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

Min-Ling Zhang, *Member, IEEE,* Qian-Wen Zhang, Jun-Peng Fang, Yu-Kun Li, and Xin Geng

Abstract—Multi-label learning deals with training examples each represented by a single instance while associated with multiple class labels, and the task is to train a predictive model which can assign a set of proper labels for the unseen instance. Existing approaches employ the common assumption of equal labeling-importance, i.e. all associated labels are regarded to be relevant to the training instance while their *relative* importance in characterizing its semantics are not differentiated. Nonetheless, this common assumption does not reflect the fact that the importance degree of each relevant label is generally different, though the importance information is not directly accessible from the training examples. In this paper, we show that it is beneficial to leverage the implicit *relative labeling-importance* (RLI) information to help induce multi-label predictive model with strong generalization performance. Specifically, RLI degrees are formalized as multinomial distribution over the label space, which can be estimated by either global label propagation procedure or local k-nearest neighbor reconstruction. Correspondingly, the multi-label predictive model is induced by fitting modeling outputs with estimated RLI degrees along with multi-label empirical loss regularization. Extensive experiments clearly validate that leveraging implicit RLI information serves as a favorable strategy to achieve effective multi-label learning.

Index Terms—Machine learning, multi-label learning, relative labeling-importance, label distribution, regularization

1 Introduction

Multi-label learning aims to model real-world objects with rich semantics, where each training example is represented by a single instance (feature vector) while associated with multiple class labels simultaneously [20], [46], [48]. Formally, 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 the multi-label training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq p\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is the d-dimensional instance and $Y_i \subseteq \mathcal{Y}$ is the set of relevant labels associated with $\boldsymbol{x}_i$, the task is to learn a multi-label predictive model $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from $\mathcal{D}$ which can assign a set of proper labels for the unseen instance. In recent years, multi-label learning techniques have been widely employed to learn from objects with rich semantics, such as text [30], image [6], audio [3], video [37], etc.

It is worth noting that the labeling information for multi-label training example $(\boldsymbol{x}_i, Y_i)$ is categorical, i.e. each class label $y \in \mathcal{Y}$ is regarded to be either relevant ($y \in Y_i$) or irrelevant ($y \notin Y_i$) for instance $\boldsymbol{x}_i$. Therefore, existing approaches learn from multi-label examples by taking the common assumption of equal labeling-importance, i.e. each relevant label contributes equally in characterizing semantics of the training example. However, for real-world multi-label learning problems, the importance degree of each associated relevant label is different by nature.

For one example, as shown in Fig. 1(a), a natural scene image may be annotated with labels *water*, *sky*, *trees* and *building* simultaneously where their (implicit) *relative labeling-importance* (RLI) for characterizing the semantics of this image are different due to varying scenery presence. Nonetheless, those RLI information are not explicitly provided by the annotator under standard multi-label learning setting. For another example, as shown in Fig. 1(b), a news document may be annotated with labels *sports*, *finance* and *venue* simultaneously where their implicit RLI for characterizing the semantics of this document are different due to varying topic length.[1] Similar scenarios arise for other types of multi-label data, such as the multiple sentiments associated with a piece of music would have different emotional presence [31], the multiple functionalities associated with a gene would have different expression levels [27], etc.

In general, it is beneficial to make use of the RLI information for multi-label learning. Specifically, the RLI information can be exploited as auxiliary supervision information to facilitate model induction, such as enforcing that the modeling output on relevant label with higher RLI degree is expected to be greater than the modeling output on relevant label with lower RLI degree. Therefore, the underlying relative importance among relevant labels should be differentiated, though these RLI information are not directly accessible from the training examples under standard multi-label learning setting.

In light of the above observations, we postulate that effective multi-label learning can be expected if the implicit RLI information is appropriately leveraged within model induction procedure. Accordingly, a novel multi-label learning approach named RELIAB, i.e. *RElative Labeling-Importance Aware multi-laBel learning*, is proposed. Firstly, the RLI degrees are formalized as multinomial distribution over the

• *Min-Ling Zhang, Qian-Wen Zhang, Jun-Peng Fang, Yu-Kun Li and Xin Geng are with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China. Email: {zhangml, zhangqw, fangjp, liyk, xgeng}@seu.edu.cn*

1. It is worth noting that the RLI marks given in Fig. 1 apply to a single object (instance). The alternative multi-label setting where each object is represented by a bag of instances, i.e. *multi-instance multi-label learning* (MIML) [48], [49], is not considered in this paper.

(a) (Implicit) RLI: *water* ≻ *sky* ≻ *trees* ≻ *building*

(b) (Implicit) RLI: *sports* ≻ *finance* ≻ *venue*

Fig. 1. Two illustrative objects each associated with multiple class labels simultaneously. The (implicit) relative labeling-importance (RLI) information are marked in either subfigure, which are not explicitly provided by the annotators under standard multi-label learning setting.

label space, which can be estimated by either invoking global label propagation procedure or conducting local k-nearest neighbor reconstruction. After that, the multi-label predictive model is induced by fitting predictive model with estimated RLI degrees along with multi-label empirical loss regularization. Comprehensive experimental studies validate the performance superiority of RELIAB against state-of-the-art compared algorithms as well as the quality of estimated RLI degrees.

The rest of this paper is organized as follows. Section 2 presents technical details of the proposed approach. Section 3 discusses existing works related to RELIAB. Section 4 reports experimental results of comparative studies. Finally, Section 5 concludes and indicates several issues for future work.

2 The Proposed Approach

In this section, we present the RELIAB approach which aims to learn from multi-label data by exploiting implicit RLI information. Firstly, the formal definition of RLI degree is introduced. After that, the two basic stages of RELIAB, i.e. *implicit RLI information estimation* and *predictive model induction*, are scrutinized respectively.

2.1 RLI Degree

As shown in Section 1, the goal of multi-label learning is to induce a multi-label predictor $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from the training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le p\}$. 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 $\boldsymbol{x}$. Conceptually, 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 for $\boldsymbol{x}$ can be determined as: $Y = \{y_l \mid \mu_{\boldsymbol{x}}^{y_l} > t(\boldsymbol{x}), 1 \le l \le q\}$, where $t(\boldsymbol{x})$ corresponds to the thresholding function which separates relevant labels from irrelevant ones for instance $\boldsymbol{x}$. Specifically, we augment the original label space $\mathcal{Y}$ into $\tilde{\mathcal{Y}} = \mathcal{Y} \bigcup \{y_0\}$, where y_0 is the complementary *virtual label* serving as an artificial bipartition point between relevant and irrelevant labels [14], [46]. In this case, $t(\boldsymbol{x})$ can be set to the thresholding-importance w.r.t. virtual label y_0, i.e. $t(\boldsymbol{x}) = \mu_{\boldsymbol{x}}^{y_0}$. 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 $\mu_{\boldsymbol{x}}^{y_l}$ $(0 \le l \le q)$, which satisfies the following constraints:

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

(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: $Y = \{y_l \mid \mu_{\boldsymbol{x}}^{y_l} > \mu_{\boldsymbol{x}}^{y_0}, 1 \le l \le q\}$.

Here, there are several issues which need to be noticed for the RLI degree formulation:

a) The RLI degree is not directly accessible from the multi-label training examples and thus *implicit* to the learning algorithm. Consequently, RLI degrees can be viewed as a refined version of the original categorical (relevant/irrelevant) labeling information and have to be derived from the given multi-label training set.

b) The RLI degree is instance-dependant which corresponds to the *relative* importance among all labels in characterizing the semantics of one particular instance. For instance, given two instances $\{\boldsymbol{x}, \boldsymbol{z}\}$ and two labels $\{y_l, y_m\}$, based on RLI degree we are only modeling and thus 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 $\mu_{\boldsymbol{z}}^{y_m}$).

c) The RLI degree for each instance, i.e. $\{\mu_{\boldsymbol{x}}^{y_l} \mid 0 \le l \le q\}$, can be viewed as a *label distribution* over the augmented label space $\tilde{\mathcal{Y}}$. For label distribution learning (LDL) [16], the label distribution information is assumed to be available for training examples. For multi-label learning, however, the RLI information needs to be further derived.

d) Under standard multi-label learning setting, it is assumed that the relevant label set Y for each training instance $\boldsymbol{x}$ is non-empty (i.e. $|Y| \geq 1$) [46], [48]. Therefore, there would be at least one relevant label $y_l \in Y$ whose RLI degree $\mu_{\boldsymbol{x}}^{y_l}$ is greater than that of the virtual label $\mu_{\boldsymbol{x}}^{y_0}$.

2.2 Implicit RLI Degree Estimation

In this paper, two simple yet effective modes are developed to show the feasibility of deriving RLI degrees from multi-label training examples. Specifically, to estimate the implicit RLI degree for all training examples, i.e. $\mathcal{U} = \{\mu_{\boldsymbol{x}_i}^{y_l} \mid 1 \leq i \leq p, 0 \leq l \leq q\}$, RELIAB employs either the global label propagation procedure or the local k-nearest neighbor reconstruction.

2.2.1 Global Label Propagation Procedure

For global style RLI degree estimation, the widely-used iterative label propagation techniques [47], [51] is adapted to fulfill the task. Let $G = (V, E)$ denote the fully-connected graph constructed over the set of training examples with vertices $V = \{\boldsymbol{x}_i \mid 1 \leq i \leq p\}$. Accordingly, a $p \times p$ symmetric similarity matrix $\mathbf{W} = [w_{ij}]_{p\times p}$ is specified for graph G as follows:

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

In this paper, the width parameter $\sigma > 0$ for similarity calculation is set to be 1.

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

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

Here, $\tau \in (0, 1)$ corresponds to the parameter of initial thresholding-importance for virtual label y_0. As shown in Eq.(2), at the initialization step, 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 according to the label propagation matrix $\mathbf{P}$:

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

Here, $\alpha \in (0, 1)$ corresponds to the parameter which balances 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:

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

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 $\mathbf{\Lambda} = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_p]$ 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.

Specifically, $\mathbf{S}$ is a stochastic matrix whose rows consist of non-negative entries and sum to one. According to the Perron-Frobenius theorem [25], [51], the absolute value of each eigenvalue of $\mathbf{S}$ satisfies $|\lambda_i| \leq 1$. 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:

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

By combining Eqs.(5) and (6), the label propagation procedure of Eq.(4) will converge to $\mathbf{F}^*$ as the number of iterations grow to infinity:

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

Thereafter, the implicit RLI degree for each label given a training example $\boldsymbol{x}_i$ is estimated by normalizing $\mathbf{F}^*$ on each row:

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

In other words, the 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 augmented label space $\tilde{\mathcal{Y}}$.

2.2.2 Local k-Nearest Neighbor Reconstruction

Other than the label propagation procedure which makes use of global relationship among all instances, another simple yet effective mode is proposed to deriving the implicit RLI degree in a local manner. Specifically, the popular k-nearest neighbor techniques are adapted to fulfill the task which have been widely used in solving multi-label learning problems [8], [36], [45].

For each multi-label training example $(\boldsymbol{x}_i, Y_i)$, let $\boldsymbol{y}_i = (y_{i0}, y_{i1}, \ldots, y_{iq})^\top$ denote the $(q+1)$-dimensional binary labeling vector w.r.t. the augmented label space $\tilde{\mathcal{Y}}$:

$$\forall_{l=0}^{q}: \quad y_{il} = \begin{cases} \tau, & \text{if } y_l = y_0 \\ 1, & \text{if } y_l \in Y_i \\ 0, & \text{if } y_l \notin Y_i \end{cases} \tag{9}$$

Furthermore, let $N(\boldsymbol{x}_i) = \{i_1, i_2, \ldots, i_k\}$ denote the index set for the k nearest neighbors of $\boldsymbol{x}_i$ identified in $\mathcal{D}$. Accordingly, let $\mathbf{X}_i = [\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_k}]$ be the $d \times k$ matrix storing all the nearest neighbors of $\boldsymbol{x}_i$, and $\mathbf{Y}_i = [\boldsymbol{y}_{i_1}, \boldsymbol{y}_{i_2}, \ldots, \boldsymbol{y}_{i_k}]$ be the $(q+1) \times k$ matrix storing all the corresponding labeling vectors.

To estimate the RLI degrees for $\boldsymbol{x}_i$, RELIAB models the local relationship among $\boldsymbol{x}_i$ and its k nearest neighbors by linear least squares reconstruction: $\min_{\boldsymbol{\beta}_i} ||\boldsymbol{x}_i - \mathbf{X}_i\boldsymbol{\beta}_i||^2$. Here, $\boldsymbol{\beta}_i$ corresponds to the reconstruction coefficients. Generally, $k \ll d$ holds and thus $\boldsymbol{\beta}_i$ can be solved as:

$$\boldsymbol{\beta}_i = \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^{-1} \mathbf{X}_i^\top \boldsymbol{x}_i \tag{10}$$

After that, a nonnegative confidence vector $\boldsymbol{g}_i = [g_{i0}, g_{i1}, \ldots, g_{iq}]^\top$ is set as:

$$\boldsymbol{g}_i = (\rho \boldsymbol{y}_i + (1-\rho)\mathbf{Y}_i\boldsymbol{\beta}_i)_+ \tag{11}$$

Here, ρ corresponds to the parameter which balances the fraction of labeling information inherited from the instance itself (i.e. $\boldsymbol{y}_i$) and the k nearest neighbors (i.e. $\mathbf{Y}_i$). Furthermore, $(\cdot)_+$ is the thresholding operator which turns negative entries of each vector into zero values.

Thereafter, the implicit RLI degrees for $\boldsymbol{x}_i$ are estimated by normalizing $\boldsymbol{g}_i$:

$$\forall_{l=0}^{q}: \quad \mu_{\boldsymbol{x}_i}^{y_l} = \frac{g_{il}}{\sum_{k=0}^{q} g_{ik}} \tag{12}$$

Similarly, the estimated RLI degrees can be regarded as a multinomial distribution over the augmented label space $\tilde{\mathcal{Y}}$.

2.3 Predictive Model Induction

In the second stage, RELIAB aims to induce the multi-label predictive model by leveraging the estimated RLI information, i.e. $\mathcal{U} = \{\mu_{\boldsymbol{x}_i}^{y_l} \mid 1 \le i \le p, 0 \le l \le q\}$. To enable exploitation of $\mathcal{U}$, we employ maximum entropy model [10] to parametrize the multi-label predictor:

$$\forall_{l=0}^{q}: \quad f(y_l \mid \boldsymbol{x}, \boldsymbol{\Theta}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\boldsymbol{\theta}_l^\top \boldsymbol{x}\right) \tag{13}$$

Here, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q]$ represents $q+1$ set of model parameters and $\boldsymbol{\theta}_l = [\theta_{l1}, \theta_{l2}, \ldots, \theta_{ld}]^\top$ is the d-dimensional 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 $\tilde{\mathcal{Y}}$, i.e. $\sum_{l=0}^{q} f(y_l \mid \boldsymbol{x}, \boldsymbol{\Theta}) = 1$. In this case, the multi-label predictor h can be derived from f by thresholding the outputs against the virtual label y_0:

$$h(\boldsymbol{x}) = \{y_l \mid f(y_l \mid \boldsymbol{x}, \boldsymbol{\Theta}) > f(y_0 \mid \boldsymbol{x}, \boldsymbol{\Theta}), 1 \le l \le q\} \tag{14}$$

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

$$V(f, \mathcal{U}, \mathcal{D}) = V_{dis}(f, \mathcal{U}) + \lambda \cdot V_{emp}(f, \mathcal{D}) \tag{15}$$

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

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

$$\begin{aligned} &V_{dis}(f, \mathcal{U}) \\ &= \sum_{i=1}^{p} \mathrm{KL}\left(\{\mu_{\boldsymbol{x}_i}^{y_l} \mid 0 \le l \le q\}, \{f(y_l \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) \mid 0 \le l \le q\}\right) \\ &= \sum_{i=1}^{p} \sum_{l=0}^{q} \left(\mu_{\boldsymbol{x}_i}^{y_l} \ln \frac{\mu_{\boldsymbol{x}_i}^{y_l}}{f(y_l \mid \boldsymbol{x}_i, \boldsymbol{\Theta})}\right) \end{aligned} \tag{16}$$

For the second term, $V_{emp}(f, \mathcal{D})$ can be measured by the *empirical loss* of the parametric model f on $\mathcal{D}$. As shown in Eq.(14), by taking the virtual label y_0 as the bipartition point, its modeling output $f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta})$ 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.(15) is instantiated as:

$$\begin{aligned} &V_{emp}(f, \mathcal{D}) \\ &= -\sum_{i=1}^{p} \left(\sum_{y_j \in Y_i} \Big(f(y_j \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) - f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big) \right. \\ &\qquad \left. + r_i \cdot \sum_{y_k \in \overline{Y}_i} \Big(f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) - f(y_k \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big) \right) \end{aligned} \tag{17}$$

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 [43]. Note that minimizing the loss in Eq.(17) can be viewed as minimizing one of the most popular multi-label metrics, namely the *ranking loss* [15], [20], [32], [46], 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.(17).

By substituting Eqs.(16) and (17) into the objective function and ignoring constant terms, Eq.(15) can then be rewritten as:

$$\begin{aligned} &V(f, \mathcal{U}, \mathcal{D}) = -\sum_{i=1}^{p} \sum_{l=0}^{q} \left(\mu_{\boldsymbol{x}_i}^{y_l} \ln f(y_l \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\right) \\ &-\lambda \cdot \sum_{i=1}^{p} \left(\sum_{y_j \in Y_i} \Big(f(y_j \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) - f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big) \right. \\ &\qquad \left. + r_i \cdot \sum_{y_k \in \overline{Y}_i} \Big(f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) - f(y_k \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big) \right) \end{aligned} \tag{18}$$

By minimizing Eq.(18), the final predictive model is obtained as: $f^* = \arg\min_f V(f, \mathcal{U}, \mathcal{D})$. To solve this 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 [26]. As a quasi-Newton algorithm, L-BFGS iteratively optimizes the objective function with resort to gradient of the function:

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

$$\frac{\partial V}{\partial \boldsymbol{\theta}_l} = -\sum_{i=1}^{p}\Big(\big(\mu_{\boldsymbol{x}_i}^{y_l} - f(y_l \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\big)\cdot \boldsymbol{x}_i\Big) - \beta \cdot \sum_{i=1}^{p}$$

$$\Bigg(f(y_l \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big(\sum_{y_j \in Y_i \setminus \{y_l\}}\Big(f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) - f(y_j \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big)$$

$$+ r_i \cdot \sum_{y_k \in \overline{Y}_i \setminus \{y_l\}}\Big(f(y_k \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) - f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big)$$

$$+\zeta(y_l, Y_i)\Big(1 - f(y_l \mid \boldsymbol{x}_i, \boldsymbol{\Theta}) + f(y_0 \mid \boldsymbol{x}_i, \boldsymbol{\Theta})\Big)\Big)\cdot \boldsymbol{x}_i\Bigg) \quad (19)$$

Here, $\zeta(y_l, Y_i)$ returns 0 if $y_l = y_0$. 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 1 summarizes the complete procedure of the proposed RELIAB approach. After incorporating the virtual label y_0 into the original label space (Step 1), the implicit RLI degrees are estimated by employing either the global label propagation procedure (Steps 3-6) or the local k-nearest neighbor reconstruction (Steps 8-13). Then, the multi-label predictive model is learned by leveraging the estimated RLI information (Steps 15-23). Finally, the predicted label set for unseen instance is determined by thresholding the modeling outputs against the virtual label (Step 24).

TABLE 1
The pseudo-code of RELIAB.

Inputs:

$\mathcal{D}$:	multi-label training set $\{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le p\}$ ($\boldsymbol{x}_i \in \mathcal{X}, Y_i \subseteq \mathcal{Y}, \mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{y_1, y_2, \ldots, y_q\}$)
mode:	mode (*global* or *local*) for RLI information estimation
τ:	initial thresholding-importance parameter $\tau \in (0, 1)$
α:	balancing parameter $\alpha \in (0, 1)$ for *global* mode
k, ρ:	number of nearest neighbors k and balancing parameter $\rho \in (0, 1)$ for *local* mode
λ:	regularization parameter for the objective function
$\boldsymbol{x}$:	unseen instance ($\boldsymbol{x} \in \mathcal{X}$)

Outputs:

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

Process:

1: Augment the original label space by introducing the virtual label y_0: $\tilde{\mathcal{Y}} = \{y_0\} \bigcup \mathcal{Y}$;
2: **if** mode = *global* **then**
3: Construct the similarity matrix $\mathbf{W} = [w_{ij}]_{p\times p}$ according to Eq.(1);
4: Construct the initial labeling-importance matrix $\boldsymbol{\Phi} = [\phi_{il}]_{p\times(q+1)}$ according to Eq.(2);
5: Conduct label propagation to yield the converged solution $\mathbf{F}^*$ according to Eq.(7);
6: Estimate the implicit RLI degrees $\{\mu_{\boldsymbol{x}_i}^{y_l} \mid 1 \le i \le p, 0 \le l \le q\}$ according to Eq.(8);
7: **else**
8: **for** $i = 1$ **to** p **do**
9: Identify the k nearest neighbors of $\boldsymbol{x}_i$ in $\mathcal{D}$ and form the neighbors matrix $\mathbf{X}_i$ and the labeling matrix $\mathbf{Y}_i$;
10: Obtain the reconstruction coefficients $\boldsymbol{\beta}_i$ according to Eq.(10);
11: Set the confidence vector $\boldsymbol{g}_i$ according to Eq.(11);
12: Estimate the implicit RLI degrees $\{\mu_{\boldsymbol{x}_i}^{y_l} \mid 0 \le l \le q\}$ according to Eq.(12);
13: **end for**
14: **end if**
15: Initialize model parameters $\boldsymbol{\Theta}^{(0)} = \frac{1}{d(q+1)} \cdot \mathbf{1}_{d\times(q+1)}$;
16: Set $t = 0$;
17: **repeat**
18: Evaluate $f(y_l \mid \boldsymbol{x}_i, \boldsymbol{\Theta}^{(t)})$ $(1 \le i \le p, 0 \le l \le q)$ according to Eq.(13);
19: Evaluate gradient $\frac{\partial V}{\partial \boldsymbol{\Theta}}\mid_{\boldsymbol{\Theta}^{(t)}}$ according to Eq.(19);
20: Update $\boldsymbol{\Theta}^{(t+1)}$ by running one L-BFGS iteration [26] with current parameters $\boldsymbol{\Theta}^{(t)}$ and gradient $\frac{\partial V}{\partial \boldsymbol{\Theta}}\mid_{\boldsymbol{\Theta}^{(t)}}$;
21: $t = t + 1$;
22: **until** convergence
23: Set the final prediction model f^* with $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}^{(t)}$;
24: Return $Y = h(\boldsymbol{x})$ according to Eq.(14).

2.4 Remarks

The RELIAB approach proposed in this paper serves as an initial attempt towards leveraging RLI information for learning from multi-label data. There are a few points which are noteworthy for the particular implementation employed by RELIAB:

a) In terms of implicit RLI degree estimation (Subsection 2.2), RELIAB relies on either the global spectral techniques of label propagation or the local similarity techniques of k-nearest neighbors. For iterative label propagation, it is originally designed for dealing with single-label examples [47], [51] while further adapted to fit multi-label scenario by introducing initial thresholding-importance (Eq.(2)) and normalization (Eq.(8)) to the confidence matrix $\mathbf{F}$. For k-nearest neighbors, it has been utilized to develop multi-label predictive model [8], [36], [45] while further adapted to help derive RLI degree via linear least squares reconstruction (Eq.(10)) and weighted aggregation (Eq.(11)).

b) The iterative label propagation procedure works in a *global* manner where the RLI degrees of each multi-label training example are estimated by synergizing information from all the other training examples. On the other hand, the k-nearest neighbor reconstruction works in a *local* manner where the RLI degrees of each multi-label training example are estimated by utilizing information from neighboring training examples. Conceptually, the former strategy has the advantage of exploiting structural information in the global feature space while may be misled by outlier examples. On the other hand, the latter strategy has the advantage of bearing the robustness of kNN estimation while may be less optimal without considering global information.

c) Intuitively, for either global label propagation procedure or local k-nearest neighbor reconstruction, the RLI information is estimated based on the *smoothness assumption* that examples close in feature space tend to share similar semantics in the label space. For instance, given multi-label examples $(\boldsymbol{x}, \{y_2, y_3, y_4\})$, $(\boldsymbol{x}', \{y_2, y_4\})$ and $(\boldsymbol{x}'', \{y_1, y_2\})$, where $\boldsymbol{x}$ is close to $\boldsymbol{x}'$ and $\boldsymbol{x}''$ in the feature space. Then, it

is reasonable to estimate the following order of RLI for $\boldsymbol{x}$: $y_2 \succ y_4 \succ y_3$, as y_2, y_4 and y_3 are shared by decreasing number of examples in $\{\boldsymbol{x}', \boldsymbol{x}''\}$. By exploiting the underlying relationships among training examples, the estimated RLI information serves as beneficial impetus for effective multi-label learning.

d) As shown in Table 1, parameters of the predictive model are optimized by invoking the iterative L-BFGS procedure (Steps 17-22). Here, the iterative procedure converges if the L-BFGS stopping criterion $\frac{||\boldsymbol{g}||}{\max(1,||\boldsymbol{\Theta}||)} < \epsilon$ ($\boldsymbol{g}$: projected gradient; $\epsilon = 0.001$) is met or the maximum number of iterations (80) is reached. Let t represent the resulting number of L-BFGS iterations, the training complexity for RELIAB corresponds to $O(p^3 \cdot (q+1) + t \cdot d \cdot (q+1))$ for *global* mode and $O(p \cdot d \cdot (p + (q+1) \cdot k \cdot d) + t \cdot d \cdot (q+1))$ for *local* mode. The testing complexity for either mode is $O(d \cdot (q+1))$.

e) Multi-label learning can be regarded as one specific instantiation of the general multi-output (or multi-target) prediction framework where each label can be assigned numerical or categorical values [1], [38], [46]. In this paper, the estimated RLI degrees are incorporated into training procedure based on the maximum entropy model. Alternatively, those RLI information can also be coupled with existing multi-output learning techniques [1], [38] for model induction. Nonetheless, it is worth noting that in multi-output learning the numerical labeling information are generally assumed to be readily available from the training examples, while in multi-label learning the RLI information are implicit and need to be estimated from the training examples.

3 RELATED WORK

Existing works related to RELIAB are briefly discussed in this section, while comprehensive reviews on multi-label learning can be found in recent surveys [20], [32], [44], [46].

Based on the *order of label correlations* being considered, most approaches to multi-label learning can be roughly grouped into three categories, i.e. first-order approaches assuming independence among class labels [2], [44], [45], second-order approaches considering pairwise correlations between class labels [13], [14], [21], and high-order approaches considering correlations among label subsets or all class labels [5], [29], [33]. For whichever order of correlations, the common modeling strategy is to treat each label categorically, 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], [7], an *ordinal scale* is assumed to characterize the membership degree and an ordinal grade is assigned for each label of the training example. In [40], 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 thus accessible to the learning algorithm. However, RELIAB differs from them fundamentally 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 $\boldsymbol{y} = (y_1, y_2, \ldots, y_q) \in \{-1, +1\}^q$ conditioned on the instance $\boldsymbol{x}$ [19], [50]. 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. In contrast, the MaxEnt model employed by RELIAB (Eq.(13)) 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.

4 EXPERIMENTS

In this section, extensive comparative studies on the proposed RELIAB approach and other state-of-the-art multi-label learning algorithms are conducted. Firstly, experimental setup including data sets, compared algorithms and evaluation metrics are introduced. Secondly, detailed experimental results are reported with statistical performance comparisons. Thirdly, properties of the proposed approaches are further investigated.

In terms of implicit RLI degree estimation, the RELIAB approach instantiated with global label propagation procedure or local k-nearest neighbor reconstruction are denoted as RELIAB-LP or RELIAB-KNN respectively.

4.1 Experimental Setup

4.1.1 Data Sets

To thoroughly evaluate the performance of compared algorithms, a total of seventeen benchmark multi-label data sets are employed for experimental studies.[2] For each multi-label data set $\mathcal{S} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq r\}$, we use $|\mathcal{S}|$, $dim(\mathcal{S})$, $L(\mathcal{S})$ and $F(\mathcal{S})$ to represent its *number of examples*, *number of features*, *total number of class labels* and *feature type* respectively. Properties of the data set can be further characterized by several multi-label statistics, including label cardinality $LCard(\mathcal{S})$, label density $LDen(\mathcal{S})$, distinct label sets $DL(\mathcal{S})$ and proportion of distinct label sets, whose definitions can be found in $PDL(\mathcal{S})$ [29], [46]:

Table 2 summarizes detailed characteristics of the experimental data sets used in this paper. Here, data sets are organized in ascending 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 2, 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.

4.1.2 Evaluation Metrics

Given the multi-label data set $\mathcal{S} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq r\}$, let $h : \mathcal{X} \rightarrow 2^{\mathcal{Y}}$ be the multi-label predictive model and $\{f_1, f_2, \ldots, f_q\}$ be the corresponding set of q real-valued functions with each $f_l : \mathcal{X} \rightarrow \mathbb{R}$ $(1 \leq l \leq q)$ determining the relevancy of class label y_l. As per the ranking nature of the maximum entropy model employed by RELIAB, four

2. Publicly available at http://mulan.sourceforge.net/datasets.html and http://meka.sourceforge.net/#datasets

TABLE 2
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

popular ranking-based multi-label metrics are used for performance evaluation [20], [46], [48]:

- *One-error*: $\frac{1}{r}\sum_{i=1}^{r}[\![\arg\max_{y_l\in\mathcal{Y}} f_l(\boldsymbol{x}_i)] \notin Y_i]\!]$
  Here, $[\![a]\!]$ returns 1 if a holds and 0 otherwise.
- *Coverage*: $\frac{1}{q}\left(\frac{1}{r}\sum_{i=1}^{r}\max_{y_l\in Y_i} rank(\boldsymbol{x}_i, y_l) - 1\right)$
  Here, $rank(\boldsymbol{x}_i, y_l) = \sum_{j=1}^{q}[\![f_j(\boldsymbol{x}_i) \geq f_l(\boldsymbol{x}_i)]\!]$
- *Ranking loss*: $\frac{1}{r}\sum_{i=1}^{r}\frac{|\mathcal{Z}_i|}{|Y_i||\bar{Y}_i|}$
  Here, $\mathcal{Z}_i = \{(y_l, y_j) \mid f_l(\boldsymbol{x}_i) \leq f_j(\boldsymbol{x}_i), (y_l, y_j) \in Y_i \times \bar{Y}_i\}$ and $\bar{Y}_i = \mathcal{Y} \setminus Y_i$
- *Average precision*: $\frac{1}{r}\sum_{i=1}^{r}\frac{1}{|Y_i|}\sum_{y_l\in Y_i}\frac{\mathcal{R}(\boldsymbol{x}_i, y_l)}{rank(\boldsymbol{x}_i, y_l)}$
  Here, $\mathcal{R}(\boldsymbol{x}_i, y_l) = \{y_j \mid f_j(\boldsymbol{x}_i) \geq f_l(\boldsymbol{x}_i),\ y_j \in Y_i\}$

Furthermore, two classification-based multi-label metrics are also used in this paper:

- *Macro-averaging F1*: $\frac{1}{q}\sum_{l=1}^{q}\frac{2\cdot TP_l}{2\cdot TP_l + FN_l + FP_l}$
  Here, $TP_l = |\{\boldsymbol{x}_i \mid y_l \in Y_i \wedge y_l \in h(\boldsymbol{x}_i),\ 1 \leq i \leq r\}|$
  $FP_l = |\{\boldsymbol{x}_i \mid y_l \notin Y_i \wedge y_l \in h(\boldsymbol{x}_i),\ 1 \leq i \leq r\}|$
  $FN_l = |\{\boldsymbol{x}_i \mid y_l \in Y_i \wedge y_l \notin h(\boldsymbol{x}_i),\ 1 \leq i \leq r\}|$
- *Micro-averaging F1*: $\frac{\sum_{l=1}^{q} 2\cdot TP_l}{\sum_{l=1}^{q}(2\cdot TP_l + FN_l + FP_l)}$

Conceptually, *one-error* considers examples whose top-ranked label is not a relevant label, *coverage* considers how many steps should be moved along the ranked label list to cover all relevant labels, *ranking loss* considers the pair of relevant-irrelevant labels which have been reversely ordered, and *average precision* considers the average fraction of relevant labels ranked higher than a particular relevant label. Furthermore, *macro-averaging F1* and *micro-averaging F1* evaluates the averaged F1 value by assuming "equal weights" for labels and examples respectively. The values for all evaluation metrics are normalized between [0,1], where we have the *smaller* the values the better the performance for *one-error*, *coverage* and *ranking loss*, and the *larger* the values the better the performance for other three metrics.

It is worth noting that there are some other multi-label metrics which can be used for performance evaluation while the corresponding results are not reported here due to limited space. Furthermore, compared to other classification-based metrics such as *hamming loss*, *zero-one loss* and *accuracy*, the employed *macro-/micro-averaging F1* is more sensitive to the inherent *class-imbalance* property of multi-label data [9], [28], [43]. For ranking-based metrics, in case of prediction ties among class labels, it is also desirable to consider using *partial ranking loss* which allows consistent convex surrogate loss function [15].

4.1.3 Compared Algorithms

In this paper, the performance of RELIAB-LP and RELIAB-KNN are compared against four well-established multi-label learning algorithms which have been widely employed for comparative studies in multi-label learning [20], [32], [44], [46]:

- *Binary relevance* (BR) [2]: A *first-order* approach which decomposes the multi-label learning problem into q independent binary classification problems, where each of them corresponds to one possible class label in $\mathcal{Y}$.
- *Calibrated label ranking* (CLR) [14]: A *second-order* approach which transforms the multi-label learning problem into a label ranking problem, where a total of $\binom{q}{2}$ binary classifiers are trained to yield the ranking among labels and further bi-partitioned via threshold calibration.
- *Ensembles of classifier chains* (ECC) [29]: A *high-order* approach which transforms the multi-label learning problem into a chain of binary classification problems, where predictions of preceding binary classifiers are used as extra features to build subsequent ones in the chain. Furthermore, ensemble learning is employed to address the randomness of chaining order.
- *Random k-labelsets* (RAKEL) [33]: A *high-order* approach which transforms the multi-label learning problem into an ensemble of *multi-class classification* problems, where each component learner in the ensemble is induced by applying label powerset techniques [46] on a random k-labelset in $\mathcal{Y}$.

TABLE 3
Predictive performance of each compared algorithm (mean±std. deviation) on the nine regular-scale data sets. The best and second best performance among all the compared algorithms are shown in • and ○ respectively.

Compared algorithm	cal500	emotions	medical	llog	msra	image	scene	yeast	slashdot
	One-error ↓								
RELIAB-LP	0.122±0.049 •	0.277±0.036	0.141±0.030 •	0.746±0.030 •	0.061±0.016 •	0.329±0.035 •	0.262±0.032 ○	0.238±0.018 ○	0.518±0.025
RELIAB-KNN	0.122±0.046 •	0.246±0.035 ○	0.143±0.033 ○	0.767±0.042 ○	0.063±0.021 ○	0.335±0.025 ○	0.257±0.030 •	0.237±0.016 •	0.495±0.018 ○
BR	0.900±0.036	0.298±0.045	0.324±0.054	0.893±0.023	0.324±0.023	0.379±0.017	0.374±0.032	0.250±0.020	0.668±0.029
CLR	0.269±0.061	0.322±0.032	0.360±0.170	0.830±0.058	0.144±0.027	0.437±0.019	0.344±0.027	0.241±0.012	0.979±0.005
ECC	0.253±0.052 ○	0.310±0.036	0.182±0.040	0.818±0.015	0.178±0.030	0.411±0.031	0.327±0.038	0.245±0.016	0.490±0.030 •
RAKEL	0.611±0.084	0.315±0.074	0.246±0.038	0.879±0.026	0.239±0.031	0.412±0.029	0.339±0.027	0.280±0.016	0.615±0.020
RANK-SVM	0.189±0.024	0.244±0.031 •	0.385±0.032	0.892±0.018	0.163±0.023	0.395±0.021	0.348±0.032	0.262±0.021	0.430±0.021
GFM	0.487±0.076	0.284±0.065	0.534±0.056	0.896±0.051	0.274±0.045	0.472±0.039	0.371±0.042	0.288±0.041	0.668±0.034
	Coverage ↓								
RELIAB-LP	0.743±0.019 •	0.299±0.029	0.044±0.014 ○	0.162±0.021 ○	0.536±0.012 ○	0.196±0.016 ○	0.106±0.014 ○	0.458±0.005 •	0.134±0.007 •
RELIAB-KNN	0.746±0.019 ○	0.284±0.032 ○	0.036±0.012 •	0.156±0.018 •	0.532±0.012 •	0.193±0.014 •	0.105±0.014 •	0.459±0.008 ○	0.135±0.007 ○
BR	0.786±0.014	0.306±0.024	0.111±0.027	0.376±0.028	0.694±0.015	0.218±0.016	0.182±0.019	0.470±0.008	0.246±0.010
CLR	0.795±0.008	0.334±0.020	0.080±0.068	0.186±0.044	0.618±0.013	0.247±0.016	0.137±0.017	0.480±0.008	0.258±0.009
ECC	0.794±0.017	0.314±0.015	0.054±0.015	0.189±0.020	0.634±0.012	0.238±0.022	0.145±0.018	0.469±0.008	0.138±0.009
RAKEL	0.964±0.006	0.348±0.021	0.089±0.019	0.340±0.023	0.670±0.010	0.253±0.009	0.174±0.015	0.564±0.008	0.218±0.012
RANK-SVM	0.743±0.013 •	0.279±0.024 •	0.047±0.020	0.187±0.033	0.599±0.031	0.229±0.008	0.111±0.012	0.479±0.012	0.174±0.016
GFM	0.884±0.065	0.342±0.031	0.112±0.029	0.321±0.024	0.637±0.046	0.278±0.012	0.153±0.028	0.582±0.033	0.252±0.022
	Ranking loss ↓								
RELIAB-LP	0.176±0.004 •	0.161±0.031	0.028±0.010 ○	0.129±0.019 ○	0.127±0.005 ○	0.177±0.018 ○	0.089±0.015 ○	0.177±0.008 ○	0.118±0.007 •
RELIAB-KNN	0.178±0.005 ○	0.145±0.032 ○	0.023±0.009 •	0.123±0.017 •	0.125±0.006 •	0.173±0.016 •	0.088±0.014 •	0.174±0.008 •	0.119±0.008 ○
BR	0.233±0.008	0.173±0.025	0.089±0.021	0.328±0.030	0.254±0.009	0.205±0.017	0.163±0.018	0.183±0.006	0.225±0.012
CLR	0.224±0.008	0.199±0.024	0.065±0.059	0.152±0.039	0.190±0.009	0.243±0.018	0.119±0.016	0.187±0.005	0.245±0.010
ECC	0.219±0.008	0.184±0.018	0.038±0.013	0.153±0.019	0.209±0.010	0.230±0.027	0.127±0.016	0.186±0.006	0.122±0.009
RAKEL	0.364±0.014	0.217±0.026	0.067±0.015	0.292±0.028	0.232±0.011	0.250±0.012	0.154±0.014	0.250±0.005	0.198±0.013
RANK-SVM	0.184±0.015	0.142±0.021 •	0.036±0.014	0.212±0.016	0.178±0.014	0.217±0.034	0.115±0.054	0.186±0.009	0.152±0.010
GFM	0.210±0.041	0.188±0.019	0.078±0.026	0.276±0.021	0.246±0.023	0.249±0.046	0.137±0.029	0.194±0.008	0.174±0.009
	Average precision ↑								
RELIAB-LP	0.513±0.017 •	0.797±0.028	0.889±0.020 •	0.392±0.034 •	0.823±0.009 ○	0.785±0.019 •	0.844±0.020 ○	0.749±0.012 ○	0.607±0.017
RELIAB-KNN	0.512±0.020 ○	0.815±0.027	0.889±0.021 •	0.377±0.040 ○	0.826±0.010 •	0.785±0.015 •	0.847±0.019 •	0.754±0.011 •	0.628±0.015
BR	0.343±0.017	0.785±0.025	0.748±0.039	0.207±0.024	0.660±0.010	0.756±0.011	0.763±0.021	0.747±0.010	0.473±0.018
CLR	0.436±0.019	0.762±0.024	0.687±0.192	0.295±0.075	0.741±0.013	0.718±0.014	0.795±0.018	0.745±0.008	0.261±0.006
ECC	0.443±0.020	0.777±0.022	0.860±0.028 ○	0.314±0.017	0.717±0.011	0.733±0.021	0.798±0.023	0.747±0.008	0.621±0.021
RAKEL	0.332±0.019	0.766±0.031	0.802±0.027	0.233±0.026	0.694±0.014	0.725±0.013	0.780±0.018	0.710±0.009	0.516±0.012
RANK-SVM	0.481±0.029	0.817±0.021 ○	0.733±0.021	0.245±0.043	0.759±0.024	0.743±0.012	0.793±0.021	0.741±0.028	0.652±0.012 ○
GFM	0.510±0.017	0.822±0.044 •	0.712±0.042	0.344±0.016	0.812±0.032	0.775±0.030 ○	0.841±0.017	0.742±0.023	0.667±0.025 •
	Macro-averaging F1 ↑								
RELIAB-LP	0.297±0.021 •	0.650±0.039	0.722±0.061 •	0.397±0.057	0.558±0.013	0.604±0.016 ○	0.677±0.021 ○	0.409±0.019	0.417±0.036 ○
RELIAB-KNN	0.284±0.023	0.671±0.045 ○	0.702±0.051	0.420±0.058 ○	0.567±0.014 ○	0.602±0.020	0.670±0.016	0.418±0.022	0.375±0.079
BR	0.166±0.017	0.613±0.039	0.622±0.066	0.271±0.029	0.488±0.016	0.549±0.016	0.610±0.024	0.391±0.015	0.359±0.032
CLR	0.211±0.025	0.601±0.038	0.600±0.129	0.395±0.062	0.499±0.017	0.525±0.022	0.620±0.025	0.400±0.018	0.165±0.035
ECC	0.231±0.024	0.615±0.046	0.706±0.061	0.398±0.057	0.487±0.018	0.531±0.020	0.643±0.027	0.397±0.013	0.416±0.047
RAKEL	0.187±0.020	0.618±0.036	0.672±0.058	0.366±0.051	0.492±0.020	0.540±0.012	0.644±0.019	0.430±0.014 ○	0.363±0.033
RANK-SVM	0.233±0.012	0.601±0.017	0.666±0.018	0.369±0.023	0.491±0.013	0.529±0.036	0.633±0.011	0.393±0.048	0.302±0.026
GFM	0.291±0.025 ○	0.680±0.031 •	0.707±0.021 ○	0.431±0.060 •	0.586±0.021 •	0.610±0.024 •	0.715±0.057 •	0.446±0.024 •	0.570±0.054 •
	Micro-averaging F1 ↑								
RELIAB-LP	0.480±0.010 ○	0.647±0.036	0.722±0.024	0.191±0.043	0.717±0.007	0.603±0.017 ○	0.663±0.022 ○	0.642±0.011	0.437±0.016
RELIAB-KNN	0.482±0.015 •	0.680±0.042 ○	0.776±0.020 ○	0.248±0.031 ○	0.723±0.008 ○	0.602±0.020	0.657±0.017	0.653±0.010 ○	0.431±0.014
BR	0.341±0.019	0.632±0.038	0.616±0.046	0.134±0.013	0.588±0.014	0.549±0.014	0.608±0.025	0.626±0.010	0.330±0.013
CLR	0.326±0.019	0.614±0.037	0.598±0.157	0.176±0.049	0.624±0.010	0.525±0.019	0.612±0.026	0.628±0.012	0.008±0.003
ECC	0.357±0.020	0.632±0.041	0.752±0.030	0.151±0.029	0.614±0.013	0.532±0.017	0.637±0.029	0.635±0.009	0.438±0.021 ○
RAKEL	0.355±0.018	0.634±0.031	0.685±0.031	0.148±0.027	0.613±0.015	0.540±0.011	0.636±0.023	0.632±0.009	0.362±0.014
RANK-SVM	0.388±0.025	0.622±0.038	0.683±0.024	0.153±0.020	0.604±0.016	0.516±0.023	0.611±0.012	0.601±0.010	0.338±0.044
GFM	0.480±0.012 ○	0.697±0.024 •	0.801±0.015 •	0.267±0.033 •	0.746±0.031 •	0.607±0.031 •	0.693±0.025 •	0.667±0.008 •	0.549±0.054 •

As shown in Eq.(13), the parametric predictor employed by RELIAB can also be viewed as multinomial logistic regression models. Accordingly, each of the four compared algorithms are implemented under the MULAN multi-label learning package [34] where their base learners are instantiated with logistic regression models. Furthermore, parameters suggested in corresponding literatures are used for ECC and RAKEL (ECC: ensemble size 30; RAKEL: ensemble size $2q$ with $k = 3$). For RELIAB, the initial threshold-importance parameter τ and the regularization parameter λ are chosen among $\{0.1, 0.15, \ldots, 0.5\}$ and $\{10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 10\}$ via cross-validation on training set.[3] In addition, for RELIAB-LP, the balancing parameter α is fixed to be 0.5. For RELIAB-KNN, the number of nearest neighbors k is fixed to be 10 and 30 for regular-scale and large-scale data sets respectively, the balancing parameter ρ is fixed to be 0.3.

Note that RELIAB makes uses of the virtual label y_0 (Eq.(14)) to help partition the ranked labels into relevant and irrelevant label set, which has been widely-used in

3. Here, τ and λ are tuned via cross-validation on the training partition for each fold of the main ten-fold cross-validation experiment.

TABLE 4
Predictive performance of each compared algorithm (mean±std. deviation) on the eight large-scale data sets. The best and second best performance among all the compared algorithms are shown in • and ○ respectively.

Compared algorithm	One-error ↓							
	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5	bibtex	mediamill
RELIAB-LP	0.718±0.018 •	0.458±0.019 ○	0.468±0.011 ○	0.475±0.028 ○	0.455±0.020 ○	0.464±0.009	0.396±0.015 •	0.191±0.009
RELIAB-KNN	0.788±0.011	0.496±0.025	0.463±0.011 •	0.462±0.028 •	0.453±0.019 •	0.460±0.018 •	0.406±0.031 ○	0.187±0.010
BR	0.925±0.009	0.723±0.020	0.728±0.025	0.740±0.015	0.728±0.015	0.748±0.013	0.856±0.012	0.159±0.007
CLR	0.741±0.018	0.501±0.027	0.507±0.019	0.533±0.037	0.499±0.017	0.503±0.018	0.470±0.028	0.146±0.007 •
ECC	0.732±0.022 ○	0.453±0.022 •	0.478±0.017	0.480±0.020	0.459±0.021	0.461±0.018 ○	0.466±0.019	0.149±0.009 ○
RAKEL	0.872±0.014	0.623±0.023	0.592±0.022	0.598±0.018	0.592±0.013	0.595±0.021	0.675±0.015	0.193±0.008
RANK-SVM	0.791±0.027	0.510±0.013	0.604±0.022	0.532±0.016	0.498±0.010	0.532±0.029	0.489±0.020	0.183±0.009
GFM	0.803±0.021	0.568±0.019	0.688±0.026	0.612±0.013	0.652±0.018	0.611±0.024	0.639±0.011	0.199±0.010
Compared algorithm	Coverage ↓							
	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5	bibtex	mediamill
RELIAB-LP	0.296±0.012 ○	0.135±0.009	0.122±0.006	0.124±0.007	0.110±0.010	0.117±0.007	0.097±0.004 •	0.195±0.003
RELIAB-KNN	0.336±0.019	0.110±0.008 •	0.102±0.006 •	0.104±0.005 •	0.088±0.010 •	0.100±0.006 •	0.108±0.009 ○	0.194±0.004
BR	0.682±0.015	0.371±0.015	0.337±0.020	0.330±0.019	0.299±0.014	0.325±0.011	0.404±0.014	0.129±0.002 ○
CLR	0.287±0.015 •	0.112±0.008 ○	0.105±0.006 ○	0.114±0.024 ○	0.095±0.007 ○	0.107±0.006 ○	0.115±0.006	0.121±0.001 •
ECC	0.434±0.017	0.166±0.009	0.152±0.008	0.152±0.005	0.130±0.012	0.149±0.008	0.224±0.007	0.151±0.010
RAKEL	0.874±0.012	0.417±0.012	0.359±0.022	0.369±0.014	0.358±0.020	0.363±0.015	0.352±0.015	0.504±0.004
RANK-SVM	0.338±0.014	0.142±0.021	0.297±0.013	0.298±0.013	0.256±0.012	0.344±0.014	0.296±0.017	0.358±0.013
GFM	0.442±0.016	0.230±0.022	0.311±0.018	0.310±0.015	0.286±0.014	0.367±0.024	0.291±0.012	0.379±0.017
Compared algorithm	Ranking loss ↓							
	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5	bibtex	mediamill
RELIAB-LP	0.128±0.005 •	0.057±0.004	0.050±0.003	0.052±0.002 ○	0.045±0.004	0.047±0.003	0.052±0.003 •	0.057±0.001
RELIAB-KNN	0.143±0.009	0.047±0.003 •	0.041±0.002 •	0.043±0.002 •	0.036±0.004 •	0.040±0.002 •	0.059±0.007 ○	0.055±0.001
BR	0.349±0.009	0.172±0.007	0.167±0.009	0.168±0.009	0.150±0.007	0.157±0.005	0.256±0.011	0.034±0.001 ○
CLR	0.131±0.008 ○	0.048±0.002 ○	0.046±0.002 ○	0.054±0.020	0.044±0.002 ○	0.046±0.003	0.066±0.003	0.031±0.001 •
ECC	0.191±0.009	0.071±0.003	0.068±0.003	0.069±0.002	0.059±0.005	0.065±0.004	0.126±0.005	0.041±0.003
RAKEL	0.586±0.011	0.233±0.007	0.209±0.012	0.222±0.009	0.224±0.013	0.209±0.012	0.211±0.010	0.190±0.001
RANK-SVM	0.144±0.009	0.067±0.008	0.101±0.022	0.095±0.014	0.054±0.018	0.044±0.014 ○	0.065±0.012	0.059±0.002
GFM	0.289±0.014	0.145±0.010	0.211±0.011	0.193±0.010	0.201±0.015	0.215±0.013	0.254±0.012	0.203±0.014
Compared algorithm	Average precision ↑							
	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5	bibtex	mediamill
RELIAB-LP	0.271±0.011 •	0.569±0.012 ○	0.592±0.009 ○	0.588±0.014 ○	0.617±0.011 ○	0.596±0.009 ○	0.592±0.013 •	0.678±0.006
RELIAB-KNN	0.247±0.010	0.574±0.016 •	0.626±0.008 •	0.626±0.016 •	0.648±0.011 •	0.622±0.008 •	0.574±0.021 ○	0.688±0.003
BR	0.137±0.005	0.377±0.012	0.386±0.014	0.385±0.011	0.402±0.010	0.384±0.008	0.216±0.007	0.749±0.004
CLR	0.247±0.009	0.564±0.012	0.579±0.011	0.554±0.050	0.589±0.013	0.576±0.012	0.515±0.018	0.764±0.003 •
ECC	0.241±0.011	0.559±0.012	0.579±0.009	0.574±0.011	0.599±0.015	0.582±0.009	0.464±0.012	0.755±0.004 ○
RAKEL	0.120±0.007	0.391±0.009	0.429±0.010	0.423±0.010	0.431±0.011	0.421±0.012	0.333±0.016	0.575±0.004
RANK-SVM	0.228±0.007	0.489±0.021	0.469±0.023	0.498±0.017	0.433±0.021	0.502±0.011	0.482±0.018	0.598±0.009
GFM	0.267±0.010 ○	0.568±0.015	0.580±0.016	0.582±0.015	0.602±0.015	0.591±0.012	0.556±0.020	0.650±0.014
Compared algorithm	Macro-averaging F1 ↑							
	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5	bibtex	mediamill
RELIAB-LP	0.304±0.024 ○	0.325±0.025 ○	0.340±0.022	0.336±0.022 ○	0.352±0.028 ○	0.346±0.017 ○	0.294±0.019 ○	0.075±0.006
RELIAB-KNN	0.296±0.019	0.316±0.022	0.343±0.027 ○	0.334±0.018	0.348±0.030	0.341±0.015	0.254±0.014	0.039±0.006
BR	0.214±0.011	0.241±0.023	0.222±0.018	0.232±0.017	0.252±0.016	0.226±0.011	0.144±0.005	0.204±0.007
CLR	0.276±0.015	0.278±0.028	0.269±0.016	0.255±0.035	0.297±0.023	0.286±0.013	0.288±0.010	0.177±0.008
ECC	0.284±0.017	0.283±0.030	0.265±0.021	0.248±0.018	0.298±0.033	0.273±0.020	0.266±0.016	0.168±0.009
RAKEL	0.257±0.013	0.266±0.029	0.237±0.024	0.243±0.023	0.256±0.020	0.255±0.016	0.217±0.007	0.212±0.008 ○
RANK-SVM	0.278±0.016	0.262±0.024	0.278±0.011	0.252±0.022	0.276±0.024	0.284±0.012	0.269±0.019	0.156±0.004
GFM	0.365±0.027 •	0.387±0.015 •	0.366±0.018 •	0.376±0.013 •	0.360±0.021 •	0.391±0.019 •	0.321±0.012 •	0.233±0.010 •
Compared algorithm	Micro-averaging F1 ↑							
	corel5k	rcv1subset1	rcv1subset2	rcv1subset3	rcv1subset4	rcv1subset5	bibtex	mediamill
RELIAB-LP	0.230±0.011 ○	0.390±0.007	0.432±0.008	0.429±0.012	0.428±0.012	0.434±0.012	0.402±0.016 ○	0.368±0.007
RELIAB-KNN	0.195±0.012	0.412±0.013 ○	0.476±0.014 •	0.477±0.010 •	0.471±0.017 •	0.473±0.015 •	0.379±0.019	0.516±0.004
BR	0.122±0.007	0.322±0.009	0.317±0.008	0.318±0.009	0.326±0.011	0.323±0.007	0.158±0.004	0.578±0.004
CLR	0.123±0.019	0.361±0.008	0.356±0.015	0.338±0.029	0.365±0.017	0.368±0.014	0.326±0.008	0.585±0.003
ECC	0.091±0.015	0.381±0.016	0.384±0.014	0.374±0.014	0.411±0.017	0.402±0.009	0.375±0.017	0.602±0.014 ○
RAKEL	0.135±0.009	0.341±0.008	0.337±0.008	0.335±0.012	0.349±0.010	0.350±0.010	0.202±0.008	0.579±0.004
RANK-SVM	0.151±0.013	0.320±0.019	0.334±0.043	0.389±0.021	0.389±0.017	0.407±0.011	0.378±0.013	0.587±0.012
GFM	0.289±0.024 •	0.489±0.023 •	0.453±0.030 ○	0.460±0.016 ○	0.467±0.013 ○	0.470±0.015 ○	0.439±0.017 •	0.621±0.022 •

ranking-based multi-label classification [14], [39]. In addition to CLR, another well-established multi-label learning algorithm RANK-SVM [13], [46] is employed for comparative studies which directly optimizes the *ranking loss* metric. Furthermore, other than the four ranking-based metrics, two classification-based metrics *macro-averaging F1* and *micro-averaging F1* are also used for performance evaluation in this paper. In addition to BR and ECC which optimize classification-based metrics, the GFM approach [11] is also employed as the compared algorithm which is one of the representative algorithms which are tailored to maximize F-measure [11], [24], [28], [35].

For each compared algorithm, ten-fold cross-validation is performed on the nine regular-scale data sets (first part of Table 2) as well as the eight large-scale data sets (second part of Table 2). Accordingly, on each data set, the mean metric value as well as the standard deviation are recorded for comparative studies.

4.2 Experimental Results

Tables 3 and 4 report the detailed experimental results of all compared algorithms on the regular-scale and large-

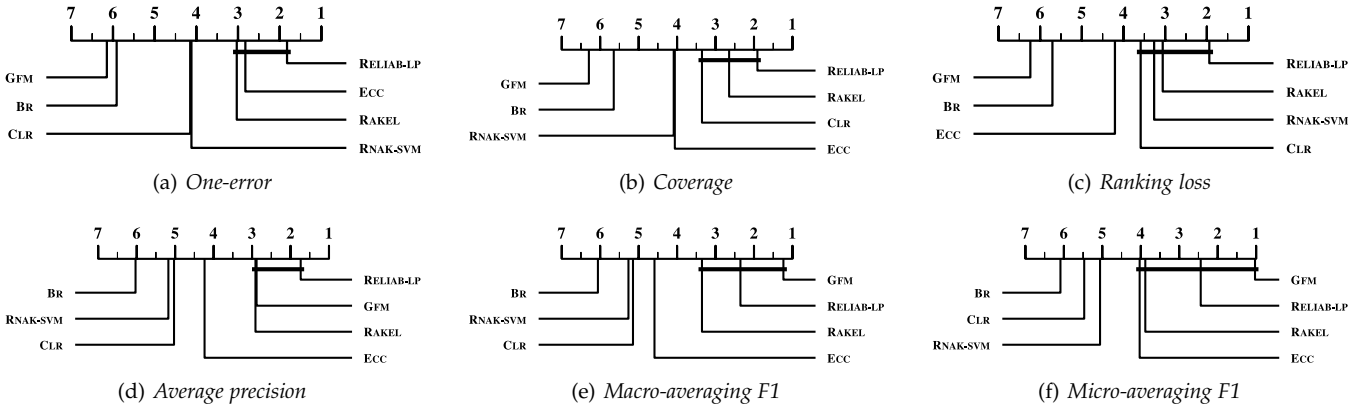


Fig. 2. Comparison of RELIAB-LP (control algorithm) against other compared algorithms with the *Bonferroni-Dunn test*. Algorithms not connected with RELIAB-LP in the CD diagram are considered to have significantly different performance from the control algorithm (CD=1.9546 at 0.05 significance level).

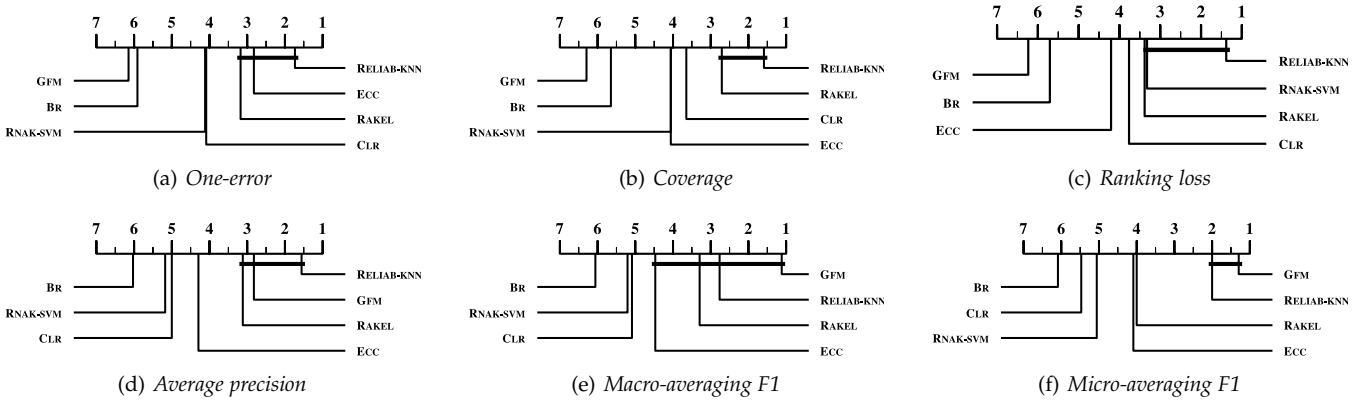


Fig. 3. Comparison of RELIAB-KNN (control algorithm) against other compared algorithms with the *Bonferroni-Dunn test*. Algorithms not connected with RELIAB-KNN in the CD diagram are considered to have significantly different performance from the control algorithm (CD=1.9546 at 0.05 significance level).

scale data sets respectively. For each evaluation metric, "↓" indicates "the smaller the better" while "↑" indicates "the larger the better".

To analyze the relative performance among the compared algorithms systematically, *Friedman test* [12] is used here which is regarded as the favorable statistical test for comparisons among *multiple algorithms* over a number of data sets. At 0.05 significance level, the Friedman statistics F_F ($F_F > 65$ on all evaluation metrics) is greater than the critical value 2.0924 (#compared algorithms $n = 8$; #data sets $N = 17$). Therefore, the null hypothesis of "equal" performance among the compared algorithms is clearly rejected.

To show the relative performance among the compared algorithms, *Bonferroni-Dunn test* [12] is employed as the post-hoc test by treating RELIAB-LP or RELIAB-KNN as the control algorithm. Here, the difference between the average ranks of control algorithm and one compared algorithm is calibrated with the *critical difference* (CD). The performance between control algorithm and one compared algorithm is deemed to be significantly different if their average ranks differ by at least one CD (CD=1.9546 with $n = 7$ and $N = 17$).

Figs. 2 and 3 illustrate the CD diagrams [12] on each evaluation metric by treating RELIAB-LP or RELIAB-KNN as the control algorithm respectively. Here, the average rank of each compared algorithm is marked along the axis with lower ranks to the right. In each subfigure, any compared algorithm whose average rank is within one CD to that of the control algorithm is interconnected to each other with a thick line. Otherwise, it is considered to have significantly different performance against the control algorithm.

The following observations can be made based on the reported experimental results:

1) On regular-scale data sets (Table 3), across all the evaluation metrics, RELIAB-LP ranks *1st* in 29.6% cases and ranks *2nd* in 40.7% cases while RELIAB-KNN ranks *1st* in 37.0% cases and ranks *2nd* in 42.6% cases. On large-scale data sets (Table 4), across all the evaluation metrics, RELIAB-LP ranks *1st* in 14.6% cases and ranks *2nd* in 39.6% cases while RELIAB-KNN ranks *1st* in 47.9% cases and ranks *2nd* in 12.5% cases.
2) Both RELIAB-LP and RELIAB-KNN achieve optimal (lowest) average rank in terms of *one-error*, *coverage*, *ranking loss* and *average precision* (Fig. 2(a)-(d), Fig. 3(a)-(d)), and significantly outperform BR on all evaluation metrics. Furthermore, RELIAB-KNN significantly outperforms CLR on all evaluation met-

TABLE 5
Quantitative analysis on the quality of estimated RLI information. On each data set, the relative ranking of compared algorithms in terms of each metric is shown in parenthesis.

Data Set	Kullback-Leibler Divergence ↓				Squared χ^2 ↓				Fidelity ↑			
	RELIAB-LP	RELIAB-KNN	RANDOM	AVERAGE	RELIAB-LP	RELIAB-KNN	RANDOM	AVERAGE	RELIAB-LP	RELIAB-KNN	RANDOM	AVERAGE
SBU_3DFE	0.2550 (2)	0.0796 (1)	8.9651 (4)	8.9580 (3)	0.2151 (2)	0.0764 (1)	0.9546 (4)	0.9484 (3)	0.9425 (2)	0.9805 (1)	0.3721 (4)	0.3739 (3)
SJAFFE	0.1787 (2)	0.0684 (1)	7.1037 (4)	7.0740 (3)	0.1574 (2)	0.0661 (1)	0.9447 (4)	0.9165 (3)	0.9586 (2)	0.9832 (1)	0.5901 (4)	0.5984 (3)
Natural Scene	0.4836 (1)	1.0237 (2)	2.9763 (4)	2.8763 (3)	0.4385 (3)	0.3153 (1)	0.4534 (4)	0.3814 (2)	0.8327 (3)	0.8831 (1)	0.8191 (4)	0.8396 (2)
Movie	0.2400 (1)	0.2618 (2)	4.6749 (4)	4.5798 (3)	0.1715 (1)	0.2148 (2)	0.6204 (4)	0.5425 (3)	0.9517 (1)	0.9406 (2)	0.7370 (4)	0.7589 (3)
Human Gene	0.1572 (1)	0.3024 (2)	4.5798 (4)	3.6307 (3)	0.1380 (1)	0.2330 (2)	0.5425 (3)	0.5970 (4)	0.9628 (1)	0.9327 (2)	0.7589 (3)	0.7476 (4)

rics.

3) Both RELIAB-LP and RELIAB-KNN are comparable to RANK-SVM in terms of *ranking loss* (Fig. 2(c), Fig. 3(c)) and achieve superior performance against RANK-SVM on the other evaluation metrics. Specifically, the comparable performance between the two variants of RELIAB and RANK-SVM on *ranking loss* is noticeable, as RANK-SVM is designed to learn from multi-label data by optimizing this particular evaluation metric [13], [46].
4) RELIAB-LP is comparable to ECC in terms of *one-error* (Fig. 2(a)) and *micro-averaging F1* (Fig. 2(f)), while RELIAB-KNN is comparable to ECC in terms of *one-error* (Fig. 3(a)) and *macro-averaging F1* (Fig. 3(e)). On all the other cases, both variants of RELIAB achieve superior performance against ECC. It is worth noting that ensemble learning techniques have been utilized by ECC to improve generalization performance, where the number of base learners employed by ECC is M-times larger than those employed by RELIAB (as specified in Subsection 4.1.3, ensemble size M for ECC is set to be 30 in this paper).
5) Both RELIAB-LP and RELIAB-KNN are comparable to GFM in terms of *macro-averaging F1* and *micro-averaging F1* (Fig. 2(e)-(f), Fig. 3(e)-(f)), though their average ranks are higher than that of GFM. It is also worth noting that GFM is designed to learn from multi-label data by optimizing the F-measure [11], [46].

4.3 Algorithmic Properties

4.3.1 Quality of Estimated RLI Information

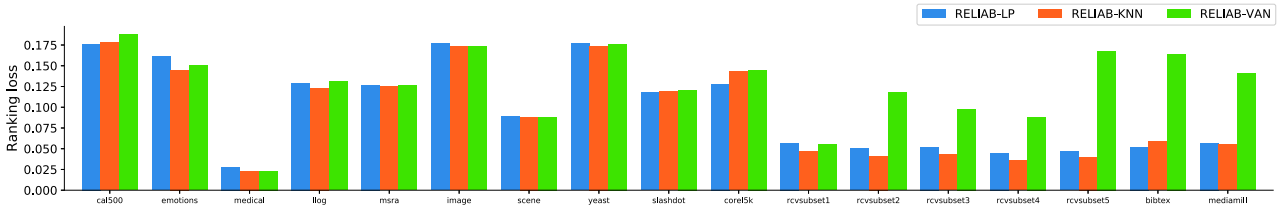
In addition to effectiveness of the proposed approach, another important issue regarding RELIAB lies in the quality of the estimated RLI information, i.e. how the RLI degrees derived from multi-label training examples coincide with the ground-truth RLI information.

To this end, other than those benchmark data sets used in Subsection 4.1.1 where the ground-truth RLI information is not available, we have tried to collect five multi-label data sets with known ground-truth RLI information for quantitative analysis:

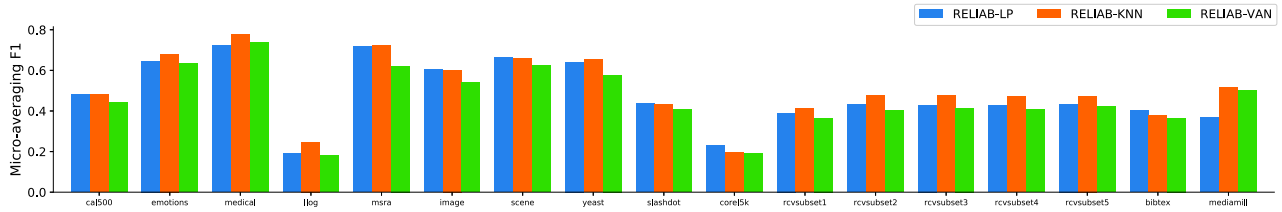
- *SBU_3DFE*: This is a 3D facial expression database [41], where each facial expression can be associated with emotions such as *happiness*, *sadness*, *surprise*, *fear*, *anger* and *disgust*. A total of 23 students are asked to annotate the level of emotion intensity (1 to 5) for each facial expression, where their averaged annotation intensities are used to yield the ground-truth RLI degrees. The resulting data set contains 2,500 examples with 243 features and 6 class labels.
- *SJAFFE*: This is a Japanese female facial expression database [23], where each facial expression can be associated with the same set of emotions as SBU_3DFE. Similarly, a total of 60 persons annotate the level of emotion intensity and their averaged annotation intensities are used to yield the ground-truth RLI degrees. The resulting data set contains 213 examples with 243 features and 6 class labels.
- *Natural Scene*: This is a natural scene image data set [18], where each image can be associated with scenes such as *plant*, *sky*, *cloud*, *snow*, *building*, *desert*, *mountain*, *water*, and *sun*. A total of 10 persons rank the relevant labels for each image, where their rankings are consolidated to yield the ground-truth RLI degrees. The resulting data set contains 2,000 examples with 294 features and 9 class labels.
- *Movie*: This is a movie pre-release ratings data set [17], where each movie can be associated with rating scales from 1 to 5 stars. A total of 54,242,292 ratings from 478,656 users over 7,755 movies are crawled from Netflix, where each movie has 6,994 ratings on average and the rating distributions are used to yield the ground-truth RLI degrees. The resulting data set contains 7,755 examples with 1869 features and 5 class labels.
- *Human Gene*: This is a human gene data set [42], where each gene can be associated with 68 kinds of possible diseases. The gene expression level for each disease is used to yield the ground-truth RLI degrees. The resulting data set contains 30,542 examples with 36 features and 68 class labels.

For each multi-label data set $\mathcal{S} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq r\}$, let $\mu_{\boldsymbol{x}_i}^{y_l}$ and $\eta_{\boldsymbol{x}_i}^{y_l}$ be the estimated and ground-truth RLI degree respectively. In this subsection, three popular measures are employed to quantify the quality of RLI information estimated by RELIAB:[4] a) *Kullback-Leibler Divergence*: $d_{KL} = \sum_{i=1}^{r}\sum_{l=1}^{q} \mu_{\boldsymbol{x}_i}^{y_l} \ln \frac{\mu_{\boldsymbol{x}_i}^{y_l}}{\eta_{\boldsymbol{x}_i}^{y_l}}$; b) *Squared* χ^2: $d_{\chi^2} = \sum_{i=1}^{r}\sum_{l=1}^{q} \frac{\left(\mu_{\boldsymbol{x}_i}^{y_l} - \eta_{\boldsymbol{x}_i}^{y_l}\right)^2}{\mu_{\boldsymbol{x}_i}^{y_l} + \eta_{\boldsymbol{x}_i}^{y_l}}$; c) *Fidelity*: $d_F = \sum_{i=1}^{r}\sum_{l=1}^{q} \sqrt{\mu_{\boldsymbol{x}_i}^{y_l} \cdot \eta_{\boldsymbol{x}_i}^{y_l}}$. For d_{KL} and d_{χ^2}, the *smaller* the values the better the performance. For d_F, the *larger* the values the better the performance.

4. In this case, the RLI information estimation procedure of RELIAB-LP and RELIAB-KNN are invoked without introducing the virtual label.



(a) *Ranking loss*



(b) *Micro-averaging F1*

Fig. 4. Performance comparison among RELIAB-LP, RELIAB-KNN and RELIAB-VAN in terms of *ranking loss* and *micro-averaging F1*.

TABLE 6
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	
	mode: *global*	mode: *local*
One-error	**win** [p=4.88e-4]	**win** [p=4.90e-3]
Coverage	**tie** [p=6.52e-2]	**tie** [p=7.11e-1]
Ranking loss	**win** [p=3.40e-3]	**tie** [p=7.50e-1]
Average precision	**win** [p=2.44e-4]	**win** [p=2.34e-2]
Macro-averaging F1	**win** [p=3.58e-2]	**win** [p=3.40e-3]
Micro-averaging F1	**tie** [p=9.85e-1]	**win** [p=2.27e-2]

TABLE 7
Wilcoxon signed-ranks test for RELIAB against its variant RELIAB-VAN in terms of each evaluation metric (at 0.05 significance level; p-values shown in the brackets).

Evaluation metric	RELIAB **against** RELIAB-VAN	
	mode: *global*	mode: *local*
One-error	**win** [p=2.03e-2]	**win** [p=1.33e-3]
Coverage	**win** [p=4.17e-2]	**win** [p=1.98e-2]
Ranking loss	**win** [p=4.66e-2]	**win** [p=9.70e-4]
Average precision	**win** [p=2.60e-2]	**win** [p=7.10e-4]
Macro-averaging F1	**win** [p=8.60e-3]	**win** [p=8.83e-2]
Micro-averaging F1	**win** [p=9.88e-3]	**win** [p=2.92e-4]

Table 5 reports the detailed metric values of each algorithm on the five multi-label data sets, which measure the quality of estimated RLI degrees w.r.t. the ground-truth ones. Due to the lack of multi-label learning algorithms which can also yield RLI degrees estimation, two baseline algorithms are utilized for comparative studies: a) RANDOM which assigns RLI degree randomly to relevant labels of each example; b) AVERAGE which assigns RLI degree uniformly (i.e. $\frac{1}{|Y_i|}$) to relevant labels of each example.

As shown in Table 5, only on the *Natural Scene* data set, RELIAB-LP has slightly worse estimation quality than AVERAGE in terms of *Squared* χ^2 and *Fidelity*. On all the other cases, the estimation quality of both RELIAB-LP and RELIAB-KNN significantly outperform RANDOM and AVERAGE. These results indicate that the two variants of RELIAB have good capability in recovering ground-truth RLI information, which would lead to their favorable generalization performance as reported in Subsection 4.2.

4.3.2 Usefulness of Regularization

As shown in Eq.(15), the parametric model is learned by fitting the estimated RLI degrees as *regularized* with multi-label empirical loss. Other than the estimated RLI degrees which serve as informative resource for designing the first term $V_{dis}(f, \mathcal{U})$, we show the helpfulness of regularization by considering a simplified version of RELIAB. Here, the regularization term $V_{emp}(f, \mathcal{D})$ in Eq.(15) is dropped from the objective function and the resulting version is denoted as RELIAB-nonREG.

Following the same evaluation protocol of Subsection 4.1.2, the performance of RELIAB-nonREG is investigated as well. For brevity, detailed experimental results of RELIAB-nonREG are not reported here. Nonetheless, to show whether RELIAB performs significantly better than its simplified version, the Wilcoxon signed-ranks test [12] is used which is a desirable statistical test for comparisons between *two algorithms* over a number of data sets. Table 6 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 6, whether the global label propagation mode or the local k-nearest neighbor mode is utilized for implicit RLI degree estimation, RELIAB achieves superior or at least comparable performance to RELIAB-nonREG across all evaluation metrics. These results clearly validate the usefulness of empirical loss regularization term for improving generalization performance.

4.3.3 Usefulness of Exploiting RLI Information

The RLI information leveraged by RELIAB is not explicit prior knowledge but needs to be estimated from the training examples. To show whether RELIAB can truly exploit the estimated RLI information to our advantage, a vanilla variant of RELIAB (termed as RELIAB-VAN) is employed here

which returns equal RLI degree over all class labels in the first stage and follows the same procedure of RELIAB in the second stage for model induction.

Fig. 4 illustrates the performance of RELIAB-LP, RELIAB-KNN and RELIAB-VAN in terms of two evaluation metrics due to limited space. As shown in Table 7, based on Wilcoxon signed-ranks tests at 0.05 significance level, both RELIAB-LP and RELIAB-KNN achieve significantly better performance against RELIAB-VAN in terms of all evaluation metrics. These results indicate that RELIAB would be appropriate for problems where the relevant labels associated with multi-label examples have inherent relative labeling-importance.

5 CONCLUSION

Existing approaches learn from multi-label examples by assuming equal labeling-importance, where each relevant label contributes equally to the learning procedure. In this paper, an extension to our earlier research [22] is presented which works by leveraging the implicit RLI information derived from the training examples for model induction. Accordingly, two simple yet effective strategies are developed for estimating RLI degrees, which are further leveraged to induce multi-label predictive model with empirical loss regularization. Comparative studies validate the quality of RLI information estimated by the proposed approach as well as the benefits of leveraging them for effective multi-label learning.

In the future, other than the label propagation and k-nearest neighbor techniques, it is interesting to explore other ways for implicit RLI information estimation. It is also interesting to further investigate the quality of RLI degrees estimated by RELIAB on data sets with ground-truth RLI information over higher number of class labels.

ACKNOWLEDGEMENTS

The authors would like to thank the associate editor and the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key R&D Program of China (2018YFB1004300), the National Science Foundation of China (61573104), and partially supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[2] 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.

[3] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.

[4] C. Brinker, E. Loza Mencia, and J. Fürnkranz, "Graded multilabel classification by pairwise comparisons," in *Proceedings of the 14th IEEE International Conference on Data Mining*, Shenzhen, China, 2014, pp. 731–736.

[5] S. Burkhardt and S. Kramer, "Online multi-label dependency topic models for text classification," *Machine Learning*, vol. 107, no. 5, pp. 859–886, 2018.

[6] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for weakly-supervised multi-label image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 121–135, 2015.

[7] 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.

[8] W. Cheng and E. Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Machine Learning*, vol. 76, no. 2-3, pp. 211–225, 2009.

[9] Z. A. Daniels and D. N. Metaxas, "Addressing imbalance in multi-label classification using structured Hellinger forests," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp. 1826–1832.

[10] 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.

[11] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "An exact algorithm for f-measure maximization," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Cambridge, MA: MIT Press, 2011, pp. 1404–1412.

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

[13] 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.

[14] 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.

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

[16] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[17] 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 Ingelligence*, Buenos Aires, Argentina, 2015, pp. 3511–3517.

[18] X. Geng and L. Luo, "Multilabel ranking with inconsistent rankers," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 3742–3747.

[19] 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.

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

[21] Y. Li, Y. Song, and J. Luo, "Improving pairwise ranking for multi-label image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hawaii, HI, 2017, pp. 1837–1845.

[22] Y.-K. Li, M.-L. Zhang, and X. Geng, "Leveraging implicit relative labeling-importance information for effective multi-label learning," in *Proceedings of the 15th IEEE International Conference on Data Mining*, Atlantic City, NJ, 2015, pp. 251–260.

[23] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[24] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012.

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

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

[27] X. Pan, Y.-X. Fan, J. Jia, and H.-B. Shen, "Identifying rna-binding proteins using multi-label deep learning," *Science China Information Sciences*, p. 62:19103, 2019.

[28] I. Pillai, G. Fumera, and F. Roli, "Designing multi-label classifiers that maximize F measures: State of the art," *Pattern Recognition*, vol. 61, pp. 394–404, 2017.

[29] 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.

[30] 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.

[31] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music by emotion," *EURASIP Journal on Audio, Speech, and Music Processing*, p. 2011:4, 2011.

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

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

[34] 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.

[35] W. Waegeman, K. Dembczyński, A. Jachnik, W. Cheng, and E. Hüllermeier, "On the bayes-optimality of F-measure maximizers," *Journal of Machine Learning Research*, vol. 15, no. Nov, pp. 3513–3568, 2014.

[36] H. Wang, C. Ding, and H. Huang, "Multi-label classification: Inconsistency and class balanced k-nearest neighbor," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, GA, 2010, pp. 1264–1266.

[37] 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.

[38] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "A survey on multi-output learning," *arXiv:1901.00248*, 2019.

[39] J. Xu, "Multi-label core vector machine with a zero label," *Pattern Recognition*, vol. 47, no. 7, pp. 2542–2557, 2012.

[40] 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.

[41] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, 2006, pp. 211–216.

[42] J.-F. Yu, D.-K. Jiang, K. Xiao, Y. Jin, J.-H. Wang, and X. Sun, "Discriminate the falsely predicted protein-coding genes in aeropyrum pernix K1 genome based on graphical representation," *MATCH Communicationsin Mathematical and in Computer Chemistry*, vol. 67, no. 3, pp. 845–866, 2012.

[43] 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.

[44] M.-L. Zhang, Y.-K. Li, Y.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview," *Frontiers of Computer Science*, vol. 12, no. 2, pp. 191–202, 2018.

[45] 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.

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

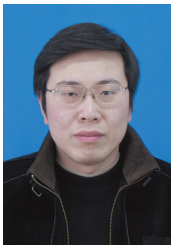
[47] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, pp. 284–291.

[48] Z.-H. Zhou and M.-L. Zhang, "Multi-label learning," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds. Berlin: Springer, 2017, pp. 875–881.

[49] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[50] 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.

[51] 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.

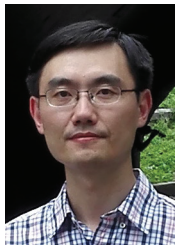


Min-Ling Zhang received the BSc, MSc, and PhD degrees in computer science from Nanjing University, China, in 2001, 2004 and 2007, respectively. Currently, he is a Professor at the School of Computer Science and Engineering, Southeast University, China. His main research interests include machine learning and data mining. In recent years, Dr. Zhang has served as the General Co-Chairs of ACML'18, Program Co-Chairs of PAKDD'19, ACML'17, CCFAI'17, PRICAI'16, Senior PC member or Area Chair of AAAI 2017-2020, IJCAI 2017-2019, ICDM 2015-2019, etc. He is also on the editorial board of ACM Transactions on Intelligent Systems and Technology, Neural Networks, Science China Information Sciences, Frontiers of Computer Science, etc. Dr. Zhang is the Steering Committee Member of PAKDD, secretary-general of the CAAI Machine Learning Society, standing committee member of the CCF Artificial Intelligence & Pattern Recognition Society.

Qian-Wen Zhang received the BSc degree in computer science from Nanjing University of Posts and Telecommunications, China, the MSc degree in computer Science from Southeast University, China in 2015 and 2018 respectively. Currently, she is an R&D enineer at the Tencent Smart Platform & Products Department. Her main research interests include machine learning and data mining, especially in learning from multi-label data.

Jun-Peng Fang received the BSc degree in computer science from North China Electric Power University, China in 2017. Currently, he is a master student at Southeast University. His main research interests include machine learning and data mining, especially in learning from multi-label data.

Yu-Kun Li received the BSc and MSc degrees in computer science from Southeast University, China, in 2012 and 2015 respectively. Currently, he is an R&D Engineer at the Baidu Inc. His main research interests include machine learning and data mining, especially in learning from multi-label data.



Xin Geng is currently a professor and the director of the PALM lab (http://palm.seu.edu.cn/) of Southeast University, China. He received the BSc (2001) and MSc (2004) degrees in computer science from Nanjing University, China, and the PhD (2008) degree in computer science from Deakin University, Australia. His research interests include pattern recognition, machine learning, and computer vision. He has published more than 50 refereed papers in these areas, including those published in prestigious journals and top international conferences.