Front.Comput.Sci.
DOI

RESEARCH ARTICLE

Compositional Metric Learning for Multi-Label Classification

Yan-Ping Sun[1,2,3], **Min-Ling Zhang** (✉)[1,2,3]

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 Collaborative Innovation Center for Wireless Communications Technology, Nanjing 211100, China

Abstract Multi-label classification aims to assign a set of proper labels for each instance, where distance metric learning can help improve the generalization ability of instance-based multi-label classification models. Existing multi-label metric learning techniques work by utilizing pairwise constraints to enforce that examples with similar label assignments should have close distance in the embedded feature space. In this paper, a novel distance metric learning approach for multi-label classification is proposed by modeling structural interactions between instance space and label space. On one hand, compositional distance metric is employed which adopts the representation of a weighted sum of rank-1 PSD matrices based on component bases. On the other hand, compositional weights are optimized by exploiting triplet similarity constraints derived from both instance and label spaces. Due to the compositional nature of employed distance metric, the resulting problem admits quadratic programming formulation with linear optimization complexity w.r.t. the number of training examples. We also derive the generalization bound for the proposed approach based on algorithmic robustness analysis of the compositional metric. Extensive experiments on sixteen benchmark data sets clearly validate the usefulness of compositional metric in yielding effective distance metric for multi-label classification.

Keywords machine learning, multi-label learning, metric learning, compositional metric, positive semidefinite matrix decomposition

Received month dd, yyyy; accepted month dd, yyyy

E-mail: *zhangml@seu.edu.cn*

1 Introduction

In multi-label classification, each instance is associated with multiple class labels simultaneously and the task is to learn a predictive model mapping from instance to the set of proper labels [1, 2]. In recent years, multi-label classification techniques have been widely applied to learn from real-world objects with rich semantics [3–7].

Distance metric learning serves as a popular strategy to facilitate supervised learning, where a positive semi-definite (PSD) matrix $\mathbf{M} \succeq 0$ is usually learned to parameterize the distance in embedded feature space [8, 9]. Some recent attempts show promising results of learning distance metric to build multi-label classification models with stronger generalization performance [10, 11]. Specifically, given training examples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $(\boldsymbol{x}_j, \boldsymbol{y}_j)$, their distance in the embedded feature space $d_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \mathbf{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)}$ should move closer if $\boldsymbol{y}_i$ is similar to $\boldsymbol{y}_j$ in the label space. This strategy can be instantiated in different ways such as large margin output coding [10, 12, 13] or pairwise similarity preservation [11, 14, 15].

In this paper, a novel multi-label distance metric learning approach named Commu, i.e. *COmpositional Metric for MUlti-label classification*, is proposed. Compared to existing approaches for multi-label metric learning, Commu considers a more advanced strategy by modeling structural interactions between instance space and label space. In Figure 1, the general framework of Commu for multi-label distance metric learning is illustrated. Specifically, the multi-label distance metric is assumed to adopt the compositional representation

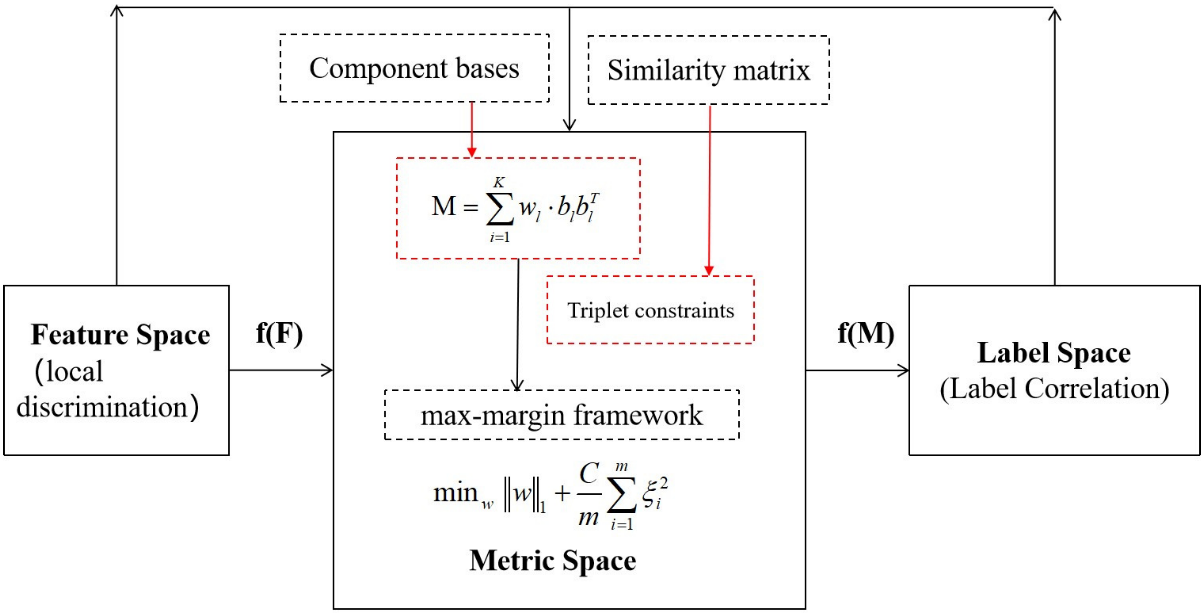


Fig. 1 The multi-label distance metric learning framework of COMMU. The original feature space is mapped into the distance metric feature space based on the compositional distance metric, based on which the prediction on multi-label output space will be made. Specifically, each component of the distance metric is generated by employing triplet constraints derived from similarity relationships in both instance and label spaces.

with a weighted sum of rank-1 PSD matrices. Here, the rank-1 PSD matrix corresponds to the outer product of component bases generated by encoding discriminative information of class labels. Furthermore, the weights forming the compositional distance metric are optimized by exploiting triplet constraints derived from similarity relationships in both instance and label spaces. Experimental studies across sixteen benchmark multi-label data sets show that COMMU is capable of significantly improving the generalization performance of instance-based multi-label classification models with the learned compositional distance metric.

The rest of this paper is organized as follows. Section 2 presents technical details of the proposed approach. Section 3 provides the corresponding theoretical analysis. Section 4 reports experimental results of comparative studies. Section 5 briefly discusses related works. Finally, Section 6 concludes this paper.

2 The COMMU Approach

Formally, let $\mathcal{X} = \mathbb{R}^d$ be the instance space and $\mathcal{Y} = \{\lambda_1, \lambda_2, \dots, \lambda_q\}$ be the label space with q class labels. Multi-label classification aims to learn a predictive function $h : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ from the training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid 1 \le i \le m\}$, where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{iq})^T$ is the labeling vector associated with $\boldsymbol{x}_i$ such that $y_{il} = 1$ if λ_l is a relevant label for $\boldsymbol{x}_i$ and $y_{il} = 0$ otherwise.

To model the distance metric $\mathbf{M}$ with enriched structural information, COMMU chooses to adopt the compositional representation with a weighted sum of K rank-1 PSD matrices [16, 17]:

$$\mathbf{M} = \sum_{l=1}^{K} w_l \cdot \boldsymbol{b}_l \boldsymbol{b}_l^\top \tag{1}$$

Here, $\boldsymbol{b}_l \in \mathbb{R}^d$ is the d-dimensional component base and $w_l \ge 0$ is the corresponding nonnegative compositional weight. In this way, one can simplify the parameterization complexity of the distance metric from $O(d^2)$ to $O(K)$ with $\boldsymbol{w} = [w_1, w_2, \dots, w_K]^\top$. More importantly, the compositional decomposition naturally enables the encoding of discriminative information into the distance metric. Specifically, COMMU generates one component base for each class label in the label space (i.e. $K = q$).

For the l-th class label $\lambda_l \in \mathcal{Y}$ $(1 \le l \le q)$, COMMU considers the difference between the mean of positive examples and negative examples w.r.t. λ_l:

$$\boldsymbol{b}_l = \frac{\sum_{\boldsymbol{u} \in \mathcal{P}_l} \boldsymbol{u}}{|\mathcal{P}_l|} - \frac{\sum_{\boldsymbol{v} \in \mathcal{N}_l} \boldsymbol{v}}{|\mathcal{N}_l|} \tag{2}$$

Here, $\mathcal{P}_l = \{\boldsymbol{x}_i \mid y_{il} = 1,\ (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{D}\}$ and $\mathcal{N}_l = \{\boldsymbol{x}_i \mid y_{il} = 0,\ (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{D}\}$ correspond to the set of positive examples and negative examples w.r.t. λ_l respectively. Conceptually, the statistic in Eq.(2) is used to reflect holistic labeling distribution of class label, which has shown to be beneficial for encoding discriminative information in the feature space [18–20].

To optimize the parameters $\boldsymbol{w}$ for distance metric $\mathbf{M}$, a set of constraints are specified to characterize the properties which $\mathbf{M}$ are expected to possess. Given the multi-label training example $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and other two reference examples $\{(\boldsymbol{x}_j, \boldsymbol{y}_j), (\boldsymbol{x}_k, \boldsymbol{y}_k)\}$, it is desirable that $d_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ should be smaller that $d_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_k)$ if $\boldsymbol{x}_i$ is semantically more similar to $\boldsymbol{x}_j$ than $\boldsymbol{x}_k$. Under traditional single-label scenario, the semantic similarity can be easily measured by considering whether two examples have the same class label [8, 21]. However, under multi-label scenario, it is impractical to measure semantic similarity by considering exact labeling equivalence due to the combinatorial nature of multiple class labels. For Commu, the semantic similarity matrix $\mathbf{S} = [s_{ij}]_{m\times m}$ is calculated by synergizing discriminative information from both input space and label space:

$$s_{ij} = \boldsymbol{y}_i^\top \mathbf{G} \boldsymbol{y}_j \tag{3}$$
$$\text{where} \quad \mathbf{G} = (\alpha\mathbf{A} + (1-\alpha)\mathbf{C})$$
$$\mathbf{A} = [a_{lh}]_{q\times q} \;\; \text{with} \;\; a_{lh} = \frac{\sum_{i=1}^m y_{il}\cdot y_{ih}}{\sum_{i=1}^m y_{il}}$$
$$\mathbf{C} = [c_{lh}]_{q\times q} \;\; \text{with} \;\; c_{lh} = \frac{\boldsymbol{b}_l^\top \boldsymbol{b}_h}{\|\boldsymbol{b}_l\|\cdot\|\boldsymbol{b}_h\|}$$

Here, a_{lh} corresponds to the fraction of examples with label y_l which also have label y_h. It is noteworthy that $a_{lh} = a_{hl}$ does not necessarily hold here to reflect the fact that correlations among class labels are usually *asymmetric* [22, 23]. Furthermore, c_{lh} corresponds to the cosine similarity between compositional bases. The coefficient α balances relative contributions from label space (i.e. $\mathbf{A}$) and instance space (i.e. $\mathbf{C}$) in calculating the semantic similarity.

Thereafter, the set of "similar" and "dissimilar" examples for training instance $\boldsymbol{x}_i$ are determined as:

$$\mathcal{Z}_i = \{\boldsymbol{x}_j \mid s_{ij} \geq \theta,\; j \neq i,\; 1 \leq j \leq m\} \tag{4}$$
$$\tilde{\mathcal{Z}}_i = \{\boldsymbol{x}_k \mid s_{ik} < \theta,\; k \neq i,\; 1 \leq k \leq m\}$$

Here, θ is used as the thresholding parameter for measuring semantic similarity. Accordingly, the following set of triplets are generated by utilizing subset $\mathcal{K}_i \subseteq \mathcal{Z}_i$ ($\tilde{\mathcal{K}}_i \subseteq \tilde{\mathcal{Z}}_i$) which consists of top k ($\tilde{k}$) instances with highest semantic similarity in $\mathcal{Z}_i$ ($\tilde{\mathcal{Z}}_i$):

$$\mathcal{R} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k) \mid 1 \leq i \leq m,\; \boldsymbol{x}_j \in \mathcal{K}_i,\; \boldsymbol{x}_k \in \tilde{\mathcal{K}}_i\} \tag{5}$$

Here, $\mathcal{R}$ contains a total of $m\cdot k\cdot \tilde{k}$ triplets. Based on Eq.(5), Commu learns the compositional distance metric by solving the following optimization problem with triplet constraints:

$$\min_w \quad \|\boldsymbol{w}\|_1 + \frac{C}{m}\sum_{i=1}^m \xi_i^2 \qquad \text{s.t.}: \tag{6}$$
$$d^2_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_k) - d^2_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq \Delta(\boldsymbol{y}_j, \boldsymbol{y}_k) - \xi_i$$
$$(\forall\; (\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k) \in \mathcal{R})$$
$$\xi_i \geq 0, \;\; w_i \geq 0 \;\; (1 \leq i \leq m)$$

Here, $d^2_{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{l=1}^q w_l\cdot(\boldsymbol{x}-\boldsymbol{x}')^\top \boldsymbol{b}_l \boldsymbol{b}_l^\top(\boldsymbol{x}-\boldsymbol{x}')$ corresponds to the distance between two instances in the embedded feature space and $\Delta(\boldsymbol{y}_j, \boldsymbol{y}_k) = \boldsymbol{y}_j^\top \left(\mathbf{1}_{q\times q} - \mathbf{G}\right)\boldsymbol{y}_k$ corresponds to the dissimilarity between two labeling vectors. Accordingly, the slack variable ξ_i corresponds to:

$$\xi_i = \max\left(0, \max_{(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_k)\in\mathcal{R}} \left(\Delta(\boldsymbol{y}_j, \boldsymbol{y}_k) - \left(d^2_{\mathbf{M}}(\boldsymbol{x}_{i,k}) - d^2_{\mathbf{M}}(\boldsymbol{x}_{i,j})\right)\right)\right) \tag{7}$$

Table 1 The pseudo-code of Commu.

Inputs:
$\mathcal{D}$: multi-label training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid 1 \leq i \leq m\}$
α: balancing parameter in Eq.(3) with $\alpha \in (0, 1)$
C: cost parameter in Eq.(6) with $C > 0$
θ: thresholding parameter in Eq.(4)
Outputs:
$\boldsymbol{w}$: compositional weight vector for the distance metric

Process:
1: **for** $l = 1$ to q **do**
2: Generate component base $\boldsymbol{b}_l$ according to Eq.(1);
3: **end for**
4: Calculate the similarity matrix $\mathbf{S}$ according to Eq.(3);
5: Form the set of triplets $\mathcal{R}$ according to Eq.(5);
6: Initialize FISTA procedure with $\boldsymbol{w}_0 = \boldsymbol{w}_1 = \frac{1}{q}\cdot\mathbf{1}_{q\times 1}$, $\tau_0 = \tau_1 = 0.01mC$, $\eta = 0.4$, and $t_0 = t_1 = 1$;
7: Set $r = 1$ and $\tilde{\boldsymbol{w}}_1 = \boldsymbol{w}_1$;
8: **repeat**
9: Set $L = \eta\cdot\tau_r$;
10: **repeat**
11: Calculate $\boldsymbol{a}^* = \tilde{\boldsymbol{w}}_r - \frac{1}{L}(\nabla f(\tilde{\boldsymbol{w}}_r) + \mathbf{1}_{q\times 1})$;
12: **if** $F(\Pi_+(\boldsymbol{a}^*)) \leq Q_L(\Pi_+(\boldsymbol{a}^*), \tilde{\boldsymbol{w}}_r)$ **then**
13: $\tau_{r+1} = L$;
14: **go to step 19**;
15: **else**
16: $L = \frac{1}{\eta}L$;
17: **end if**
18: **until** false
19: $\boldsymbol{w}_{r+1} = \Pi_+(\boldsymbol{a}^*)$;
20: $t_{r+1} = \frac{1+\sqrt{1+4t_r^2}}{2}$;
21: $r = r + 1$;
22: $\tilde{\boldsymbol{w}}_r = \boldsymbol{w}_{r-1} + \frac{t_{i-1}-1}{t_i}\cdot(\boldsymbol{w}_{r-1} - \boldsymbol{w}_{r-2})$;
23: **until** convergence
24: Return $\boldsymbol{w} = \boldsymbol{w}_r$;

Table 2 Characteristics of the benchmark multi-label data sets.

Data set	$\vert\mathcal{D}\vert$	$dim(\mathcal{D})$	$CL(\mathcal{D})$	$LCard(\mathcal{D})$	$LDen(\mathcal{D})$	$DL(\mathcal{D})$	$PDL(\mathcal{D})$	Domain
genbase	662	1186	27	1.252	0.046	32	0.048	biology
Society	2000	636	27	1.692	0.063	329	0.165	text
Social	2000	1047	39	1.283	0.033	137	0.069	text
Reference	2000	793	33	1.169	0.035	132	0.066	text
Health	2000	612	32	1.662	0.052	164	0.082	text
Education	2000	550	33	1.461	0.044	200	0.1	text
Computers	2000	681	33	1.508	0.046	148	0.074	text
Business	2000	438	30	1.588	0.053	96	0.048	text
Arts	2000	462	26	1.636	0.063	254	0.127	text
yeast	2417	103	14	4.237	0.303	198	0.082	biology
corel5k	5000	499	374	3.522	0.009	3175	0.635	images
rcv1-subset1	6000	944	101	2.88	0.029	1028	0.171	text
corel16k001	13766	500	153	2.859	0.019	4937	0.359	images
eurlex-dc	19348	100	412	1.292	0.003	1615	0.083	text
eurlex-sm	19348	100	201	2.213	0.011	2504	0.129	text
eurlex	19314	1854	815	4.273	0.0052	14763	0.764	text

Therefore, the solution to Eq.(6) can be obtained by optimizing the following equivalent problem:

$$\min_{\boldsymbol{w}} \quad F(\boldsymbol{w}) \equiv f(\boldsymbol{w}) + g(\boldsymbol{w}) \tag{8}$$

Here, $f(\boldsymbol{w}) = \frac{C}{m}\sum_{i=1}^{m}\xi_i^2$ whose gradient ∇f is Lipschitz continuous w.r.t. $\boldsymbol{w}$ [10, 24] and $g(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$ is convex. For optimization problem admitting such decomposition, its solution can be obtained by employing the *FISTA (Fast Iterative Shrinkage-Thresholding Algorithm)* procedure [25, 26]. Specifically, given the current solution $\boldsymbol{w}$, the solution at next iteration is solved by minimizing the following quadratic programming problem:

$$Q_L(\boldsymbol{a}, \boldsymbol{w}) = f(\boldsymbol{w}) + \langle \nabla f(\boldsymbol{w}), \boldsymbol{a} - \boldsymbol{w} \rangle + \frac{L}{2}\|\boldsymbol{a} - \boldsymbol{w}\|^2 + g(\boldsymbol{a}) \tag{9}$$

where $L > 0$ is the Lipschitz constant for $f(\boldsymbol{w})$. By setting the gradient of Eq.(9) to zero, one can obtain the minimizer $\boldsymbol{a}^* = \boldsymbol{w} - \frac{1}{L}(\nabla f(\boldsymbol{w}) + \mathbf{1}_{q\times 1})$. To ensure nonnegativity of compositional weights for the distance metric $\mathbf{M}$, the iterative solution $\boldsymbol{a}^*$ will be mapped to $\Pi_+(\boldsymbol{a}^*)$ by setting negative elements in $\boldsymbol{a}^*$ to zero.

Table 1 summarizes the complete procedure of COMMU. Firstly, a set of compositional bases are generated by discriminative information encoding (steps 1 to 3). After that, the set of triplet constraints are specified by considering semantic similarity among training examples (steps 4-5). Thirdly, the compositional weights are learned by invoking the FISTA iterative optimization procedure (steps 6-24).[1)]

3 Theoretical Analysis

In this section, we provide a theoretical analysis of our approach in the form of a generalization bound based on algorithmic robustness analysis for metric learning [27].

Given a multi-label dataset $\mathcal{S} = \{z = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ drawn i.i.d. from a distribution P over the labelled space $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, where the label vector $\boldsymbol{y}_i$ simultaneously contains multiple labels. Assume that $\|\boldsymbol{x}\| \leq R$ (for some convenient norm), $\forall \boldsymbol{x} \in \mathcal{X}$. Different from the single-label setting, COMMU defines the multi-label semantic similarity matrix to construct the triplet (z, z', z''), where $\boldsymbol{y}$ is similar to $\boldsymbol{y}'$ and dissimilar to $\boldsymbol{y}''$. Let $\mathcal{S}_R$ be the set of all admissible triplets built from $\mathcal{S}$ and $L(\boldsymbol{w}, z, z', z'') = [\Delta(\boldsymbol{y}', \boldsymbol{y}'') + d_w(\boldsymbol{x}, \boldsymbol{x}') - d_w(\boldsymbol{x}, \boldsymbol{x}'')]_+$ denote the multi-label triple loss function in Eq.(6), which is uniformly upper-bounded by a constant U.

The empirical loss $R_{emp}^{\mathcal{S}_R}(\boldsymbol{w})$ of $\boldsymbol{w}$ on $\mathcal{S}_R$ is defined as

$$R_{emp}^{\mathcal{S}_R}(\boldsymbol{w}) = \frac{1}{|\mathcal{S}_R|}\sum_{(z,z',z'')\in\mathcal{S}_R} L(\boldsymbol{w}, z, z', z''),$$

and its expected loss $R(\boldsymbol{w})$ over distribution P as

$$R(\boldsymbol{w}) = \mathbb{E}_{z,z',z''\sim P} L(\boldsymbol{w}, z, z', z'').$$

[1)] The FISTA procedure terminates when the value of the objective function in Eq.(8) does not significantly decrease for two consecutive solutions $\boldsymbol{w}_r$ and $\boldsymbol{w}_{r+1}$, i.e. $F(\boldsymbol{w}_r) - F(\boldsymbol{w}_{r+1}) \leq \epsilon \cdot F(\boldsymbol{w}_r)$ with $\epsilon = 0.001$.

Table 3 Predictive performance (mean ± std. deviation) of each distance metric learning approach in terms of *ranking loss*.

	*k*NN-				Mlknn-			
Data set	Commu	Lm	Nje	Original	Commu	Lm	Nje	Original
genbase	0.008±0.008	0.146±0.071	**0.006±0.007**	0.007±0.008	**0.006±0.007**	0.009±0.007	0.008±0.009	0.007±0.007
Society	0.261±0.017	0.306±0.019	**0.215±0.021**	0.263±0.02	0.142±0.007	0.164±0.013	0.202±0.016	**0.138±0.006**
Social	0.165±0.016	0.260±0.023	**0.137±0.023**	0.160±0.02	**0.065±0.008**	0.084±0.007	0.099±0.013	0.065±0.007
Reference	0.182±0.032	0.303±0.034	**0.143±0.017**	0.246±0.032	**0.063±0.009**	0.103±0.010	0.093±0.010	0.083±0.011
Health	0.154±0.031	0.184±0.036	**0.115±0.015**	0.199±0.03	**0.057±0.010**	0.062±0.010	0.075±0.010	0.063±0.009
Education	0.210±0.023	0.299±0.02	**0.168±0.020**	0.209±0.024	**0.087±0.006**	0.105±0.011	0.131±0.014	0.087±0.006
Computers	0.186±0.029	0.285±0.017	**0.132±0.013**	0.192±0.041	**0.082±0.008**	0.093±0.011	0.114±0.011	0.082±0.006
Business	**0.090±0.012**	0.122±0.015	0.089±0.015	0.092±0.012	**0.037±0.005**	0.043±0.007	0.075±0.013	0.038±0.006
Arts	0.270±0.028	0.315±0.026	**0.202±0.025**	0.302±0.031	**0.149±0.013**	0.168±0.012	0.181±0.023	0.153±0.015
yeast	0.197±0.006	0.322±0.015	**0.188±0.014**	0.195±0.009	0.176±0.008	0.182±0.009	0.197±0.016	**0.175±0.008**
corel5k	**0.456±0.016**	0.675±0.011	-	0.578±0.027	**0.120±0.006**	0.129±0.006	-	0.132±0.006
rcv1-subset1	**0.191±0.030**	0.307±0.013	-	0.226±0.010	**0.073±0.008**	0.098±0.007	-	0.080±0.004
corel16k	**0.495±0.008**	0.680±0.004	-	0.537±0.017	**0.169±0.002**	0.173±0.002	-	0.175±0.001
eurlex-dc	**0.376±0.034**	0.637±0.023	-	0.376±0.035	**0.094±0.009**	0.125±0.008	-	0.094±0.009
eurlex-sm	**0.191±0.003**	0.402±0.006	-	0.191±0.003	**0.051±0.001**	0.073±0.001	-	0.051±0.001
eurlex	0.923±0.000	**0.922±0.000**	-	0.985±0.000	0.320±0.000	0.326±0.000	-	**0.316±0.000**

Table 4 Predictive performance (mean ± std. deviation) of each distance metric learning approach in terms of *coverage*.

	*k*NN-				Mlknn-			
Data set	Commu	Lm	Nje	Original	Commu	Lm	Nje	Original
genbase	0.019±0.011	0.106±0.053	0.020±0.017	**0.014±0.005**	**0.021±0.012**	0.025±0.011	0.230±0.019	0.021±0.013
Society	**0.272±0.025**	0.284±0.008	0.307±0.030	0.277±0.001	0.208±0.015	0.228±0.017	0.289±0.029	**0.203±0.001**
Social	**0.120±0.015**	0.144±0.053	0.175±0.029	0.120±0.001	**0.087±0.013**	0.107±0.012	0.129±0.017	0.088±0.001
Reference	0.153±0.014	**0.138±0.011**	0.165±0.017	0.150±0.001	0.104±0.011	0.117±0.012	0.108±0.009	**0.097±0.001**
Health	0.142±0.018	**0.138±0.011**	0.198±0.022	0.169±0.001	**0.098±0.013**	0.106±0.012	0.131±0.015	0.104±0.001
Education	**0.167±0.011**	0.192±0.016	0.228±0.026	0.168±0.001	**0.116±0.009**	0.135±0.013	0.173±0.013	0.116±0.001
Computers	**0.153±0.019**	0.167±0.003	0.182±0.021	0.160±0.001	**0.118±0.013**	0.131±0.016	0.156±0.017	0.118±0.001
Business	**0.090±0.012**	0.098±0.013	0.149±0.016	0.095±0.001	**0.071±0.007**	0.080±0.010	0.130±0.020	0.073±0.001
Arts	**0.281±0.028**	0.295±0.004	0.287±0.031	0.304±0.001	**0.209±0.020**	0.226±0.016	0.253±0.032	0.213±0.001
yeast	**0.462±0.010**	0.560±0.011	0.470±0.020	0.473±0.011	0.456±0.010	0.469±0.007	0.483±0.022	**0.454±0.012**
corel5k	**0.359±0.017**	0.792±0.011	-	0.739±0.019	**0.280±0.013**	0.294±0.012	-	0.302±0.014
rcv1-subset1	**0.244±0.023**	0.290±0.016	-	0.267±0.011	**0.165±0.016**	0.202±0.012	-	0.178±0.010
corel16k	**0.511±0.006**	0.586±0.003	-	0.548±0.009	**0.328±0.003**	0.335±0.003	-	0.339±0.002
eurlex-dc	**0.276±0.026**	0.449±0.013	-	0.276±0.026	**0.114±0.010**	0.150±0.009	-	0.114±0.010
eurlex-sm	**0.233±0.003**	0.429±0.004	-	0.233±0.004	**0.092±0.001**	0.126±0.002	-	0.092±0.001
eurlex	0.642±0.000	**0.640±0.000**	-	0.640±0.000	**0.600±0.000**	0.609±0.000	-	0.600±0.000

The goal of the theoretical analysis is to bound the deviation between $R(\boldsymbol{w})$ and $R_{emp}^{\mathcal{S}_R}(\boldsymbol{w})$, where $\boldsymbol{w}$ is the metric coefficient to learn.

Theorem 1. Let $\boldsymbol{w}^*$ be the optimal solution to Commu with K basis elements, $C > 0$ and the triplet $\mathcal{S}_R$ constructed from $\mathcal{S} = \{\boldsymbol{z} = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$. Let $K^* \leq K$ be the number of nonzero entries in $\boldsymbol{w}^*$. Assume the norm of any instance bounded by some constant R and the loss L uniformly upper-bounded by some constant U. Then for any $\delta > 0$, with probability at least $1 - \delta$ we have:

$$\left|R(w^*) - R_{emp}^{\mathcal{S}_R}(w^*)\right| \leq 16\gamma R K^* \Delta_0 C + \theta + 3U\sqrt{\frac{2^{q+1}\mathcal{N}(\gamma, \mathcal{X}, \|\cdot\|_1)\ln 2 + 2\ln\frac{1}{\delta}}{n}},$$

where $2^q\mathcal{N}(\gamma, \mathcal{X}, \|\cdot\|_1)$ is the size of an γ-cover of $\mathcal{Z}$, Δ_0 is the maximum distance of the dissimialrity set and θ is the theshold of the triplet construction in Eq.(4). This bound has a standard $O(1/\sqrt{n})$ asymptotic convergence rate.[2)] The detailed proofs can be found in the Appendixes.

2) In robustness bounds, the cover radius γ can be made arbitrarily close to zero at the expense of increasing $\mathcal{N}(\gamma, \mathcal{Z}, \rho)$. Since $\mathcal{N}(\gamma, \mathcal{Z}, \rho) = 2^q\mathcal{N}(\gamma, \mathcal{X}, \|\cdot\|_1)$ appears in the second term, the right hand side of the bound indeed goes to zero when $n \to \infty$. This is in accordance with other similar learning bounds.

Table 5 Predictive performance (mean ± std. deviation) of each distance metric learning approach in terms of *average precision*.

	*k*NN-				Mlknn-			
Data set	Commu	Lm	Nje	Original	Commu	Lm	Nje	Original
genbase	**0.991±0.007**	0.841±0.067	0.921±0.015	0.986±0.010	**0.988±0.011**	0.983±0.008	0.933±0.016	0.984±0.012
Society	0.544±0.016	**0.555±0.026**	0.131±0.013	0.539±0.018	0.584±0.011	0.527±0.031	0.148±0.020	**0.586±0.014**
Social	**0.120±0.015**	**0.662±0.025**	0.068±0.005	0.664±0.027	0.705±0.018	0.674±0.017	0.139±0.117	**0.706±0.021**
Reference	0.579±0.032	**0.592±0.024**	0.078±0.012	0.435±0.025	**0.659±0.040**	0.585±0.020	0.087±0.028	0.627±0.031
Health	0.588±0.066	**0.703±0.041**	0.106±0.010	0.474±0.031	0.690±0.046	**0.726±0.036**	0.090±0.013	0.659±0.040
Education	0.516±0.022	0.514±0.020	0.124±0.009	**0.520±0.027**	0.571±0.015	0.534±0.034	0.126±0.013	**0.572±0.019**
Computers	**0.625±0.030**	0.600±0.027	0.092±0.006	0.613±0.024	**0.653±0.020**	0.643±0.022	0.102±0.035	0.652±0.017
Business	**0.872±0.016**	0.846±0.019	0.088±0.005	0.864±0.018	**0.882±0.015**	0.870±0.017	0.077±0.008	0.878±0.019
Arts	0.473±0.033	**0.526±0.028**	0.173±0.016	0.422±0.029	**0.533±0.018**	0.494±0.026	0.157±0.017	0.513±0.029
yeast	**0.752±0.009**	0.659±0.011	0.381±0.012	0.746±0.011	**0.754±0.010**	0.746±0.010	0.457±0.051	0.753±0.009
corel5k	**0.251±0.011**	0.191±0.012	-	0.154±0.013	**0.303±0.013**	0.288±0.009	-	0.252±0.013
rcv1-subset1	0.522±0.028	**0.527±0.017**	-	0.488±0.014	**0.554±0.017**	0.449±0.011	-	0.539±0.013
corel16k	**0.212±0.004**	0.185±0.002	-	0.184±0.006	0.293±0.004	**0.303±0.002**	-	0.279±0.002
eurlex-dc	**0.440±0.027**	0.310±0.018	-	0.440±0.027	**0.464±0.027**	0.371±0.018	-	0.464±0.027
eurlex-sm	**0.609±0.004**	0.510±0.006	-	0.609±0.004	**0.652±0.004**	0.560±0.003	-	0.652±0.004
eurlex	0.023±0.000	**0.030±0.000**	-	0.011±0.000	0.032±0.000	**0.040±0.000**	-	0.033±0.000

Table 6 Predictive performance (mean ± std. deviation) of each distance metric learning approach in terms of *micro-F1*.

	*k*NN-				Mlknn-			
Data set	Commu	Lm	Nje	Original	Commu	Lm	Nje	Original
genbase	**0.957±0.020**	0.848±0.066	0.806±0.245	0.950±0.025	0.942±0.029	**0.951±0.030**	0.843±0.125	0.945±0.031
Society	0.377±0.011	0.404±0.032	**0.415±0.028**	0.381±0.019	0.318±0.020	0.345±0.020	**0.410±0.020**	0.312±0.020
Social	0.506±0.035	0.556±0.032	**0.587±0.026**	0.507±0.031	0.517±0.020	0.502±0.019	**0.583±0.135**	0.519±0.018
Reference	0.373±0.041	0.493±0.028	**0.530±0.033**	0.295±0.032	0.388±0.021	0.441±0.015	**0.514±0.015**	0.385±0.011
Health	0.451±0.058	0.588±0.043	**0.625±0.029**	0.367±0.033	0.439±0.004	0.562±0.011	**0.618±0.020**	0.388±0.004
Education	0.367±0.025	0.384±0.027	**0.430±0.027**	0.376±0.031	0.252±0.016	0.355±0.008	**0.426±0.135**	0.243±0.010
Computers	0.470±0.020	**0.475±0.032**	0.471±0.026	0.453±0.020	0.406±0.002	0.474±0.005	**0.478±0.015**	0.376±0.001
Business	0.715±0.018	**0.726±0.021**	0.339±0.307	0.674±0.017	**0.696±0.029**	0.718±0.024	0.603±0.135	0.693±0.029
Arts	0.322±0.031	0.376±0.027	**0.413±0.040**	0.273±0.032	0.195±0.004	0.304±0.008	**0.404±0.015**	0.143±0.004
yeast	0.610±0.013	0.572±0.016	0.408±0.176	**0.641±0.012**	0.634±0.011	0.623±0.015	0.417±0.174	**0.635±0.011**
corel5k	**0.229±0.008**	0.193±0.020	-	0.122±0.015	0.067±0.012	**0.113±0.011**	-	0.030±0.007
rcv1-subset1	0.434±0.024	**0.446±0.016**	-	0.398±0.009	0.301±0.018	**0.321±0.011**	-	0.285±0.012
corel16k	**0.168±0.005**	0.047±0.002	-	0.021±0.006	0.015±0.003	**0.046±0.007**	-	0.009±0.002
eurlex-dc	**0.363±0.027**	0.287±0.023	-	0.363±0.027	**0.324±0.032**	0.254±0.030	-	0.324±0.032
eurlex-sm	**0.537±0.004**	0.476±0.005	-	0.506±0.003	**0.554±0.004**	0.475±0.007	-	0.554±0.004
eurlex	**0.009±0.000**	0.002±0.000	-	0.002±0.000	0.002±0.000	**0.013±0.000**	-	0.002±0.000

4 Experiments

4.1 Experimental Setup

4.1.1 Data Sets

To comprehensively evaluate the effectiveness of Commu, sixteen benchmark multi-label data sets are collected for experimental studies.[3] Given a multi-label data set $\mathcal{S}$, we use $|\mathcal{S}|$, $dim(\mathcal{S})$ and $L(\mathcal{S})$ to represent its number of examples, number of features and number of class labels respectively. In addition, properties of $\mathcal{S}$ are further characterized by several useful multi-label statistics [28], including label cardinality $LCard(\mathcal{S})$, label density $LDen(\mathcal{S})$, distinct labelsets $DL(\mathcal{S})$ and proportion of distinct label sets $PDL(\mathcal{S})$.

Table 2 summarizes characteristics of the benchmark multi-label data sets, which are roughly ordered according to $|\mathcal{S}|$. As shown in Table 2, these data sets serve as a solid basis for comparative studies which exhibit diversified properties in terms of different multi-label statistics.

[3] Publicly available at http://mulan.sourceforge.net/datasets.html, http://waikato.github.io/meka/datasets/ and http://manikvarma.org/downloads/XC/XMLRepository.

Table 7 Predictive performance (mean ± std. deviation) of each distance metric learning approach in terms of *macro-F1*.

	*k*NN-				Mlknn-			
Data set	Commu	Lm	Nje	Original	Commu	Lm	Nje	Original
genbase	**0.896±0.055**	0.848±0.069	0.439±0.156	0.864±0.062	0.820±0.081	**0.836±0.078**	0.485±0.123	0.836±0.083
Society	0.255±0.040	0.150±0.024	0.131±0.014	**0.268±0.031**	0.200±0.085	**0.266±0.033**	0.110±0.012	0.203±0.034
Social	**0.403±0.066**	0.127±0.028	0.089±0.014	0.401±0.068	**0.384±0.085**	0.365±0.056	0.082±0.017	0.382±0.061
Reference	**0.538±0.052**	0.133±0.020	0.134±0.017	0.475±0.061	0.486±0.038	**0.500±0.058**	0.122±0.017	0.466±0.067
Health	**0.546±0.049**	0.209±0.039	0.216±0.026	0.536±0.042	0.520±0.014	**0.565±0.048**	0.201±0.025	0.486±0.038
Education	0.454±0.052	0.108±0.013	0.133±0.018	**0.461±0.052**	0.408±0.030	**0.448±0.062**	0.118±0.015	0.409±0.067
Computers	**0.347±0.052**	0.137±0.022	0.095±0.015	0.327±0.051	0.291±0.002	**0.339±0.044**	0.095±0.011	0.294±0.043
Business	0.376±0.042	0.167±0.023	0.094±0.027	**0.398±0.044**	0.374±0.025	**0.402±0.050**	0.135±0.018	0.364±0.051
Arts	**0.255±0.038**	0.150±0.018	0.157±0.021	0.240±0.035	0.187±0.013	**0.233±0.039**	0.146±0.020	0.173±0.045
yeast	0.473±0.014	0.394±0.022	0.438±0.024	**0.468±0.017**	0.381±0.024	0.355±0.031	**0.429±0.019**	0.381±0.023
corel5k	0.237±0.009	**0.339±0.014**	-	0.325±0.013	0.328±0.014	**0.329±0.014**	-	0.321±0.013
rcv1-subset1	0.307±0.030	**0.324±0.032**	-	0.286±0.020	0.213±0.028	**0.223±0.020**	-	0.205±0.022
corel16k	**0.055±0.003**	0.017±0.002	-	0.005±0.001	0.011±0.002	**0.022±0.002**	-	0.008±0.002
eurlex-dc	**0.325±0.025**	0.278±0.027	-	0.325±0.025	**0.296±0.025**	0.268±0.023	-	0.296±0.025
eurlex-sm	**0.252±0.012**	0.205±0.011	-	0.182±0.009	**0.224±0.013**	0.145±0.009	-	0.224±0.013
eurlex	**8.784e-4±0.000**	2.993e4±0.000	-	2.914e-4±0.000	3.400e-4±0.000	**0.002±0.000**	-	1.831e-4±0.000

Table 8 Win/tie/loss counts (pairwise *t*-test at 0.05 significance level) between Commu and the comparing approaches.

	*k*NN-Commu **against**			Mlknn-Commu **against**		
	*k*NN-Lm	*k*NN-Nje	*k*NN-Original	Mlknn-Lm	Mlknn-Nje	Mlknn-Original
ranking loss	14/1/0	1/2/7	5/10/0	12/3/0	8/2/0	3/12/0
coverage	9/6/0	7/3/0	11/4/0	12/3/0	8/2/0	11/4/0
average precision	7/5/3	10/0/0	6/9/0	9/5/1	10/0/0	3/12/0
micro-F1	6/4/5	2/2/6	7/8/0	2/4/9	3/7/0	5/10/0
macro-F1	12/3/0	9/1/0	2/13/0	2/9/4	9/1/0	1/14/0
In Total	**48/19/8**	**29/8/13**	**31/44/0**	**37/24/14**	**38/12/0**	**23/52/0**

4.1.2 Comparing Algorithms

Based on the learned distance metric, it is desirable to show whether the performance of instance-based multi-label classification models can be improved along with the distance measure in embedded feature space. Accordingly, the vanilla *k*NN method and the Mlknn method [29] are utilized as two natural choices for instance-based multi-label classification models. In this paper, the effectiveness of Commu is compared against two state-of-the-art multi-label metric learning approaches:

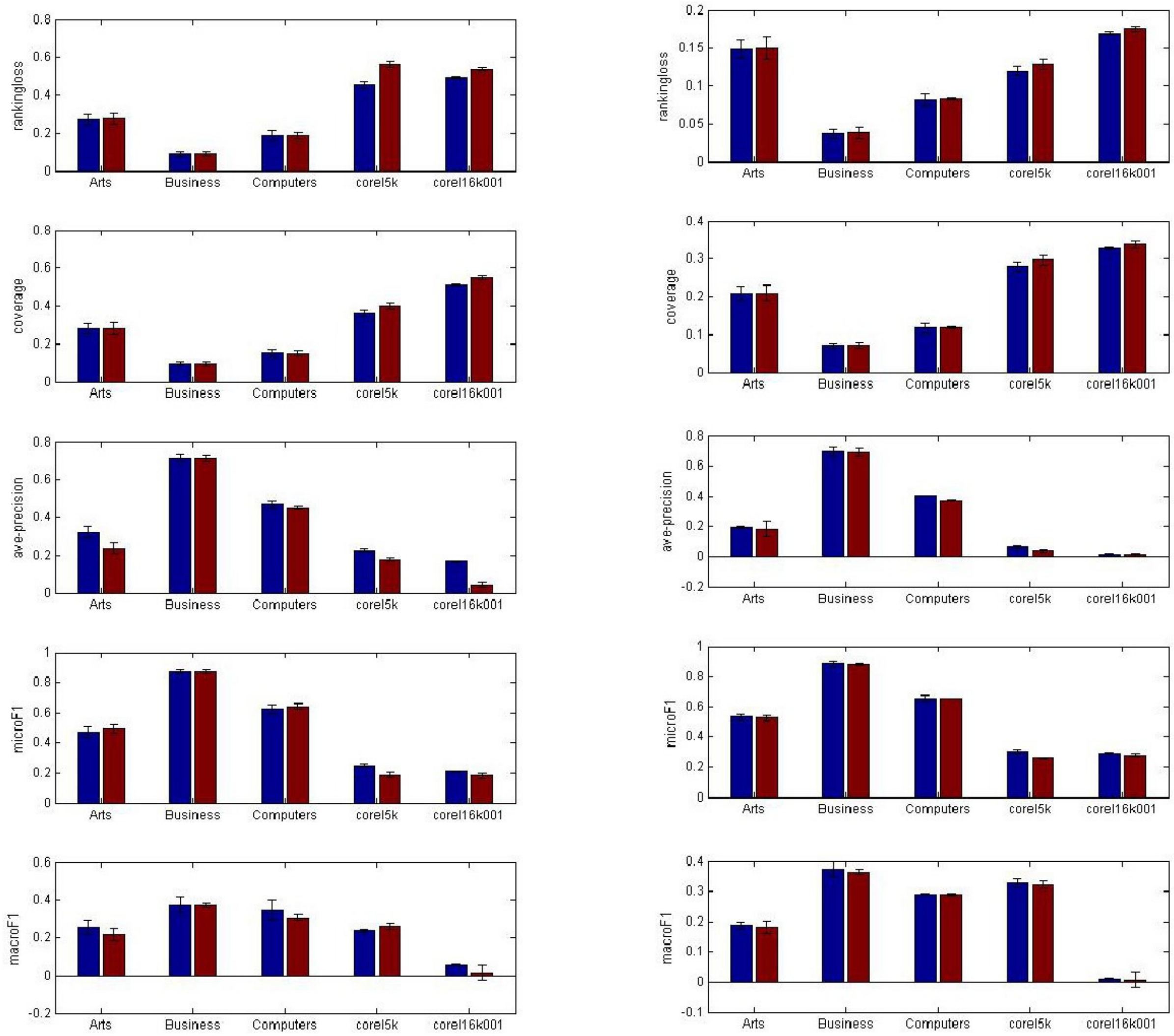
- Lm [10]: Based on the maximum margin output coding formulation [12], Lm learns the distance metric by maximizing the margin of embedded feature vectors and labeling vectors.
- Nje [11]: Based on the Jaccard distance between labeling vectors, Nje learns the distance metric by preserving the similarity of instances in the embedded feature space w.r.t. the labeling Jaccard distance.

Given the multi-label classification model $\mathcal{A}$ ($\mathcal{A} \in \{k$NN, Mlknn$\}$), its improved version by incorporating the learned distance metric is denoted as $\mathcal{A}$-Commu, $\mathcal{A}$-Lm and $\mathcal{A}$-Nje respectively.

Parameters suggested in the literatures are used to instantiate Lm and Nje. As shown in Table 1, the balancing parameter α, cost parameter C and thresholding parameter θ for Commu are chosen among $\{0.1, 0.2, ..., 1\}$, $\{1, 2, ..., 10\}$ and $\{0.1, 0.2, ..., 1\}$ with cross-validation on the training set. In addition, the number of nearest neighbors used by *k*NN and Mlknn are set to be 10.

4.2 Experimental Results

In this paper, the classification performance is evaluated in terms of five popular multi-label evaluation criteria including *ranking loss*, *coverage*, *average precision*, *micro-F1* and *macro-F1* [1, 2]. For *ranking loss* and *coverage*, the smaller the criterion value the better the performance. For *average precision*, *micro-F1* and *macro-F1*, the greater the criterion value the better the performance. Tables 3-7 report the de-



(a) Performance of Commu (blue bar) and Commu-Pca (brown bar) based on *k*NN

(b) Performance of Commu (blue bar) and Commu-Pca (brown bar) based on Mlknn

Fig. 2 Performance of Commu and Commu-Pca based on *k*NN and Mlknn in terms of *ranking loss*, *coverage*, *average precision*, *macro-F1* and *micro-F1* (top to bottom) on five data sets.

tailed experimental results of each comparing approach in terms of *ranking loss*,*coverage*, *average precision*, *micro-F1* and *macro-F1* when the learned distance metric is incorporated with *k*NN and Mlknn for multi-label prediction. On each data set, ten-fold cross-validation is performed where the mean criterion value as well as the standard deviation are recorded.[4)]

Given the experimental data set and evaluation criterion, pairwise *t*-test at 0.05 significance level is conducted to show whether the performance of Commu is significantly different to the comparing approaches. Table 8 summarizes the win/tie/loss counts between Commu and the comparing approaches in terms of each evaluation criterion.

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

- Across all evaluation metrics, *k*NN-Commu ranks 1st in 53.8% cases and ranks 2nd in 28.8% cases while Mlknn-Commu ranks 1st in 52.5% cases and ranks 2nd in 31.3% cases. It is impressive that whenever *k*NN or Mlknn are utilized to make multi-label prediction, their counterpart versions (*k*NN-Commu or Mlknn-Commu) always achieve significantly better or at least comparable performance

4) One exception is the *eurlex* dataset from extreme multi-label classification repository, where the predefined training and testing split are used for performance evaluation.

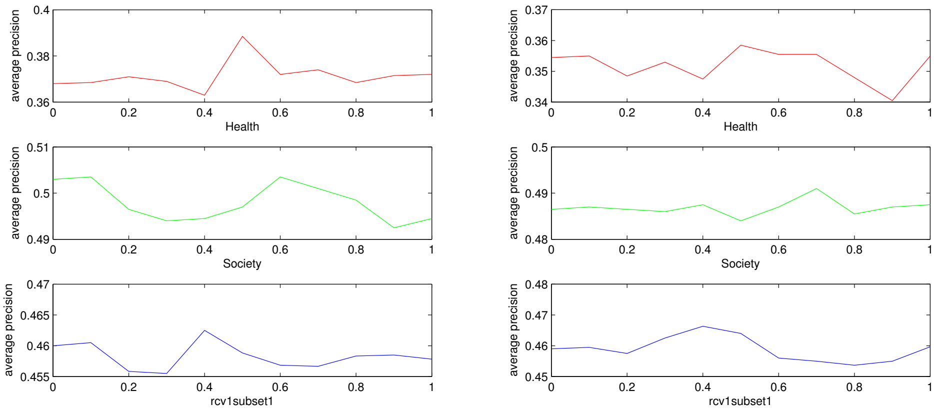


Fig. 3 Performance of COMMU (in terms of *average precision*) changes with varying value of parameter α based on kNN (left column) and MLKNN (right column) on three data sets.

after employing the learned distance metric (Table 8).

- As shown in Table 8, compared to LM, COMMU can lead to superior performance in 64.0% cases for kNN (kNN-COMMU against kNN-LM) and 49.3% cases for MLKNN (MLKNN-COMMU against MLKNN-LM). Compared to NJE, COMMU can lead to superior performance in 58.0% cases for kNN (kNN-COMMU against kNN-NJE) and 76.0% cases for MLKNN (MLKNN-COMMU against MLKNN-NJE).
- As shown in Tables 3, 4, 5, 6,7, the performance advantage of COMMU is more pronounced than the comparing approaches on data sets with larger number of class labels (i.e. `corel5k`, `rcv1-subset1`, `corel16k`, `eurlex-dc` and `eurlex-sm`). This desirable merit might be attributed to the compostional nature of the distance metric employed by COMMU, where the L_1 regularization term $||\boldsymbol{w}||_1$ in Eq.(6) can help identify component bases whose corresponding class labels do bring beneficial information for distance metric learning.
- As shown in Tables 3-7, the performance of kNN-NJE and MLKNN-NJE are not reported on data sets at large scale ($|S| \geq 5,000$) due to its quadratic training complexity w.r.t. the number of training examples. Specifically, let m, q and d denote the number of training examples, number of class labels and number of features, the training complexities for COMMU, LM and NJE correspond to $O((d + q^2)m)$, $O(q^3 + mdq^2)$ and $O(m^2 q + qdm\log(m))$ respectively. For LM, the main computation is the *SVD* operation of the PSD matrix in each iteration. For NJE, firstly the target vectors is solved at $O(m^2 t)$ (t is the dimension of the target vector which is unfixed) and then the embedder is learned at $O(tdmlog(m))$. For COMMU, to achieve an ϵ-solution, the number of iterations needed by *FISTA* update is $O(\frac{1}{\sqrt{\epsilon}})$. At each iteration, projections onto the positive semi-definite cone are performed to solve the coefficient vector $\mathbf{w}$. Therefore, the training stage complexity for each iteration is $O((d+q^2)mk\tilde{k})$ with k, $\tilde{k}$ being the values specified in Eq.(5).

4.3 Further Analysis

4.3.1 Effectiveness of Component Bases Generation

We further investigate the effectiveness of COMMU's strategy in generating component bases by encoding discriminative information in label space (Eq.(2)). Specifically, we derive a variant of COMMU (COMMU-PCA) by setting the component bases to the principal components yielded with top q eigenvalues by conducting PCA over the training instances. Figure 2 compares the performance of COMMU and COMMU-PCA based on kNN and MLKNN in terms of *ranking loss*, *coverage*,*average precision*, *macro-F1* and *micro-F1* on five data sets, which clearly show the benefits of exploiting discriminative information in generating component bases for COMMU.

4.3.2 Parameter Sensitivity

The parameter α in Eq.(3) represents the relative contributions from label space and instance space in calculating the semantic similarity. In Figure 3, the performance of COMMU (in terms of *average precision*) on three data sets are illustrated as the parameter α increases from 0 to 1 with stepsize 0.1 (left column: kNN; right column: MLKNN). It is obvious that the parameter setting of α has significant influence on classification performance of the COMMU approach. Therefore, the value of α are chosen are chosen among $\{0.1, 0.2, ..., 1\}$

Table 9 Training time of comparing algorithms on five data sets (in seconds).

	*k*NN-				MLKNN-			
Training time	COMMU	LM	NJE	ORIGINAL	COMMU	LM	NJE	ORIGINAL
Arts	960.583	106.042	383.512	18.069	1005.790	176.966	347.050	84.309
Business	650.475	123.689	377.070	17.122	690.933	188.033	378.474	76.942
Computers	2699.429	211.573	435.263	36.847	2872.315	478.411	646.539	298.919
yeast	93.605	50.947	364.206	5.519	102.203	71.077	365.362	18.366
genbase	409.308	16.15	182.503	9.158	484.188	130.709	295.765	122.345
corel16k	25341.622	23128.438	-	18585.445	16032.433	8695.786	-	4111.488

with cross-validation on the training set in the experimental studies.

4.3.3 Cost of Training Time

Table 9 reports the training time of comparing algorithms on five data sets. For COMMU, the cost of training time is generally higher than ORIGINAL and comparable to LM and NJE.

5 Related Works

In Section 4, the performance of COMMU is compared against LM and NJE which to the best of our knowledge are the only two available works on multi-label metric learning. LM [10] adapts the maximum margin output coding formulation [12] for distance metric learning, where the encoding projections are optimized by maximizing the margin of embedded feature vectors and labeling vectors. NJE [11] learns the distance metric by preserving pairwise similarity of labeling vectors in the embedded feature space, where the Jaccard distance is utilized for similarity measurement. Other than the single instance representation, there have been some works on distance metric learning for multi-instance multi-label data [18, 30, 31].

Exploitation of label correlations plays a key role for the success of multi-label classification, where numerous multi-label learning techniques have been proposed by considering different orders of label correlations [1, 2, 32]. Full-order label correlations are considered by LM via linear projection of the labeling vector, while first-order label correlations are considered by NJE via bitwise Jaccard distance measurement. For the proposed COMMU approach, label correlations are brought into the compositional structure of distance metric with label-dependent component bases.

Distance metric learning plays an important role in real-world applications (such as Person Re-ID [33]) in measuring similarity between objects. Generally, distance metric learning can be viewed as feature manipulation techniques where the embedded feature vector $\mathbf{V}^\top \boldsymbol{x}$ naturally follows from the mapping induced by the learned PSD matrix $\mathbf{M} = \mathbf{V}\mathbf{V}^\top$. Correspondingly, dimensionality reduction serves as the most popular techniques for manipulating multi-label features [34, 35]. There are some other strategies to manipulate the feature space for multi-label learning such as label-specific features [19, 20, 36, 37], meta-level features [38, 39] and multi-view features [40–43].

6 Conclusion

In this paper, the problem of distance metric learning for multi-label classification is studied. A novel multi-label metric learning approach named COMMU is proposed, which assumes compositional representation for distance metric. Specifically, component bases as well as triplet constraints are generated by exploiting semantic similarity in label space, and the resulting optimization problem is iteratively solved with linear complexity w.r.t. the number of training examples. Theoretical analysis as well as extensive experiments clearly validate the effectiveness of the proposed compositional distance metric for multi-label classification.

In the future, it is interesting to leverage auxiliary information such as domain knowledge [44] to facilitate multi-label distance metric learning. Furthermore, it is worthwhile to investigate strategies of combining distance metric learning with other popular mechanisms such as feature selection [45–47] for multi-label classification.

Appendix

Given a multi-label dataset $\mathcal{S} = \{z = (\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ drawn i.i.d. from a distribution P over the labelled space $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, where the label vector y_i simultaneously contains multiple labels. Assume that $||\boldsymbol{x}|| \leq R$ (for some convenient norm), $\forall \boldsymbol{x} \in \mathcal{X}$. Different from the single-label setting, COMMU defines the multi-label semantic similarity matrix to construct the triplet (z, z', z''), where $\boldsymbol{y}$ is similar to $\boldsymbol{y}'$ and dissimilar to $\boldsymbol{y}''$. Let $\mathcal{S}_R$ be the set of all admissible triplets built from instances in $\mathcal{S}$.

Let $L(h, z, z', z'')$ be the loss suffered by some hypothesis h on triplet (z, z', z'') with the convention that L returns 0 for non-admissible triplets. Assume L to be uniformly upper-bounded by a constant U. The empirical loss $R^{\mathcal{S}_R}_{emp}(h)$ of h on $\mathcal{S}_R$ is defined as

$$R^{\mathcal{S}_R}_{emp}(h) = \frac{1}{|\mathcal{S}_R|} \sum_{(z,z',z'')\in\mathcal{S}_R} L(h, z, z', z''),$$

and its expected loss $R(h)$ over distribution P as

$$R(h) = \mathbb{E}_{z,z',z''\sim P} L(h, z, z', z'').$$

The goal of the theoretical analysis is to bound the deviation between $R(\mathcal{A}_{\mathcal{S}_R})$ and $R^{\mathcal{S}_R}_{emp}(\mathcal{A}_{\mathcal{S}_R})$, where $\mathcal{A}_{\mathcal{S}_R}$ is the hypothesis learned by algorithm $\mathcal{A}$ on $\mathcal{S}_R$.

Theoretical Basis

To derive the generalization bounds of Commu,we use the recent framework of algorithmic robustness in metric learning [16,27]. Algorithmic robustness is the ability of an algorithm to perform "similarly" on a training example and on a test example that are "close". The proximity of points is based on a partitioning of the space $\mathcal{Z}$: two examples are close to each other if they lie in the same region. The partition is based on the notion of covering number.

Definition 1 (Covering number). For a metric space $(\mathcal{M}, \rho)$ and $\nu \subset \mathcal{M}$, we say that $\hat{\nu} \subset \nu$ is a γ-cover of ν if $\forall \boldsymbol{t} \in \nu$, $\exists \hat{\boldsymbol{t}} \in \hat{\nu}$ such that $\rho(\boldsymbol{t}, \hat{\boldsymbol{t}}) \leq \gamma$. The γ-covering number of ν is

$$\mathcal{N}(\gamma, \mathcal{X}, \rho) = min\ \{|\hat{\nu}| :\ \hat{\nu}\, is\, a\, \gamma - cover\, of\, \nu\}.$$

In particular, when $\mathcal{X}$ is compact, $\mathcal{N}(\gamma, \mathcal{X}, \rho)$is finite, leading to a finite cover. Then, $\mathcal{Z}$ can be partitioned into $|\mathcal{Y}|\, \mathcal{N}(\gamma, \mathcal{X}, \rho)$ subsets such that if two examples $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{z}' = (\boldsymbol{x}', \boldsymbol{y}')$ belong to the same subset, then $\boldsymbol{y} = \boldsymbol{y}'$ and $\rho(\boldsymbol{x}, \boldsymbol{x}') \leq \gamma$. The definition of robustness for tripletwise loss functions is as follows.

Definition 2 (Robustness for metric learning) [27]. An algorithm $\mathcal{A}$ is $(N, \epsilon(\cdot))$ robust for $N \in \mathbb{N}$ and $\epsilon(\cdot) : (\mathcal{Z} \times \mathcal{Z})^n \to \mathbb{R}$ if $\mathcal{Z}$ can be partitioned into N disjoints sets, denoted by $\{Q_i\}_{i=1}^N$, such that the following holds for all $\mathcal{S} \in \mathcal{Z}^n$: $\forall (z_1, z_2, z_3) \in \mathcal{S}_R, \forall z, z', z'' \in \mathcal{Z}, \forall i, j, k \in [N] : if\ z_1, z \in Q_i, z_2, z' \in Q_i, z_3, z'' \in Q_i$ then

$$\left|L(\mathcal{A}_{\mathcal{S}_R}, z_1, z_2, z_3) - L(\mathcal{A}_{\mathcal{S}_R}, z, z', z'')\right| \leq \epsilon(\mathcal{S}_R),$$

where $\mathcal{A}_{\mathcal{S}_R}$ is the hypothesis learned by $\mathcal{A}$ on $\mathcal{S}_R$.

N and $\epsilon(\cdot)$ quantify the robustness of the algorithm and depend on the training data. The work [27] showed that a metric learning algorithm that satisfies Definition 2 has the following generalization guarantees.

Theorem 2. If a learning algorithm $\mathcal{A}$ is $(N, \epsilon(\cdot))$–robust and the training data consists of the triplets $\mathcal{S}_R$ obtained from a sample $\mathcal{S}$ generated by n i.i.d draws from P, then for any $\delta > 0$, with probability at least $1 - \delta$ we have :

$$\left|L(\mathcal{A}_{\mathcal{S}_R}, z_1, z_2, z_3) - L(\mathcal{A}_{\mathcal{S}_R}, z, z', z'')\right| \leq \epsilon(\mathcal{S}_R) + 3U\sqrt{\frac{2N\ln 2 + 2\ln\frac{1}{\delta}}{n}}.$$

Additionally, as shown in [27], the following theorem, which basically says that if a metric learning algorithm has approximately the same loss for triplets that are close to each other and then it is robust, can be used to determine the robustness of the algorithm more conveniently.

Theorem 3. Fix $\gamma > 0$ and a metric ρ of $\mathcal{Z}$. Suppose that $\forall z_1, z_2, z_3, z, z', z'' : (z_1, z_2, z_3) \in \mathcal{S}_R, \rho(z_1, z) \leq \gamma, \rho(z_2, z') \leq \gamma, \rho(z_3, z'') \leq \gamma$, $\mathcal{A}$ satisfies

$$\left|L(\mathcal{A}_{\mathcal{S}_R}, z_1, z_2, z_3) - L(\mathcal{A}_{\mathcal{S}_R}, z, z', z'')\right| \leq \epsilon(\mathcal{S}_R),$$

and $\mathcal{N}(\frac{\gamma}{2}, \mathcal{Z}, \rho) < \infty$. Then the algorithm $\mathcal{A}$ is $(\mathcal{N}(\frac{\gamma}{2}, \mathcal{Z}, \rho), \epsilon(\mathcal{S}_R))$-robust.

Generalization Bounds for Commu

To derive the generalization bound of Commu, the main work is to prove its robustness, which contains the computation for N and $\epsilon(\mathcal{S}_R)$.

The loss function of Commu is defined as:

$$L(\boldsymbol{w}, z, z', z'') = \left[\Delta(\boldsymbol{y}', \boldsymbol{y}'') + d_w(\boldsymbol{x}, \boldsymbol{x}') - d_w(\boldsymbol{x}, \boldsymbol{x}'')\right]_+.$$

Let $\boldsymbol{w}^*$ be the optimal solution to Commu. By optimality of $\boldsymbol{w}^*$ we have:

$$L(\boldsymbol{w}^*, z, z', z'') + \frac{1}{C}\|\boldsymbol{w}^*\|_1 \leq L(\boldsymbol{0}, z, z', z'') + \frac{1}{C}\|\boldsymbol{0}\|_1 = \Delta(\boldsymbol{y}', \boldsymbol{y}'') = \boldsymbol{y}'^T\boldsymbol{y}'' - \boldsymbol{y}'^T G\boldsymbol{y}'',$$

where the second item $\boldsymbol{y}'^T G\boldsymbol{y}'' \leq \theta$ because of the dissimilarity between $\boldsymbol{y}'$ and $\boldsymbol{y}''$. Let $\Delta_0 = \boldsymbol{y}'^T\boldsymbol{y}''$, thus $\Delta_0 - \theta \leq \Delta(\boldsymbol{y}', \boldsymbol{y}'') \leq \Delta_0$ and $\|\boldsymbol{w}^*\|_1 \leq \Delta_0 C$.

$M^* = \sum_{i=1}^K w_i^* \boldsymbol{b}_i \boldsymbol{b}_i^T$ is the corresponding metric. The norm of the basis element $\boldsymbol{b}_i$ is bounded by 1. Based on Holder's inequality and the bound on $\boldsymbol{w}^*$ and $\boldsymbol{b}$'s, the bound for M^* is derived.

$$\|\boldsymbol{M}^*\|_1 = \left\|\sum_{i=1}^K w_i^* \boldsymbol{b}_i \boldsymbol{b}_i^T\right\|_1 = \left\|\sum_{i:w_i\neq 0}^K w_i^* \boldsymbol{b}_i \boldsymbol{b}_i^T\right\|_1 \leq \|\boldsymbol{w}^*\|_1 \sum_{i:w_i\neq 0} \|\boldsymbol{b}_i\|_\infty \|\boldsymbol{b}_i\|_\infty \leq K^* \Delta_0 C,$$

$$
\begin{aligned}
&\left|[\Delta(\boldsymbol{y}_2,\boldsymbol{y}_3)+d_w^*(\boldsymbol{x}_1,\boldsymbol{x}_2)-d_w^*(\boldsymbol{x}_1,\boldsymbol{x}_3)]_+-[\Delta(\boldsymbol{y}_2',\boldsymbol{y}_3')+d_w^*(\boldsymbol{x}_1',\boldsymbol{x}_2')-d_w^*(\boldsymbol{x}_1',\boldsymbol{x}_3')]_+\right|\\
\leq&\left|\Delta(\boldsymbol{y}_2,\boldsymbol{y}_3)-\Delta(\boldsymbol{y}_2',\boldsymbol{y}_3')+d_w^*(\boldsymbol{x}_1,\boldsymbol{x}_2)-d_w^*(\boldsymbol{x}_1',\boldsymbol{x}_2')+d_w^*(\boldsymbol{x}_1',\boldsymbol{x}_3')-d_w^*(\boldsymbol{x}_1,\boldsymbol{x}_3)\right|\\
\leq&\left|\Delta(\boldsymbol{y}_2,\boldsymbol{y}_3)-\Delta(\boldsymbol{y}_2',\boldsymbol{y}_3')\right|+\\
&\left|(\boldsymbol{x}_1-\boldsymbol{x}_2)^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_2)+(\boldsymbol{x}_1-\boldsymbol{x}_2)^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_2')-(\boldsymbol{x}_1-\boldsymbol{x}_2)^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_2')-(\boldsymbol{x}_1'-\boldsymbol{x}_2')^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_2')\right|+\\
&\left|(\boldsymbol{x}_1'-\boldsymbol{x}_3')^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_3')-(\boldsymbol{x}_1'-\boldsymbol{x}_3')^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_3)+(\boldsymbol{x}_1'-\boldsymbol{x}_3')^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_3)-(\boldsymbol{x}_1-\boldsymbol{x}_3)^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_3)\right|\\
=&\left|\Delta(\boldsymbol{y}_2,\boldsymbol{y}_3)-\Delta(\boldsymbol{y}_2',\boldsymbol{y}_3')\right|+\left|(\boldsymbol{x}_1-\boldsymbol{x}_2)^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_2-(\boldsymbol{x}_1'-\boldsymbol{x}_2'))+(\boldsymbol{x}_1-\boldsymbol{x}_2-(\boldsymbol{x}_1'-\boldsymbol{x}_2'))^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_2')\right|+\\
&\left|(\boldsymbol{x}_1'-\boldsymbol{x}_3')^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_3'-(\boldsymbol{x}_1-\boldsymbol{x}_3))+(\boldsymbol{x}_1'-\boldsymbol{x}_3'-(\boldsymbol{x}_1-\boldsymbol{x}_3))^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_3)\right|\\
\leq&\left|\Delta(\boldsymbol{y}_2,\boldsymbol{y}_3)-\Delta(\boldsymbol{y}_2',\boldsymbol{y}_3')\right|+\left|(\boldsymbol{x}_1-\boldsymbol{x}_2)^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_1')\right|+\left|(\boldsymbol{x}_1-\boldsymbol{x}_2)^TM^*(\boldsymbol{x}_2'-\boldsymbol{x}_2)\right|+\\
&\left|(\boldsymbol{x}_1-\boldsymbol{x}_1')^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_2')\right|+\left|(\boldsymbol{x}_2'-\boldsymbol{x}_2)^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_2')\right|+\left|(\boldsymbol{x}_1'-\boldsymbol{x}_3')^TM^*(\boldsymbol{x}_1'-\boldsymbol{x}_1)\right|+\\
&\left|(\boldsymbol{x}_1'-\boldsymbol{x}_3')^TM^*(\boldsymbol{x}_3-\boldsymbol{x}_3')\right|+\left|(\boldsymbol{x}_1'-\boldsymbol{x}_1)^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_3)\right|+\left|(\boldsymbol{x}_3-\boldsymbol{x}_3')^TM^*(\boldsymbol{x}_1-\boldsymbol{x}_3)\right|\\
\leq&\left|\Delta(\boldsymbol{y}_2,\boldsymbol{y}_3)-\Delta(\boldsymbol{y}_2',\boldsymbol{y}_3')\right|+\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_\infty\|M^*\|_1\left\|\boldsymbol{x}_1-\boldsymbol{x}_1'\right\|_1+\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_\infty\|M^*\|_1\left\|\boldsymbol{x}_2'-\boldsymbol{x}_2\right\|_1+\\
&\left\|\boldsymbol{x}_1-\boldsymbol{x}_1'\right\|_1\|M^*\|_1\left\|\boldsymbol{x}_1'-\boldsymbol{x}_2'\right\|_\infty+\left\|\boldsymbol{x}_2'-\boldsymbol{x}_2\right\|_1\|M^*\|_1\left\|\boldsymbol{x}_1'-\boldsymbol{x}_2'\right\|_\infty+\left\|\boldsymbol{x}_1'-\boldsymbol{x}_3'\right\|_\infty\|M^*\|_1\left\|\boldsymbol{x}_1'-\boldsymbol{x}_1\right\|_1+\\
&\left\|\boldsymbol{x}_1'-\boldsymbol{x}_3'\right\|_\infty\|M^*\|_1\left\|\boldsymbol{x}_3-\boldsymbol{x}_3'\right\|_1+\left\|\boldsymbol{x}_3-\boldsymbol{x}_3'\right\|_1\|M^*\|_1\|\boldsymbol{x}_1-\boldsymbol{x}_3\|_\infty+\left\|\boldsymbol{x}_1'-\boldsymbol{x}_1\right\|_1\|M^*\|_1\|\boldsymbol{x}_1-\boldsymbol{x}_3\|_\infty\\
\leq&16\gamma RK^*\Delta_0C+\theta
\end{aligned}
$$

where $K^* \leq K$ is the number of nonzero entries in w^*.

According to Definition 1, $\mathcal{Z}$ can be partitioned into $2^q\mathcal{N}(\gamma,\mathcal{X},\rho)$ subsets, where q is the size of label vector. Commu constructs the triplet (z,z',z'') by the multi-label semantic similarity matrix. For $z_1,z_2,z_3,z_1',z_2',z_3'' \in \mathcal{Z}$, if $\boldsymbol{y}_1$ is similar to $\boldsymbol{y}_1'$, $\left\|x_1-x_1'\right\|_1 \leq \gamma$, $\boldsymbol{y}_2$ is similar to $\boldsymbol{y}_2'$, $\left\|x_2-x_2'\right\|_1 \leq \gamma$, $\boldsymbol{y}_3$ is similar to $\boldsymbol{y}_3'$, $\left\|x_3-x_3'\right\|_1 \leq \gamma$, then (z_1,z_2,z_3) and (z_1',z_2',z_3'') are either both admissible or non-admissible triplets.

In the non-admissible case, it can be seen from definition that their respective loss is 0 and so is the deviation between the losses. In the admissible case we have the above result, by the property that the hinge loss is 1-Lipschitz (the first $\leq$), Holder's inequality(the 2th-4th $\leq$) and $\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|_\infty \leq 2R$ ($\|\boldsymbol{x}\| \leq R, \forall \boldsymbol{x}\in\mathcal{X}$) , $\left|\Delta(\boldsymbol{y}_2,\boldsymbol{y}_3)-\Delta(\boldsymbol{y}_2',\boldsymbol{y}_3')\right| \leq \theta$ ($\Delta_0-\theta \leq \Delta(\boldsymbol{y}',\boldsymbol{y}'') \leq \Delta_0$) in the last $\leq$. Thus Commu is $(2^q\mathcal{N}(\gamma,\mathcal{X},\|\cdot\|_1), 16\gamma RK^*\Delta_0C+\theta)$-robust and the generalization bound follows.

$$
\left|R(\mathcal{A}_{S_R})-R_{emp}^{S_R}(\mathcal{A}_{S_R})\right| \leq 16\gamma RK^*\Delta_0C+\theta+
3U\sqrt{\frac{2^{q+1}\mathcal{N}(\gamma,\mathcal{X},\|\cdot\|_1)\ln 2+2\ln\frac{1}{\delta}}{n}}.
$$

References

1. 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.
2. E. Gibaja and S. Ventura. A tutorial on multilabel learning. *ACM Computing Surveys*, 47(3):Article 52, 2015.
3. F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012.
4. R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):121–135, 2015.
5. J. Liu, W.-C. Chang, Y. Wu, and Y. Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124, Tokyo, Japan, 2017.
6. X. Pan, Y.-X. Fan, J. Jia, and H.-B. Shen. Identifying rna-binding proteins using multi-label deep learning. *Science China Information Sciences*, page 62:19103, 2019.
7. L. Sun, H. Ge, and W. Kang. Non-negative matrix factorization based modeling and training algorithm for multi-label learning. *Frontiers of Computer Science*, 13(6):1243–1254, 2019.
8. A. Bellet, A. Habrard, and M. Sebban. Metric learning. *Synthesis*

Lectures on Artificial Intelligence and Machine Learning, 9(1):1–151, 2015.

9. F. Wang and J. Sun. Survey on distance metric learning and dimensionality reduction in data mining. *Data Mining and Knowledge Discovery*, 29(2):534–564, 2015.
10. W. Liu and I. W. Tsang. Large margin metric learning for multi-label prediction. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2800–2806, Austin, TX, 2015.
11. H. Goukand B. Pfahringer and M. 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.
12. Y. Zhang and J. Schneider. Maximum margin output coding. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1575–1582, Edinburgh, UK, 2012.
13. Y. Verma and C. V. Jawahar. Image annotation by propagating labels from semantic neighbourhoods. *International Journal of Computer Vision*, 121(1):126–148, 2017.
14. H Gouk, B. Pfahringer, and M. Cree. Learning similarity metrics by factorising adjacency matrices. *CoRR*, abs/1511.06442, 2015.
15. J. Ni, J. Liu, C. Zhang, D. Ye, and Z. Ma. Fine-grained patient similarity measuring using deep metric learning. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, pages 1189–1198, Singapore, 2017.
16. Y. Shi, A. Bellet, and F. Sha. Sparse compositional metric learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 2078–2084, Québec City, Canada, 2014.
17. J. St.Amand and J. Huan. Sparse compositional local metric learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1097–1104, Halifax, Canada, 2017.
18. Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.
19. M.-L. Zhang and L. Wu. LIFT: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):107–120, 2015.
20. J. Huang, G. Li, Q. Huang, and X. Wu. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3309–3323, 2016.
21. K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.
22. S.-J. Huang and Z.-H. Zhou. Multi-label learning by exploiting label correlations locally. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 949–955, Toronto, Canada, 2012.
23. Y. Zhu, J. Kwok, and Z.-H. Zhou. Multi-label learning with global and local correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2018.
24. G.-X. Yuan, C.-H. Ho, and C.-J. Lin. An improved GLMNET for L1-regularized logistic regression. *Journal of Machine Learning Research*, 13:1999–2030, 2012.
25. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *Siam Journal on Imaging Sciences*, 2(1):183–202, 2009.
26. K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6(3):615–640, 2010.
27. A. Bellet and A. Habrard. Robustness and generalization for metric learning. *Neurocomputing*, 151(14):259–267, 2015.
28. J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.
29. M.-L. Zhang and Z.-H. Zhou. ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.
30. J. Rong, S. Wang, and Z.-H. Zhou. Learning a distance metric from multi-instance multi-label data. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 896–902, Miami, FL, 2009.
31. Y. Verma and C. V. Jawahar. A robust distance with correlated metric learning for multi-instance multi-label data. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 441–445, Amsterdam, The Netherlands, 2016.
32. M.-L. Zhang, Y.-K. Li, Y.-Y. Liu, and X. Geng. Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science*, 12(2):191–202, 2018.
33. Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang. Progressive learning for person re-identification with one example. *IEEE Transactions on Image Processing*, 28(6):2872–2881, 2019.
34. L. Sun, S. Ji, and J. Ye. *Multi-label Dimensionality Reduction*. Chapman and Hall/CRC, Boca Ration, FL, 2013.
35. R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review*, 49(1):57–78, 2018.
36. J. Zhang, C. Li, D. Cao, Y. Lin, S. Su, L. Dai, and S. Li. Multi-label learning with label-specific features by resolving label correlations. *Knowledge-Based Systems*, 159:148–157, 2018.
37. Z.-S. Chen and M.-L. Zhang. Multi-label learning with regularization enriched label-specific features. In *Proceedings of the 11th Asian Conference on Machine Learning*, pages 411–424, Nagoya, Japan, 2019.
38. Y. Yang and S. Gopal. Multilabel classification with meta-level features in a learning-to-rank framework. *Machine Learning*, 88(1-2):47–68, 2012.
39. S. Canuto, M. A. Gonçalves, and F. 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.
40. X. Zhu, X. Li, and S. Zhang. Block-row sparse multiview multilabel learning for image classification. *IEEE Transactions on Cybernetics*, 46(2):450–461, 2016.
41. C. Zhang, Z. Yu, Q. Hu, P. Zhu, X. Liu, and X. Wang. Latent semantic aware multi-view multi-label classification. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4414–4421, New Orleans, LA, 2018.
42. X. Wu, Q.-G. Chen, Y. Hu, D.-B. Wang, X. Chang, X. Wang, and M.-L.

Zhang. Multi-view multi-label learning with view-specific information extraction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3884–3890, Macau, China, 2019.
43. R. Zhang, F. Nie, X. Li, and X. Wei. Feature selection with multi-view data: A survey. *Information Fusion*, 50:158–167, 2019.
44. Z.-H. Zhou. Abductive learning: Towards bridging machine learning and logical reasoning. *Science China Information Sciences*, page 62:076101, 2019.
45. Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Transactions on Multimedia*, 15(3):661–669, 2013.
46. R. Zhang, F. Nie, and X. Li. Self-weighted supervised discriminative feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3913–3918, 2018.
47. R. Zhang, F. Nie, Y. Wang, and X. Li. Unsupervised feature selection via adaptive multimeasure fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2886–2892, 2019.