Implicit Relative Labeling-Importance Aware Multi-Label Metric Learning

Jun-Xiang Mao[1,2,3], **Yong Rui**[4], **Min-Ling Zhang**[1,2*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
[3]Information Technology and Data Management Department of China Mobile Communications Group Zhejiang Co., Ltd
[4]Lenovo Research, Lenovo Group Ltd., Beijing, China
maojx@seu.edu.cn, yongrui@lenovo.com, zhangml@seu.edu.cn

Abstract

Multi-label metric learning, as an extension of metric learning to multi-label scenarios, aims to learn better similarity metrics for objects with rich semantics. Existing multi-label metric learning approaches employ the common assumption of equal labeling-importance, i.e., all associated labels are considered relevant to the training instance, while there is no differentiation in the relative importance of their semantics. However, this common assumption does not reflect the fact that the importance of each relevant label is generally different, even though such importance information is not directly accessible from the training examples. In this paper, we claim that it is beneficial to leverage the implicit ***R**elative **L**abeling-**I**mportance* (RLI) information to facilitate multi-label metric learning. Specifically, the manifold structure within the feature space is exploited by local linear reconstruction, and then the RLIs are recovered by transferring such structure to the label space. Subsequently, a discrimiative multi-label metric learning framework is introduced to align the predictive modeling outputs with the recovered RLIs, under which instances with similar RLI are implicitly pulled closer to each other, while those with dissimilar RLI are pushed further apart. Comprehensive experiments on benchmark multi-label datasets validate the superiority of our proposed approach in learning effective similarity metrics between multi-label examples.

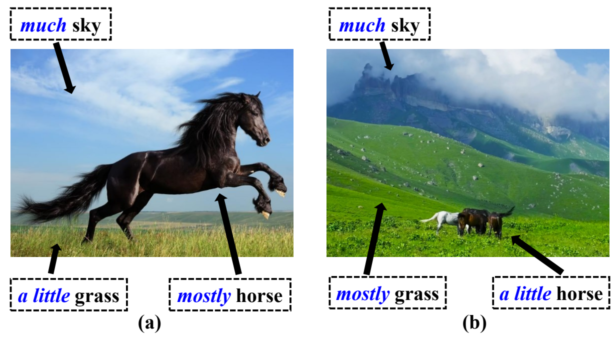

Figure 1: Two landscape images both annotated with the labels '*horse*', '*grass*', and '*sky*' simultaneously. For each image, the implicit relative labeling-importance (RIL) (a): '*horse*' > '*sky*' > '*grass*' and (b): '*grass*' > '*sky*' > '*horse*'.

Introduction

Similarity between objects plays an important role in both human cognitive processes and the recognition capabilities of intelligent systems. Appropriately measuring such similarity for a given task is crucial to the performance of many machine learning algorithms, such as k-nearest neighbor (KNN), k-means, etc. Metric learning, as a solution to this problem, aims to learn task-specific similarity metrics by leveraging side information such as linkages and comparisons derived from examples (Xing et al. 2002; Weinberger, Blitzer, and Saul 2005). The learned similarity metrics align with the inherent relations between examples, ensuring that similar instances exhibit proximity while distances between dissimilar instances are sufficiently large. With its powerful ability to characterize similarities, metric learning has been widely applied in real-world applications, including face recognition (Uzun, Cevikalp, and Saribas 2022), person re-identification (Liao and Shao 2022), information retrieval (Warburg et al. 2023), and recommender systems (Yu et al. 2023).

Despite the tremendous success of metric learning, the vast majority of research has focused on single-label scenarios where each instance is associated with only one label (Ye et al. 2020; Yang, Wang, and Zhang 2023; Ren et al. 2024). However, in the face of more prevalent and practical multi-label scenarios, where each instance is associated with multiple labels, existing single-label metric learning techniques are not applicable due to the complicated semantics of multi-label examples. Therefore, *multi-label metric learning*, which aims to assess the more intricate semantic similarities among objects with rich semantics, has emerged as a new research hotspot in recent years (Liu and Tsang 2015; Gouk, Pfahringer, and Cree 2016; Sun and Zhang 2021; Mao, Wang, and Zhang 2023; Mao, Hang, and Zhang 2024).

It is worth noting that the labeling information for multi-label training examples is categorical, i.e., each label is regarded to be either relevant or irrelevant for each multi-label instance. Therefore, existing multi-label metric learning 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 multi-label examples. However, for real-world multi-label examples, the importance of each associated relevant label is

*Corresponding author.

different by nature. For example, as shown in Figure 1, both landscape images (a) and (b) are annotated with the labels '*horse*', '*grass*', and '*sky*' simultaneously, while the implicit ***R**elative **L**abeling-**I**mportance* (RLI) that characterizes their semantics is different due to varying scenery presence. Nevertheless, such RLI information is not explicitly provided by annotators under standard multi-label learning setting (Zhang and Zhou 2014; Liu et al. 2021).

In light of the above observations, we postulate that more effective similarity metrics between multi-label examples can be expected if the implicit RLI information is appropriately leveraged within multi-label metric learning procedure. Accordingly, a novel multi-label metric learning approach named ILIA, i.e., ***I**mplicit relative **L**abeling-**I**mportance **A**ware multi-label metric learning*, is proposed. Specifically, ILIA begins by leveraging local linear reconstruction to exploit the manifold structure within the feature space, and then the implicit RLIs are recovered by transferring such structure to the label space. After that, a discriminative multi-label metric learning framework is introduced to align the predictive modeling outputs with the recovered RLIs, under which instances with similar RLI are implicitly pulled closer to each other, while those with dissimilar RLI are pushed further apart. Comprehensive experiments on benchmark multi-label datasets validate the superiority of ILIA in learning effective similarity metrics between multi-label examples.

The rest of this paper is organized as follows. Section 2 briefly reviews related works. Section 3 presents the details of the proposed ILIA approach. Section 4 reports the experimental results of comparative studies on benchmark multi-label datasets. Section 5 concludes the paper.

Related Work

Multi-Label Learning. Unlike multi-class classification that deals with single-label examples (Gong, Demmel, and You 2024; Jia et al. 2023), the purpose of multi-label learning is to train a predictive model that can assign a set of proper labels for unseen instances (Zhang and Zhou 2014). To address the challenge of an exponential-sized output space, modeling label correlations has become a mainstream strategy to solve this problem. Generally speaking, these approaches can be grouped into three categories, differing in the order of label correlations under consideration. The order of label correlations can be considered in a first-order manner by treating each label independently (Boutell et al. 2004; Zhang and Zhou 2007), a second-order manner by exploiting pairwise interactions between labels (Zhu, Kwok, and Zhou 2017; Yu and Zhang 2021), and a high-order manner by exploring relations among a subset or all labels (Zhang et al. 2021; Si et al. 2023). BRKNN (Boutell et al. 2004) and MLKNN (Zhang and Zhou 2007), as the most classic first-order approaches in multi-label learning, extend classic KNN to multi-label scenarios and have achieved certain outcomes in multi-label learning tasks. However, their performance seriously relies on the chosen similarity metrics. In the absence of prior knowledge, the commonly used predefined Euclidean metric may not be sufficiently effective in utilizing the label correlations among multiple labels, often leading to inferior performance compared to second-order and high-order approaches.

Metric Learning. To address the limitations of predefined metrics in characterizing the similarity between objects, metric learning has been proposed to obtain task-specific similarity metrics through a learning process (Xing et al. 2002; Hadsell, Chopra, and LeCun 2006; Weinberger and Saul 2009). By utilizing various types of supervision, such as linkages and comparisons derived from examples, metric learning aims to align the learned similarity metrics with the intrinsic relations between examples, i.e., similar instances are close to each other and dissimilar instances are far apart. In metric learning, the Mahalanobis metric is extensively employed as a substitute for the Euclidean metric due to its broad applicability as a general form of the Euclidean metric and its efficient optimization capabilities (Zhao and Yang 2023; Bansal et al. 2023; Xu et al. 2023). The Mahalanobis distance between instances is essentially equivalent to the Euclidean distance in the learned metric space. The superiority of metric learning has been substantiated in improving classic KNN classifiers (Ye et al. 2019, 2020; Li et al. 2022; Chen et al. 2023; Ren et al. 2024). With the effective modeling of semantic similarities among examples accomplished by metric learning, there is the potential for simple KNN classifiers to achieve state-of-the-art classification performance. Nevertheless, although metric learning has achieved great success, most research has concentrated on single-label scenarios. In more prevalent and practical multi-label scenarios, where each instance is associated with multiple labels, existing single-label metric learning techniques are not applicable due to the complicated semantics of multi-label examples.

Multi-Label Metric Learning. To compensate for the inapplicability of metric learning in multi-label scenarios, multi-label metric learning has been introduced in recent years. To the best of our knowledge, there are five available multi-label metric learning approaches: LM(Liu and Tsang 2015) employs a large margin formulation to establish a unified metric space, maintaining the correlation between feature and label spaces; LJE(Gouk, Pfahringer, and Cree 2016) learns a metric that projects instances into a space where the Euclidean distance closely mirrors the Jaccard similarity of multiple labels; COMMU(Sun and Zhang 2021) constructs a compositional metric by modeling structural interactions between feature and label spaces, exploring the integrated semantics of all labels; The core idea of both LIMIC(Mao, Wang, and Zhang 2023) and LSMM(Mao, Hang, and Zhang 2024) encompasses learning label-specific metrics for each label, incorporating a global metric to exploit label correlations. However, the above approaches learn from multi-label examples by assuming equal labeling-importance, which might be suboptimal because, in reality, the importance of each associated relevant label is inherently different. In this paper, we make the first attempt to recover and leverage such implicit RLI information in multi-label metric learning. The proposed ILIA approach will be introduced in the next section.

The ILIA approach

Preliminaries

Let $\mathcal{X} = \mathbb{R}^d$ be the feature space and $\mathcal{Y} = \{l_1, l_2, \ldots, l_q\}$ denote the label space with q labels. A multi-label example is denoted as $(\boldsymbol{x}, Y)$, where $\boldsymbol{x} \in \mathcal{X}$ is its feature vector and $Y \subseteq \mathcal{Y}$ corresponds to the set of its relevant labels. Here, a q-dimensional indicator vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_q]^\top \in \{0,1\}^q$ is utilized to denote Y, where $y_p = 1$ when $l_p \in Y$ and $y_p = 0$ otherwise. The task of multi-label metric learning is to learn a function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ from the multi-label training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid 1 \leq i \leq n\}$, which can reflect the semantic similarities between multi-label examples.

In metric learning, the Mahalanobis metric is extensively employed as an instantiation of the similarity metrics to be learned (Xu and Davenport 2020; Bellet, Habrard, and Sebban 2015). Let $\mathbb{S}_+^d$ denotes the cone of positive semi-definite $d \times d$ matrices. Given a Mahalanobis metric $\mathbf{M} \in \mathbb{S}_+^d$, the (squared) Mahalanobis distance between a pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is

$$\begin{aligned}\mathrm{Dis}^2_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) &= (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \mathbf{M}(\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &= ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2_{\mathbf{M}}. \end{aligned} \quad (1)$$

In this manner, examples exhibiting shorter Mahalanobis distances indicate higher similarity, while those with longer distances suggest lower similarity.

Implicit RLI Recovery

Following the ideas of locally linear embedding (Roweis and Saul 2000; Wang and Zhang 2006), each instance $\boldsymbol{x}$ can be reconstructed via linear combination of its k nearest neighbors, and this manifold structure also holds in the label space. For each training multi-label instance $\boldsymbol{x}_i (1 \leq i \leq m)$, the combination coefficients for its k nearest neighbors can be determined by solving the following optimization problem:

$$\begin{aligned} \min_{s_{ii_1}, s_{ii_2}, \ldots, s_{ii_k}} & \left\| \boldsymbol{x}_i - \sum_{j \in \mathcal{N}_k(\boldsymbol{x}_i)} s_{ij} \boldsymbol{x}_j \right\|_2^2 \\ \text{s.t.} & \sum_{j \in \mathcal{N}_k(\boldsymbol{x}_i)} s_{ij} = 1. \end{aligned} \quad (2)$$

Here, $\mathcal{N}_k(\boldsymbol{x}_i) = \{i_r | 1 \leq r \leq k\}$ denotes the set of indices for $\boldsymbol{x}_i$'s k nearest neighbors. Let $\hat{\boldsymbol{s}}_i = [s_{ii_1}, s_{ii_2}, \ldots, s_{ii_k}]^\top$ be the neighborhood coefficient vector of $\boldsymbol{x}_i$, then Eq.(2) can be easily reformulated as the following matrix form:

$$\begin{aligned} \min_{\hat{\boldsymbol{s}}_i} & \ \hat{\boldsymbol{s}}_i^\top \mathbf{G}_i \hat{\boldsymbol{s}}_i \\ \text{s.t.} & \ \mathbf{1}_k^\top \hat{\boldsymbol{s}}_i = 1, \end{aligned} \quad (3)$$

where $\mathbf{G}_i = \mathbf{D}_i^\top \mathbf{D}_i \in \mathbb{R}^{k \times k}$ is the Gram matrix, $\mathbf{D}_i = [\boldsymbol{x}_i - \boldsymbol{x}_{i_1}, \boldsymbol{x}_i - \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_i - \boldsymbol{x}_{i_k}] \in \mathbb{R}^{d \times k}$, and $\mathbf{1}_k$ is an all 1 column vector with size k.

To solve the above problem Eq.(3), we construct a Lagrange function:

$$\mathcal{L}(\hat{\boldsymbol{s}}_i, \lambda) = \hat{\boldsymbol{s}}_i^\top \mathbf{G}_i \hat{\boldsymbol{s}}_i + \lambda(\mathbf{1}_n^\top \hat{\boldsymbol{s}}_i - 1). \quad (4)$$

Then setting the first-order derivatives of $\mathcal{L}(\hat{\boldsymbol{s}}_i, \lambda)$ w.r.t $\hat{\boldsymbol{s}}_i$ and λ to 0, respectively, we have

$$\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{s}}_i} = 2\mathbf{G}_i \hat{\boldsymbol{s}}_i + \lambda \mathbf{1}_k \overset{\text{set}}{=} 0 \Longrightarrow \hat{\boldsymbol{s}}_i = -\frac{\lambda}{2} \mathbf{G}_i^{-1} \mathbf{1}_k, \quad (5)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \mathbf{1}_k^\top \hat{\boldsymbol{s}}_i - 1 \overset{\text{set}}{=} 0 \Longrightarrow \mathbf{1}_k^\top \hat{\boldsymbol{s}}_i = 1. \quad (6)$$

Substituting Eq.(5) into Eq.(6), we have

$$-\frac{\lambda}{2} \mathbf{1}_k^\top \mathbf{G}_i^{-1} \mathbf{1}_k = 1 \Longrightarrow \lambda = -\frac{2}{\mathbf{1}_k^\top \mathbf{G}_i^{-1} \mathbf{1}_k}. \quad (7)$$

Using Eq.(5) and Eq.(7), we can achieve the closed-form solution of the optimization problem Eq.(2):

$$\hat{\boldsymbol{s}}_i = \frac{\mathbf{G}_i^{-1} \mathbf{1}_k}{\mathbf{1}_k^\top \mathbf{G}_i^{-1} \mathbf{1}_k}. \quad (8)$$

Let $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]^\top \in \mathbb{R}^{n \times q}$ denotes the label matrix and $\mathbf{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n]^\top \in \mathbb{R}^{n \times q}$ represents the recovered RLI matrix of $\mathbf{Y}$. After all $\hat{\boldsymbol{s}}_i (1 \leq i \leq n)$ have been determined by Eq.(8), $\mathbf{F}$ can be generated by transferring the exploited manifold structure of the feature space to the label space, which is formalized as follows:

$$\min_{\mathbf{F}} \frac{1}{n} \sum_{i=1}^{n} \left\| \boldsymbol{f}_i - \sum_{j \in \mathcal{N}_k(\boldsymbol{x}_i)} s_{ij} \boldsymbol{f}_j \right\|_2^2 + \mu \left\| \mathbf{F} - \mathbf{Y} \right\|_F^2, \quad (9)$$

where μ is a trade-off parameter. The first term ensures that the similar manifold structure to the feature space is maintained in the label space, and the second term ensures that the recovered RLI matrix $\mathbf{F}$ should also be similar to the original logical label matrix $\mathbf{Y}$. For ease of solution, Eq.(9) can be equivalently reformulated as follows:

$$\min_{\mathbf{F}} \frac{1}{n} \mathrm{tr}(\mathbf{F}^\top (\mathbf{I}_n - \mathbf{S})(\mathbf{I}_n - \mathbf{S})^\top \mathbf{F}) + \mu \left\| \mathbf{F} - \mathbf{Y} \right\|_F^2. \quad (10)$$

Here, $\mathrm{tr}(\cdot)$ computes the trace of a matrix, $\mathbf{I}_n$ represents an $n \times n$ identity matrix, $\mathbf{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n] \in \mathbb{R}^{n \times n}$, and $\boldsymbol{s}_i = [s_{i1}, s_{i2}, \ldots, s_{in}]^\top$, where s_{ij} is determined by Eq.(2) if $j \in \mathcal{N}_k(\boldsymbol{x}_i)$ and $s_{ij} = 0$ otherwise.

Let $\mathcal{G}(\mathbf{F})$ denotes the objective function of Eq.(10), the first-order derivative of $\mathcal{G}(\mathbf{F})$ w.r.t $\mathbf{F}$ is

$$\frac{\partial \mathcal{G}}{\partial \mathbf{F}} = \frac{2}{n} (\mathbf{I}_n - \mathbf{S})(\mathbf{I}_n - \mathbf{S})^\top \mathbf{F} + 2\mu \mathbf{F} - 2\mu \mathbf{Y}. \quad (11)$$

Then we can achieve a closed-form solution of Eq.(10) through setting Eq.(11) to 0:

$$\mathbf{F} = \left(\frac{1}{n} (\mathbf{I}_n - \mathbf{S})(\mathbf{I}_n - \mathbf{S})^\top + \mu \mathbf{I_n} \right)^{-1} (\mu \mathbf{Y}). \quad (12)$$

In this way, the implicit RLIs $\mathbf{F}$ of multi-label examples can be recovered through the above procedure. Then, $\mathbf{F}$ will be leveraged as more comprehensive and complete supervision information to guide the following discriminative multi-label metric learning procedure.

Discriminative Multi-Label Metric Learning

It is worth noting that the recovered implicit RLI information $\mathbf{F}$ is numerical rather than logical. Therefore, it is natural to tackle the resulting multi-label learning problem with multi-output regression techniques (Borchani et al. 2015) in a discriminative manner. Specifically, we can assign a simple ridge regression model to each label space over a new multi-label training set $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \boldsymbol{f}_i) \mid 1 \leq i \leq n\}$:

$$\min_{\mathbf{W}, \boldsymbol{b}} \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{W}^\top \phi(\boldsymbol{x}_i) + \boldsymbol{b} - \boldsymbol{f}_i \right\|_2^2 + \eta \left\| \mathbf{W} \right\|_F^2. \quad (13)$$

Here, η is a trade-off parameter, $\phi(\cdot)$ is a nonlinear mapping implemented by kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $\phi(\boldsymbol{x}_i) \in \mathbb{R}^{d'}$. $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_q] \in \mathbb{R}^{d' \times q}$ is the predictive modeling coefficients, and $\boldsymbol{b} = [b_1, b_2, \ldots, b_q]^\top \in \mathbb{R}^q$ is the intercept to be determined. Let $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}$ represents the instance matrix. The intercept term $\boldsymbol{b}$ in Eq.(13) can then be omitted by centering the instance matrix $\mathbf{X}$ and the recovered RLI matrix $\mathbf{F}$:

$$\min_{\mathbf{W}} \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right\|_2^2 + \eta \left\| \mathbf{W} \right\|_F^2, \tag{14}$$

where $\hat{\boldsymbol{x}}_i$ and $\hat{\boldsymbol{f}}_i$ denote the i-th centered instance vector and RLI vector, respectively. Furthermore, we denote $\hat{\mathbf{X}} = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_n]^\top \in \mathbb{R}^{n \times d}$ be the centered instance matrix and $\hat{\mathbf{F}} = [\hat{\boldsymbol{f}}_1, \hat{\boldsymbol{f}}_2, \ldots, \hat{\boldsymbol{f}}_n]^\top \in \mathbb{R}^{n \times q}$ be the centered RLI matrix. However, the above predictive model Eq.(14) actually deals with the q labels independently. To exploit the intrinsic label correlations among multi-label examples, we employ a Mahalanobis metric $\mathbf{M}$ to measure the distance between $\mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i)$ and $\hat{\boldsymbol{f}}_i$:

$$\min_{\mathbf{W}} \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right\|_{\mathbf{M}}^2 + \eta \left\| \mathbf{W} \right\|_F^2. \tag{15}$$

Here, $\mathbf{M}$ can be viewed as a discriminative metric for multi-label examples, which enforces a shorter distance between $\boldsymbol{x}_i$'s encoding $\mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i)$ and its corresponding RLI $\hat{\boldsymbol{f}}_i$. To further enhance the discriminability of $\mathbf{M}$, we penalize encodings and RLIs that are not consistent with each other. Consequently, $\mathbf{M}$ can be determined by solving the following optimization problem (Zadeh, Hosseini, and Sra 2016):

$$\begin{aligned}
\min_{\mathbf{M} \succ 0} & \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right\|_{\mathbf{M}}^2 \\
& + \frac{1}{nk} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_k(\hat{\boldsymbol{f}}_i)} \left\| \mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_j \right\|_{\mathbf{M}^{-1}}^2 \\
& + \frac{1}{nk} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_k(\hat{\boldsymbol{x}}_i)} \left\| \hat{\boldsymbol{f}}_i - \mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_j) \right\|_{\mathbf{M}^{-1}}^2 \\
& + \gamma D(\mathbf{M}, \mathbf{I}_q)
\end{aligned} \tag{16}$$

where $\mathcal{N}_k(\hat{\boldsymbol{f}}_i)$ denotes the set of indices for $\hat{\boldsymbol{f}}_i$'s k nearest neighbors in $\{\hat{\boldsymbol{f}}_1, \hat{\boldsymbol{f}}_2, \ldots, \hat{\boldsymbol{f}}_n\} \setminus \hat{\boldsymbol{f}}_i$, the definition of $\mathcal{N}_k(\hat{\boldsymbol{x}}_i)$ is similar to that of $\mathcal{N}_k(\hat{\boldsymbol{f}}_i)$, γ is a trade-off parameter, $D(\mathbf{M}, \mathbf{I}_q) = \mathrm{tr}(\mathbf{M}\mathbf{I}_q^{-1}) + \mathrm{tr}(\mathbf{M}^{-1}\mathbf{I}_q) - 2q$ is the symmetrized LogDet divergence, and $\mathbf{I}_q$ is a $q \times q$ identity matrix. Here, the first term enforces the distance between $\mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i)$ and the corresponding $\hat{\boldsymbol{f}}_i$ closer. The second term ensures $\mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i)$ stay away from targets that are not $\hat{\boldsymbol{f}}_i$, but are similar to $\hat{\boldsymbol{f}}_i$. The third term pushes $\hat{\boldsymbol{f}}_i$ futher away from targets that are not $\mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i)$, but are similar to the $\hat{\boldsymbol{x}}_i$'s encoding. The fourth term penalizes the complexity of $\mathbf{M}$ to avoid overfitting. Consequently, by optimizing Eq.(16), instances with similar RLI are implicitly pulled closer to each other, while those with dissimilar RLI are pushed further apart.

Optimization

Obviously, $\mathbf{M}$ should be known when solving the optimization problem w.r.t $\mathbf{W}$ in Eq.(15). Conversely, $\mathbf{W}$ should be known when solving the optimization problem w.r.t $\mathbf{M}$ in Eq.(16). The interaction between $\mathbf{W}$ and $\mathbf{M}$ prevents them from being calculated simultaneously. Consequently, in this paper, we alternately calculate one of them while the remaining one is fixed until convergence.

Calculating $\mathbf{W}$ when $\mathbf{M}$ is fixed. It is worth noting that, in the optimization problem Eq.(15), $\phi(\cdot)$ is a nonlinear mapping implemented by kernel function κ. Therefore, we cannot obtain an explicit solution of $\mathbf{W}$. According to the Representer Theorem (Schölkopf and Smola 2002), under fairly general conditions, the predictive model can be expressed as a linear combination of the training instances. Let $\mathbf{\Phi} = [\phi(\hat{\boldsymbol{x}}_1), \phi(\hat{\boldsymbol{x}}_2), \ldots, \phi(\hat{\boldsymbol{x}}_n)]^\top \in \mathbb{R}^{n \times d'}$ be the nonlinear mapping centered instance matrix, for the multi-output regression problem in Eq.(15), we have $\boldsymbol{w}_i = \sum_{j=1}^{n} \theta_{ij} \phi(\hat{\boldsymbol{x}}_j) = \mathbf{\Phi}^\top \boldsymbol{\theta}_i$ and then $\mathbf{W} = \mathbf{\Phi}^\top \mathbf{\Theta}$, where $\mathbf{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_q] \in \mathbb{R}^{n \times q}$ is the combination coefficients to be determined. By substituting $\mathbf{W} = \mathbf{\Phi}^\top \mathbf{\Theta}$ into the objective function in Eq.(15) which is denoted as $\mathcal{H}(\mathbf{W})$, we have

$$\begin{aligned}
\mathcal{H}(\mathbf{W}) &= \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{\Theta}^\top \mathbf{\Phi} \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right\|_{\mathbf{M}}^2 + \eta \left\| \mathbf{\Phi}^\top \mathbf{\Theta} \right\|_F^2 \\
&= \frac{1}{n} \left\| \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{\Theta} - \hat{\mathbf{F}} \right\|_{\mathbf{M}}^2 + \eta \left\| \mathbf{\Phi}^\top \mathbf{\Theta} \right\|_F^2 \\
&= \frac{1}{n} \mathrm{tr} \left(\left(\mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{\Theta} - \hat{\mathbf{F}} \right) \mathbf{M} \left(\mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{\Theta} - \hat{\mathbf{F}} \right)^\top \right) \\
&\quad + \eta \mathrm{tr} \left(\mathbf{\Theta}^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{\Theta} \right) \\
&\triangleq \mathcal{H}(\mathbf{\Theta}).
\end{aligned} \tag{17}$$

Let $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^\top \in \mathbb{R}^{n \times n}$ represents the kernel matrix with (i, j)-th element $K_{ij} = \kappa(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j)$, then the first-order derivative of $\mathcal{H}(\mathbf{\Theta})$ w.r.t $\mathbf{\Theta}$ is

$$\frac{\partial \mathcal{H}(\mathbf{\Theta})}{\partial \mathbf{\Theta}} = \frac{2}{n} \left(\mathbf{K}^\top \mathbf{K} \mathbf{\Theta} \mathbf{M} - \mathbf{K}^\top \hat{\mathbf{F}} \mathbf{M} \right) + 2\eta \mathbf{K} \mathbf{\Theta}. \tag{18}$$

Setting the above Eq.(18) to 0, we have

$$n\eta \left(\mathbf{K}^\top \mathbf{K} \right)^{-1} \mathbf{K} \mathbf{\Theta} + \mathbf{\Theta} \mathbf{M} = \left(\mathbf{K}^\top \mathbf{K} \right)^{-1} \mathbf{K}^\top \hat{\mathbf{F}} \mathbf{M}, \tag{19}$$

which is a Sylvester equation w.r.t $\mathbf{\Theta}$ and can be solved by any off-the-shelf solvers (Wei, Dobigeon, and Tourneret 2015).

Calculating $\mathbf{M}$ when $\mathbf{W}$ is fixed. The optimization problem in Eq.(16) can be equivalently reformulated as

$$\min_{\mathbf{M} \succ 0} \mathrm{tr}\left(\mathbf{M} \mathbf{U} \right) + \mathrm{tr}\left(\mathbf{M}^{-1} \mathbf{V} \right) + \gamma D(\mathbf{M}, \mathbf{I}_q). \tag{20}$$

Here,

$$\begin{aligned}
\mathbf{U} &= \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right) \left(\mathbf{W}^\top \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right)^\top \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{\Theta}^\top \mathbf{\Phi} \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right) \left(\mathbf{\Theta}^\top \mathbf{\Phi} \phi(\hat{\boldsymbol{x}}_i) - \hat{\boldsymbol{f}}_i \right)^\top \\
&= \frac{1}{n} \left(\mathbf{K} \mathbf{\Theta} - \hat{\mathbf{F}} \right)^\top \left(\mathbf{K} \mathbf{\Theta} - \hat{\mathbf{F}} \right),
\end{aligned} \tag{21}$$

$$\mathbf{V} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_k(\hat{\boldsymbol{f}}_i)}\left(\mathbf{W}^\top\phi(\hat{\boldsymbol{x}}_i)-\hat{\boldsymbol{f}}_j\right)\left(\mathbf{W}^\top\phi(\hat{\boldsymbol{x}}_i)-\hat{\boldsymbol{f}}_j\right)^\top$$
$$+\frac{1}{nk}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_k(\hat{\boldsymbol{x}}_i)}\left(\hat{\boldsymbol{f}}_i-\mathbf{W}^\top\phi(\hat{\boldsymbol{x}}_j)\right)\left(\hat{\boldsymbol{f}}_i-\mathbf{W}^\top\phi(\hat{\boldsymbol{x}}_j)\right)^\top$$
$$=\frac{1}{nk}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_k(\hat{\boldsymbol{f}}_i)}\left(\boldsymbol{\Theta}^\top\boldsymbol{\Phi}\phi(\hat{\boldsymbol{x}}_i)-\hat{\boldsymbol{f}}_j\right)\left(\boldsymbol{\Theta}^\top\boldsymbol{\Phi}\phi(\hat{\boldsymbol{x}}_i)-\hat{\boldsymbol{f}}_j\right)^\top$$
$$+\frac{1}{nk}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_k(\hat{\boldsymbol{x}}_i)}\left(\hat{\boldsymbol{f}}_i-\boldsymbol{\Theta}^\top\boldsymbol{\Phi}\phi(\hat{\boldsymbol{x}}_j)\right)\left(\hat{\boldsymbol{f}}_i-\boldsymbol{\Theta}^\top\boldsymbol{\Phi}\phi(\hat{\boldsymbol{x}}_j)\right)^\top$$
$$=\frac{1}{nk}\sum_{r=1}^{k}\left(\mathbf{K}\boldsymbol{\Theta}-\hat{\mathbf{F}}_r\right)^\top\left(\mathbf{K}\boldsymbol{\Theta}-\hat{\mathbf{F}}_r\right)$$
$$+\frac{1}{nk}\sum_{r=1}^{k}\left(\hat{\mathbf{F}}-\mathbf{K}_r\boldsymbol{\Theta}\right)^\top\left(\hat{\mathbf{F}}-\mathbf{K}_r\boldsymbol{\Theta}\right), \tag{22}$$

where $\hat{\mathbf{F}}_r = [\hat{\boldsymbol{f}}_{1_r}, \hat{\boldsymbol{f}}_{2_r}, \ldots, \hat{\boldsymbol{f}}_{n_r}]^\top \in \mathbb{R}^{n\times q}$, $\mathbf{K}_r = \boldsymbol{\Phi}_r\boldsymbol{\Phi}_r^\top$, and $\boldsymbol{\Phi}_r = [\phi(\hat{\boldsymbol{x}}_{1_r}), \phi(\hat{\boldsymbol{x}}_{2_r}), \ldots, \phi(\hat{\boldsymbol{x}}_{n_r})]^\top \in \mathbb{R}^{n\times d'}$. Following (Zadeh, Hosseini, and Sra 2016), the optimization problem in Eq.(20) is strictly convex, then its global minimum can be obtained when the gradient of the objective function vanishes. Specifically, by calculating the first-order derivative w.r.t $\mathbf{M}$ and setting it to 0, we have

$$\begin{aligned}&(\mathbf{U}+\gamma\mathbf{I}_q)-\mathbf{M}^{-1}(\mathbf{V}+\gamma\mathbf{I}_q)\mathbf{M}^{-1}=0\\ \Longrightarrow\ &\mathbf{M}(\mathbf{U}+\gamma\mathbf{I}_q)\mathbf{M}=(\mathbf{V}+\gamma\mathbf{I}_q).\end{aligned}\tag{23}$$

Eq.(23) is a Riccati equation (Bhatia 2009) and its unique solution corresponds to the midpoint of the geodesic joining $(\mathbf{U}+\gamma\mathbf{I}_q)^{-1}$ to $(\mathbf{V}+\gamma\mathbf{I}_q)$, i.e.,

$$\mathbf{M}=(\mathbf{U}+\gamma\mathbf{I}_q)^{-1}\#_{1/2}(\mathbf{V}+\gamma\mathbf{I}_q), \tag{24}$$

where $\mathbf{A}\#_{1/2}\mathbf{B}=\mathbf{A}^{1/2}\left(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}\right)^{1/2}\mathbf{A}^{1/2}$.

The complete procedure of the proposed ILIA approach is summarized in Appendix A. After the above two alternating optimization steps converge, we can obtain the predictive modeling coefficients $\mathbf{W}$ and the discriminative metric $\mathbf{M}$. Subsequently, the semantic similarity between a pair multi-label instances $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ can be explicitly formalized in the form of the following Mahalanobis distance:

$$\begin{aligned}\mathrm{Dis}(\boldsymbol{x}_i,\boldsymbol{x}_j)&=\left\|\mathbf{W}^\top\phi(\hat{\boldsymbol{x}}_i)-\mathbf{W}^\top\phi(\hat{\boldsymbol{x}}_j)\right\|_{\mathbf{M}}\\&=\left\|\boldsymbol{\Theta}^\top\boldsymbol{\Phi}\phi(\hat{\boldsymbol{x}}_i)-\boldsymbol{\Theta}^\top\boldsymbol{\Phi}\phi(\hat{\boldsymbol{x}}_j)\right\|_{\mathbf{M}}\\&=\left\|\boldsymbol{\Theta}^\top\mathbf{K}^i-\boldsymbol{\Theta}^\top\mathbf{K}^j\right\|_{\mathbf{M}},\end{aligned}\tag{25}$$

where $\mathbf{K}^i \in \mathbb{R}^n$ with r-th element $K_r^i = \kappa(\hat{\boldsymbol{x}}_r, \hat{\boldsymbol{x}}_i)(1 \le r \le n)$. A shorter Mahalanobis distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ indicates higher similarity, while a longer Mahalanobis distance indicates lower similarity.

Experiments

Experimental Setup

Datasets. In this paper, ten real-world multi-label datasets with diversified properties are employed for comparative studies. Table 1 summarizes the detailed characteristics of each benchmark dataset $\mathcal{D}$, including the number of examples $|\mathcal{D}|$, number of features $dim(\mathcal{D})$, number of labels $L(\mathcal{D})$, label cardinality $LCard(\mathcal{D})$, and domain of datasets.

Dataset	$\lvert\mathcal{D}\rvert$	$dim(\mathcal{D})$	$L(\mathcal{D})$	$LCard(\mathcal{D})$	Domain
CAL500	502	68	174	26.044	Music[1]
emotions	593	72	6	1.869	Music[1]
medical	978	1449	45	1.245	Text[1]
image	2000	294	5	1.236	Image[2]
scene	2407	294	6	1.074	Image[1]
arts	5000	462	26	1.636	Text[1]
corel5k	5000	499	374	3.522	Image[1]
education	5000	550	33	1.461	Text[1]
health	8116	1483	32	1.649	Text[1]
entertainment	8166	545	21	1.438	Text[1]

[1] http://mulan.sourceforge.net/datasets.html
[2] http://palm.seu.edu.cn/zhangml/Resources.htm#data

Table 1: Characteristics of experimental datasets.

Evaluation protocols. To validate the effectiveness of the proposed ILIA approach in learning similarity metrics between multi-label examples, following (Mao, Hang, and Zhang 2024), we employ BRKNN (Boutell et al. 2004) and MLKNN (Zhang and Zhou 2007) as subsequent multi-label learning methods after learning metrics. If the learned metrics can well characterize the semantic similarities between multi-label examples, simple KNN-based multi-label learning algorithms can achieve good classification performance.

Evaluation metrics. Six widely used evaluation metrics for multi-label learning are utilized for performance evaluation, including *Hamming loss*, *Ranking loss*, *Coverage*, *Average precision*, *Macro-F1*, and *Macro-averaging AUC*. Detailed definitions can be found in (Zhang and Zhou 2014).

Compared methods. We compare ILIA with five state-of-the-art multi-label metric learning methods, including LM(Liu and Tsang 2015), LJE(Gouk, Pfahringer, and Cree 2016), COMMU(Sun and Zhang 2021), LIMIC (Mao, Wang, and Zhang 2023), and LSMM (Mao, Hang, and Zhang 2024). More details about these compared methods can be found in Appendix B.1. Denote $\mathcal{A} \in \{\text{BRKNN}, \text{MLKNN}\}$ as a KNN-based multi-label learning algorithm, $\mathcal{B} \in \{\text{ILIA}, \text{LM}, \text{LJE}, \text{COMMU}, \text{LIMIC}, \text{LSMM}\}$ as a multi-label metric learning algorithm, and $\mathcal{A}$-$\mathcal{B}$ as the coupling version of them. The classification performance of $\mathcal{A}$-ILIA is compared against other state-of-the-art multi-label metric learning algorithms coupled with $\mathcal{A}$ to manifest whether ILIA does learn better similarity metrics between multi-label examples.

Configuration. For the proposed ILIA approach, we use the Polynomial kernel and set the parameters as follows: the trade-off parameters $\mu = 10^{-3}, \eta = 10^{-2}, \gamma = 10^{-2}$, and the number of nearest neighbors $k = 20$. Detailed discussion about the choice of these parameters can be found in 'Sensitivity analysis' paragraph and Appendix D. For KNN and MLKNN, the number of nearest neighbors is fixed to 10 for

Compared Algorithms	Datasets									
	CAL500	emotions	medical	image	scene	arts	corel5k	education	health	entertainment
	Hamming Loss ↓									
BRKNN	$.145_{\pm.003}$	$.263_{\pm.023}\bullet$	$.016_{\pm.002}\bullet$	$.170_{\pm.017}\bullet$	$.091_{\pm.007}\bullet$	$.075_{\pm.002}\bullet$	$\underline{.010}_{\pm.000}$	$\underline{.038}_{\pm.001}$	$.047_{\pm.001}\bullet$	$\underline{.065}_{\pm.002}$
BRKNN-LM	$.150_{\pm.003}\bullet$	$.270_{\pm.019}\bullet$	$\mathbf{.010}_{\pm.002}$	$.175_{\pm.016}\bullet$	$.090_{\pm.009}\bullet$	$\underline{.056}_{\pm.001}\bullet$	$\underline{.010}_{\pm.000}$	$\underline{.038}_{\pm.001}$	$.046_{\pm.002}$	$.068_{\pm.002}\bullet$
BRKNN-LJE	$.146_{\pm.004}\bullet$	$.219_{\pm.022}\bullet$	$.022_{\pm.003}\bullet$	$.184_{\pm.018}\bullet$	$.110_{\pm.011}\bullet$	$.061_{\pm.001}\bullet$	$\underline{.010}_{\pm.000}$	$.043_{\pm.002}\bullet$	$.054_{\pm.003}\bullet$	$.067_{\pm.001}\bullet$
BRKNN-COMMU	$\underline{.144}_{\pm.003}$	$.263_{\pm.023}\bullet$	$.016_{\pm.002}\bullet$	$.171_{\pm.016}\bullet$	$.091_{\pm.007}\bullet$	$.056_{\pm.001}\bullet$	$\mathbf{.009}_{\pm.000}$	$\underline{.038}_{\pm.001}$	$.047_{\pm.001}\bullet$	$\underline{.065}_{\pm.002}$
BRKNN-LIMIC	$.145_{\pm.005}$	$.212_{\pm.008}\bullet$	$.012_{\pm.002}$	$\underline{.161}_{\pm.016}$	$.081_{\pm.007}$	$.058_{\pm.002}\bullet$	$\underline{.010}_{\pm.000}$	$.039_{\pm.001}\bullet$	$.046_{\pm.001}\bullet$	$\underline{.065}_{\pm.002}$
BRKNN-LSMM	$.145_{\pm.004}$	$\underline{.207}_{\pm.020}$	$.014_{\pm.003}\bullet$	$.162_{\pm.013}$	$\underline{.080}_{\pm.006}$	$.060_{\pm.001}\bullet$	$\mathbf{.009}_{\pm.000}$	$\mathbf{.037}_{\pm.002}$	$\underline{.045}_{\pm.002}$	$\mathbf{.064}_{\pm.002}$
BRKNN-ILIA (Ours)	$\mathbf{.142}_{\pm.004}$	$\mathbf{.203}_{\pm.018}$	$\underline{.011}_{\pm.002}$	$\mathbf{.157}_{\pm.015}$	$\mathbf{.078}_{\pm.007}$	$\mathbf{.053}_{\pm.001}$	$\underline{.010}_{\pm.000}$	$\mathbf{.037}_{\pm.001}$	$\mathbf{.044}_{\pm.002}$	$\mathbf{.064}_{\pm.001}$
MLKNN	$\mathbf{.139}_{\pm.005}$	$.262_{\pm.022}\bullet$	$.015_{\pm.002}\bullet$	$.174_{\pm.013}\bullet$	$\underline{.085}_{\pm.009}$	$.060_{\pm.001}\bullet$	$.009_{\pm.000}$	$\underline{.038}_{\pm.001}$	$.047_{\pm.001}\bullet$	$\mathbf{.064}_{\pm.002}$
MLKNN-LM	$\mathbf{.139}_{\pm.004}\circ$	$.254_{\pm.017}\bullet$	$\underline{.012}_{\pm.002}$	$.176_{\pm.014}\bullet$	$.088_{\pm.008}$	$\underline{.055}_{\pm.001}\bullet$	$\mathbf{.009}_{\pm.000}$	$\underline{.038}_{\pm.001}$	$\mathbf{.044}_{\pm.002}$	$\mathbf{.064}_{\pm.001}$
MLKNN-LJE	$\mathbf{.139}_{\pm.005}$	$.227_{\pm.022}\bullet$	$.023_{\pm.003}\bullet$	$.184_{\pm.017}\bullet$	$.109_{\pm.009}\bullet$	$.060_{\pm.001}\bullet$	$\underline{.010}_{\pm.000}$	$.042_{\pm.002}\bullet$	$.053_{\pm.002}\bullet$	$\underline{.066}_{\pm.002}\bullet$
MLKNN-COMMU	$\mathbf{.139}_{\pm.004}\circ$	$.262_{\pm.022}\bullet$	$.016_{\pm.002}\bullet$	$.174_{\pm.013}\bullet$	$.086_{\pm.009}$	$.060_{\pm.001}\bullet$	$\mathbf{.009}_{\pm.000}$	$\underline{.038}_{\pm.001}$	$.046_{\pm.002}$	$\mathbf{.064}_{\pm.002}$
MLKNN-LIMIC	$\mathbf{.139}_{\pm.004}$	$.236_{\pm.009}\bullet$	$\underline{.012}_{\pm.003}$	$.161_{\pm.018}$	$\mathbf{.081}_{\pm.004}\circ$	$.057_{\pm.001}\bullet$	$\mathbf{.009}_{\pm.000}$	$\underline{.038}_{\pm.001}$	$.048_{\pm.001}\bullet$	$\mathbf{.064}_{\pm.002}$
MLKNN-LSMM	$\underline{.140}_{\pm.005}$	$\underline{.225}_{\pm.016}$	$\underline{.012}_{\pm.002}$	$\underline{.159}_{\pm.014}$	$.087_{\pm.007}$	$\underline{.055}_{\pm.001}\bullet$	$\mathbf{.009}_{\pm.000}$	$\mathbf{.037}_{\pm.001}$	$.048_{\pm.001}\bullet$	$\mathbf{.064}_{\pm.002}$
MLKNN-ILIA (Ours)	$.142_{\pm.004}$	$\mathbf{.218}_{\pm.014}$	$\mathbf{.011}_{\pm.002}$	$\mathbf{.156}_{\pm.014}$	$\underline{.085}_{\pm.008}$	$\mathbf{.053}_{\pm.001}$	$\underline{.010}_{\pm.000}$	$\mathbf{.037}_{\pm.002}$	$\underline{.045}_{\pm.002}$	$\mathbf{.064}_{\pm.001}$
	Average precision ↑									
BRKNN	$.463_{\pm.009}\bullet$	$.700_{\pm.049}\bullet$	$.778_{\pm.027}\bullet$	$.788_{\pm.023}\bullet$	$.850_{\pm.012}\bullet$	$.400_{\pm.025}\bullet$	$.151_{\pm.013}\bullet$	$.573_{\pm.014}\bullet$	$.605_{\pm.013}\bullet$	$.491_{\pm.013}\bullet$
BRKNN-LM	$.451_{\pm.007}\bullet$	$.711_{\pm.038}\bullet$	$.848_{\pm.029}$	$.789_{\pm.020}\bullet$	$.847_{\pm.013}\bullet$	$.576_{\pm.016}\bullet$	$\mathbf{.272}_{\pm.012}\circ$	$.599_{\pm.016}\bullet$	$.631_{\pm.011}\bullet$	$.515_{\pm.014}\bullet$
BRKNN-LJE	$.453_{\pm.013}\bullet$	$.773_{\pm.041}\bullet$	$.782_{\pm.041}\bullet$	$.769_{\pm.021}\bullet$	$.812_{\pm.022}\bullet$	$.536_{\pm.020}\bullet$	$.188_{\pm.008}\bullet$	$.561_{\pm.013}\bullet$	$.580_{\pm.014}\bullet$	$.488_{\pm.013}\bullet$
BRKNN-COMMU	$.467_{\pm.010}\bullet$	$.700_{\pm.049}\bullet$	$.793_{\pm.029}\bullet$	$.789_{\pm.023}\bullet$	$.850_{\pm.012}\bullet$	$.546_{\pm.016}\bullet$	$.228_{\pm.015}\bullet$	$.586_{\pm.013}\bullet$	$.605_{\pm.014}\bullet$	$.492_{\pm.013}\bullet$
BRKNN-LIMIC	$.464_{\pm.010}\bullet$	$.783_{\pm.032}$	$.854_{\pm.025}$	$.808_{\pm.026}$	$\underline{.859}_{\pm.013}$	$.527_{\pm.019}\bullet$	$\underline{.254}_{\pm.010}\circ$	$.598_{\pm.020}\bullet$	$.647_{\pm.015}\bullet$	$\underline{.532}_{\pm.012}\bullet$
BRKNN-LSMM	$.463_{\pm.010}\bullet$	$\underline{.788}_{\pm.039}$	$\mathbf{.867}_{\pm.036}$	$\underline{.810}_{\pm.023}$	$.857_{\pm.012}\bullet$	$\underline{.583}_{\pm.020}\bullet$	$.248_{\pm.013}\circ$	$\underline{.612}_{\pm.015}\bullet$	$\underline{.652}_{\pm.016}\bullet$	$.528_{\pm.010}\bullet$
BRKNN-ILIA (Ours)	$\mathbf{.482}_{\pm.012}$	$\mathbf{.790}_{\pm.022}$	$\underline{.856}_{\pm.022}$	$\mathbf{.819}_{\pm.021}$	$\mathbf{.868}_{\pm.016}$	$\mathbf{.616}_{\pm.012}$	$.239_{\pm.010}$	$\mathbf{.631}_{\pm.012}$	$\mathbf{.661}_{\pm.013}$	$\mathbf{.554}_{\pm.008}$
MLKNN	$\underline{.494}_{\pm.008}\circ$	$.712_{\pm.042}\bullet$	$.819_{\pm.020}\bullet$	$.789_{\pm.021}\bullet$	$.867_{\pm.017}$	$.525_{\pm.021}\bullet$	$.246_{\pm.006}$	$.616_{\pm.015}\bullet$	$.653_{\pm.012}\bullet$	$.564_{\pm.012}\bullet$
MLKNN-LM	$.493_{\pm.007}$	$.719_{\pm.019}\bullet$	$\mathbf{.864}_{\pm.027}$	$.789_{\pm.017}\bullet$	$.857_{\pm.015}\bullet$	$.606_{\pm.016}\bullet$	$\mathbf{.303}_{\pm.011}\circ$	$.630_{\pm.013}$	$\underline{.671}_{\pm.010}$	$\underline{.570}_{\pm.015}$
MLKNN-LJE	$.491_{\pm.006}$	$.767_{\pm.043}$	$.778_{\pm.041}\bullet$	$.765_{\pm.022}\bullet$	$.819_{\pm.024}\bullet$	$.557_{\pm.020}\bullet$	$.229_{\pm.006}\bullet$	$.588_{\pm.008}\bullet$	$.626_{\pm.012}\bullet$	$.547_{\pm.010}\bullet$
MLKNN-COMMU	$\underline{.494}_{\pm.009}\circ$	$.712_{\pm.042}\bullet$	$.810_{\pm.024}\bullet$	$.790_{\pm.022}\bullet$	$.867_{\pm.016}$	$.519_{\pm.019}\bullet$	$.239_{\pm.006}\bullet$	$.618_{\pm.016}\bullet$	$.653_{\pm.011}\bullet$	$.564_{\pm.012}\bullet$
MLKNN-LIMIC	$\mathbf{.496}_{\pm.007}\circ$	$.773_{\pm.023}$	$.862_{\pm.022}$	$\underline{.813}_{\pm.027}$	$\mathbf{.874}_{\pm.011}$	$.580_{\pm.015}\bullet$	$\underline{.257}_{\pm.009}\circ$	$.628_{\pm.018}$	$.658_{\pm.012}\bullet$	$.568_{\pm.010}\bullet$
MLKNN-LSMM	$\underline{.494}_{\pm.009}\circ$	$\underline{.777}_{\pm.036}$	$\underline{.860}_{\pm.038}$	$.812_{\pm.019}\bullet$	$.869_{\pm.016}$	$\underline{.597}_{\pm.013}\bullet$	$.250_{\pm.010}$	$\underline{.632}_{\pm.015}$	$.659_{\pm.011}\bullet$	$\underline{.570}_{\pm.012}$
MLKNN-ILIA (Ours)	$.485_{\pm.009}$	$\mathbf{.780}_{\pm.026}$	$.862_{\pm.020}$	$\mathbf{.821}_{\pm.017}$	$\underline{.871}_{\pm.011}$	$\mathbf{.620}_{\pm.014}$	$.248_{\pm.008}$	$\mathbf{.634}_{\pm.012}$	$\mathbf{.673}_{\pm.015}$	$\mathbf{.573}_{\pm.008}$

Table 2: Predictive performance (mean$_{\pm \text{std}}$) of $\mathcal{A} \in \{\text{BRKNN}, \text{MLKNN}\}$ coupled with ILIA and state-of-the-art multi-label metric learning approaches in terms of *Hamming Loss* and *Macro-averaging AUC*. ↑ (↓) indicates the larger (smaller) the value, the better the performance. The best and second best results are highlighted in **boldface** and underline, respectively. In addition, •/○ indicates whether $\mathcal{A}$-ILIA achieves significantly superior/inferior to other compared approaches on each dataset in terms of different evaluation metrics (pairwise t-test at 5% significance level).

fair comparisons. Ten-fold cross-validation is employed to evaluate the above approaches.

Empirical Results

Table 2 reports detailed empirical results in terms of *Hamming loss* and *Average precision*. The results on other evaluation metrics can be found in Appendix B.2. Furthermore, pairwise t-test (Dietterich 1998) at 5% significance level is conducted to demonstrate whether the performance difference between $\mathcal{A}$-ILIA and other compared methods is significant statistically, where the resulting win/tie/loss counts are reported in Appendix B.2. The results clearly demonstrate that our proposed ILIA approach has achieved significant improvements in classification performance compared to other multi-label metric learning methods. For example, in terms of BRKNN, ILIA is significantly superior (or comparable) to methods LM, LJE, COMMU, LIMIC, and LSMM in 81.7% (16.7%), 76.7% (18.3%), 96.7% (3.3%), 56.7% (39.3%), and 41.7% (53.3%) of cases, respectively. The superior performance provides persuasive evidence for ILIA in learning effective similarity metrics between multi-label examples.

Additional Comparison

To underscore the significance of learning similarity metrics for multi-label examples, in Appendix C, we compare ILIA-enhanced BRKNN and MLKNN against four well-established metric-free multi-label learning approaches that consider different orders of label correlations, including LIFT (Zhang and Wu 2014), RELIAB (Zhang et al. 2021), WRAP (Yu and Zhang 2021), and HOMI (Si et al. 2023). Detailed empirical results are reported in Appendix C. The results demonstrate that although the performance of BRKNN and MLKNN are inferior to that of second-order and high-order multi-label learning methods, the ILIA-enhanced versions have the potential to approach or even surpass state-of-the-art multi-label learning methods. This outcome not only reaffirms the superiority of ILIA in characterizing the similarity of multi-label examples, but also emphasizes the significance of learning similarity metrics for multi-label examples.

Further Analysis

Ablation study. We study the effects of the two critical algorithmic designs in our ILIA approach: (1) Implicit RLI recovery; (2) Discriminative multi-label metric learning. Ac-

Evaluation Metrics	BRKNN-ILIA **against** BRKNN-DeV1	BRKNN-ILIA **against** BRKNN-DeV2
Hamming Loss	6/3/1	5/5/0
Ranking Loss	7/3/0	6/4/0
Coverage	9/1/0	8/1/1
Average precision	8/2/0	7/3/0
Macro-F1	9/1/0	9/1/0
Macro-averaging AUC	7/3/0	6/3/1
In Total	**46/13/1**	**41/17/2**

Evaluation Metrics	MLKNN-ILIA **against** MLKNN-DeV1	MLKNN-ILIA **against** MLKNN-DeV2
Hamming Loss	3/7/0	4/5/1
Ranking Loss	7/3/0	6/4/0
Coverage	8/2/0	9/1/0
Average precision	9/1/0	10/0/0
Macro-F1	7/2/1	9/1/0
Macro-averaging AUC	6/3/1	7/2/1
In Total	**40/18/2**	**45/13/2**

Table 3: Win/tie/loss counts (pairwise t-test at 5% significant level) for $\mathcal{A}$-ILIA against $\mathcal{A}$-variants.

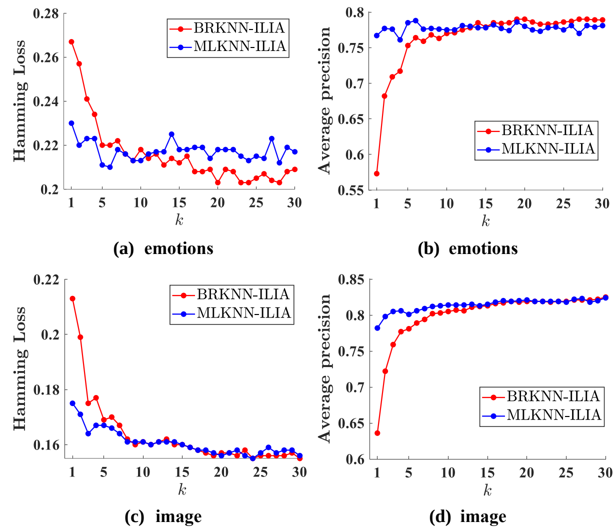

Figure 2: Performance of $\mathcal{A}$-ILIA changes as the number of nearest neighbor k varies in the range of $\{1, 2, \dots, 30\}$.

cordingly, two degenerate variants named DeV1 and DeV2 are implemented for performance comparison:

- DeV1: DeV1 is implemented by removing the implicit RLI recovery procedure in ILIA, which corresponds to the degenerate case considering equal labeling-importance for multi-label metric learning.
- DeV2: DeV2 employs Eq.(14) instead of Eq.(15) for predictive model training, which corresponds to the degenerate case without introducing the discriminative metric $\mathbf{M}$ for similarity characterization. In this case, $\mathbf{M}$ in Eq.(25) is degenerated to an identity matrix.

Table 3 summarizes the win/tie/loss counts (pairwise t-test at 5% significant level) for $\mathcal{A}$-ILIA against $\mathcal{A}$-variants on each evaluation metric. Compared with the two variants, we can observe ILIA achieves statistically superior performance against them in terms of each evaluation metric, demonstrating the usefulness of the two critical algorithmic designs in ILIA for similarity characterization.

Sensitivity analysis. Figure 2 illustrates how the performance of $\mathcal{A}$-ILIA fluctuates with different values of k, i.e, the number of nearest neighbors mentioned in Eq.(2,9,16). (Datasets: emotions, image; Evaluation metrics: *Hamming loss*, *Average precision*). The other parameters are fixed as the same in the 'Configuration' paragraph. It is shown that the performance of $\mathcal{A}$-ILIA gradually improves before $k = 20$ and then tends to stabilize. Therefore, we take $k = 20$ as a fixed parameter in this paper. We also perform sensitivity analyses on the kernel function κ and the trade-off parameters μ, η, and γ, which can be found in Appendix D.

Complexity analysis. There are two critical procedures included in our ILIA approach, i.e., (1) Implicit RLI recovery and (2) discriminative multi-label metric learning. The training complexity of the former procedure is $\mathcal{O}(n \cdot d \cdot logn + n^2 \cdot k + n^3))$. For the latter, the complexity arises from its alternating optimization process. We denote t as the number of iterations, and then the training complexity of the latter procedure is $\mathcal{O}(t \cdot (n \cdot d + n^3))$. Due to the fact that the former procedure is executed only once, the overall training complexity of ILIA is approximately equivalent to the complexity of the latter procedure. To further enhance computational efficiency, following (Zadeh, Hosseini, and Sra 2016), we employ the Cholesky-Schur method (Iannazzo 2016) in this paper to speed up the calculation of Riemannian geodesics for symmetric positive definite matrices in Eq.(24).

Conclusion

In this paper, the first attempt towards leveraging implicit RLI information of multi-label examples for similarity characterization is presented. Different from existing multi-label metric learning approaches learning from multi-label examples by taking the common assumption of equal labeling-importance, we propose a novel approach ILIA, which takes different labeling-importance into consideration. ILIA encompasses two critical procedures, i.e., (1) Implicit RLI recovery and (2) discriminative multi-label metric learning. In (1), the manifold structure within the feature space is exploited by local linear reconstruction, and then the implicit RLIs are recovered by transferring such structure to the label space. In (2), a discriminative multi-label metric learning framework is introduced, which implicitly pulls similar instances closer while pushing dissimilar instances further apart. Comprehensive experiments validate the superiority of ILIA in learning effective similarity metrics between multi-label examples. In the future, it will be interesting to investigate how to recover more accurate RLIs for multi-label examples and how to enable multi-label metric learning to utilize such information for better similarity characterization. Furthermore, it is promising to extend our proposed ILIA approach to weakly supervised scenarios (Xia et al. 2024; Tang, Zhang, and Zhang 2024).

References

Bansal, D.; Chen, R. T.; Mukadam, M.; and Amos, B. 2023. Taskmet: Task-driven metric learning for model learning. *Advances in Neural Information Processing Systems*, 36: 46505–46519.

Bellet, A.; Habrard, A.; and Sebban, M. 2015. *Metric Learning*. Morgan & Claypool Publishers.

Bhatia, R. 2009. *Positive Definite Matrices*. Princeton University Press.

Borchani, H.; Varando, G.; Bielza, C.; and Larranaga, P. 2015. A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(5): 216–233.

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

Chen, S.; Gong, C.; Li, X.; Yang, J.; Niu, G.; and Sugiyama, M. 2023. Boundary-restricted metric learning. *Machine Learning*, 112(12): 4723–4762.

Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7): 1895–1923.

Gong, C.; Demmel, J.; and You, Y. 2024. Distributed and joint evidential k-nearest neighbor classification. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 5972–5985.

Gouk, H.; Pfahringer, B.; and Cree, M. 2016. Learning distance metrics for multi-label classification. In *Proceedings of the 8th Asian Conference on Machine Learning*, 318–333. Hamilton, New Zealand.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 17th IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1735–1742. New York, NY.

Iannazzo, B. 2016. The geometric mean of two matrices from a computational viewpoint. *Numerical Linear Algebra with Applications*, 23(2): 208–229.

Jia, B.-B.; Liu, J.-Y.; Hang, J.-Y.; and Zhang, M.-L. 2023. Learning label-specific features for decomposition-based multi-class classification. *Frontiers of Computer Science*, 17(6): 176348.

Li, P.; Li, Y.; Xie, H.; and Zhang, L. 2022. Neighborhood-adaptive structure augmented metric learning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 1367–1375. Virtual Event.

Liao, S.; and Shao, L. 2022. Graph sampling based deep metric learning for generalizable person re-identification. In *Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7359–7368. New Orleans, LA.

Liu, W.; and Tsang, I. W. 2015. Large margin metric learning for multi-label prediction. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2800–2806. Austin, Tex.

Liu, W.; Wang, H.; Shen, X.; and Tsang, I. W. 2021. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8): 1–19.

Mao, J.-X.; Hang, J.-Y.; and Zhang, M.-L. 2024. Learning label-specific multiple local metrics for multi-label classification. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 4742–4750. Jeju, South Korea.

Mao, J.-X.; Wang, W.; and Zhang, M.-L. 2023. Label specific multi-semantics metric learning for multi-label classification: Global consideration helps. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 4055–4063. Macao, China.

Ren, L.; Chen, C.; Wang, L.; and Hua, K. 2024. Towards improved proxy-based deep metric learning via data-augmented domain adaptation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 14811–14819. Vancouver, Canada.

Roweis, S. T.; and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326.

Schölkopf, B.; and Smola, A. J. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Si, C.; Jia, Y.; Wang, R.; Zhang, M.-L.; Feng, Y.; and Qu, C. 2023. Multi-label classification with high-rank and high-order label correlations. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 4076–4088.

Sun, Y.-P.; and Zhang, M.-L. 2021. Compositional metric learning for multi-label classification. *Frontiers of Computer Science*, 15(5): 1–12.

Tang, W.; Zhang, W.; and Zhang, M.-L. 2024. Multi-instance partial-label learning: Towards exploiting dual inexact supervision. *Science China Information Sciences*, 67(3): 132103.

Uzun, B.; Cevikalp, H.; and Saribas, H. 2022. Deep discriminative feature models (DDFMs) for set based face recognition and distance metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5594–5608.

Wang, F.; and Zhang, C. 2006. Label propagation through linear neighborhoods. In *Proceedings of the 23rd International Conference on Machine Learning*, 985–992. Pittsburgh, PA.

Warburg, F.; Miani, M.; Brack, S.; and Hauberg, S. 2023. Bayesian metric learning for uncertainty quantification in image retrieval. *Advances in Neural Information Processing Systems*, 36: 69178–69190.

Wei, Q.; Dobigeon, N.; and Tourneret, J.-Y. 2015. Fast fusion of multi-band images based on solving a Sylvester equation. *IEEE Transactions on Image Processing*, 24(11): 4109–4121.

Weinberger, K. Q.; Blitzer, J.; and Saul, L. 2005. Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, 18: 1473–1480.

Weinberger, K. Q.; and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2): 207–244.

Xia, X.; Lu, P.; Gong, C.; Han, B.; Yu, J.; and Liu, T. 2024. Regularly truncated m-estimators for learning with noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3522–3536.

Xing, E.; Jordan, M.; Russell, S. J.; and Ng, A. 2002. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems*, 15: 521–528.

Xu, A.; and Davenport, M. 2020. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33: 454–465.

Xu, A.; McRae, A.; Wang, J.; Davenport, M.; and Pananjady, A. 2023. Perceptual adjustment queries and an inverted measurement paradigm for low-rank metric learning. *Advances in Neural Information Processing Systems*, 36: 17969–18000.

Yang, L.; Wang, P.; and Zhang, Y. 2023. Stop-gradient softmax loss for deep metric learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 3164–3172. Washington, DC.

Ye, H.-J.; Zhan, D.-C.; Li, N.; and Jiang, Y. 2020. Learning multiple local metrics: Global consideration helps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7): 1698–1712.

Ye, H.-J.; Zhan, D.-C.; Si, X.-M.; Jiang, Y.; and Zhou, Z.-H. 2019. What makes objects similar: A unified multi-metric learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5): 1257–1270.

Yu, J.; Yin, H.; Xia, X.; Chen, T.; Li, J.; and Huang, Z. 2023. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 335–355.

Yu, Z.-B.; and Zhang, M.-L. 2021. Multi-label classification with label-specific feature generation: A wrapped approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5199–5210.

Zadeh, P.; Hosseini, R.; and Sra, S. 2016. Geometric mean metric learning. In *Proceedings of the 33nd International Conference on Machine Learning*, 2464–2471. New York, NY.

Zhang, M.-L.; and Wu, L. 2014. LIFT: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1): 107–120.

Zhang, M.-L.; Zhang, Q.-W.; Fang, J.-P.; Li, Y.-K.; and Geng, X. 2021. Leveraging implicit relative labeling-importance information for effective multi-Label Learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(5): 2057–2070.

Zhang, M.-L.; and Zhou, Z.-H. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7): 2038–2048.

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

Zhao, Y.; and Yang, L. 2023. Distance metric learning based on the class center and nearest neighbor relationship. *Neural Networks*, 164: 631–644.

Zhu, Y.; Kwok, J. T.; and Zhou, Z.-H. 2017. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6): 1081–1094.