq j \leq q) \quad (16)$$

The set of relevant labels Y_i for $\boldsymbol{x}_i$ is determined by thresholding the RLI degree against the actual mixture component responsible for generating $\boldsymbol{x}_i$.

Table II
THE KL-DIVERGENCE BETWEEN THE ESTIMATED AND THE GROUND-TRUTH RLI DEGREES (DENOTED AS "LP"), AS WELL AS THAT BETWEEN THE PRIOR AND THE GROUND-TRUTH RLI DEGREES (DENOTED AS "NONLP"). RESULTS ARE REPORTED FOR DIFFERENT SETTINGS OF m (# SYNTHETIC INSTANCES) AND q (# CLASS LABELS).

	$m = 1000$					
	$q = 5$	$q = 6$	$q = 7$	$q = 8$	$q = 9$	$q = 10$
LP	0.073	0.098	0.095	0.106	0.110	0.100
nonLP	1.587	1.583	1.638	1.584	1.680	1.622
	$m = 5000$					
	$q = 5$	$q = 6$	$q = 7$	$q = 8$	$q = 9$	$q = 10$
LP	0.088	0.079	0.094	0.111	0.104	0.115
nonLP	1.409	1.584	1.636	1.604	1.584	1.621

To evaluate the quality of the RLI degrees estimated by RELIAB's LP procedure, Table II reports the average KL-divergence between the *estimated* and the *ground-truth* RLI degrees over each example in the data set. Furthermore, to illustrate the helpfulness of the LP procedure, the KL-divergence between the *prior* (i.e. setting the RLI degree to each relevant label as $\frac{1}{|Y_i|}$) and the *ground-truth* RLI degrees is also reported.

As shown in Table II, it is intriguing to see that the LP procedure has good capability in recovering the ground-truth RLI degrees, where the KL-divergence has been much improved compared to the prior RLI degrees and is shown to take small values (around 0.1). Next, extensive experiments are conducted to validate the effectiveness of the proposed RELIAB approach.

B. Experimental Setup

1) Data Sets: For comprehensive performance evaluation, a total of seventeen benchmark multi-label data sets have been collected for experimental studies.[3] For each multi-label data set $\mathcal{S}$, we use $|\mathcal{S}|$, $dim(\mathcal{S})$, $L(\mathcal{S})$ and $F(\mathcal{S})$ to represent its number of examples, number of features, number of class labels and feature type respectively. In addition, several multi-label statistics [20] are further used to characterize properties of the data set, including label cardinality $LCard(\mathcal{S})$, label density $LDen(\mathcal{S})$, distinct label sets $DL(\mathcal{S})$ and proportion of distinct label sets $PDL(\mathcal{S})$. Detailed definitions on these properties can be found in [20].

Table III summarizes detailed characteristics of the benchmark data sets, which are roughly organized in ascending

[3]Publicly available at http://mulan.sourceforge.net/datasets.html and http://meka.sourceforge.net/#datasets

Table III
CHARACTERISTICS OF THE BENCHMARK MULTI-LABEL DATA SETS.

Data set	$\|\mathcal{S}\|$	$dim(\mathcal{S})$	$L(\mathcal{S})$	$F(\mathcal{S})$	$LCard(\mathcal{S})$	$LDen(\mathcal{S})$	$DL(\mathcal{S})$	$PDL(\mathcal{S})$	Domain
cal500	502	68	174	numeric	26.044	0.150	502	1.000	audio
emotions	593	72	6	numeric	1.868	0.311	27	0.046	audio
medical	978	1,449	45	nominal	1.245	0.028	94	0.096	text
llog	1,460	1,004	75	nominal	1.180	0.016	304	0.208	text
msra	1,868	898	19	numeric	6.315	0.332	947	0.507	image
image	2,000	294	5	numeric	1.236	0.247	20	0.010	image
scene	2.407	294	5	numeric	1.074	0.179	15	0.006	image
yeast	2.417	103	14	numeric	4.237	0.303	198	0.082	biology
slashdot	3,782	1,079	22	nominal	1.181	0.054	156	0.041	text
corel5k	5,000	499	374	nominal	3.522	0.009	3,175	0.635	image
rcv1-s1	6,000	500	101	nominal	2.880	0.029	1,028	0.171	text
rcv1-s2	6,000	500	101	nominal	2.634	0.026	954	0.159	text
rcv1-s3	6,000	500	101	nominal	2.614	0.026	939	0.156	text
rcv1-s4	6,000	500	101	nominal	2.484	0.025	816	0.136	text
rcv1-s5	6,000	500	101	nominal	2.642	0.026	946	0.158	text
bibtex	7,395	1836	159	nominal	2.402	0.015	2,856	0.386	text
mediamill	43,907	120	101	numeric	4.376	0.043	6,555	0.149	video

order of $|\mathcal{S}|$, with nine of them being regular-scale (first part, $|\mathcal{S}| < 5,000$) and eight of them being large-scale (second part, $|\mathcal{S}| \geq 5,000$). As shown in Table III, the seventeen data sets cover a broad range of cases with diversified multi-label properties and thus serve as a solid basis for thorough comparative studies.

2) Comparing Algorithms: In this paper, we choose to compare the performance of RELIAB against four well-established multi-label learning algorithms [30], including first-order approach *binary relevance* (BR) [1], second-order approach *calibrated label ranking* (CLR) [8], and high-order approaches *ensemble of classifier chains* (ECC) [20] and *random k-labelsets* (RAKEL) [23].

As shown in Eq.(9), the parametric predictor employed by RELIAB can be viewed equivalently as multinomial logistic regression models. Accordingly, each of the four comparing algorithms are implemented under the MULAN multi-label learning package [24] by instantiating their base learners with logistic regression models. Furthermore, parameters suggested in the literatures are used for ECC and RAKEL (ECC: ensemble size 30; RAKEL: ensemble size $2q$ with $k = 3$). For RELIAB, the balancing parameter α is fixed to be 0.5 which yields stable performance across the experimental data sets. In addition, the initial threshold-importance parameter τ and the regularization parameter β are chosen among $\{0.1, 0.15, \ldots, 0.5\}$ and $\{10^{-3}, 10^{-2}, \ldots, 10\}$ respectively by conducting cross-validation on training set.

3) Evaluation Protocol: A number of evaluation metrics specific to multi-label learning have been proposed, which can be generally categorized into two groups [22], [30], i.e. *example-based* metrics and *label-based* metrics. Example-based metrics work by evaluating the predictor's performance on each test example separately and then returning the *mean value* across all test examples. On the other hand, label-based metrics work by evaluating the predictor's performance on each label separately and then returning the *macro/micro-averaged value* across all class labels.

In this paper, six widely-used multi-label metrics are employed for performance evaluation, including four example-based metrics: *one-error*, *coverage*, *ranking loss*, *average precision*, and two label-based metrics: *macro-averaging F1*, *micro-averaging F1*. These evaluation metrics consider the performance of multi-label predictor from various aspects, whose values all vary between [0,1].[4] For *one-error*, *coverage* and *ranking loss*, the *smaller* the values the better the performance. For the other three metrics, the *larger* the values the better the performance.

For each comparing algorithm, ten-fold cross-validation is performed on regular-scale data sets (first part of Table III) while five-fold cross-validation is performed on large-scale data sets (second part of Table III). Accordingly, the mean metric value as well as the standard deviation are recorded for comparative studies.

C. Experimental Results

Tables IV and V report the detailed experimental results of all comparing algorithms on the regular-scale and large-scale data sets respectively. For each evaluation metric, "$\downarrow$" indicates "the smaller the better" while "$\uparrow$" indicates "the larger the better". Furthermore, the best performance among the five comparing algorithms is shown in boldface.

To analyze the relative performance among the comparing algorithms systematically, *Friedman test* [6] is used here which is regarded as the favorable statistical test for comparisons among *multiple algorithms* over a number of data sets. Table VI summarizes the Friedman statistics F_F

[4]Concrete metric definitions can be found in [30]. In addition, the *coverage* metric is normalized by the number of class labels (i.e. q).

Table IV
PREDICTIVE PERFORMANCE OF EACH COMPARING ALGORITHM (MEAN±STD. DEVIATION) ON THE NINE REGULAR-SCALE DATA SETS.

Comparing algorithm	cal500	emotions	medical	llog	msra	image	scene	yeast	slashdot
	One-error ↓								
RELIAB	**0.129±0.019**	**0.273±0.019**	**0.160±0.012**	**0.745±0.007**	**0.066±0.014**	**0.348±0.016**	**0.248±0.007**	**0.223±0.011**	0.509±0.014
BR	0.906±0.025	0.375±0.027	0.306±0.031	0.885±0.013	0.362±0.013	0.527±0.011	0.472±0.016	0.284±0.010	0.731±0.014
CLR	0.375±0.118	0.356±0.030	0.706±0.149	0.883±0.023	0.152±0.009	0.502±0.016	0.367±0.017	0.272±0.012	0.978±0.003
ECC	0.255±0.028	0.353±0.040	0.187±0.016	0.794±0.011	0.211±0.011	0.475±0.011	0.378±0.015	0.261±0.010	**0.476±0.015**
RAKEL	0.672±0.029	0.394±0.027	0.252±0.025	0.876±0.015	0.288±0.014	0.498±0.013	0.440±0.016	0.297±0.012	0.596±0.011
Comparing algorithm	cal500	emotions	medical	llog	msra	image	scene	yeast	slashdot
	Coverage ↓								
RELIAB	**0.744±0.008**	**0.304±0.014**	**0.045±0.007**	**0.156±0.005**	**0.545±0.012**	**0.204±0.005**	**0.099±0.003**	**0.453±0.007**	**0.138±0.002**
BR	0.877±0.009	0.364±0.015	0.117±0.018	0.380±0.006	0.716±0.004	0.297±0.009	0.209±0.010	0.479±0.007	0.261±0.009
CLR	0.792±0.014	0.351±0.016	0.134±0.026	0.234±0.019	0.636±0.004	0.285±0.009	0.119±0.004	0.496±0.006	0.271±0.004
ECC	0.796±0.008	0.356±0.013	0.052±0.007	0.195±0.006	0.665±0.004	0.271±0.008	0.144±0.008	0.479±0.006	**0.138±0.006**
RAKEL	0.958±0.003	0.386±0.016	0.113±0.012	0.360±0.007	0.698±0.006	0.293±0.008	0.190±0.009	0.573±0.008	0.219±0.005
Comparing algorithm	cal500	emotions	medical	llog	msra	image	scene	yeast	slashdot
	Ranking loss ↓								
RELIAB	**0.179±0.003**	**0.165±0.011**	**0.030±0.006**	**0.121±0.004**	**0.134±0.008**	**0.185±0.006**	**0.081±0.002**	**0.171±0.006**	0.122±0.002
BR	0.266±0.005	0.233±0.016	0.089±0.013	0.329±0.005	0.287±0.004	0.309±0.010	0.230±0.012	0.191±0.005	0.242±0.009
CLR	0.248±0.029	0.222±0.014	0.114±0.024	0.197±0.017	0.207±0.003	0.291±0.010	0.125±0.005	0.200±0.005	0.258±0.005
ECC	0.218±0.004	0.227±0.017	0.036±0.006	0.156±0.005	0.238±0.004	0.273±0.010	0.154±0.008	0.193±0.005	**0.121±0.006**
RAKEL	0.342±0.003	0.260±0.016	0.087±0.009	0.309±0.006	0.260±0.004	0.303±0.009	0.209±0.010	0.254±0.006	0.198±0.005
Comparing algorithm	cal500	emotions	medical	llog	msra	image	scene	yeast	slashdot
	Average precision ↑								
RELIAB	**0.503±0.007**	**0.796±0.011**	**0.876±0.010**	**0.394±0.009**	**0.816±0.012**	**0.774±0.008**	**0.853±0.004**	**0.760±0.007**	0.613±0.010
BR	0.301±0.006	0.730±0.015	0.756±0.025	0.214±0.014	0.626±0.005	0.656±0.007	0.692±0.012	0.733±0.007	0.427±0.013
CLR	0.383±0.048	0.742±0.016	0.403±0.051	0.209±0.019	0.722±0.003	0.672±0.010	0.781±0.008	0.729±0.008	0.251±0.007
ECC	0.431±0.005	0.740±0.021	0.856±0.011	0.335±0.009	0.684±0.004	0.690±0.008	0.763±0.010	0.738±0.007	**0.631±0.012**
RAKEL	0.323±0.006	0.713±0.017	0.782±0.017	0.228±0.012	0.661±0.005	0.670±0.008	0.713±0.011	0.697±0.006	0.529±0.009
Comparing algorithm	cal500	emotions	medical	llog	msra	image	scene	yeast	slashdot
	Macro-averaging F1 ↑								
RELIAB	0.171±0.007	**0.642±0.009**	0.419±0.049	**0.128±0.032**	**0.565±0.015**	**0.586±0.014**	**0.664±0.031**	0.409±0.013	0.324±0.047
BR	0.172±0.003	0.564±0.022	0.422±0.032	0.110±0.022	0.454±0.005	0.473±0.006	0.541±0.011	0.392±0.006	0.290±0.011
CLR	0.108±0.037	0.575±0.018	0.175±0.048	0.105±0.032	0.481±0.007	0.472±0.007	0.581±0.008	0.398±0.008	0.104±0.003
ECC	0.116±0.005	0.557±0.022	**0.464±0.039**	0.121±0.026	0.455±0.007	0.473±0.012	0.575±0.015	0.393±0.006	**0.399±0.012**
RAKEL	**0.174±0.004**	0.569±0.021	0.443±0.040	0.119±0.020	0.435±0.010	0.486±0.011	0.556±0.014	**0.420±0.006**	0.346±0.009
Comparing algorithm	cal500	emotions	medical	llog	msra	image	scene	yeast	slashdot
	Micro-averaging F1 ↑								
RELIAB	**0.468±0.006**	**0.642±0.008**	0.695±0.013	**0.182±0.014**	**0.683±0.012**	**0.577±0.016**	**0.644±0.029**	**0.637±0.004**	0.430±0.010
BR	0.331±0.004	0.574±0.023	0.643±0.028	0.130±0.007	0.546±0.005	0.474±0.006	0.536±0.010	0.613±0.006	0.281±0.012
CLR	0.286±0.084	0.581±0.018	0.270±0.136	0.101±0.043	0.604±0.006	0.472±0.007	0.568±0.007	0.610±0.006	0.011±0.002
ECC	0.353±0.005	0.566±0.024	**0.751±0.017**	0.149±0.015	0.575±0.003	0.472±0.012	0.568±0.014	0.617±0.006	**0.480±0.015**
RAKEL	0.353±0.007	0.576±0.020	0.689±0.022	0.148±0.010	0.576±0.006	0.486±0.012	0.546±0.012	0.613±0.007	0.378±0.012

and the corresponding critical values on each evaluation metric. As shown in Table VI, at 0.05 significance level, the null hypothesis of indistinguishable performance among the comparing algorithms is clearly rejected on each evaluation metric. Consequently, *Bonferroni-Dunn test* [6] is employed as the post-hoc test to show the relative performance among the comparing algorithms, where RELIAB is treated as the control algorithm. Here, the average rank difference between RELIAB and one comparing algorithm is calibrated with the *critical difference* (CD). Accordingly, the performance between RELIAB and one comparing algorithm is deemed to be significantly different if their average ranks differ by at least one CD (CD=1.3547 in this paper: # comparing algorithms $k = 5$, # data sets $N = 17$).

Fig. 2 illustrates the CD diagrams [6] 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 algorithm whose average rank is within one CD to that of RELIAB is interconnected to each other with a thick line. Otherwise, it is considered to have significantly different performance against RELIAB.

Table VI
SUMMARY OF THE FRIEDMAN STATISTICS F_F IN TERMS OF EACH EVALUATION METRIC AND THE CRITICAL VALUE AT 0.05 SIGNIFICANCE LEVEL (# COMPARING ALGORITHMS $k = 5$, # DATA SETS $N = 17$).

Evaluation metric	F_F	critical value
One-error	25.3600	2.5153
Coverage	21.1110	
Ranking loss	22.0890	
Average precision	18.8190	
Macro-averaging F1	6.9365	
Micro-averaging F1	11.1360	

Based on the above experimental results, the following observations can be apparently made:

1) On regular-scale data sets (Table IV), across all the evaluation metrics, RELIAB ranks *1st* in 83.3% cases and ranks *2nd* in 11.1% cases; On large-scale data sets

Table V
PREDICTIVE PERFORMANCE OF EACH COMPARING ALGORITHM (MEAN±STD. DEVIATION) ON THE EIGHT LARGE-SCALE DATA SETS.

Comparing algorithm	***One-error*** ↓							
	corel5k	rcv1-s1	rcv1-s2	rcv1-s3	rcv1-s4	rcv1-s5	bibtex	mediamill
RELIAB	0.795±0.009	0.510±0.005	**0.479±0.006**	**0.487±0.007**	**0.466±0.008**	**0.467±0.012**	**0.418±0.007**	0.192±0.007
BR	0.921±0.004	0.736±0.006	0.758±0.008	0.755±0.003	0.737±0.010	0.763±0.008	0.880±0.004	0.185±0.004
CLR	**0.748±0.011**	0.503±0.006	0.549±0.006	0.549±0.025	0.584±0.076	0.678±0.092	0.514±0.003	**0.147±0.002**
ECC	0.911±0.004	**0.490±0.005**	0.515±0.007	0.512±0.006	0.485±0.004	0.495±0.005	0.907±0.003	0.158±0.002
RAKEL	0.867±0.004	0.626±0.008	0.622±0.008	0.637±0.008	0.618±0.010	0.614±0.013	0.779±0.015	0.200±0.003
Comparing algorithm	***Coverage*** ↓							
	corel5k	rcv1-s1	rcv1-s2	rcv1-s3	rcv1-s4	rcv1-s5	bibtex	mediamill
RELIAB	0.342±0.008	0.158±0.002	0.128±0.004	**0.130±0.004**	**0.118±0.005**	**0.123±0.004**	**0.113±0.003**	0.198±0.002
BR	0.757±0.007	0.411±0.004	0.377±0.006	0.366±0.003	0.314±0.005	0.366±0.004	0.434±0.007	0.136±0.001
CLR	**0.311±0.011**	**0.123±0.002**	**0.122±0.004**	**0.130±0.018**	0.152±0.044	0.204±0.041	0.136±0.002	**0.127±0.001**
ECC	0.889±0.004	0.176±0.002	0.168±0.006	0.166±0.003	0.148±0.003	0.160±0.004	0.460±0.006	0.132±0.001
RAKEL	0.855±0.005	0.457±0.011	0.387±0.009	0.370±0.005	0.354±0.009	0.380±0.010	0.401±0.008	0.503±0.001
Comparing algorithm	***Ranking loss*** ↓							
	corel5k	rcv1-s1	rcv1-s2	rcv1-s3	rcv1-s4	rcv1-s5	bibtex	mediamill
RELIAB	0.152±0.005	0.069±0.001	**0.054±0.002**	**0.055±0.002**	**0.050±0.002**	**0.051±0.001**	**0.063±0.002**	0.058±0.001
BR	0.416±0.006	0.214±0.002	0.213±0.004	0.207±0.002	0.169±0.004	0.204±0.004	0.280±0.002	0.036±0.001
CLR	**0.147±0.007**	**0.052±0.001**	0.055±0.002	0.063±0.015	0.083±0.037	0.125±0.035	0.080±0.002	**0.033±0.001**
ECC	0.600±0.005	0.079±0.00	0.079±0.003	0.078±0.002	0.070±0.001	0.074±0.002	0.307±0.006	0.036±0.001
RAKEL	0.547±0.004	0.245±0.008	0.225±0.007	0.216±0.003	0.204±0.007	0.220±0.005	0.250±0.006	0.190±0.001
Comparing algorithm	***Average precision*** ↑							
	corel5k	rcv1-s1	rcv1-s2	rcv1-s3	rcv1-s4	rcv1-s5	bibtex	mediamill
RELIAB	0.221±0.007	0.532±0.003	**0.583±0.006**	**0.583±0.005**	**0.607±0.002**	**0.589±0.007**	**0.562±0.003**	0.676±0.003
BR	0.122±0.003	0.334±0.003	0.340±0.008	0.340±0.002	0.372±0.007	0.342±0.007	0.186±0.005	0.738±0.001
CLR	**0.222±0.007**	**0.555±0.004**	0.542±0.004	0.527±0.040	0.459±0.013	0.312±0.014	0.469±0.002	**0.758±0.001**
ECC	0.093±0.004	0.528±0.004	0.536±0.004	0.538±0.005	0.565±0.001	0.547±0.004	0.151±0.004	0.750±0.001
RAKEL	0.125±0.002	0.371±0.005	0.401±0.006	0.398±0.004	0.425±0.006	0.405±0.003	0.249±0.007	0.573±0.001
Comparing algorithm	***Macro-averaging F1*** ↑							
	corel5k	rcv1-s1	rcv1-s2	rcv1-s3	rcv1-s4	rcv1-s5	bibtex	mediamill
RELIAB	**0.089±0.008**	**0.253±0.003**	**0.260±0.009**	**0.266±0.021**	**0.258±0.015**	**0.271±0.006**	**0.300±0.009**	0.053±0.001
BR	0.073±0.006	0.187±0.004	0.167±0.006	0.171±0.008	0.170±0.006	0.167±0.004	0.127±0.003	0.197±0.003
CLR	0.074±0.012	0.233±0.008	0.221±0.006	0.213±0.032	0.157±0.073	0.088±0.079	0.247±0.003	0.171±0.002
ECC	0.062±0.009	0.198±0.009	0.174±0.004	0.174±0.015	0.185±0.013	0.184±0.009	0.101±0.002	0.163±0.002
RAKEL	0.079±0.007	0.194±0.007	0.174±0.005	0.174±0.005	0.180±0.009	0.188±0.003	0.177±0.007	**0.206±0.002**
Comparing algorithm	***Micro-averaging F1*** ↑							
	corel5k	rcv1-s1	rcv1-s2	rcv1-s3	rcv1-s4	rcv1-s5	bibtex	mediamill
RELIAB	**0.178±0.008**	**0.428±0.012**	**0.459±0.007**	**0.449±0.010**	**0.472±0.005**	**0.462±0.007**	**0.378±0.015**	0.502±0.005
BR	0.120±0.002	0.291±0.002	0.282±0.005	0.279±0.002	0.298±0.002	0.289±0.005	0.128±0.003	0.576±0.001
CLR	0.113±0.023	0.392±0.005	0.365±0.004	0.358±0.027	0.305±0.010	0.182±0.121	0.260±0.003	**0.585±0.001**
ECC	0.102±0.005	0.359±0.005	0.338±0.006	0.337±0.006	0.368±0.002	0.364±0.009	0.102±0.003	0.568±0.001
RAKEL	0.134±0.003	0.311±0.002	0.309±0.003	0.306±0.005	0.326±0.004	0.320±0.005	0.174±0.007	0.576±0.001

(Table V), across all the evaluation metrics, RELIAB ranks *1st* in 68.7% cases and ranks *2nd* in 16.7% cases.

2) RELIAB achieves optimal (lowest) average rank in terms of each evaluation metric (Fig. 2(a)-(f)). Furthermore, RELIAB significantly outperforms BR on all the evaluation metrics.
3) RELIAB is comparable to RAKEL in terms of *macro-averaging F1* (Fig. 2(e)), comparable to CLR in terms of *coverage* (Fig. 2(b)) and *ranking loss* (Fig, 2(c)), and significantly outperforms RAKEL and CLR on all the other cases. The comparable performance between RELIAB and CLR on *ranking loss* is also noticeable, as CLR is designed to learn from multi-label data by optimizing this particular evaluation metric [8], [30].
4) RELIAB is comparable to ECC in terms of example-based evaluation metrics (Fig. 2(a)-(d)), and significantly outperforms ECC in terms of label-based evaluation metrics (Fig. 2(e)-(f)). It is worth noting that ensemble learning techniques has been utilized by ECC to improve generalization, and the number of base learners employed by ECC is M-times larger than those employed by RELIAB (as specified in Subsection IV-B2, ensemble size M for ECC is set to be 30 in this paper).

To summarize, RELIAB achieves rather competitive performance against the well-established multi-label learning algorithms across extensive benchmark data sets and diverse evaluation metrics, which validate the effectiveness of leveraging implicit RLI information to learn from multi-label data.

D. Further Analysis

In this subsection, one variant of RELIAB is implemented to further analyze certain properties of the proposed approach. As shown in Subsection II-B, the parametric model is learned by fitting the estimated labeling-importance as *regularized* with multi-label empirical loss. To show the

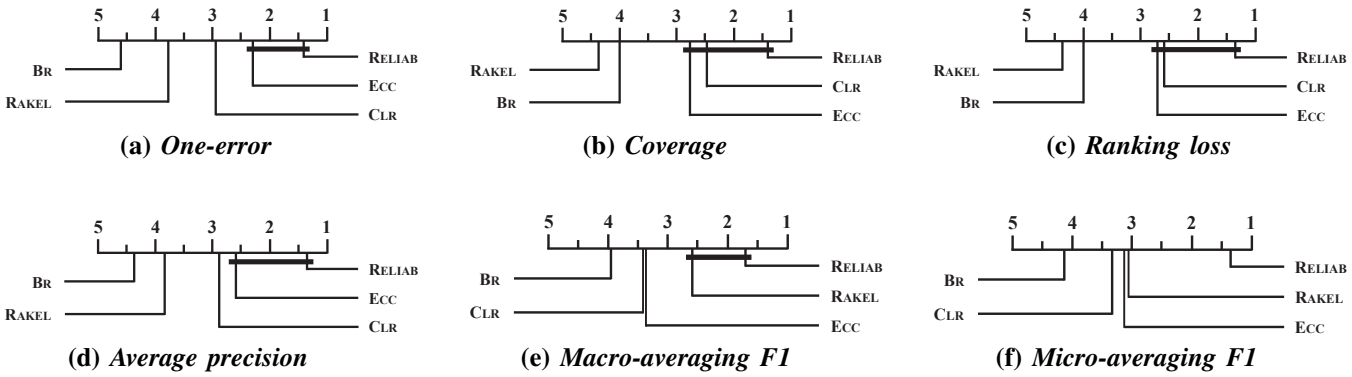

Figure 2. Comparison of RELIAB (control algorithm) against other comparing algorithms with the *Bonferroni-Dunn test*. Algorithms not connected with RELIAB in the CD diagram are considered to have significantly different performance from the control algorithm (CD=1.3547 at 0.05 significance level).

Table VII
WILCOXON SIGNED-RANKS TEST FOR RELIAB AGAINST ITS VARIANT RELIAB-NONREG IN TERMS OF EACH EVALUATION METRIC (AT 0.05 SIGNIFICANCE LEVEL; p-VALUES SHOWN IN THE BRACKETS).

Evaluation metric	RELIAB **against** RELIAB-nonReg
One-error	**win** [p=1.00e-3]
Coverage	**tie** [p=2.10e-1]
Ranking loss	**win** [p=9.13e-3]
Average precision	**win** [p=1.91e-2]
Macro-averaging F1	**tie** [p=6.19e-1]
Micro-averaging F1	**tie** [p=6.87e-1]

helpfulness of regularization, another variant of RELIAB is designed by dropping the regularization term $V_{emp}(f, \mathcal{D})$ (Eq.(11)) from the objective function of RELIAB. Thereafter, the resulting variant is denoted as RELIAB-nonREG.

Accordingly, the performance of RELIAB-nonREG is evaluated following the same protocol of Subsection IV-B3. Due to space limit, detailed experimental results of the variant are not reported here. Nonetheless, to show whether RELIAB performs significantly better than its variant, the Wilcoxon signed-ranks test [6], [26] is used here which is a desirable statistical test for comparisons between *two algorithms* over a number of data sets. Table VII summarizes the statistical test results at 0.05 significance level, where the p-values for the corresponding tests are also shown in the brackets.

As shown in Table VII, RELIAB achieves comparable performance against RELIAB-nREG on *Coverage*, *Macro-averaging F1* and *Micro-averaging F1*, while significantly outperforms RELIAB-nREG on all the other evaluation metrics. These results indicate that the regularization term based on multi-label empirical loss does help induce robust parametric models. Actually, the implicit RLI information exploited in the first objective term $V_{dis}(f, \mathcal{U})$ (Eq.(11)) are only estimations instead of being ground-truth values. In this case, optimizing objective function without necessary regularization is prone to produce unstable prediction models.

V. CONCLUSION

In this paper, the problem of multi-label learning is addressed by taking into account the fact that the relative labeling-importance is different for each label associated with the multi-label data. Accordingly, a novel multi-label learning approach named RELIAB is proposed, which works by leveraging the implicit RLI information derived from the training examples for model induction. Extensive comparative studies clearly validate the superiority of RELIAB against state-of-the-art multi-label learning approaches. In the future, we will explore if there exist better ways to estimate and make use of the implicit RLI information.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Science Foundation of China (61175049, 61222309, 61273300, 61232007), the Jiangsu Natural Science Funds for Distinguished Young Scholar (BK20140022), and the MOE Program for New Century Excellent Talents in University (NCET-13-0130).

REFERENCES

[1] 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.

[2] B. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for MIML instance annotation," in *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 534–542.

[3] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Cambridge, MA: MIT Press, 2011, pp. 190–198.

[4] W. Cheng, K. Dembczyński, and E. Hüllermeier, "Graded multilabel classification: The ordinal case," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp. 223–230.

[5] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, 1997.

[6] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. Jan, pp. 1–30, 2006.

[7] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, pp. 681–687.

[8] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, 2008.

[9] W. Gao and Z.-H. Zhou, "On the consistency of multi-label learning," *Artificial Intelligence*, vol. 199-200, pp. 22–44, 2013.

[10] X. Geng and P. Hou, "Pre-release prediction of crowd opinion on movies by label distribution learning," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 3511–3517.

[11] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1837–1842.

[12] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[13] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany, 2005, pp. 195–200.

[14] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Computing Surveys*, vol. 47, no. 3, p. Article 52, 2015.

[15] 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, 2010, Article 8.

[16] 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.

[17] G. Madjarov, D. Gjorgjevikj, and S. Džeroski, "Two stage architecture for multi-label learning," *Pattern Recognition*, vol. 45, no. 3, pp. 1019–1034, 2012.

[18] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: SIAM, 2000.

[19] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. Berlin: Springer, 2006.

[20] 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.

[21] 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.

[22] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*. Berlin: Springer, 2010, pp. 667–686.

[23] ——, "Random k-labelsets for multi-label classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[24] 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.

[25] J. Wang, Y. Zhao, X. Wu, and X.-S. Hua, "A transductive multi-label learning approach for video concept detection," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2274–2286, 2011.

[26] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, pp. 80–83, 1945.

[27] M. Xu, Y.-F. Li, and Z.-H. Zhou, "Multi-label learning with PRO loss," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue, WA, 2013, pp. 998–1004.

[28] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu, "Towards class-imbalance aware multi-label learning," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 4041–4047.

[29] 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.

[30] ——, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[31] 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 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, pp. 284–291.

[32] S. Zhu, X. Ji, W. Xu, and Y. Gong, "Multi-labelled classification using maximum entropy method," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, 2005, pp. 274–281.

[33] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," in *Synthesis Lectures to Artificial Intelligence and Machine Learning*, R. J. Brachman and T. G. Dietterich, Eds. San Francisco, CA: Morgan & Claypool Publishers, 2009, pp. 1–130.