# Label Specific Multi-Semantics Metric Learning for Multi-Label Classification: Global Consideration Helps

**Jun-Xiang Mao**[1,2], **Wei Wang**[3*] and **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]The University of Tokyo, Japan
maojx@seu.edu.cn, wangw@g.ecc.u-tokyo.ac.jp, zhangml@seu.edu.cn

## Abstract

In multi-label classification, it is critical to capitalize on complicated data structures and semantic relationships. Metric learning serves as an effective strategy to provide a better measurement of distances between examples. Existing works on metric learning for multi-label classification mainly learn *one single global metric* that characterizes latent semantic similarity between multi-label instances. However, such single-semantics metric exploitation approaches can not capture the intrinsic properties of multi-label data possessed of rich semantics. In this paper, the first attempt towards multi-semantics metric learning for multi-label classification is investigated. Specifically, the proposed LIMIC approach simultaneously learns *one global and multiple label-specific local metrics* by exploiting label-specific side information. The global metric is learned to capture the commonality across all the labels and label-specific local metrics characterize the individuality of each semantic space. The combination of global metric and label-specific local metrics is utilized to construct latent semantic space for each label, in which similar intra-class instances are pushed closer and inter-class instances are pulled apart. Furthermore, a metric-based label correlation regularization is constructed to maintain similarity between correlated label spaces. Extensive experiments on benchmark multi-label data sets validate the superiority of our proposed approach in learning effective distance metrics for multi-label classification.

## 1 Introduction

Multi-label classification deals with the problem where an instance can be associated with multiple labels simultaneously [Zhang and Zhou, 2014; Liu *et al.*, 2021]. As a practical learning paradigm involving instances with multiple semantics, multi-label classification has been widely driven by real-world applications, such as multimedia annotation [You *et al.*, 2020], text categorization [Tang *et al.*, 2020], bioinformatics analysis [Chen *et al.*, 2017], information retrieval [Gopal and Yang, 2010], and so on.

Similarity elucidates the closeness of connections between instances and is usually reflected by their distances. The predefined distance measurement between instances, e.g. Euclidean distance, is difficult to be adapted for all scenarios [Bellet *et al.*, 2015]. Therefore, metric learning [Xing *et al.*, 2002; Shental *et al.*, 2002; Goldberger *et al.*, 2004] was proposed to take advantage of side information like linkages and comparisons between instances. It automatically learns better distance measurement than the Euclidean one, with which the distance between a pair of instances is consistent with their given relationship, i.e. similar intra-class instances are close to each other, and the distances between dissimilar inter-class instances are large enough. With an adaptively learned distance measurement, the superiority of metric learning has been verified on single-label data for improving similarity/distance-based classification models [Davis *et al.*, 2007; Weinberger and Saul, 2009; Schroff *et al.*, 2015]. Under the single-label scenario, semantic similarity/dissimilarity is easily available by considering whether two instances have the same label. Nevertheless, in the multi-label scenario, it is impractical to measure semantic similarity/dissimilarity by considering each label as an equivalent contribution due to the complicated semantics of multi-label data. As a result, it is much more difficult to learn appropriate metrics to characterize the latent semantic similarity/dissimilarity between multi-label data than single-label ones.

In view of the powerful representation capability of latent semantic space, metric learning has been applied to multi-label classification in recent years [Liu and Tsang, 2015; Gouk *et al.*, 2016; Sun and Zhang, 2021], a.k.a. *multi-label metric learning*, for characterizing more complicated semantic similarity between multi-semantics instances. Existing multi-label metric learning approaches have focused on fusing multi-semantic label information to construct a common distance measurement across all the labels, and then learn one single global metric to characterize the underlying semantic similarity/dissimilarity. Nevertheless, such single-semantics metric exploitation strategies are not consistent with the intrinsic properties of multi-label data possessed of rich semantics. To capitalize on the inherent multi-semantic properties

*The work was done when Wei Wang was with Southeast University.
†Corresponding author.

of multi-label data, it is reasonable and important to consider multiple label-specific distance measurements to demonstrate data structures and relationships when constructing latent label semantic space.

Based on the above observations, this paper presents the first attempt towards *multi-semantics metric learning* for multi-label classification. A novel approach named LIMIC, i.e. ***L**abel Spec**I**cs **M**ulti-Semant**I**cs Metri**C** Learning for Multi-Label Classification*, is proposed accordingly. Different from existing multi-label metric learning approaches considering only one single global metric, LIMIC learns multiple label-specific semantic metrics on the shoulder of the global one. The global metric plays a fundamental role in LIMIC, which considers the side information generated across all the labels to reveal the common characteristics of multi-label data, while each label-specific local metric is a local bias, which depicts the individuality of the corresponding semantic space. In this way, the relationship between instances w.r.t each specific label can be measured in the latent semantic space formed by the combination of the global metric and the corresponding label-specific local metric. To take label correlation into consideration, a metric-based label correlation regularization is further introduced based on label co-occurrence to maintain consistency between correlated label spaces. Extensive experiments on benchmark multi-label data sets validate the superiority of LIMIC in learning effective distance metrics for multi-label classification.

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

## 2 Related Work

**Multi-Label Classification.** As a practical and challenging machine learning paradigm, multi-label classification has been studied extensively in recent years [Zhang and Zhou, 2014; Liu *et al.*, 2021]. To tackle the challenge of exponential-sized output space, label correlation exploitation has been adopted as the most popular strategy. Generally speaking, these approaches can be roughly grouped into three categories, which differ in the order of label correlation considered. The order of label correlation 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 [Fürnkranz *et al.*, 2008; Zhu *et al.*, 2017], and a high-order manner by exploiting relations among a subset or all labels [Tsoumakas *et al.*, 2010; Feng *et al.*, 2019]. In addition to label correlation exploitation, another effective strategy to facilitate multi-label classification is to manipulate the feature space. Dimensionality reduction [Siblini *et al.*, 2021] and feature selection [Pereira *et al.*, 2018] over the original feature space serve as the most common strategies for feature manipulation. Furthermore, there are other feature manipulation strategies such as generating discriminative meta-level features from original features [Canuto *et al.*, 2016], aligning latent space for features and labels [Yeh *et al.*, 2017; Chen *et al.*, 2019a], and exploiting multi-view representation [Xing *et al.*, 2018] or label-specific features [Zhang and Wu, 2014; Hang and Zhang, 2021; Hang *et al.*, 2022] for multi-label data.

**Metric Learning.** Different from the traditional feature manipulation strategies mentioned above, metric learning has been proposed as an alternative feature manipulation strategy. The superiority of metric learning has been verified on single-label data for improving similarity/distance-based classification approaches [Niu *et al.*, 2014; Ye *et al.*, 2019; Ye *et al.*, 2020]. With supervision from various types of side information like linkages and comparisons between instances, metric learning resorts to a suitable similarity or distance measure between instances [Xing *et al.*, 2002; Weinberger and Saul, 2009]. In metric learning, Mahalanobis metric is widely used to replace the Euclidean measurement since it generalizes the Euclidean measurement and can be optimized efficiently [Kulis and others, 2013; Bellet *et al.*, 2015]. Furthermore, Euclidean distance in the transformed space can be viewed as Mahalanobis distance in the original space equivalently. By leveraging such a metric, similar instances tend to have small distances while dissimilar ones are pushed away from each other. Besides, the metric contributes to discovering semantic relationships between instances effectively. Generally speaking, the single metric characterizes the *average* of data [Weinberger *et al.*, 2005; Davis *et al.*, 2007], which represents correlation patterns making relationships between instances in accordance with the provided side information.

**Multi-Label Metric Learning.** To the best of our knowledge, there are three multi-label metric learning approaches available, namely LM [Liu and Tsang, 2015], LJE [Gouk *et al.*, 2016], and COMMU [Sun and Zhang, 2021]. LM exploits a large margin formulation to construct a common metric space, in which the similarity relationship in the input space should be preserved in the output space. LJE aims at learning a metric that can project instances into the feature space where the Euclidean distance provides an estimation of the Jaccard distance between corresponding label vectors. COMMU constructs a compositional metric by modeling structural interactions between feature and label space to explore the integrated semantics of multiple labels. The multi-label metric learning approaches above all focus on fusing multi-semantic label information to construct a common distance measurement across all the labels, and then learn one single global metric to characterize the underlying semantic similarity/dissimilarity. However, these strategies are not consistent with the intrinsic properties of multi-label data possessed of rich semantics and might lead to suboptimal performance in learning distance metrics. In the next section, the first attempt towards multi-semantics metric learning for multi-label classification is introduced

## 3 The LIMIC Approach

### 3.1 Preliminaries

Let $\mathcal{X} = \mathbb{R}^d$ denote the input 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 vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_q] \in \{0,1\}^q$ can be utilized to denote $Y$, where $y_p = 1$ when $l_p \in Y$ and $y_p = 0$ otherwise. Generally speaking, multi-label classification aims to induce a multi-label prediction function $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from a multi-label data set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq n\}$. Given an unseen instance $\boldsymbol{x'} \in \mathcal{X}$, its associated label set is predicted as $h(\boldsymbol{x'}) \subseteq \mathcal{Y}$.

Meanwhile, let $\mathbb{S}_+^d$ denotes the space of $d \times d$ Positive Semi-Definite (PSD) matrices. Given a metric $\mathbf{M} \in \mathbb{S}_+^d$, the (squared) Mahalanobis distance between pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is $(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \mathbf{M} (\boldsymbol{x}_i - \boldsymbol{x}_j) = \langle \mathbf{M}, \mathbf{A}_{ij} \rangle = \mathrm{Tr}(\mathbf{M}\mathbf{A}_{ij})$. The outer product of the pair difference is $\mathbf{A}_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \in \mathbb{S}_+^d$. [1] By decomposing the metric into the inner product of transformations $\mathbf{L}$ ($\mathbf{L} \in \mathbb{R}^{d \times d'}, d' \leq d$) as $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$, the (square) Mahalanobis distance between two instances is equal to their Euclidean distance in a projected space:

$$\begin{aligned} & \mathrm{Dis}^2_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \mathbf{M} (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ \Longleftrightarrow \ & \mathrm{Dis}^2_{\mathbf{L}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \mathbf{L}\mathbf{L}^\top (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ & \qquad = ||\mathbf{L}^\top (\boldsymbol{x}_i - \boldsymbol{x}_j)||_2^2. \end{aligned} \tag{1}$$

Generally speaking, there are several advantages in learning transformation $\mathbf{L}$ rather than metric $\mathbf{M}$. Since there is no PSD constraint on transformation $\mathbf{L}$, no PSD projection step is required, which can accelerate the optimization procedure. In addition, the transformation decomposition often leads to low-rank metrics, which can be advantageous in many real-world applications, such as information retrieval. It is also noteworthy that although the decomposition leads to non-convex problems, satisfactory solutions can be obtained as well [Weinberger and Saul, 2009; Parameswaran and Weinberger, 2010; Ye *et al.*, 2020]. Based on the above consideration, we learn the transformation $\mathbf{L}$ rather than the metric $\mathbf{M}$ in this paper.

### 3.2 Label-Specific Multi-Semantics Metric Learning

For the $p$-th label $l_p$, the set of positive training instances $\mathcal{P}_p$ as well as the set of negative training instances $\mathcal{N}_p$ are determined by considering the relevance of each example to $l_p$:

$$\begin{aligned} \mathcal{P}_p &= \{\boldsymbol{x}_i | (\boldsymbol{x}_i, Y_i) \in \mathcal{D}, l_p \in Y_i\}, \\ \mathcal{N}_p &= \{\boldsymbol{x}_i | (\boldsymbol{x}_i, Y_i) \in \mathcal{D}, l_p \notin Y_i\}. \end{aligned} \tag{2}$$

The label-specific local side information w.r.t the label $l_p$ is composed of all pairwise combinations between all the training examples as $\mathcal{T}_p = \{(\boldsymbol{x}_i, \boldsymbol{x}_j, \theta_{ij}^p)\}$ where $\theta_{ij}^p \in \{-1, +1\}$ indicates whether $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ have the same relevance w.r.t the label $l_p$. Concretely, $\theta_{ij}^p = 1$ means the instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are similar because they are (or not) in possession of the label $l_p$ simultaneously, i.e. $(\boldsymbol{x}_i \in \mathcal{P}_p \wedge \boldsymbol{x}_j \in \mathcal{P}_p) \vee (\boldsymbol{x}_i \in \mathcal{N}_p \wedge \boldsymbol{x}_j \in \mathcal{N}_p)$ and $\theta_{ij}^p$ equals -1 otherwise. Let $(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p$ denote the enumeration of totally $n(n-1)$ pairs from the label-specific local side information $\mathcal{T}_p$. Let $T_p$ denote the number of pairs in $\mathcal{T}_p$, and $T_p = n(n-1)$. In practice, there is no need to compute all the tuples of side information, which may suffer from a severe computational burden. A reasonable amount of targets and imposters selected among the nearest neighbors can retrench computation and facilitate the training procedure, where targets indicate similar instances w.r.t. the anchor and imposters otherwise [Weinberger and Saul, 2009; Chen *et al.*, 2019b; Ye *et al.*, 2019; Ye *et al.*, 2020].

To take both global and local semantic relationships into consideration, we take advantage of the sum of global transformation $\mathbf{L}_0$ and local bias $\mathbf{L}_p$ as the distance metric for the label space of the $p$-th label $l_p$. In this combination, the global transformation represents a common view of semantic measurement, while the local bias conducts adaptation for the individuality of each label-specific semantic space. Specifically, given the global transformation $\mathbf{L}_0$ which construct the commonality across all the labels, the $p$-th local label-specific transformation $\mathbf{L}_p$ can be determined by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{L}_p} \frac{1}{T_p} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p} & \ell(\theta_{ij}^p (\gamma - \mathrm{Dis}^2_{\mathbf{L}_0 + \mathbf{L}_p}(\boldsymbol{x}_i, \boldsymbol{x}_j))) \\ & + \lambda ||\mathbf{L}_p||_F^2. \end{aligned} \tag{3}$$

Here, $\gamma$ is a pre-defined non-negative threshold value and can be different for similar and dissimilar pairs. $\ell(\cdot)$ is a convex and non-increasing loss function. If two instances $\boldsymbol{x}_i, \boldsymbol{x}_j$ are (or not) in possession of the label $l_p$ simultaneously, i.e. $\theta_{ij}^p = 1$, then the loss equals 0 if their distance with $\mathbf{L}_p$ is smaller than $\gamma$. On the other hand, when they are dissimilar ($\theta_{ij}^p = -1$), their distance should be larger than $\gamma$. By optimizing Eq.(3), the learned transformation requires similar instances to have small distances, while instances from different classes are far away enough. Besides, $\lambda$ is a non-negative weight to balance the influence of the regularization term. In this paper, the smooth hinge loss is used to instantiate $\ell(\cdot)$, which is defined as

$$\ell(x) = \begin{cases} 0, & \text{if} \quad x > 1 \\ \frac{1}{2}(x-1)^2, & \text{if} \quad 0 \leq x \leq 1 \\ \frac{1}{2} - x, & \text{if} \quad x < 0. \end{cases} \tag{4}$$

The smoothness property of this loss function will facilitate the optimization process. Besides, $\ell(\cdot)$ also keeps a small margin, which further improves the generalization of $\mathbf{L}_p$.

Based on the above modeling procedures for a single label, we can easily extend the objective function Eq.(3) to the whole label space with $q$ labels. The global transformation $\mathbf{L}_0$ and $q$ label-specific local biases $\mathbf{L}_1, \mathbf{L}_2, \ldots$, and $\mathbf{L}_q$ can be determined by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{L}_0, \mathbf{L}_1, \ldots, \mathbf{L}_q} \sum_{p=1}^{q} \frac{1}{T_p} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p} & \ell(\theta_{ij}^p (\gamma - \mathrm{Dis}^2_{\mathbf{L}_0 + \mathbf{L}_p}(\boldsymbol{x}_i, \boldsymbol{x}_j))) \\ & + \lambda \sum_{p=0}^{q} ||\mathbf{L}_p||_F^2. \end{aligned} \tag{5}$$

[1] In the following descriptions, we do not differentiate "metric" and "transformation", which can be determined from the context.

### 3.3 Label Correlation Exploitation

Label co-occurrence is an essential semantic relationship in multi-label classification, which has been proven to be effective prior information for label correlation exploitation [Kurata *et al.*, 2016; Hang and Zhang, 2021]. In this section, we capitalize on the label co-occurrence information to establish the relationship of label-specific metrics, where labels with strong co-occurrence possess similar transformations.

Specifically, we construct a label relation graph based on statistics of label co-occurrence. Let $G = (V, E)$ denote such a label relation graph, where $V$ denotes the set of labels and $E$ denotes the set of edges between label pairs. The adjacency matrix $\mathbf{W}$ denotes the weights associated with each edge, representing the strength of the co-occurrence relationship between pairs of labels. We formulate the adjacency matrix $\mathbf{W}$ as the symmetric conditional probability matrix.[2] Each element in $\mathbf{W}$ is calculated as

$$w_{ij} = \frac{1}{2}[P(l_j|l_i) + P(l_i|l_j)], \tag{6}$$

where $P(l_j|l_i)$ is the probability that label $l_j$ appears when label $l_i$ appears and the diagonal elements of adjacency matrix $\mathbf{W}$ are set to 0. We calculate the adjacency matrix $\mathbf{W}$ from the training set.

Intuitively, the stronger the correlation between two labels $l_i, l_j$ is, the more similar the corresponding two label-specific transformations $\mathbf{L}_i, \mathbf{L}_j$ are. On the contrary, if two labels $l_i, l_j$ have a low frequency of co-occurrence, the corresponding label transformations $\mathbf{L}_i, \mathbf{L}_j$ should be dissimilar. Therefore, the objective function for modeling the relation between different label-specific local transformations $\mathbf{L}_1, \mathbf{L}_2, \ldots$, and $\mathbf{L}_q$ can be formulated as

$$\min_{\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_q} \sum_{i=1}^{q} \sum_{j=1}^{q} w_{ij} ||\mathbf{L}_i - \mathbf{L}_j||_F^2. \tag{7}$$

By introducing the label correlation exploitation strategy shown above into Eq.(5), the overall LIMIC framework can be achieved as follows:

$$\begin{aligned} \min_{\mathbf{L}_0, \mathbf{L}_1, \ldots, \mathbf{L}_q} & \sum_{p=1}^{q} \frac{1}{T_p} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p} \ell(\theta_{ij}^p(\gamma - \mathrm{Dis}^2_{\mathbf{L}_0 + \mathbf{L}_p}(\boldsymbol{x}_i, \boldsymbol{x}_j))) \\ & + \lambda_1 \sum_{p=0}^{q} ||\mathbf{L}_p||_F^2 + \lambda_2 \sum_{i=1}^{q} \sum_{j=1}^{q} w_{ij} ||\mathbf{L}_i - \mathbf{L}_j||_F^2. \end{aligned} \tag{8}$$

### 3.4 Optimization Procedure

Let $\mathcal{L}$ denote the objective of the LIMIC framework in Eq.(8). Following traditional implementation methods of metric learning algorithms [de Vazelhes *et al.*, 2020; Suárez *et al.*, 2020], we optimize the LIMIC framework with gradient-based optimization strategies. The detailed procedures are discussed as follows.

[2] Actually, the adjacency matrix $\mathbf{W}$ can be constructed in numerous alternative ways or even can be implemented in a learnable formulation. We attempt to focus on the label-specific multi-semantics exploitation process and will leave it for further work.

**Algorithm 1** The pseudo-code of LIMIC.

**Input**:
$\mathcal{D}$: a multi-label training set $\{(\boldsymbol{x}_i, Y_i) \mid 1 \le i \le n\}$ $(\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{l_1, l_2, \ldots, l_q\}, \boldsymbol{x}_i \in \mathcal{X}, Y_i \subseteq \mathcal{Y})$
$\lambda_1, \lambda_2$: regularization parameters in Eq.(8)
**Output**:
$\mathbf{L}_0, \mathbf{L}_1, \ldots, \mathbf{L}_q$: the learned global and label-specific local transformations
**Process**:
1: Initialize $\mathbf{L}_0$ with $\mathbf{I}$ or employ multi-label metric learning approaches;
2: Initialize $\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_q$ with $\mathbf{0}$;
3: Compute adjacency matrix $\mathbf{W}$ according to Eq.(6);
4: **for** $p = 1$ to $q$ **do**
5: Generate positive set $\mathcal{P}_p$ and negative set $\mathcal{N}_p$ according to Eq.(2);
6: Generate label-specific local side information tuples $\mathcal{T}_p$;
7: **end for**
8: **repeat**
9: Optimize Eq.(8) over $\mathbf{L}_0, \mathbf{L}_2, \ldots, \mathbf{L}_q$ with accelerated gradient descent according to Eq.(9) and Eq.(10);
10: **until** convergence or maximum number of iterations being reached
11: Return $\mathbf{L}_0, \mathbf{L}_1, \ldots, \mathbf{L}_q$

**Initialization.** For the global transformation $\mathbf{L}_0$, it can be initialized in two ways. We can instantiate the global transformation as an identity matrix, i.e. $\mathbf{L}_0 = \mathbf{I}$, which equals the Euclidean distance metric. It can also be initialized using existing multi-label metric learning approaches such as LM. In our experiments, we adopt the first strategy. Label-specific local transformations $\mathbf{L}_1, \mathbf{L}_2, \ldots$, and $\mathbf{L}_q$ can be initialized as zero metrics since they are only complementary components based on $\mathbf{L}_0$ for label-specific local distance metrics.

**Gradient-based optimization.** We utilize accelerated gradient descent [Nesterov, 2003; Boyd *et al.*, 2004] to optimize the global transformation $\mathbf{L}_0$ and local transformations $\mathbf{L}_1, \mathbf{L}_2, \ldots$, and $\mathbf{L}_q$ simultaneously. Let $\delta_{i,j,p} = \theta_{ij}^p(\gamma - \mathrm{Dis}^2_{\mathbf{L}_0 + \mathbf{L}_p}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and the derivations of $\mathcal{L}$ w.r.t $\mathbf{L}_0$ and $\mathbf{L}_p$ respectively are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{L}_0} = & \sum_{p=1}^{q} \frac{2}{T_p} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p} \sigma_{i,j,p} \theta_{ij}^p \mathbf{A}_{ij} (\mathbf{L}_0 + \mathbf{L}_p) \\ & + 2\lambda_1 \mathbf{L}_0, \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{L}_p} = & \frac{2}{T_p} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p} \sigma_{i,j,p} \theta_{ij}^p \mathbf{A}_{ij} (\mathbf{L}_0 + \mathbf{L}_p) \\ & + 2\lambda_1 \mathbf{L}_p + 4\lambda_2 \sum_{i=1}^{q} w_{pi} (\mathbf{L}_p - \mathbf{L}_i), \end{aligned} \tag{10}$$

where $\sigma_{i,j,p}$ is a piecewise function defined as follows:

$$\sigma_{i,j,p} = \begin{cases} 0, & \text{if} \quad \delta_{i,j,p} > 1 \\ 1 - \delta_{i,j,p}, & \text{if} \quad 0 \le \delta_{i,j,p} \le 1 \\ 1, & \text{if} \quad \delta_{i,j,p} < 0. \end{cases} \tag{11}$$

| Dataset | $\lvert\mathcal{S}\rvert$ | $dim(\mathcal{S})$ | $L(\mathcal{S})$ | $F(\mathcal{S})$ | $LCard(\mathcal{S})$ | $LDen(\mathcal{S})$ | $DL(\mathcal{S})$ | $PDL(\mathcal{S})$ | Domain |
|---|---|---|---|---|---|---|---|---|---|
| CAL500 | 502 | 68 | 174 | Numeric | 26.044 | 0.150 | 502 | 1.000 | Music[1] |
| emotions | 593 | 72 | 6 | Numeric | 1.869 | 0.311 | 27 | 0.046 | Music[1] |
| birds | 645 | 260 | 19 | Numeric | 1.014 | 0.053 | 133 | 0.206 | Audio[1] |
| genbase | 662 | 1186 | 27 | Nominal | 1.252 | 0.046 | 32 | 0.048 | Biology[1] |
| medical | 978 | 1449 | 45 | Nominal | 1.245 | 0.028 | 94 | 0.096 | Text[1] |
| image | 2000 | 294 | 5 | Numeric | 1.236 | 0.247 | 20 | 0.010 | Image[2] |
| scene | 2407 | 294 | 6 | Numeric | 1.074 | 0.179 | 15 | 0.006 | Image[1] |
| yeast | 2417 | 103 | 14 | Numeric | 4.237 | 0.303 | 198 | 0.082 | Biology[1] |

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

Table 1: Characteristics of experimental data sets.

The complete procedure of LIMIC is summarized in Algorithm 1. Firstly, global and label-specific local transformations $\mathbf{L}_0, \mathbf{L}_1, \ldots$, and $\mathbf{L}_q$ are initialized (Step 1-2). After that, the adjacency matrix $\mathbf{W}$ is computed for label correlation exploitation (Step 3) and label-specific side information is generated from the label semantic information (Step 4-7). Finally, an accelerated gradient descent procedure is invoked to optimize $\mathbf{L}_0, \mathbf{L}_1, \ldots$, and $\mathbf{L}_q$ iteratively (Step 8-10).

After learning the global transformation $\mathbf{L}_0$ and multiple label-specific local transformations $\mathbf{L}_1, \mathbf{L}_2, \ldots$, and $\mathbf{L}_q$ with the LIMIC approach, it is straightforward to calculate the distance between instances in the semantic space of each label according to Eq.(1). Therefore, the associated label set of an unseen instance can be predicted by resorting to classic similarity/distance-based classification strategies such as the $k$-nearest neighbor (KNN) algorithm.

# 4 Experiments

## 4.1 Experimental Setup

**Data sets.** In this paper, eight benchmark multi-label data sets have been employed for comprehensive performance evaluation. Table 1 summarizes the characteristics of each experimental data set $S$, including the number of examples $|S|$, number of features $dim(S)$, number of class labels $L(S)$, feature type $F(S)$, label cardinality $LCard(S)$, i.e. the average number of labels per instance, label density (label cardinality over $L(S)$) $LDen(S)$, number of distinct label sets $DL(S)$, and proportion of distinct label sets $PDL(S)$.

**Evaluation metrics.** For performance evaluation, six widely-used evaluation metrics are utilized for multi-label classification, including *Hamming loss*, *Ranking loss*, *Coverage*, *Average precision*, *Macro-F1*, and *Macro-averaging AUC*. Detailed definitions of these metrics can be found in [Zhang and Zhou, 2014].

## 4.2 Comparative Studies

To validate the effectiveness of the proposed LIMIC approach in learning effective distance metrics for multi-label classification, three similarity/distance-based multi-label classification strategies are introduced as subsequent classification methods after learning the distance metric:

- BR-KNN [Boutell *et al.*, 2004]: A classic multi-label classification approach that decomposes the multi-label classification into a set of binary KNN classification tasks [parameter configuration: K = 10].
- ML-KNN [Zhang and Zhou, 2007]: A popular lazy learning approach for multi-label classification with Bayesian inference. [parameter configuration: K = 10].
- RELIAB-KNN [Zhang *et al.*, 2021]: A KNN-based multi-label classification approach that leverages the implicit relative labeling-importance information with local KNN reconstruction [parameter configuration: K = 10, $\rho = 0.3$, $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 10\}$].

Given a similarity/distance-based multi-label classification strategy $\mathcal{A} \in \{\text{BR-KNN}, \text{ML-KNN}, \text{RELIAB-KNN}\}$ and a multi-label metric learning algorithm $\mathcal{B}$, the coupling version of them is denoted as $\mathcal{A}$-$\mathcal{B}$. The predictive performance of $\mathcal{A}$-LIMIC is compared with other multi-label metric learning algorithms coupled with $\mathcal{A}$ to manifest whether the proposed multi-label metric learning technique does learn effective distance metrics and improve the generalization performance for multi-label classification.

In this paper, three well-established multi-label metric learning algorithms are employed to instantiate $\mathcal{B}$ with suggested configurations in respective literature:

- LM [Liu and Tsang, 2015]: A margin-based multi-label metric learning approach that learns a common semantic metric by employing a large margin formulation [parameter configuration: $\eta = 0.4$, $C = 10$].
- LJE [Gouk *et al.*, 2016]: An integration-based multi-label metric learning approach that employs Jaccard distance between label vectors to provide more fine-grained side information [parameter configuration: $t = 32$, $e = 5$].
- COMMU [Sun and Zhang, 2021]: A composition-based multi-label metric learning approach that learns a compositional metric by modeling structural interactions between feature and label space [parameter configuration: $\alpha, \theta \in \{0.2, 0.4, \ldots, 0.8\}$ and $C = 10$].

For the proposed LIMIC approach, regularization parameters $\lambda_1$ and $\lambda_2$ are searched in $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. The number of targets and imposters is fixed to 10 and $\gamma$ in Eq.(8) is set to 2 which is consistent with conventional metric learning approaches [Weinberger and Saul, 2009; Ye *et al.*, 2020]. Ten-fold cross-validation is employed to evaluate the above approaches on the 8 benchmark multi-label data sets.

| Compared Algorithms | Data Sets | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | CAL500 | emotions | birds | genbase | medical | image | scene | yeast |
| | *Hamming Loss* ↓ | | | | | | | |
| BR-KNN | **0.145±0.003** | 0.263±0.023• | 0.056±0.007• | 0.003±0.001• | 0.016±0.002• | 0.170±0.017• | 0.091±0.007• | 0.198±0.006• |
| BR-KNN-LM | 0.150±0.003• | 0.270±0.019• | 0.065±0.009• | **0.001±0.001**○ | **0.011±0.002**○ | 0.175±0.016• | 0.090±0.009 | 0.212±0.013• |
| BR-KNN-LJE | **0.145±0.004** | 0.221±0.017• | 0.055±0.006• | 0.003±0.001• | 0.021±0.003• | 0.186±0.016• | 0.108±0.009• | 0.205±0.010• |
| BR-KNN-COMMU | **0.145±0.003** | 0.263±0.023• | 0.056±0.007• | 0.003±0.001 | 0.016±0.002• | 0.171±0.016• | 0.091±0.007 | 0.198±0.006• |
| BR-KNN-LIMIC | **0.145±0.005** | **0.207±0.012** | **0.050±0.006** | 0.003±0.001 | 0.012±0.002 | **0.160±0.017** | **0.087±0.011** | **0.192±0.012** |
| ML-KNN | 0.139±0.005 | 0.262±0.022• | 0.054±0.006 | 0.005±0.002 | 0.016±0.002• | 0.174±0.013• | 0.085±0.009• | 0.195±0.009• |
| ML-KNN-LM | 0.139±0.004 | 0.254±0.017• | 0.054±0.007 | **0.003±0.001**○ | **0.013±0.002** | 0.176±0.014• | 0.088±0.008• | 0.205±0.012• |
| ML-KNN-LJE | **0.138±0.005** | 0.224±0.015• | 0.054±0.006 | 0.005±0.001 | 0.022±0.002• | 0.186±0.016• | 0.107±0.007• | 0.204±0.010• |
| ML-KNN-COMMU | 0.139±0.004 | 0.262±0.022• | 0.054±0.006 | 0.005±0.002 | 0.015±0.002• | 0.174±0.014• | 0.085±0.009 | 0.195±0.009• |
| ML-KNN-LIMIC | **0.138±0.004** | **0.209±0.017** | **0.052±0.006** | 0.004±0.001 | **0.013±0.002** | **0.161±0.019** | **0.083±0.006** | **0.191±0.011** |
| RELIAB-KNN | 0.118±0.007 | 0.238±0.030• | 0.036±0.005• | 0.002±0.002 | 0.013±0.004• | 0.147±0.025• | 0.065±0.015• | 0.167±0.010• |
| RELIAB-KNN-LM | 0.121±0.005• | 0.245±0.028• | 0.042±0.006• | **0.001±0.001** | **0.010±0.003** | 0.153±0.023• | 0.060±0.012• | 0.175±0.009• |
| RELIAB-KNN-LJE | **0.115±0.006**○ | 0.213±0.026• | 0.033±0.004• | 0.002±0.002 | 0.014±0.003• | 0.159±0.020• | 0.078±0.013• | 0.172±0.012• |
| RELIAB-KNN-COMMU | 0.118±0.007 | 0.238±0.030• | 0.036±0.005• | 0.002±0.002 | 0.012±0.004• | 0.148±0.025• | 0.065±0.015• | 0.167±0.010• |
| RELIAB-KNN-LIMIC | 0.116±0.008 | **0.176±0.026** | **0.028±0.004** | **0.001±0.002** | **0.010±0.003** | **0.132±0.018** | **0.058±0.014** | **0.161±0.008** |
| | *Ranking Loss* ↓ | | | | | | | |
| BR-KNN | 0.255±0.011 | 0.272±0.048• | 0.477±0.051• | 0.005±0.005 | 0.081±0.028• | 0.185±0.020• | **0.096±0.011**• | 0.184±0.013• |
| BR-KNN-LM | 0.258±0.007 | 0.256±0.030• | 0.491±0.045• | 0.010±0.009• | 0.106±0.037• | 0.187±0.020 | 0.107±0.012 | 0.202±0.020• |
| BR-KNN-LJE | 0.260±0.012 | 0.197±0.034• | 0.455±0.055• | 0.005±0.005 | 0.123±0.035• | 0.203±0.021• | 0.123±0.014• | 0.195±0.012• |
| BR-KNN-COMMU | 0.253±0.011 | 0.272±0.048• | 0.477±0.052• | 0.005±0.005 | 0.082±0.027 | 0.185±0.020 | **0.096±0.011** | 0.184±0.013 |
| BR-KNN-LIMIC | **0.251±0.015** | **0.172±0.035** | **0.391±0.039** | **0.003±0.005** | **0.078±0.024** | **0.177±0.026** | 0.100±0.015 | **0.181±0.017** |
| ML-KNN | 0.183±0.005• | 0.258±0.038• | 0.295±0.035• | 0.004±0.004 | **0.032±0.009**• | 0.176±0.019 | **0.077±0.010**• | **0.167±0.013**• |
| ML-KNN-LM | 0.184±0.004 | 0.240±0.026• | 0.298±0.047• | 0.003±0.003 | 0.034±0.016 | 0.176±0.018• | 0.084±0.010• | 0.178±0.020• |
| ML-KNN-LJE | 0.184±0.005• | 0.196±0.028• | 0.275±0.050• | **0.002±0.003** | 0.056±0.012• | 0.194±0.021• | 0.111±0.012• | 0.177±0.012• |
| ML-KNN-COMMU | **0.182±0.004** | 0.258±0.038• | 0.294±0.035• | 0.004±0.004• | 0.033±0.009 | 0.176±0.019• | **0.077±0.010** | **0.167±0.014** |
| ML-KNN-LIMIC | **0.182±0.004** | **0.177±0.030** | **0.254±0.039** | **0.002±0.002** | **0.032±0.013** | **0.161±0.024** | 0.078±0.011 | **0.167±0.014** |
| RELIAB-KNN | 0.126±0.003• | 0.167±0.032• | 0.272±0.032• | **0.003±0.002** | 0.028±0.011• | 0.143±0.018• | **0.061±0.008** | 0.139±0.016• |
| RELIAB-KNN-LM | **0.122±0.004** | 0.145±0.027• | **0.226±0.035** | 0.004±0.001• | 0.032±0.012• | 0.156±0.015• | 0.068±0.007• | 0.137±0.015 |
| RELIAB-KNN-LJE | 0.138±0.005• | 0.122±0.030• | 0.296±0.032• | **0.003±0.002** | 0.035±0.016• | 0.148±0.016• | 0.072±0.006• | **0.134±0.013** |
| RELIAB-KNN-COMMU | 0.126±0.003• | 0.167±0.032• | 0.272±0.034• | **0.003±0.002** | **0.027±0.011** | 0.143±0.018• | **0.061±0.008** | 0.139±0.016• |
| RELIAB-KNN-LIMIC | **0.122±0.005** | **0.117±0.025** | **0.226±0.030** | **0.003±0.002** | **0.027±0.010** | **0.126±0.013** | 0.062±0.005 | 0.135±0.017 |
| | *Coverage* ↓ | | | | | | | |
| BR-KNN | 0.840±0.025 | 0.378±0.032• | 0.216±0.037• | 0.015±0.008 | 0.075±0.025• | 0.197±0.021• | 0.087±0.011• | 0.452±0.015• |
| BR-KNN-LM | **0.833±0.015** | 0.364±0.032• | 0.220±0.036• | 0.022±0.010• | 0.082±0.026• | 0.194±0.021 | 0.095±0.009• | 0.473±0.018• |
| BR-KNN-LJE | **0.833±0.021** | 0.324±0.022• | 0.202±0.039• | 0.015±0.005 | 0.114±0.031• | 0.209±0.020• | 0.110±0.010• | 0.461±0.017• |
| BR-KNN-COMMU | 0.837±0.024 | 0.378±0.032• | 0.216±0.037• | 0.015±0.008 | 0.076±0.024 | 0.197±0.021• | 0.087±0.011 | 0.453±0.015• |
| BR-KNN-LIMIC | 0.838±0.025 | **0.300±0.027** | **0.176±0.024** | **0.012±0.005** | **0.066±0.019** | **0.184±0.026** | **0.085±0.011** | **0.443±0.016** |
| ML-KNN | **0.745±0.018** | 0.377±0.026 | 0.191±0.029• | 0.016±0.008 | 0.047±0.011• | 0.195±0.021• | **0.078±0.010**• | **0.448±0.015**• |
| ML-KNN-LM | 0.755±0.018 | 0.360±0.033• | 0.190±0.033• | 0.015±0.006 | 0.049±0.019 | 0.194±0.020• | 0.084±0.009 | 0.465±0.019• |
| ML-KNN-LJE | 0.753±0.016 | 0.328±0.016 | 0.178±0.040• | 0.014±0.006 | 0.075±0.013• | 0.208±0.021• | 0.107±0.011• | 0.460±0.016• |
| ML-KNN-COMMU | 0.746±0.016○ | 0.377±0.026• | 0.191±0.029• | 0.016±0.008 | 0.047±0.011 | 0.195±0.021• | **0.078±0.010** | **0.448±0.015** |
| ML-KNN-LIMIC | 0.755±0.016 | **0.305±0.024** | **0.164±0.026** | **0.013±0.006** | **0.046±0.014** | **0.182±0.024** | 0.080±0.011 | 0.450±0.015 |
| RELIAB-KNN | 0.672±0.023• | 0.304±0.022• | 0.189±0.030• | 0.015±0.008• | 0.038±0.015• | 0.164±0.023• | **0.062±0.011** | 0.387±0.016• |
| RELIAB-KNN-LM | 0.662±0.019 | 0.287±0.020• | 0.168±0.028• | **0.012±0.005** | 0.051±0.017• | 0.158±0.022• | 0.071±0.010• | 0.394±0.012• |
| RELIAB-KNN-LJE | **0.658±0.017**○ | 0.253±0.026• | 0.173±0.025• | 0.013±0.006 | 0.047±0.012• | 0.178±0.025• | 0.069±0.012• | 0.399±0.013• |
| RELIAB-KNN-COMMU | 0.671±0.023• | 0.304±0.022• | 0.188±0.030• | 0.015±0.008• | 0.038±0.015• | 0.164±0.023• | **0.062±0.011** | 0.388±0.016• |
| RELIAB-KNN-LIMIC | 0.667±0.020 | **0.229±0.023** | **0.144±0.031** | **0.012±0.007** | **0.029±0.013** | **0.147±0.022** | **0.062±0.013** | **0.365±0.015** |

Table 2: Predictive performance of each compared approach (mean±std) in terms of *Hamming Loss*, *Ranking Loss*, and *Coverage*. ↑ (↓) indicates the larger (smaller) the value, the better the performance. The best results are highlighted in **boldface**. In addition, •/○ indicates whether $\mathcal{A}$-LIMIC ($\mathcal{A} \in \{$BR-KNN, ML-KNN, RELIAB-KNN$\}$) achieves significantly superior/inferior to the compared approach on each data set in terms of different evaluation metrics (pairwise t-test at 0.05 significance level).

Due to page limit, Table 2 reports detailed experimental results in terms of *Hamming loss*, *Ranking loss*, and *Coverage*. The results on other metrics can be found in the supplementary material. Our proposed LIMIC approach coupled with $\mathcal{A} \in \{$BR-KNN, ML-KNN, RELIAB-KNN$\}$ are compared with other coupling versions of multi-label metric learning approaches $\mathcal{B} \in \{$LM, LJE, COMMU$\}$ respectively. Meanwhile, the original version of BR-KNN, ML-KNN, and RELIAB-KNN which adopt the Euclidean metric are also included as compared approaches. For each evaluation, "↓" indicates "the smaller the better" while "↑" indicates "the larger the better". The best performance among compared algorithms is shown in boldface. Furthermore, pairwise $t$-test at 0.05 significance level is conducted to demonstrate whether the performance difference between $\mathcal{A}$-LIMIC and $\mathcal{A}$-$\mathcal{B}$ is significant statistically, where the resulting win/tie/loss counts are reported in the supplementary material. Based on the experimental results reported in Table 2, it is impressive to observe that:

- For similarity/distance-based multi-label classification strategies $\mathcal{A} \in \{\text{BR-KNN}, \text{ML-KNN}, \text{RELIAB-KNN}\}$, the generalization performance has been greatly improved after coupling multi-label metric learning algorithms. Especially, BR-KNN-LIMIC, ML-KNN-LIMIC, and RELIAB-KNN-LIMIC achieve better performance than BR-KNN, ML-KNN, RELIAB-KNN in 89.6%, 81.3% and 87.5% cases respectively. The results validate the effectiveness of metric learning for improving similarity/distance-based classification methods.
- Across all evaluation metrics, BR-KNN-LIMIC, ML-KNN-LIMIC, and RELIAB-KNN-LIMIC achieve the best performance in 87.5%, 81.3%, 83.3% respectively over all the multi-label data sets. Meanwhile, BR-KNN-LIMIC (ML-KNN-LIMIC, RELIAB-KNN-LIMIC) significantly outperforms corresponding coupled versions of other multi-label metric learning algorithms $\mathcal{B} \in \{\text{LM}, \text{LJE}, \text{COMMU}\}$ in 70.8% (62.5%, 83.3%), 79.2% (72.9%, 87.5%), and 47.9% (54.2%, 79.2%) cases respectively. The superior performance of LIMIC provides persuasive evidence for the effectiveness of the label-specific multi-semantics metric exploitation strategy in learning distance metrics for multi-label classification.

### 4.3 Further Analysis

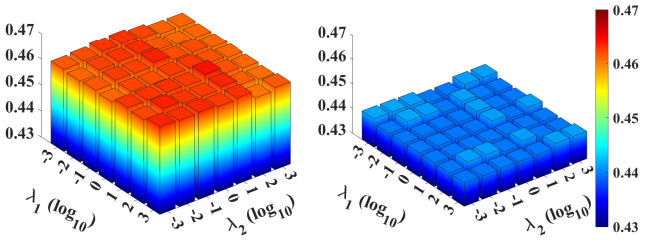
**Ablation study.** In this subsection, the ablation study on several variants of LIMIC is further conducted to analyze the effectiveness of each constituent part. To validate the usefulness of global metric consideration, LIMIC-NG is implemented by learning label-specific local metrics directly. In addition, we implement a variant named LIMIC-NL by removing the label correlation regularization in Eq.(8). Furthermore, LIMIC-NGL is implemented by ignoring both the global metric and label correlation regularization. Table 3 reports detailed experimental results coupled with BR-KNN in terms of *Average precision*. Pairwise $t$-test results can be found in the supplementary material. Based on these results, the significant usefulness of global metric consideration and label-correlation exploitation procedures can be validated.

**Sensitivity analysis.** As shown in Algorithm 1, $\lambda_1$ and $\lambda_2$ serve as hyperparameters for LIMIC which indicate the relative importance of the semantic metric and label correlation exploitation parts. To investigate the performance sensitivity of LIMIC approach, Figure 1 gives an illustrative example of how the performance of BR-KNN-LIMIC changes with varying configurations of the hyperparameters $\lambda_1$ and $\lambda_2$ (data set: CAL500 and birds; evaluation metric: *Average precision*). As shown in Figure 1, the performance of BR-KNN-LIMIC is relatively stable as $\lambda_1$ and $\lambda_2$ change within a reasonable range, which demonstrates the stability of our approach.

**Complexity analysis.** Let $k$, $k_t$, $k_i$, and $t$ denote the number of selected neighbors, number of targets, number of imposters, and dimension of projection. The training complexity of one iteration for LM, LJE, COMMU, and LIMIC are $\mathcal{O}(q^3 + kndq^2)$, $\mathcal{O}(tn^2 + tdnlog(n))$, $\mathcal{O}((d+q^2)nk_tk_i)$, and $\mathcal{O}(q(dn^2 + nd^2 + d^3))$. For LIMIC, the main computation lies in the gradient calculation and update for the global metric and label-specific local metrics. It is noteworthy that LIMIC learns multiple metrics of which the number equals $q+1$, which may be slow when applied to date sets with a large number of labels. This is inevitable if considering label-specific metrics. We will leave efficiency improvement for future work.

| Data sets | *Average precision* ↑ | | | |
|---|---|---|---|---|
| | LIMIC | LIMIC-NG | LIMIC-NL | LIMIC-NGL |
| CAL500 | **0.464±0.010** | **0.464±0.010** | 0.462±0.011• | 0.454±0.014• |
| emotions | **0.800±0.040** | 0.753±0.040• | 0.792±0.029• | 0.748±0.031• |
| birds | **0.441±0.065** | 0.436±0.062• | 0.441±0.066• | 0.436±0.062• |
| genbase | **0.996±0.005** | 0.995±0.005• | **0.996±0.005** | 0.995±0.005• |
| medical | **0.881±0.024** | 0.871±0.031• | 0.873±0.020• | 0.868±0.026• |
| image | **0.807±0.024** | 0.782±0.021• | 0.793±0.025• | 0.718±0.024• |
| scene | **0.857±0.019** | 0.817±0.013• | 0.856±0.017• | 0.776±0.018• |
| yeast | **0.767±0.022** | 0.760±0.018• | **0.767±0.023** | 0.760±0.022• |

Table 3: Predictive performance of LIMIC and its variants coupled with BR-KNN (mean±std) in terms of *Average precision*. In addition, •/◦ indicates whether BR-KNN-LIMIC achieves significantly superior/inferior to the variants on each data set in terms of *Average precision* (pairwise t-test at 0.05 significance level).




(a) CAL500 (b) birds

Figure 1: Performance of BR-KNN-LIMIC with varying parameter configurations in terms of *Average precision*.

## 5 Conclusion

In this paper, the first attempt towards multi-semantics metric learning for multi-label classification is investigated. Different from existing single-metric exploitation strategies, we propose a novel approach LIMIC which exploits multiple latent label-specific semantics for multi-label classification. LIMIC learns one global and multiple label-specific local metrics simultaneously to characterize label-specific semantic space, in which similar intra-class instances are closer while inter-class distances are far away. Comprehensive experiments demonstrate that LIMIC outperforms other well-established multi-label metric learning approaches in learning effective distance metrics for multi-label classification. However, LIMIC learns #labels+1 metrics, which could be hard to generalize to extreme multi-label learning. It is inevitable if considering label-specific metrics. It is interesting to investigate towards this dilemma to achieve better performance and tolerable scalability for multi-label metric learning. Furthermore, it is promising to extend our approach to weakly supervised and open-environment scenarios [Zhou, 2022].

## References


[Bellet *et al.*, 2015] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[Boyd *et al.*, 2004] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Canuto *et al.*, 2016] Sérgio Canuto, Marcos André Gonçalves, and Fabrício Benevenuto. Exploiting new sentiment-based meta-level features for effective sentiment analysis. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 53–62, San Francisco, CA, 2016.

[Chen *et al.*, 2017] Di Chen, Yexiang Xue, Daniel Fink, Shuo Chen, and Carla P. Gomes. Deep multi-species embedding. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3639–3646, Melbourne, Australia, 2017.

[Chen *et al.*, 2019a] Chen Chen, Haobo Wang, Weiwei Liu, Xingyuan Zhao, Tianlei Hu, and Gang Chen. Two-stage label embedding via neural factorization machine for multi-label classification. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 3304–3311, Honolulu, HI, 2019.

[Chen *et al.*, 2019b] Shuo Chen, Lei Luo, Jian Yang, Chen Gong, Jun Li, and Heng Huang. Curvilinear distance metric learning. In *Advances in Neural Information Processing Systems*, pages 4225–4234, Vancouver, Canada, 2019.

[Davis *et al.*, 2007] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, New York, NY, 2007.

[de Vazelhes *et al.*, 2020] William de Vazelhes, CJ Carey, Yuan Tang, Nathalie Vauquier, and Aurélien Bellet. Metric-learn: Metric learning algorithms in python. *Journal of Machine Learning Research*, 21(138):1–6, 2020.

[Feng *et al.*, 2019] Lei Feng, Bo An, and Shuo He. Collaboration based multi-label learning. In *Proceedings of the 19th AAAI Conference on Artificial Intelligence*, pages 3550–3557, Honolulu, HI, 2019.

[Fürnkranz *et al.*, 2008] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[Goldberger *et al.*, 2004] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, pages 513–520, Vancouver, Canada, 2004.

[Gopal and Yang, 2010] Siddharth Gopal and Yiming Yang. Multilabel classification with meta-level features. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–322, Geneva, Switzerland, 2010.

[Gouk *et al.*, 2016] Henry Gouk, Bernhard Pfahringer, and Michael Cree. Learning distance metrics for multi-label classification. In *Proceedings of the 8th Asian Conference on Machine Learning*, pages 318–333, Hamilton, New Zealand, 2016.

[Hang and Zhang, 2021] Jun-Yi Hang and Min-Ling Zhang. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9860–9871, 2021.

[Hang *et al.*, 2022] Jun-Yi Hang, Min-Ling Zhang, Yanghe Feng, and Xiaocheng Song. End-to-end probabilistic label-specific feature learning for multi-label classification. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 6847–6855, Online, 2022.

[Kulis and others, 2013] Brian Kulis et al. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[Kurata *et al.*, 2016] Gakuto Kurata, Bing Xiang, and Bowen Zhou. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–526, San Diego, California, 2016.

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

[Liu *et al.*, 2021] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W. Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):1–19, 2021.

[Nesterov, 2003] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003.

[Niu *et al.*, 2014] Gang Niu, Bo Dai, Makoto Yamada, and Masashi Sugiyama. Information-theoretic semi-supervised metric learning via entropy regularization. *Neural computation*, 26(8):1717–1762, 2014.

[Parameswaran and Weinberger, 2010] Shibin Parameswaran and Kilian Q Weinberger. Large margin multi-task metric learning. In *Advances in Neural Information Processing Systems*, pages 1867–1875, Vancouver, Canada, 2010.

[Pereira *et al.*, 2018] Rafael B Pereira, Alexandre Plastino, Bianca Zadrozny, and Luiz HC Merschmann. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review*, 49(1):57–78, 2018.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 815–823, Boston, MA, 2015.

[Shental *et al.*, 2002] Noam Shental, Tomer Hertz, Daphna Weinshall, and Misha Pavel. Adjustment learning and relevant component analysis. In *Proceedings of the 7th European Conference on Computer Vision*, pages 776–790, Copenhagen, Denmark, 2002.

[Siblini *et al.*, 2021] Wissam Siblini, Pascale Kuntz, and Frank Meyer. A review on dimensionality reduction for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 33(3):839–857, 2021.

[Suárez *et al.*, 2020] Juan Luis Suárez, Salvador García, and Francisco Herrera. Pydml: A python library for distance metric learning. *Journal of Machine Learning Research*, 21(96):1–7, 2020.

[Sun and Zhang, 2021] Yan-Ping Sun and Min-Ling Zhang. Compositional metric learning for multi-label classification. *Frontiers of Computer Science*, 15(5):1–12, 2021.

[Tang *et al.*, 2020] Pingjie Tang, Meng Jiang, Bryan (Ning) Xia, Jed W. Pitera, Jeffrey Welser, and Nitesh V. Chawla. Multi-label patent categorization with non-local attention-based graph convolutional network. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 9024–9031, New York, NY, 2020.

[Tsoumakas *et al.*, 2010] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089, 2010.

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

[Weinberger *et al.*, 2005] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, page 1473–1480, Vancouver, Canada, 2005.

[Xing *et al.*, 2002] Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, page 521–528, Cambridge, MA, 2002.

[Xing *et al.*, 2018] Yuying Xing, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Multi-label co-training. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2882–2888, Stockholm, Sweden, 2018.

[Ye *et al.*, 2019] Han-Jia Ye, De-Chuan Zhan, Xue-Min Si, Yuan Jiang, and Zhi-Hua Zhou. What makes objects similar: A unified multi-metric learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 41(5):1257–1270, 2019.

[Ye *et al.*, 2020] Han-Jia Ye, De-Chuan Zhan, Nan Li, and Yuan Jiang. Learning multiple local metrics: Global consideration helps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1698–1712, 2020.

[Yeh *et al.*, 2017] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent space for multi-label classification. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, pages 2838–2844, San Francisco, CA, 2017.

[You *et al.*, 2020] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 12709–12716, New York, NY, 2020.

[Zhang and Wu, 2014] Min-Ling Zhang and Lei Wu. Lift: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):107–120, 2014.

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

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

[Zhang *et al.*, 2021] Min-Ling Zhang, Qian-Wen Zhang, Jun-Peng Fang, Yu-Kun Li, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):2057–2070, 2021.

[Zhou, 2022] Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.

[Zhu *et al.*, 2017] Yue Zhu, James T Kwok, and Zhi-Hua Zhou. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2017.