Learning Label-Specific Multiple Local Metrics for Multi-Label Classification

Jun-Xiang Mao[1,2] , **Jun-Yi Hang**[1,2] , **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
{maojx, hangjy, zhangml}@seu.edu.cn

Abstract

Multi-label metric learning serve as an effective strategy to facilitate multi-label classification, aiming to learn better similarity metrics from multi-label examples. Existing multi-label metric learning approaches learn consistent metrics across all multi-label instances in the label space. However, such consistent metric learning approaches are suboptimal as they neglect the nonlinear distribution characteristics of multi-label instances. In this paper, we present LSMM, *a **L**abel-**S**pecific **M**ulti-**M**etric learning framework for multi-label classification*, where nonlinear distribution characteristics of multi-label examples are considered by learning label-specific multiple local metrics for different instances on the shoulder of a global one. Specifically, multi-label instances within each label space can be divided into several disjoint partitions through either semantic-based or cluster-based partition strategies, in each of which a local metric is trained to separate the instances locally. Besides, a global metric is introduced to implicitly exploit high-order label correlations across all labels. The combination of the global metric and label-specific local metrics is utilized to measure the semantic similarities between multi-label instances in each label space, under which similar intra-class instances are pushed closer and inter-class instances are pulled apart. Comprehensive experiments on benchmark multi-label datasets validate the superiority of LSMM in learning effective similarity metrics for multi-label classification.

1 Introduction

Different from multi-class classification[Jia *et al.*, 2023], multi-label classification aims to model real-world objects with rich semantics, where each multi-label example is represented by an instance while associated with multiple labels simultaneously [Zhang and Zhou, 2014; Liu *et al.*, 2021]. As a practical machine learning paradigm, multi-label classification has been widely applied in various real-world application, such as text categorization [Xu *et al.*, 2023], bioinformatics analysis [Li *et al.*, 2023], etc.

Similarity is a crucial concept in machine learning that reveals the degree of connectedness between objects. The predefined similarity metric, such as the Euclidean distance, face challenges in universal adaptation across diverse scenarios. In response, Distance metric learning [Xing *et al.*, 2002] has emerged as a solution to learn task-specific distance metrics by leveraging side information such as linkages and comparisons derived from examples. This approach autonomously refines similarity measurements beyond the limitations of the predefined metric. The learned distance metrics align with the inherent relationships between examples, ensuring that similar intra-class instances exhibit proximity, while distances between dissimilar inter-class instances are sufficiently large. Through the utilization of adaptively learned distance metrics, the efficacy of distance metric learning has been substantiated in enhancing k-nearest neighbor (KNN) classifiers on single-label examples[Song *et al.*, 2021; Li and Lu, 2022; Chen *et al.*, 2023]. Given the effective modeling of semantic similarity among examples achieved through distance metric learning, there exist the potential for simple KNN classifiers to achieve state-of-the-art classification performance.

In light of the remarkable capability for measuring semantic similarities, distance metric learning has been extended to multi-label scenarios in recent years, a.k.a. *multi-label metric learning* [Liu and Tsang, 2015; Gouk *et al.*, 2016; Sun and Zhang, 2021; Mao *et al.*, 2023]. Multi-label metric learning aims to assess the more intricate semantic similarities among examples with rich semantics. Experimental evidence has demonstrated its capability to enhance the performance of KNN-based multi-label classifiers, such as BR-KNN [Boutell *et al.*, 2004] and ML-KNN [Zhang and Zhou, 2007], thereby facilitating multi-label classification. Existing multi-label metric learning approaches learn consistent metrics across all multi-label instances in the label space. Although such consistent multi-label metric learning framework has achieved favorable outcomes, it might be suboptimal as they neglect the fact that multi-label instances exhibit nonlinear distribution characteristics in each label space. For example, in multimedia annotation, the label "sun" is applicable to images depicting forests as well as those representing urban environments. Likewise, in text categorization, the label "president" manifests in both political and economic textual

[*] Corresponding author.

contexts. This implies that in each label space, the distribution of multi-label instances exhibits a high degree of nonlinearity. Consequently, in such cases, employing a single consistent metric encounters formidable challenges in simultaneously pushing similar instances closer while pulling dissimilar instances further away.

With the above observations, we present LSMM, *a **L**abel-**S**pecific **M**ulti-**M**etric learning framework for multi-label classification*, where nonlinear distribution characteristics of multi-label instances are considered by learning label-specific multiple local metrics for different instances on the shoulder of a global one. Specifically, multi-label instances within each label space can be divided into several disjoint partitions through either semantic-based or cluster-based partition strategies, in each of which a local metric is trained to separate the instances locally. To take label correlation into consideration, a global metric is introduced to implicitly exploit high-order label correlations across all labels. The global metric serves as the foundation to reveal the common characteristics among multi-label examples, while the label-specific local metrics act as local biases, depicting the individuality of different instances in each label space. The combination of the global metric and label-specific local metrics is utilized to measure the semantic similarities between multi-label instances in each label space, under which similar intra-class instances are pushed closer and inter-class instances are pulled apart. Comprehensive experiments on benchmark multi-label datasets validate the superiority of LSMM in learning effective similarity 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 LSMM framework. Section 4 reports experimental results of comparative studies over benchmark multi-label datasets. Section 5 concludes this paper.

2 Related Work

Multi-Label Classification. In the last decades, numerous approaches have been proposed to tackle multi-label classification problem [Zhang and Zhou, 2014; Liu *et al.*, 2021]. To deal with 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 *et al.*, 2017; Yu and Zhang, 2021], and a high-order manner by exploiting relations among a subset or all labels [Zhang *et al.*, 2021; Si *et al.*, 2023]. BR-KNN [Boutell *et al.*, 2004] and ML-KNN [Zhang and Zhou, 2007], as the most classic first-order multi-label classification methods, extend KNN to the multi-label scenario and have achieved certain results in multi-label classification tasks. However, due to the neglect of label correlations, their performance is usually inferior to second-order and high-order methods. In addition, another significant reason for their insufficient performance is their heavy reliance on the chosen similarity metric. In the absence of prior knowledge, the predefined Euclidean metric used in BR-KNN and ML-KNN is often unsuitable for all scenarios.

Distance Metric Learning. To address the limitations imposed by predefined metrics, distance metric learning is proposed to learn task-specific similarity metrics, mainly for single-label examples [Xing *et al.*, 2002; Weinberger *et al.*, 2005; Schroff *et al.*, 2015]. By leveraging various types of supervision extracted from examples, such as linkages and comparisons, distance metric learning aims to align the learned distance metrics with the intrinsic relationships between examples, i.e. similar instances are close to each other and dissimilar instances are far away from each other. In distance 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 [Ye *et al.*, 2019; Ye *et al.*, 2020; Ren *et al.*, 2022; Zhao and Yang, 2023]. The Mahalanobis distance between instances is essentially equivalent to the Euclidean distance within the metric space. The superiority of distance metric learning has been substantiated in improving KNN classifiers on single-label examples [Liao *et al.*, 2021; Song *et al.*, 2021; Li *et al.*, 2022; Chen *et al.*, 2023]. With the effective modeling of semantic similarity among examples accomplished by distance metric learning, there is the potential for simple KNN classifiers to achieve state-of-the-art classification performance.

Multi-Label Metric Learning. In recent years, distance metric learning has been extended to the multi-label scenario to measure the more complicated semantic similarities between multi-label examples. To the best of our knowledge, there are four available multi-label metric learning approaches, namely LM [Liu and Tsang, 2015], LJE [Gouk *et al.*, 2016], COMMU [Sun and Zhang, 2021], and LIMIC [Mao *et al.*, 2023]. LM utilizes a large margin formulation to create a shared metric space, preserving the similarity relationships from the feature space to the label space. LJE learns a metric that projects instances into a metric space, where the Euclidean distance is a good approximation of the Jaccard distance between labels. COMMU constructs a compositional metric by modeling structural interactions between the feature and label spaces, exploring the integrated semantics of all labels. LIMIC learns a consistent metric in each label space and employs label co-occurrence to constrain the dependencies between multiple metrics. The multi-label metric learning approaches above all learn consistent metrics across all instances in the label space, which might be suboptimal as they neglect the fact that multi-label instances exhibit nonlinear distribution characteristics in each label space. In the next section, our proposed LSMM framework is introduced.

3 The LSMM Framework

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 indicator vector $\boldsymbol{y} = [y^1, y^2, \ldots, y^q] \in \{0, 1\}^q$

is 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 $h(\boldsymbol{x'}) \subseteq \mathcal{Y}$.

Let $\mathbb{S}_+^d$ denotes the cone of Positive Semi-Definite (PSD) $d \times d$ matrices. Given a metric $\mathbf{M} \in \mathbb{S}_+^d$, the (squared) Mahalanobis distance between a 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 $\mathbf{A}_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \in \mathbb{S}_+^d$ [Bellet *et al.*, 2015]. 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, i.e.

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

Generally speaking, there are several advantages to learning transformation $\mathbf{L}$ rather than metric $\mathbf{M}$. On one hand, transformation $\mathbf{L}$ does not have a PSD constraint, eliminating the need for a PSD projection step, in turn, accelerates the optimization process [Kulis, 2013]. In addition, the transformation decomposition often leads to low-rank metrics, proving advantageous in many real-world applications [Zhang and Zhang, 2017]. It is also noteworthy that although the decomposition leads to non-convex problems, satisfactory solutions can still be obtained. Based on the above consideration, we learn transformation $\mathbf{L}$ rather than metric $\mathbf{M}$ in this paper.

3.2 Label-Specific Multi-Metric Learning

To begin with, we introduce the construction procedure of label-specific side information, which is used to direct the learning process of metrics.

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

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

The label-specific side information w.r.t the label l_p is constructed from all possible pairwise combinations of training examples, denoted as $\mathcal{T}_p = \{(\boldsymbol{x}_i, \boldsymbol{x}_j, \theta^p_{ij})\}$, where $\theta^p_{ij} \in \{-1, 1\}$ indicates whether $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ share relevance w.r.t the label l_p. Specifically, $\theta^p_{ij} = 1$ implies that instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are considered similar due to their simultaneous presence (or absence) of the label l_p, i.e. $(\boldsymbol{x}_i \in \mathcal{P}_p \wedge \boldsymbol{x}_j \in \mathcal{P}_p) \vee (\boldsymbol{x}_i \in \mathcal{N}_p \wedge \boldsymbol{x}_j \in \mathcal{N}_p)$, while θ^p_{ij} equals -1 otherwise. Let T_p represent the number of pairs in $\mathcal{T}_p$, and $(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p$ denote the enumeration of a total of T_p pairs from the label-specific side information $\mathcal{T}_p$. In practice, there is no need to compute all pairs of side information, as this may lead to substantial computational burden. Opting for a judicious selection of targets (denoted as k_t) and imposters (denoted as k_i) among the nearest neighbors can mitigate computational load and enhance the efficiency of the training procedure, where targets indicate instances similar to an anchor and imposters otherwise. [Weinberger and Saul, 2009; Ye *et al.*, 2019; Ye *et al.*, 2020]. In this paper, k_t and k_i are set to 20.

Next, we present two simple yet effective partition strategies, i.e. semantic-based partition and cluster-based partition, designed to divide each label space into several disjoint partitions. In each partition, a local metric is trained to separate the instances locally. Furthermore, a global metric is introduced to implicitly exploit high-order label correlations across all labels. In this way, different instances can be measured with the combination of a global metric and corresponding label-specific local metrics, as opposed to a consistent metric.

Semantic-based Partition. The first strategy is to assign different local metrics to positive and negative examples in each label space. For the p-th label l_p, the (squared) distance between a pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is calculated as

$$\mathrm{Dis}^2_{\mathbf{L}_0 + \mathbf{L}_p^{s(\boldsymbol{x}_i)}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||(\mathbf{L}_0 + \mathbf{L}_p^{s(\boldsymbol{x}_i)})^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)||_2^2. \tag{3}$$

Here, $s(\boldsymbol{x}_i) = y_i^p \in \{0, 1\}$ is a semantic indicator function. $\mathbf{L}_0$ serves as the foundation to reveal the common characteristics among multi-label examples, while the label-specific local metric $\mathbf{L}_p^{s(\boldsymbol{x}_i)}$ act as a local bias, depicting the individuality of $\boldsymbol{x}_i$ in the p-th label space. Our purpose is to push similar instances closer, while dissimilar instances far away from each other. Therefore, we formulate the following optimization problem:

$$\ell^p_{\mathrm{pos}} = \frac{1}{|\mathcal{P}_p|} \sum_{\substack{(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p \\ \boldsymbol{x}_i \in \mathcal{P}_p}} \ell\left(\theta^p_{ij}\left(\gamma - \alpha \mathrm{Dis}^2_{\mathbf{L}_0 + \mathbf{L}^1_p}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)\right)$$

$$\ell^p_{\mathrm{neg}} = \frac{1}{|\mathcal{N}_p|} \sum_{\substack{(\boldsymbol{x}_i, \boldsymbol{x}_j) \sim \mathcal{T}_p \\ \boldsymbol{x}_i \in \mathcal{N}_p}} \ell\left(\theta^p_{ij}\left(\gamma - \alpha \mathrm{Dis}^2_{\mathbf{L}_0 + \mathbf{L}^0_p}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)\right)$$

$$\begin{aligned} \min_{\mathbf{L}_0, \{\mathbf{L}^1_p, \mathbf{L}^0_p\}_{p=1}^q} \frac{1}{q} \sum_{p=1}^q \left(\ell^p_{\mathrm{pos}} + \ell^p_{\mathrm{neg}}\right) + \frac{\lambda_1}{2q} \sum_{p=1}^q \sum_{t=0}^1 ||\mathbf{L}^t_p||_F^2 \\ + \lambda_2 ||\mathbf{L}_0||_F^2. \end{aligned} \tag{4}$$

Here, γ and α are predefined non-negative threshold values, which can differ for similar and dissimilar pairs. $\ell(\cdot)$ is a convex and non-increasing loss function. If two instances $\boldsymbol{x}_i, \boldsymbol{x}_j$ possess (or do not possess) the label l_p simultaneously, i.e. $\theta^p_{ij} = 1$, then the loss equals 0 when their distance is smaller than γ/α. Conversely, when they are dissimilar ($\theta^p_{ij} = -1$), their distance should exceed γ/α. By optimizing Eq.(4), the learned metrics require that similar instances have small distances, while dissimilar instances are sufficiently far apart. Besides, λ_1 and λ_2 serve as non-negative weights to balance the impact of the regularization terms. Intuitively, for a fixed value of λ_2, a large value of the ratio λ_1/λ_2 tends to make the model learn a single global metric $\mathbf{L}_0$ across all labels ($\mathbf{L}^t_p$ are nearly equal to zero). Conversely, for a fixed λ_1, a small value of the ratio λ_1/λ_2 tends to make all label-specific local metrics unrelated ($\mathbf{L}_0$ nearly equal to zero). In this paper,

γ and α are fixed to 2 and 0.4 respectively, and the smooth hinge loss is used to instantiate $\ell(\cdot)$, which is defined as

$$\ell(x) = \begin{cases} 0, & \text{if} \quad x > 1 \\ \frac{1}{2}(x-1)^2, & \text{if} \quad 0 \leq x \leq 1 \\ \frac{1}{2} - x, & \text{if} \quad x < 0. \end{cases} \tag{5}$$

The smoothness property of this loss function will facilitate the optimization process. Besides, $\ell(\cdot)$ also keeps a small margin besides γ and α, which further improves the generalization ability of the learned metrics.

Cluster-based Partition. The second strategy is to assign different metrics to different clusters. In this paper, the popular k-means algorithm [Jain *et al.*, 1999] is adopted due to its effectiveness and simplicity. Assuming that all multi-label instances are divided into C disjoint clusters $\mathcal{D} = \mathcal{C}^1 \cup \mathcal{C}^2 \ldots \cup \mathcal{C}^C$, the (squared) distance between a pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in the p-th label space is calculated as

$$\mathrm{Dis}^2_{\mathbf{L}_0+\mathbf{L}_p^{c(\boldsymbol{x}_i)}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||(\mathbf{L}_0 + \mathbf{L}_p^{c(\boldsymbol{x}_i)})^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)||_2^2. \tag{6}$$

Here, $c(\boldsymbol{x}_i) \in \{0, 1, \ldots, C\}$ is a cluster indicator function. The optimization problem is formulated as follows:

$$\ell_m^p = \frac{1}{|\mathcal{C}^m|} \sum_{\substack{(\boldsymbol{x}_i,\boldsymbol{x}_j)\sim\mathcal{T}_p \\ \boldsymbol{x}_i\in\mathcal{C}^m}} \ell\left(\theta_{ij}^p\left(\gamma - \alpha\mathrm{Dis}^2_{\mathbf{L}_0+\mathbf{L}_p^m}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)\right)$$

$$\min_{\mathbf{L}_0,\left\{\mathbf{L}_p^1,\mathbf{L}_p^2,\ldots,\mathbf{L}_p^C\right\}_{p=1}^q} \frac{1}{q}\sum_{p=1}^q\sum_{m=1}^C \ell_m^p + \frac{\lambda_1}{Cq}\sum_{p=1}^q\sum_{t=1}^C ||\mathbf{L}_p^t||_F^2 + \lambda_2||\mathbf{L}_0||_F^2. \tag{7}$$

In this paper, C is set to 3. Besides, the similar instantiations and properties have been elaborated on in the previous paragraph. Therefore, we do not repeat them here.

3.3 Optimization Procedure

In order to solve the unconstrained nonlinear optimization problems as shown in Eq.(4) and Eq.(7), we employ the *Limited-memory Broyde-Fletcher-Golfarb-Shanno* (L-BFGS) algorithm [Liu and Nocedal, 1989], which is particularly suited for problems with a large number of optimization variables. Let $\mathcal{L}_{se}$ and $\mathcal{L}_{cl}$ denote the objective functions of the semantic-based and cluster-based partition strategies in Eq.(4) and Eq.(7) respectively. L-BFGS iteratively optimizes the objective functions with resort to the first-order derivatives of the functions.

For the semantic-based partition strategy, the first-order derivatives of $\mathcal{L}_{se}$ w.r.t $\mathbf{L}_0$ and $\left\{\mathbf{L}_p^1, \mathbf{L}_p^0\right\}_{p=1}^q$ respectively are

$$\frac{\partial\mathcal{L}_{se}}{\partial\mathbf{L}_0} = \frac{2}{q}\sum_{p=1}^q\left(\frac{1}{|\mathcal{P}_p|}\sum_{\substack{(\boldsymbol{x}_i,\boldsymbol{x}_j)\sim\mathcal{T}_p \\ \boldsymbol{x}_i\in\mathcal{P}_p}} \sigma_{ij}^p\theta_{ij}^p\mathbf{A}_{ij}\left(\mathbf{L}_0 + \mathbf{L}_p^1\right) + \frac{1}{|\mathcal{N}_p|}\sum_{\substack{(\boldsymbol{x}_i,\boldsymbol{x}_j)\sim\mathcal{T}_p \\ \boldsymbol{x}_i\in\mathcal{N}_p}} \sigma_{ij}^p\theta_{ij}^p\mathbf{A}_{ij}\left(\mathbf{L}_0 + \mathbf{L}_p^0\right)\right) + 2\lambda_2\mathbf{L}_0 \tag{8}$$

$$\frac{\partial\mathcal{L}_{se}}{\partial\mathbf{L}_p^1} = \frac{2}{q|\mathcal{P}_p|}\sum_{\substack{(\boldsymbol{x}_i,\boldsymbol{x}_j)\sim\mathcal{T}_p \\ \boldsymbol{x}_i\in\mathcal{P}_p}} \sigma_{ij}^p\theta_{ij}^p\mathbf{A}_{ij}\left(\mathbf{L}_0 + \mathbf{L}_p^1\right) + \frac{\lambda_1}{q}\mathbf{L}_p^1 \tag{9}$$

$$\frac{\partial\mathcal{L}_{se}}{\partial\mathbf{L}_p^0} = \frac{2}{q|\mathcal{N}_p|}\sum_{\substack{(\boldsymbol{x}_i,\boldsymbol{x}_j)\sim\mathcal{T}_p \\ \boldsymbol{x}_i\in\mathcal{N}_p}} \sigma_{ij}^p\theta_{ij}^p\mathbf{A}_{ij}\left(\mathbf{L}_0 + \mathbf{L}_p^0\right) + \frac{\lambda_1}{q}\mathbf{L}_p^0 \tag{10}$$

where σ_{ij}^p is a piecewise function defined as follows:

$$\sigma_{ij}^p = \begin{cases} 0, & \text{if} \quad \delta_{ij}^p > 1 \\ \alpha\left(1 - \delta_{ij}^p\right), & \text{if} \quad 0 \leq \delta_{ij}^p \leq 1 \\ \alpha, & \text{if} \quad \delta_{ij}^p < 0. \end{cases} \tag{11}$$

$$\delta_{ij}^p = \theta_{ij}^p\left(\gamma - \alpha\mathrm{Dis}^2_{\mathbf{L}_0+\mathbf{L}_p^{s(\boldsymbol{x}_i)}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right) \tag{12}$$

For the cluster-based partition strategy, the first-order derivatives of $\mathcal{L}_{cl}$ w.r.t $\mathbf{L}_0$ and $\left\{\mathbf{L}_p^1, \mathbf{L}_p^2, \ldots, \mathbf{L}_p^C\right\}_{p=1}^q$ respectively are

$$\frac{\partial\mathcal{L}_{cl}}{\partial\mathbf{L}_0} = \frac{2}{q}\sum_{p=1}^q\sum_{m=1}^C\frac{1}{|\mathcal{C}^m|}\sum_{\substack{(\boldsymbol{x}_i,\boldsymbol{x}_j)\sim\mathcal{T}_p \\ \boldsymbol{x}_i\in\mathcal{C}^m}} \zeta_{ij}^p\theta_{ij}^p\mathbf{A}_{ij}\left(\mathbf{L}_0 + \mathbf{L}_p^m\right) + 2\lambda_2\mathbf{L}_0 \tag{13}$$

$$\frac{\partial\mathcal{L}_{cl}}{\partial\mathbf{L}_p^m} = \frac{2}{q|\mathcal{C}^m|}\sum_{\substack{(\boldsymbol{x}_i,\boldsymbol{x}_j)\sim\mathcal{T}_p \\ \boldsymbol{x}_i\in\mathcal{C}^m}} \zeta_{ij}^p\theta_{ij}^p\mathbf{A}_{ij}\left(\mathbf{L}_0 + \mathbf{L}_p^m\right) + \frac{2\lambda_1}{Cq}\mathbf{L}_p^m \tag{14}$$

where ζ_{ij}^p is a piecewise function defined as follows:

$$\zeta_{ij}^p = \begin{cases} 0, & \text{if} \quad \xi_{ij}^p > 1 \\ \alpha\left(1 - \xi_{ij}^p\right), & \text{if} \quad 0 \leq \xi_{ij}^p \leq 1 \\ \alpha, & \text{if} \quad \xi_{ij}^p < 0. \end{cases} \tag{15}$$

$$\xi_{ij}^p = \theta_{ij}^p\left(\gamma - \alpha\mathrm{Dis}^2_{\mathbf{L}_0+\mathbf{L}_p^{c(\boldsymbol{x}_i)}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right) \tag{16}$$

The complete procedure of LSMM can be found in Appendix A. After learning the global metric and label-specific multiple local metrics with the above LSMM framework, the distances between multi-label instances can be measured with the combination of the global metric and corresponding label-specific local metrics according to Eq.(3) and Eq.(6). Therefore, the associated label set of an unseen multi-label instance can be easily predicted by resorting to KNN-based multi-label classifiers, such as BR-KNN and ML-KNN.

Dataset	$\lvert\mathcal{D}\rvert$	$dim(\mathcal{D})$	$L(\mathcal{D})$	$LCard(\mathcal{D})$	Domain
emotions	593	72	6	1.869	Music[1]
birds	645	258	19	1.014	Audio[1]
medical	978	1449	45	1.245	Text[1]
enron	1702	1001	53	3.378	Text[1]
image	2000	294	5	1.236	Image[2]
scene	2407	294	6	1.074	Image[1]
slashdot	3782	1079	22	1.177	Text[3]
arts	5000	462	26	1.636	Text[1]
education	5000	550	33	1.461	Text[1]

[1] http://mulan.sourceforge.net/datasets.html
[2] http://palm.seu.edu.cn/zhangml/Resources.htm#data
[3] https://waikato.github.io/meka/datasets/

Table 1: Characteristics of experimental datasets.

4 Experiments

4.1 Experimental Setup

Datasets. Nine benchmark multi-label datasets with diversified properties are employed for comprehensive performance evaluation. Table 1 summarizes the characteristics of each experimental 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. For all the datasets except emotions and birds, we use principal component analysis as a preprocessing to reduce the dimensionality and retain 95% of the information.

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 LSMM framework in learning effective similarity metrics for multi-label classification, BR-KNN [Boutell *et al.*, 2004] and ML-KNN [Zhang and Zhou, 2007] are introduced as subsequent multi-label classification methods after learning similarity metrics. Our semantic-based and cluster-based partition approaches are denoted as LSMM-SE and LSMM-CL respectively.

Given a KNN-based multi-label classification approach $\mathcal{A} \in \{\text{BR-KNN}, \text{ML-KNN}\}$ 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 $\mathcal{A}$-LSMM-SE and $\mathcal{A}$-LSMM-CL are compared against other state-of-the-art multi-label metric learning algorithms coupled with $\mathcal{A}$ to manifest whether the proposed LSMM framework does learn effective similarity metrics and improve the generalization performance of KNN-based multi-label classification.

In this paper, four state-of-the-art multi-label metric learning algorithms are employed to instantiate $\mathcal{B}$ with suggested parameter configurations in respective literature:

- LM [Liu and Tsang, 2015]: A margin-based multi-label metric learning approach that learns a shared metric by preserving the similarity relation from the feature to label space [suggested configuration: η = 0.4, C = 10].
- LJE [Gouk *et al.*, 2016]: An integration-based multi-label metric learning approach that employs Jaccard distance between labels to provide more fine-grained side information [suggested 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 the feature and label space [suggested configuration: $\alpha, \theta \in \{0.2, 0.4, \ldots, 0.8\}$, C = 10].
- LIMIC [Mao *et al.*, 2023]: A decomposition-based multi-label metric learning approach that learns a consistent metric in each label space and employs label co-occurrence to constrain the dependencies between multiple metrics [suggested configuration: γ = 2, $\lambda_1, \lambda_2 \in \{10^{-3}, 10^{-1}, \ldots, 10^{3}\}$].

For the proposed LSMM-SE and LSMM-CL approaches, regularization parameters λ_1 and λ_2 are searched in $\{10^{-1}, 1, \ldots, 10^{3}\}$ and $\{10^{-3}, 10^{-2}, \ldots, 10\}$ respectively. The number of nearest neighbors (denoted as k) in KNN and ML-KNN is set to 10, which is consistent with previous multi-label metric learning works for fair comparisons. We take out 10% examples in each dataset as a hold out validation set for hyperparameter searching and perform ten-fold cross-validation on the remaining 90% examples to evaluate the above approaches on the nine benchmark multi-label datasets.

Due to page limit, Table 2 reports detailed experimental results in terms of *Hamming loss* and *Ranking loss*. The results on other metrics can be found in Appendix B.1. Furthermore, pairwise t-test [Dietterich, 1998] at 0.05 significance level is conducted to demonstrate whether the performance difference between LSMM-SE (LSMM-CL) and other comparing multi-label metric learning algorithms coupled with $\mathcal{A} \in \{\text{BR-KNN}, \text{ML-KNN}\}$ is significant statistically, where the resulting win/tie/loss counts are also reported in Appendix B.1. Based on these results, it is impressive to observe that:

- The prediction performance of BR-KNN and ML-KNN is significantly improved after coupling multi-label metric learning algorithms. In particular, our proposed LSMM-SE (LSMM-CL) approach significantly enhances the performance of BR-KNN and ML-KNN in 88.9% (92.6%) and 94.4% (92.6%) of cases, respectively. The results clearly demonstrate the effectiveness of multi-label metric learning for improving KNN-based multi-label classification methods.
- Across all evaluation metrics, our proposed LSMM-SE and LSMM-CL approaches have demonstrated significantly improved performance compared to other multi-label metric learning algorithms. One one hand, in terms of BR-KNN, LSMM-SE (LSMM-CL) significantly outperforms LM, LJE, COMMU, and LIMIC in 75.9% (87%), 94.4% (100%), 87% (88.9%), and 77.8% (85.2%) of cases, respectively. On the other hand, in terms of ML-KNN, LSMM-SE (LSMM-CL) significantly outperforms them in 81.5% (85.2%), 100% (100%), 94.4% (94.4%), and 72.2% (75.9%) of cases, respectively. The superior

Compared Algorithms	Datasets								
	emotions	birds	medical	enron	image	scene	slashdot	arts	education
					Hamming Loss ↓				
BR-KNN	0.263±0.023	0.056±0.007	0.016±0.002	0.055±0.002	0.167±0.016	0.088±0.008	0.065±0.001	0.073±0.002	0.038±0.001
BR-KNN-LM	0.270±0.019	0.065±0.009	**0.011±0.002**	0.048±0.003	0.180±0.016	0.088±0.012	0.044±0.003	0.055±0.002	0.038±0.001
BR-KNN-LJE	0.219±0.022	0.055±0.007	0.022±0.003	0.059±0.003	0.184±0.018	0.110±0.011	0.060±0.002	0.061±0.001	0.043±0.002
BR-KNN-COMMU	0.263±0.023	0.056±0.007	0.016±0.002	0.055±0.002	0.167±0.016	0.088±0.008	0.056±0.001	0.072±0.002	0.038±0.001
BR-KNN-LIMIC	0.212±0.008	0.053±0.006	0.014±0.002	0.049±0.003	0.165±0.016	0.083±0.008	0.045±0.002	0.058±0.002	0.039±0.001
BR-KNN-LSMM-SE	0.204±0.018	0.051±0.006	0.014±0.003	0.045±0.003	0.162±0.013	0.081±0.007	**0.035±0.002**	0.054±0.001	0.037±0.002
BR-KNN-LSMM-CL	**0.202±0.017**	**0.050±0.007**	0.012±0.002	**0.044±0.003**	**0.157±0.017**	**0.079±0.009**	0.037±0.002	**0.052±0.002**	0.038±0.001
ML-KNN	0.262±0.022	0.054±0.006	0.016±0.002	0.052±0.003	0.171±0.013	0.084±0.009	0.058±0.001	0.060±0.001	0.038±0.001
ML-KNN-LM	0.254±0.017	0.054±0.007	0.013±0.002	0.048±0.002	0.174±0.015	0.086±0.010	0.045±0.003	0.054±0.001	0.038±0.001
ML-KNN-LJE	0.227±0.022	0.054±0.007	0.023±0.003	0.058±0.002	0.184±0.017	0.109±0.009	0.056±0.001	0.060±0.001	0.042±0.002
ML-KNN-COMMU	0.262±0.022	0.054±0.006	0.015±0.002	0.052±0.003	0.171±0.013	0.084±0.009	0.050±0.001	0.059±0.001	0.038±0.001
ML-KNN-LIMIC	0.215±0.009	0.053±0.006	0.013±0.002	0.050±0.003	0.164±0.017	0.082±0.007	0.050±0.003	0.057±0.002	0.038±0.002
ML-KNN-LSMM-SE	**0.204±0.016**	**0.050±0.005**	**0.012±0.003**	**0.043±0.002**	0.159±0.014	0.080±0.007	0.041±0.002	0.053±0.001	**0.037±0.001**
ML-KNN-LSMM-CL	0.209±0.018	0.051±0.006	**0.012±0.002**	0.045±0.003	**0.155±0.013**	**0.078±0.007**	**0.039±0.003**	**0.052±0.001**	0.038±0.001
					Ranking Loss ↓				
BR-KNN	0.272±0.048	0.516±0.064	0.081±0.028	0.228±0.023	0.180±0.020	0.095±0.014	0.417±0.026	0.354±0.029	0.244±0.013
BR-KNN-LM	0.256±0.030	0.527±0.066	0.105±0.033	0.201±0.011	0.191±0.020	0.097±0.016	0.221±0.021	**0.257±0.015**	0.222±0.014
BR-KNN-LJE	0.202±0.035	0.497±0.067	0.114±0.037	0.244±0.027	0.205±0.020	0.124±0.019	0.345±0.024	0.265±0.015	0.234±0.012
BR-KNN-COMMU	0.272±0.048	0.516±0.064	0.088±0.031	0.230±0.023	0.180±0.020	0.095±0.014	0.407±0.027	0.357±0.033	0.244±0.013
BR-KNN-LIMIC	0.182±0.028	0.513±0.066	0.077±0.023	0.245±0.020	0.171±0.024	0.102±0.011	0.393±0.023	0.348±0.021	0.220±0.015
BR-KNN-LSMM-SE	**0.170±0.033**	**0.472±0.061**	**0.065±0.018**	0.194±0.026	0.157±0.021	0.092±0.011	0.201±0.018	0.263±0.010	0.210±0.010
BR-KNN-LSMM-CL	0.178±0.034	0.487±0.058	0.071±0.023	**0.192±0.022**	**0.150±0.025**	**0.090±0.013**	**0.190±0.025**	0.258±0.027	**0.205±0.012**
ML-KNN	0.258±0.038	0.295±0.035	0.032±0.009	0.091±0.010	0.173±0.018	0.075±0.011	0.182±0.013	0.148±0.008	0.097±0.004
ML-KNN-LM	0.240±0.026	0.298±0.047	0.034±0.015	0.090±0.008	0.177±0.018	0.078±0.012	0.112±0.008	0.129±0.007	0.078±0.005
ML-KNN-LJE	0.201±0.030	0.299±0.047	0.053±0.017	0.111±0.011	0.201±0.022	0.110±0.017	0.198±0.013	0.145±0.009	0.089±0.003
ML-KNN-COMMU	0.258±0.038	0.295±0.035	0.038±0.013	0.091±0.010	0.173±0.018	0.075±0.011	0.159±0.012	0.147±0.008	0.097±0.004
ML-KNN-LIMIC	0.184±0.020	0.263±0.039	**0.031±0.012**	0.088±0.010	0.163±0.022	0.072±0.007	0.133±0.009	0.137±0.006	0.079±0.004
ML-KNN-LSMM-SE	**0.166±0.033**	**0.256±0.034**	0.034±0.015	0.075±0.007	0.157±0.021	0.072±0.004	0.095±0.008	0.123±0.006	0.078±0.004
ML-KNN-LSMM-CL	0.170±0.036	0.260±0.038	0.038±0.013	**0.073±0.009**	**0.146±0.020**	**0.069±0.005**	**0.083±0.007**	**0.120±0.005**	**0.074±0.004**

Table 2: Predictive performance (mean±std) of $\mathcal{A}$ ($\mathcal{A} \in \{\text{BR-KNN}, \text{ML-KNN}\}$) coupled with our proposed approaches and state-of-the-art multi-label metric learning approaches in terms of *Hamming Loss* and *Ranking Loss*. ↑ (↓) indicates the larger (smaller) the value, the better the performance. The best and second best results are highlighted in **boldface** and underline respectively.

performance provides persuasive evidence for our proposed LSMM framework in learning effective similarity metrics for multi-label classification.

In addition to comparing our proposed LSMM-SE and LSMM-CL with state-of-the-art multi-label metric learning algorithms, we further evaluate LSMM-enhanced BR-KNN and ML-KNN against four state-of-the-art non-metric learning multi-label classification approaches with different order of label correlations under consideration:

- LIFT [Zhang and Wu, 2014]: A first-order multi-label classification approach that exploits label-specific features [suggested configuration: $r = 0.1$].
- RELIAB [Zhang *et al.*, 2021]: A high-order multi-label classification approach that leverages implicit relative labeling-importance information [suggested configuration: mode = *global*, $\tau \in \{0.1, 0.2, \ldots, 0.5\}$, and $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10\}$].
- WRAP [Yu and Zhang, 2021]: A second-order multi-label classification approach that jointly performs label-specific feature generation and classification model induction [suggested configuration: $\lambda_1, \lambda_2 \in \{0, 1, \ldots, 10\}$, $\lambda_3 = 10$, and $\alpha = 0.9$].
- HOMI [Si *et al.*, 2023]:A high-order multi-label classification approach that constrains the high-rank of the label matrix and depict high-order label correlations explicitly [suggested configuration: $\beta, \gamma, \in \{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$, $\lambda = 1$, $s = 10$].

Detailed experimental results and resulting win/tie/loss counts of pairwise t-test at 0.05 significance in terms of six evaluation metrics are reported in Table 3 and Appendix B.2. The results clearly demonstrate that, although the performance of simple BR-KNN and ML-KNN are inferior to state-of-the-art multi-label classification approaches, the LSMM-enhanced version achieve statistically superior or at least comparable performance against state-of-the-art multi-label classification approaches. The results once again validate the superiority of our proposed LSMM framework in learning effective similarity metrics for multi-label classification.

4.3 Further Analysis

Ablation study. To validate the usefulness of the introduced global metric when learning label-specific multiple local metrics, two variants called LSMM-SE-N and LSMM-CL-N are implemented by learning label-specific multiple local metrics directly. Table 4 reports detailed experimental results of BR-KNN coupled with LSMM-SE, LSMM-CL, and variants in terms of *Average precision*. We can observe clear performance degradation with variants, demonstrating the usefulness of the global metric in LSMM framework. Similar results can be observed on other evaluation metrics.

Compared Algorithms	Datasets								
	emotions	birds	medical	enron	image	scene	slashdot	arts	education
					Hamming Loss ↓				
LIFT	0.267±0.015	0.074±0.026	0.012±0.002	0.046±0.002	0.159±0.014	0.079±0.007	0.042±0.001	0.053±0.001	**0.037±0.001**
RELIAB	0.260±0.031	0.074±0.007	0.016±0.001	0.059±0.003	0.178±0.015	0.110±0.015	0.051±0.003	0.055±0.003	**0.037±0.002**
WRAP	0.242±0.019	0.053±0.008	**0.010±0.001**	0.047±0.002	0.186±0.009	0.114±0.007	0.037±0.001	0.054±0.002	**0.037±0.001**
HOMI	0.238±0.018	0.073±0.025	0.013±0.002	0.045±0.002	0.162±0.017	0.085±0.006	0.044±0.003	0.057±0.002	0.038±0.001
BR-KNN-LSMM-SE	0.204±0.018	0.051±0.006	0.014±0.003	0.045±0.003	0.162±0.013	0.081±0.007	**0.035±0.002**	0.054±0.001	**0.037±0.002**
BR-KNN-LSMM-CL	**0.202±0.017**	**0.050±0.007**	0.012±0.002	0.044±0.003	0.157±0.017	0.079±0.009	0.037±0.002	**0.052±0.002**	0.038±0.001
ML-KNN-LSMM-SE	0.204±0.016	**0.050±0.005**	0.012±0.003	**0.043±0.002**	0.159±0.014	0.080±0.007	0.041±0.002	0.053±0.001	**0.037±0.001**
ML-KNN-LSMM-CL	0.209±0.018	0.051±0.006	0.012±0.002	0.045±0.003	**0.155±0.013**	**0.078±0.007**	0.039±0.003	**0.052±0.001**	0.038±0.001
					Ranking Loss ↓				
LIFT	0.254±0.046	0.323±0.050	0.028±0.011	0.079±0.006	0.156±0.020	0.074±0.009	0.096±0.008	0.124±0.006	0.084±0.003
RELIAB	0.256±0.040	0.427±0.036	**0.014±0.009**	0.085±0.009	0.181±0.025	0.091±0.009	0.117±0.017	0.145±0.006	0.097±0.008
WRAP	0.229±0.029	0.322±0.030	0.017±0.008	0.090±0.009	0.173±0.023	0.085±0.008	0.092±0.012	0.128±0.009	0.086±0.006
HOMI	0.237±0.032	0.329±0.046	0.041±0.012	0.098±0.013	0.196±0.021	0.084±0.010	0.103±0.009	0.186±0.010	0.132±0.004
BR-KNN-LSMM-SE	0.170±0.033	0.472±0.061	0.065±0.018	0.194±0.026	0.157±0.021	0.092±0.011	0.201±0.018	0.263±0.010	0.210±0.010
BR-KNN-LSMM-CL	0.178±0.034	0.487±0.058	0.071±0.023	0.192±0.022	**0.150±0.025**	0.090±0.013	0.190±0.025	0.258±0.027	0.205±0.012
ML-KNN-LSMM-SE	**0.166±0.033**	**0.256±0.034**	0.034±0.015	0.075±0.007	0.157±0.021	0.072±0.004	0.095±0.008	0.123±0.006	0.078±0.004
ML-KNN-LSMM-CL	0.170±0.036	0.260±0.038	0.038±0.013	**0.073±0.009**	**0.146±0.020**	**0.069±0.005**	**0.083±0.007**	**0.120±0.005**	**0.074±0.004**

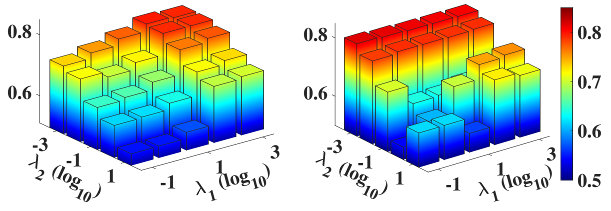
Table 3: Predictive performance (mean±std) of $\mathcal{A}$ ($\mathcal{A} \in \{$BR-KNN, ML-KNN$\}$) coupled with our proposed approaches and state-of-the-art non-metric learning multi-label classification approaches in terms of *Hamming Loss* and *Ranking Loss*. ↑ (↓) indicates the larger (smaller) the value, the better the performance. The best and second best results are highlighted in **boldface** and underline respectively.

Data sets	*Average precision* ↑			
	LSMM-SE	LSMM-SE-N	LSMM-CL	LSMM-CL-N
emotions	**0.795±0.039**	0.772±0.035●	**0.797±0.038**	0.768±0.032●
birds	**0.438±0.063**	0.397±0.057●	**0.427±0.058**	0.386±0.061●
medical	**0.882±0.036**	0.833±0.031●	**0.878±0.030**	0.836±0.027●
enron	**0.637±0.029**	0.624±0.022●	**0.648±0.032**	0.637±0.030●
image	**0.817±0.023**	0.805±0.021●	**0.820±0.018**	0.808±0.023●
scene	**0.857±0.012**	0.855±0.013	**0.868±0.017**	0.860±0.020●
slashdot	**0.678±0.021**	0.596±0.019●	**0.677±0.025**	0.611±0.024●
arts	**0.593±0.020**	0.521±0.017●	**0.602±0.018**	0.535±0.023●
education	**0.612±0.015**	0.587±0.014●	**0.619±0.017**	0.591±0.019●

Table 4: Predictive performance (mean±std) of BR-KNN coupled with LSMM-SE, LSMM-CL, and variants in terms of *Average precision*. ●/○ indicates whether BR-KNN-LSMM-SE and BR-KNN-LSMM-CL achieve significantly superior/inferior to the variants on each dataset (pairwise t-test at 0.05 significance level).

Sensitivity analysis. λ_1 and λ_2 serve as primary trade-off parameters in LSMM, balancing the impact of the global and label-specific multiple local metrics. Figure 1 provides an illustrative example of how the performance of BR-KNN-LSMM-SE changes with varying values of λ_1 and λ_2 (dataset: emotions, image; evaluation metric: *Average precision*). As shown in Figure 1, the performance of BR-KNN-LSMM-SE is quite sensitive to the values of λ_1 and λ_2. This demonstrates again the effectiveness of the global metric in exploiting label correlation. Similar results can be found in other cases.

Complexity analysis. In LSMM, the main computation lies in the gradient calculation and update for the global and label-specific multiple local metrics. The training complexity of one iteration for LSMM-SE, LSMM-CL are $\mathcal{O}(q(nk_tk_id^2 + 2dn^2 + 2d^3))$, $\mathcal{O}(q(nk_tk_id^2 + Cdn^2 + Cd^3))$ respectively. It is noteworthy that LSMM learns multiple metrics of which the number equals $2q + 1$ (for LSMM-SE) or $Cq + 1$ (for LSMM-CL), which could be hard to generalize to datasets with huge labels, i.e. extreme multi-label learning [Liu *et al.*, 2021]. This is unavoidable if considering label-specific metrics. We will leave efficiency improvement for future work.



(a) emotions (b) image

Figure 1: Performance of BR-KNN-LSMM-SE with varying values of λ_1 and λ_2 in terms of *Average precision*.

5 Conclusion

In this paper, we present a novel label-specific multi-metric learning framework LSMM for multi-label classification, where nonlinear distribution characteristics of multi-label instances are considered by learning label-specific multiple local metrics for different instances on the shoulder of a global one. Comprehensive experiments verify LSMM outperforms existing consistent multi-label metric learning algorithms in learning effective similarity metrics. Furthermore, we observe that when coupled with LSMM, simple BR-KNN and ML-KNN also have the potential to approach or even surpass state-of-the-art multi-label classification methods. However, LSMM learns multiple metrics, which could be hard to generalize to extreme multi-label learning. This is unavoidable 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.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (62176055), and the Big Data Computing Center of Southeast University.

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.

[Chen *et al.*, 2023] Shuo Chen, Chen Gong, Xiang Li, Jian Yang, Gang Niu, and Masashi Sugiyama. Boundary-restricted metric learning. *Machine Learning*, 112(12):4723–4762, 2023.

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

[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.

[Jain *et al.*, 1999] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys*, 31(3):264–323, 1999.

[Jia *et al.*, 2023] Bin-Bin Jia, Jun-Ying Liu, Jun-Yi Hang, and Min-Ling Zhang. Learning label-specific features for decomposition-based multi-class classification. *Frontiers of Computer Science*, 17(6):176348, 2023.

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

[Li and Lu, 2022] Yangyang Li and Ruqian Lu. Curvature flow learning: algorithm and analysis. *Science China Information Sciences*, 65(9):192105, 2022.

[Li *et al.*, 2022] Pandeng Li, Yan Li, Hongtao Xie, and Lei Zhang. Neighborhood-adaptive structure augmented metric learning. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, pages 1367–1375, Virtual Event, 2022.

[Li *et al.*, 2023] Jiaqi Li, Zhuofeng Li, Lei Wei, and Xuegong Zhang. Machine learning in lung cancer radiomics. *Machine Intelligence Research*, 20(6):753–782, 2023.

[Liao *et al.*, 2021] Tingting Liao, Zhen Lei, Tianqing Zhu, Shan Zeng, Yaqin Li, and Cao Yuan. Deep metric learning for k nearest neighbor classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):264–275, 2021.

[Liu and Nocedal, 1989] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[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.

[Mao *et al.*, 2023] Jun-Xiang Mao, Wei Wang, and Min-Ling Zhang. Label specific multi-semantics metric learning for multi-label classification: global consideration helps. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 4055–4063, Macao, SAR, China, 2023.

[Ren *et al.*, 2022] QiangQiang Ren, Chao Yuan, Yifeng Zhao, and Liming Yang. A multi-birth metric learning framework based on binary constraints. *Neural Networks*, 154:165–178, 2022.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the 18th IEEE conference on computer vision and pattern recognition*, pages 815–823, Boston, MA, 2015.

[Si *et al.*, 2023] Chongjie Si, Yuheng Jia, Ran Wang, Min-Ling Zhang, Yanghe Feng, and Chongxiao Qu. Multi-label classification with high-rank and high-order label correlations. *IEEE Transactions on Knowledge and Data Engineering*, (1):1–13, 2023.

[Song *et al.*, 2021] Kun Song, Junwei Han, Gong Cheng, Jiwen Lu, and Feiping Nie. Adaptive neighborhood metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4591–4604, 2021.

[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.

[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. *Advances in Neural Information Processing Systems*, 18:1473–1480, 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. *Advances in Neural Information Processing Systems*, 15:521–528, 2002.

[Xu *et al.*, 2023] Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. Label-specific feature augmentation for long-tailed multi-label text classification. In *Proceedings of the 37th AAAI Conference on Artificial Intelli-*

gence, volume 37, pages 10602–10610, Washington, DC, 2023.

[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.

[Yu and Zhang, 2021] Ze-Bang Yu and Min-Ling Zhang. Multi-label classification with label-specific feature generation: A wrapped approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5199–5210, 2021.

[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 Zhang, 2017] Jie Zhang and Lijun Zhang. Efficient stochastic optimization for low-rank distance metric learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, volume 31, San Francisco, CA, 2017.

[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] Min-Ling Zhang and Zhi-Hua 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.

[Zhao and Yang, 2023] Yifeng Zhao and Liming Yang. Distance metric learning based on the class center and nearest neighbor relationship. *Neural Networks*, 164:631–644, 2023.

[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.