Beyond Similar and Dissimilar Relations: A Kernel Regression Formulation for Metric Learning

Pengfei Zhu, Ren Qi, Qinghua Hu*, Qilong Wang, Changqing Zhang, Liu Yang
School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
zhupengfei@tju.edu.cn, huqinghua@tju.edu.cn

Abstract

Most existing metric learning methods focus on learning a similarity or distance measure relying on similar and dissimilar relations between sample pairs. However, pairs of samples cannot be simply identified as similar or dissimilar in many real-world applications, e.g., multi-label learning, label distribution learning and tasks with continuous decision values. To this end, in this paper we propose a novel relation alignment metric learning (RAML) formulation to handle the metric learning problem in those scenarios. Since the relation of two samples can be measured by the difference degree of the decision values, motivated by the consistency of the sample relations in the feature space and decision space, our proposed RAML utilizes the sample relations in the decision space to guide the metric learning in the feature space. In this way, our RAML method formulates metric learning as a kernel regression problem, which can be efficiently optimized by the standard regression solvers. We carry out several experiments on the single-label classification, multi-label classification, and label distribution learning tasks, to demonstrate that our method achieves favorable performance against the state-of-the-art methods.

1 Introduction

In many computer vision and pattern recognition tasks, *e.g.*, face recognition [Guillaumin *et al.*, 2009], image classification [Mensink *et al.*, 2012], and person re-identification [Liao *et al.*, 2015], it is crucial to learn a discriminative distance metric to measure the similarity between pairs of samples. Intuitively, metric learning aims to learn a discriminative similarity or dissimilarity metric by pushing the dissimilar samples away and pulling the similar samples together. Typical distance metrics include Euclidean distance, cosine distance, and Mahalanobis distance. Most existing metric learning methods focus on learning a discriminative Mahalanobis distance. Beyond Mahalanobis distance, generalized distance metric learning methods are presented by learning high-order discriminant functions [Li *et al.*, 2012].

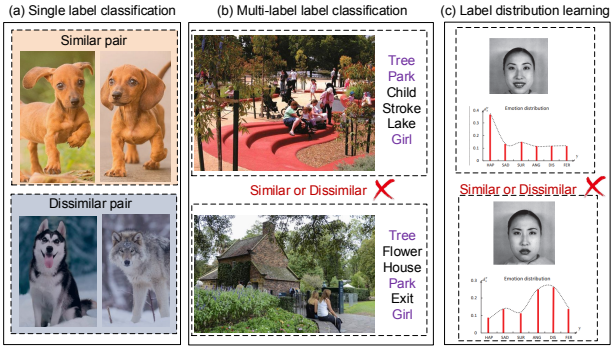

Figure 1: Examples of sample pairs in different learning tasks. (a) In single-label classification, two images of dogs construct a similar pair while one image of wolf and the other image of Husky construct a dissimilar pair. (b) For multi-label classification, one scene image (*tree, park, child, stroke, lake, girl*) and the other scene image (*tree, flower, house, park, exit, girl*) are hard to be identified as similar or dissimilar. (c) For label distribution learning, the labels of two face images are label distributions rather than discrete labels. Two face images are difficult to be identified as similar or dissimilar as well.

According to the availability of the label information, metric learning can be partitioned into three categories, *i.e.*, the unsupervised, semi-supervised and supervised methods. To deal with the heterogeneous data, multi-modal [McFee and Lanckriet, 2011] and cross-modal [Wang *et al.*, 2016] metric learning algorithms are developed. Because of the diversity of the feature space, linear, kernel and tensor distance metrics are learned for different data structures. Different from shallow metric learning, deep learning based methods learn the feature and metric jointly and achieve superior performance [Oh Song *et al.*, 2016].

One of the key steps in existing metric learning methods is to generate doublet [Davis *et al.*, 2007], triplet [Weinberger and Saul, 2009] or even quadruplet [Law *et al.*, 2013] constraints using the label information. Doublet constraints are the most commonly used in metric learning methods. Similar and dissimilar sample pairs are generated in the k-nearest neighbors or ε-neighborhood by measuring whether two samples belong to the same class. In some applications, *e.g.*, weakly supervised learning [Mu *et al.*, 2010] or social networks [Shaw *et al.*, 2011], sample pairs are generated from

*The corresponding author

connectivity information or other side information. Generally, there are two sets of sample pairs, *i.e.*, one contains the similar sample pairs and the other one contains the dissimilar ones.

However, for some learning tasks, *e.g.*, multi-label learning [Zhang and Wu, 2015] and label distribution learning [Geng, 2016], relations between sample pairs cannot be simply identified as similar or dissimilar. Thus, the existing metric learning methods cannot work on the above tasks. As shown in Figure 1(a), for single-label classification, it is very simple to identify two images as similar or dissimilar. Figure 1(b) shows there are two scene images that share part of the labels (*tree*, *park*, and *girl*). The problem arises that it is difficult to classify two images into similar or dissimilar sample pair. Figure 1(c) is an example of label distribution learning, where decision space is modeled by label distributions rather than discrete labels. Above discussions encourage us to propose a generalized metric learning method, which can be flexibly adopted to various kinds of tasks.

In machine learning community, one of the basic assumptions is that samples should keep with the same relations in different spaces, especially in the feature space and label space. The principle of metric learning is to encourage samples in the feature space to satisfy the expected relations induced by supervised information. Manifold learning emphasizes locality preserving, which requires that the nearest neighbors of samples should be close to each other in the projected low-dimensional feature space [Yang *et al.*, 2006]. For kernel learning machines, the kernel matrix can be considered as the similarity relation of all samples. Kernel alignment exploits the similarity between kernel matrices for learning kernels [Cortes *et al.*, 2012] and matrix completion [Bhadra *et al.*, 2017]. For multi-modal learning, the sample relation in feature spaces of different modalities should be consistent with that in the label space. For metric learning, as long as the sample relations in the decision space are modeled, the distance metric can be learned by minimizing the difference between sample relations in feature space and decision space.

In this paper, we propose a metric learning formulation, namely relation alignment metric learning (RAML). Our RAML aims to restrain that sample relations measured by metric in the feature space tend to be consistent with those in the decision space during learning. The contributions of this paper are summarized as follows.

- A novel metric learning formulation is proposed to learn distance metrics for different learning tasks, including single-label learning, multi-label learning, and label distribution learning.
- The proposed RAML is formulated as a kernel regression model. Based on RAML formulation, two metric learning methods are developed, *i.e.*, support vector regression metric learning and ridge regression metric learning. Furthermore, we analyze the generalization error bound for the proposed metric learning methods.
- Experiments on single-label classification, multi-label classification and label distribution learning tasks show that our RAML method achieves superior performance against the state-of-the-art methods.

2 Problem Statement

Most existing metric learning methods focus on learning the Mahalanobis distance. Given two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, where $\mathbf{x}_i \in \mathbb{R}^d$ and d is the dimension of the feature space, the Mahalanobis distance between the two samples $\mathbf{x}_i$ and $\mathbf{x}_j$, is defined as

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) \quad (1)$$

where $\mathbf{M} \in \mathbb{R}^{\mathbf{d}\times\mathbf{d}}$ is a positive semi-definite (PSD) matrix. The parameter $\mathbf{M}$ can be predefined, *i.e.*, the inverse of the covariance matrix, or learned by distance metric learning methods.

Generally, metric learning enlarges the distances between dissimilar samples while reduces the distances between similar ones. Two samples with the same label form a positive pair, and ones with the different labels form a negative pair. In supervised learning, sample pairs are previously given in [Guillaumin *et al.*, 2009] or generated based on the available label information [Weinberger and Saul, 2009]. However, in some applications, *e.g.*, multi-label learning, and label distribution learning tasks, two samples cannot be simply classified as positive or negative ones. Let $\mathbf{y} = [y_1, ..., y_k, ..., y_c]$ be the label vector of $\mathbf{x}$. For the traditional single-label classification task, $\mathbf{x}$ can only belong to one class. If $\mathbf{x}$ belongs to the k-th class, $y_k = 1$; otherwise, $y_k = 0$. Different from the traditional single-label classification task, multi-label classification assumes that $\mathbf{x}$ can belong to multiple classes. For label distribution learning, there are two constraints for $\mathbf{y}$, *i.e.*, $y_k \geq 0, k = 1, 2, ...c$, and $\sum_{k=1}^{c} y_k = 1$. *Different from the existing metric learning methods only designed for traditional single-label classification task, we aims at designing a novel metric learning formulation to accommodate different learning tasks.*

3 A Kernel Regression Formulation

In this section, we present a kernel regression formulation for metric learning. Different from the existing metric learning methods, this method can be used for various kinds of tasks.

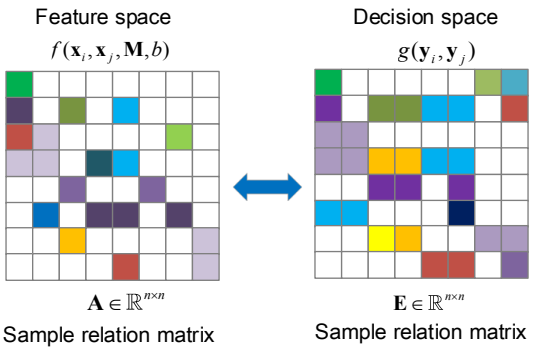

Figure 2: The consistency of the sample relations in feature and decision spaces. The sample relation matrix $\mathbf{E}$ is calculated via $g(\mathbf{x}_i, \mathbf{x}_j)$. Then we can use $\mathbf{E}$ to guide metric learning (*i.e.*, learning of $\mathbf{M}$, b) in the feature space to get a more discriminative sample relation $\mathbf{A}$.

3.1 Relation Alignment Learning

For metric learning, doublet constraint is a kind of description of relationship between a pair of samples in the decision space. As shown in Figure 2, $f(\mathbf{x}_i, \mathbf{x}_j, \mathbf{M}, b)$ is used to measure the sample relations in feature space while $g(\mathbf{y}_i, \mathbf{y}_j)$ is used to measure the sample relations in decision space. $g(\mathbf{y}_i, \mathbf{y}_j)$ is specially designed for different tasks. Let $\mathbf{A} \in \mathbb{R}^{\mathbf{n}\times\mathbf{n}}$ and $\mathbf{E} \in \mathbb{R}^{\mathbf{n}\times\mathbf{n}}$ be the sample relation matrix in feature and decision spaces, respectively. In general, sample relation in the feature space should be consistent with that in the decision space, *i.e.*,

$$\begin{bmatrix} a_{11} & \dots & a_{i1} & \dots & a_{n1} \\ \dots & \dots & \dots & \dots & \dots \\ a_{1i} & \dots & a_{ii} & \dots & a_{ni} \\ \dots & \dots & \dots & \dots & \dots \\ a_{1n} & \dots & a_{in} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} e_{11} & \dots & e_{i1} & \dots & e_{n1} \\ \dots & \dots & \dots & \dots & \dots \\ e_{1i} & \dots & e_{ii} & \dots & e_{ni} \\ \dots & \dots & \dots & \dots & \dots \\ e_{1n} & \dots & e_{in} & \dots & e_{nn} \end{bmatrix}$$

where a_{ij} and e_{ij} represent the sample relation of $\mathbf{x}_i$ and $\mathbf{x}_j$ in the feature space and decision space, respectively. Here, to keep consistency, we require that

$$f(\mathbf{x}_i, \mathbf{x}_j, \mathbf{M}, b) = g(\mathbf{y}_i, \mathbf{y}_j). \tag{2}$$

where $g(\mathbf{y}_i, \mathbf{y}_j)$ is the difference degree of two samples in the decision space. $g(\mathbf{y}_i, \mathbf{y}_j)$ reflects the sample relation in the decision space, and guides the learning of $(\mathbf{M}, b)$ in feature space.

$$\begin{aligned} f(\mathbf{x}_i, \mathbf{x}_j, \mathbf{M}, b) &= (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) + b \\ &= \langle \mathbf{M}, \mathbf{T}_{ij} \rangle + b \end{aligned} \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is defined as the Frobenius inner product of two matrices, b is the bias item, and $\mathbf{T}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$. Then we rewrite (2) to

$$g(\mathbf{y}_i, \mathbf{y}_j) = \langle \mathbf{M}, \mathbf{T}_{ij} \rangle + b \tag{4}$$

Once the relation function $g(\mathbf{y}_i, \mathbf{y}_j)$ is chosen, (4) can be considered as a linear regression problem. Hence, the metric learning problem is converted to solve a sample pair regression problem with the scaled second sample moment $\mathbf{T}_{ij}$ of sample pair $(\mathbf{x}_i, \mathbf{x}_j)$ as the input.

3.2 Sample Pair Kernel

To formulate the sample pair regression problem in (4), we introduce a 2-degree polynomial kernel for sample pairs. Let $\mathbf{z}_i$ denote the sample pair $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$. Then the 2-degree polynomial kernel is defined as

$$\begin{aligned} &k(\mathbf{z}_i, \mathbf{z}_j) = \langle \mathbf{T}_i, \mathbf{T}_j \rangle \\ &= tr\left((\mathbf{x}_{i1} - \mathbf{x}_{i2})(\mathbf{x}_{i1} - \mathbf{x}_{i2})^T (\mathbf{x}_{j1} - \mathbf{x}_{j2})(\mathbf{x}_{j1} - \mathbf{x}_{j2})^T \right) \\ &= \left((\mathbf{x}_{i1} - \mathbf{x}_{i2})^T (\mathbf{x}_{j1} - \mathbf{x}_{j2}) \right)^2 \end{aligned} \tag{5}$$

With the sample pair kernel, given a sample pair $\mathbf{z} = (\mathbf{x}_1, \mathbf{x}_2)$, the regression function can be rewritten as

$$f(\mathbf{z}) = \sum\nolimits_{i=1}^{n} \beta_i \langle \mathbf{T}, \mathbf{T}_i \rangle + b = \langle \mathbf{M}, \mathbf{T} \rangle + b \tag{6}$$

where $\mathbf{T} = (\mathbf{x}_1 - \mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)^T$ and $\mathbf{T}_i = (\mathbf{x}_{i1} - \mathbf{x}_{i2})(\mathbf{x}_{i1} - \mathbf{x}_{i2})^T$. Here $\mathbf{M} = \sum_{i=1}^{n} \beta_i \mathbf{T}_i$. $\mathbf{M}$ is actually a linear combination of the scaled sample moments of the difference between two samples in one pair.

3.3 Regression Metric Learning

We aim to learn the discriminative $\mathbf{M}$ for Mahalanobis distance with a kernel regression formulation, then our Relation Alignment Metric Learning (RAML) can be formulated as

$$\min_{\mathbf{M}} \lambda r(\mathbf{M}) + \sum\nolimits_{i=1}^{n} l(\mathbf{M}, \mathbf{z}_i, g(\mathbf{z}_i)) \tag{7}$$

where $r(\mathbf{M})$ is the regularization item imposed on $\mathbf{M}$, λ is a positive constant and $l(\mathbf{M}, \mathbf{z}_i, g(\mathbf{z}_i))$ is the regression loss function. The parameter $\mathbf{M}$ is learned using (7) for enhancing the consistency of the sample relation in the feature space and decision space. The combinations of different regularization on $\mathbf{M}$ and loss functions will lead to different metric learning models. In this paper, we investigate two widely used regression models, *i.e.*, support vector regression (SVR) [Drucker *et al.*, 1997] and ridge regression [Gu *et al.*, 2016].

4 Support Vector Regression Metric Learning

We first extend our RAML formulation (7) to develop a SVR-like distance metric method:

$$\begin{aligned} &\min_{\mathbf{M},\xi,\xi^*} \ \lambda r(\mathbf{M}) + \rho(\xi, \xi^*) \\ &s.t. \begin{cases} g(\mathbf{z}_i) - (\langle \mathbf{M}, \mathbf{T}_i \rangle + b) \le \varepsilon + \xi_i \\ (\langle \mathbf{M}, \mathbf{T}_i \rangle + b) - g(\mathbf{z}_i) \le \varepsilon + \xi_i^* \\ \xi_i^*, \xi_i \ge 0 \end{cases} \end{aligned} \tag{8}$$

where ξ_i and ξ_i^* are slack variables, and $\rho(\xi, \xi^*)$ is the margin loss item. If we adopt the Frobenius norm to regularize $\mathbf{M}$ and ε-sensitive loss function, then this is a standard SVR model. We can also choose other regularizer and loss functions, *e.g.*, sparse regularizer, the Laplacian loss function or the Huberis loss function. As there are a lot of large-scale and efficient SVR solvers, we tend to develop a SVR-like model. Additionally, the variants of SVR have been well investigated, including semi-supervised SVR and multi-kernel SVR. Therefore, the proposed model can be used for semi-supervised metric learning, multi-modal metric learning and other specific learning tasks.

In this section, we will discuss how to learn the distance metric via support vector regression. By using Frobenius norm regularization for $r(\mathbf{M})$ and ε-sensitive loss function for $\rho(\xi, \xi^*)$, the metric learning problem in (8) can be formulated as:

$$\begin{aligned} &\min_{\mathbf{M},\xi,\xi^*} \ \tfrac{1}{2} \|\mathbf{M}\|_F^2 + \lambda \textstyle\sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ &s.t. \begin{cases} g(\mathbf{z}_i) - (\langle \mathbf{M}, \mathbf{T}_i \rangle + b) \le \varepsilon + \xi_i \\ (\langle \mathbf{M}, \mathbf{T}_i \rangle + b) - g(\mathbf{z}_i) \le \varepsilon + \xi_i^* \\ \xi_i^*, \xi_i \ge 0 \end{cases} \end{aligned} \tag{9}$$

where $\|\mathbf{M}\|_F^2$ is the Frobenius norm of $\mathbf{M}$, and λ is a trade-off constant. By using the Lagrange multipliers, we have

$$\mathrm{L} = \left\{ \begin{array}{l} \frac{1}{2} \|\mathbf{M}\|_F^2 + \lambda \sum_{i=1}^{n} (\xi_i + \xi_i^*) - \\ \sum_{i=1}^{n} a_i (\varepsilon + \xi_i - g(\mathbf{z}_i) + \langle \mathbf{M}, \mathbf{T}_i \rangle + b) - \\ \sum_{i=1}^{n} a_i^* (\varepsilon + \xi_i^* + g(\mathbf{z}_i) - \langle \mathbf{M}, \mathbf{T}_i \rangle - b) - \\ \sum_{i=1}^{n} (\eta_i \xi_i + \eta_i^* \xi_i^*) \end{array} \right\} \tag{10}$$

Algorithm 1 The algorithms of our proposed RAML-SVR and RAML-KRR

Input:

Training data $\mathbf{X} \in \mathbb{R}^{\mathbf{d}\times\mathbf{m}}$, where d and m are the numbers of feature dimension and samples, respectively.

1: Generate sample pairs $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$,$i = 1, 2, ..., n$.
2: Compute sample relation $g(\mathbf{x}_{i1}, \mathbf{x}_{i2})$,$i = 1, 2, ..., n$.
3: RAML-SVR: Solve (9) by SVR solvers
RAML-KRR: Solve (17) by (20)
4: RAML-SVR: $\mathbf{M} = \sum_{i=1}^{n} (a_i - a_i^*)\, \mathbf{T}_i$.
RAML-KRR: $\mathbf{M} = \sum_{i=1}^{n} \beta_i \mathbf{T}_i$.

Output:

Distance metric matrix $\mathbf{M}$

All dual variables should satisfy the positivity constraints, *i.e.*, $a_i, a_i^*, \eta_i, \eta_i^* \geq 0$. According to the saddle point condition, the partial derivatives of L with respect to the primal variables will be vanishing, *i.e.*,

$$\frac{\partial \mathrm{L}}{\partial b} = \sum\nolimits_{i=1}^{n} (a_i - a_i^*) = 0 \tag{11}$$

$$\frac{\partial \mathrm{L}}{\partial \mathbf{M}} = \mathbf{M} - \sum\nolimits_{i=1}^{n} (a_i - a_i^*)\mathbf{T}_i = 0 \tag{12}$$

$$\frac{\partial \mathrm{L}}{\partial \xi_i^*} = \lambda - a_i^* - \eta_i^* \tag{13}$$

Substituting (11), (12) and (13) into (10), we get the dual optimization problem of (9) with

$$\max \left\{ \begin{array}{l} -\frac{1}{2} \sum\limits_{i,j=1} (a_i - a_i^*)\left(a_j - a_j^*\right) \langle \mathbf{T}_i, \mathbf{T}_j \rangle \\ -\varepsilon \sum\limits_{i=1}^{n} (a_i + a_i^*) + \sum\limits_{i=1}^{n} g(\mathbf{z}_i)\,(a_i - a_i^*) \end{array} \right\} \tag{14}$$
$$s.t. \sum_{i=1}^{n} g(\mathbf{z}_i)\,(a_i - a_i^*) = 0, a_i, a_i^* \in [0, \lambda]$$

Similar to the solution of SVR, we can get the solution for (14), *i.e.*,

$$\mathbf{M} = \sum\nolimits_{i=1}^{n} (a_i - a_i^*)\, \mathbf{T}_i \tag{15}$$

Then, the corresponding regression function can be rewritten as

$$f(\mathbf{z}) = \sum\nolimits_{i=1}^{n} (a_i - a_i^*) \langle \mathbf{T}_i, \mathbf{T} \rangle + b \tag{16}$$

For the metric learning task, $\mathbf{M}$ is required to be positive semi-definite. Whereas, the solution for (14) cannot ensure that $\mathbf{M}$ is a PSD matrix. We compute the singular value decomposition of $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$ and only keep the positive part of $\mathbf{\Lambda}$ to form a new matrix $\mathbf{\Lambda}_+$. Finally, we obtain the PSD matrix $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}_+\mathbf{V}$. The corresponding metric learning (RAML-SVR) algorithm is summarized in Algorithm 1.

5 Ridge Regression Metric Learning

Besides support vector regression, we also incorporate kernel ridge regression into our RAML formulation (7) for metric learning. As $\mathbf{M} = \sum_{i=1}^{n} \beta_i \mathbf{T}_i$, it is equivalent to regularize $\mathbf{M}$ by regularizing β. The metric learning problem in (7) can be formulated as:

$$\min_{\beta} \sum\nolimits_{j=1}^{n} \left(f(\mathbf{z}_j) - g(\mathbf{z}_j)\right)^2 + r\,(\beta) \tag{17}$$

where $f(\mathbf{z}_j) = \sum_{i=1}^{n} \beta_i \langle \mathbf{T}_j, \mathbf{T}_i \rangle + b$, and $r\,(\beta)$ is the regularization item for β . When l_2-norm regularization is imposed on β, the objective function becomes

$$\begin{aligned} L\,(\beta) &= \textstyle\sum_{j=1}^{n} \left(\langle \mathbf{T}_j, \mathbf{M} \rangle - g\,(z_j)\right)^2 + \lambda \left\|\beta\right\|_2^2 \\ &= \textstyle\sum_{j=1}^{n} \left(\sum_{i=1}^{n} \beta_i \langle \mathbf{T}_j, \mathbf{T}_i \rangle - g\,(z_j)\right)^2 + \lambda \left\|\beta\right\|_2^2 \\ &= \textstyle\sum_{j=1}^{n} \left(\sum_{i=1}^{n} \beta_i K_{ij} - g\,(z_j)\right)^2 + \lambda \left\|\beta\right\|_2^2 \\ &= \left\|\beta\mathbf{K} - g\,(z)\right\|_2^2 + \lambda \left\|\beta\right\|_2^2 \end{aligned} \tag{18}$$

By setting the partial derivatives of L with respect to the variable β to 0, we have

$$\frac{\partial L\,(\beta)}{\partial \beta} = 2\left(\beta\mathbf{K} - g\,(z)\right)\mathbf{K}^T + 2\lambda\beta = 0 \tag{19}$$

Then we can get

$$\beta = g\,(z)\,\mathbf{K}^T\left(\mathbf{K}\mathbf{K}^T + \lambda\mathbf{I}\right)^{-1} \tag{20}$$

Furthermore, with β we can compute $\mathbf{M} = \sum_{i=1}^{n} \beta_i \mathbf{T}_i$. The corresponding metric learning (RAML-KRR) algorithm is summarized in Algorithm 1.

6 Discussion

6.1 Generalization Error Analysis

Let $\mathcal{P}$ be a fixed (but unknown) distribution over $\mathcal{X}\times\mathcal{Y}$. Let each training point $(\mathbf{x}_i, \mathbf{y}_i)\epsilon\mathcal{D}$ be sampled i.i.d. from $\mathcal{P}$. Metric learning aims to learn a discriminative matrix $\mathbf{M}$ which can preserve the sample relations in the label space. Then metric learning can be formulated as the following stochastic optimization problem:

$$\min_{\mathbf{M}\succeq\mathbf{0}} \mathcal{L}(\mathbf{M}) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}),\,(\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}) \sim \mathcal{P}} \mathbf{l}(\mathbf{M}; (\mathbf{x}, \mathbf{y}), (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})), \tag{21}$$

In fact, (7) minimizes a regularized empirical estimation of (21), *i.e.*, regularized Empirical Risk Minimization (ERM) w.r.t. (21). Following the theoretical analysis in [Bhatia *et al.*, 2015], using techniques from the AUC maximization literature, the excess risk bound for the problem in (7) is given as follows:

Theorem 1. *Assume that all data points are confined to a ball of radius R,* i.e., $\|\mathbf{x}\|_2 \leq R$ *for all* $\mathbf{x} \in \mathcal{X}$. *With probability at least* 1-δ *over the sampling of the dataset* $\mathcal{D}$, *the solution* $\hat{\mathbf{M}}$ *to the optimization problem* (7) *satisfies*

$$\mathcal{L}(\hat{\mathbf{M}}) \leq \inf_{\mathbf{M}^*\in\mathcal{M}} \left\{ \begin{array}{l} \mathcal{L}(\mathbf{M}^*) + \\ \overbrace{(\bar{L}^2 + (v^2 + \|\mathbf{M}^*\|_F^2)R^4)\sqrt{\frac{1}{n} log\frac{1}{\delta}}}^{E-Risk(n)} \end{array} \right\} \tag{22}$$

where $\hat{\mathbf{M}}$ *is the minimizer of* (7), $v = \frac{\overline{L}}{\lambda}$ *and* $\mathcal{M}$ *:=*$\{\mathbf{M} \in \mathbb{R}^{d\times d} : \mathbf{M} \succeq 0\}$. $\bar{L}$ *is the average number of labels active in a data point.*

We can see(22) shows that the optimal solution $\hat{\mathbf{M}}$ to (7) minimizes (21) with an additive approximation error. Additionally, the generalization error bound is independent of the feature dimension.

Data	S/F/C	ITML	LDML	LMNN	DSVM	GMML	DML	RAML-SVR	RAML-KRR
iris	150/5/3	0.9400±0.0668	0.9633±0.0562	0.9528±0.0843	0.9400±0.0584	0.9467±0.0689	**0.9667±0.0471**	**0.9667±0.0423**	**0.9667±0.0471**
wpbc	198/34/2	0.6919±0.0886	0.7629±0.0819	0.7432±0.1052	0.7329±0.0555	0.6968±0.0714	0.7476±0.0776	**0.7632±0.0608**	**0.7879±0.0520**
wine	178/14/3	0.9607±0.0300	**0.9833±0.0268**	0.9722±0.0393	0.9611±0.0375	0.9604±0.0379	0.9764±0.0425	0.9667±0.0468	**0.9833±0.0268**
sonar	208/61/2	0.8558±0.0615	0.7638±0.0903	0.8462±0.0705	0.8657±0.0583	0.8560±0.0749	0.8369±0.0840	**0.8900±0.0591**	**0.8850±0.0454**
glass	214/10/6	0.6636±0.0997	0.5740±0.1266	0.6556±0.0797	0.6258±0.1109	0.6308±0.1100	0.5724±0.1032	**0.6735±0.1072**	0.6395±0.1040
wdbc	569/31/2	0.9666±0.0263	0.9685±0.0284	0.9719±0.0223	0.9561±0.0355	0.9631±0.0315	0.9631±0.0193	**0.9789±0.0162**	**0.9755±0.0237**
credit	690/15/2	0.8043±0.0496	0.8422±0.0517	**0.8436±0.0525**	0.8013±0.0540	0.7941±0.0508	0.8261±0.0531	0.8246±0.0461	**0.8479±0.0320**

Table 1: Classification accuracy on UCI datasets

Data	S/F/C	ITML	LDML	LMNN	DSVM	GMML	DML	RAML-SVR	RAML-KRR
binalpha	1404/320/36	0.6303±0.0501	0.6542±0.0317	0.6112±0.0358	0.5625±0.0322	0.5338±0.1986	0.5063±0.0251	**0.7250±0.0348**	**0.6850±0.0351**
caltech101	8641/256/101	**0.5803±0.0162**	0.5528±0.0157	0.5795±0.0126	0.5584±0.0159	0.5500±0.0117	0.3936±0.0123	**0.5855±0.0095**	**0.5803±0.0147**
MnistDat	3495/784/10	0.8695±0.0142	0.8858±0.0124	0.8721±0.0255	0.8848±0.0194	0.8589±0.0171	0.8323±0.0239	**0.9019±0.0175**	**0.9087±0.0137**
Mpeg7	1400/6000/70	0.8214±0.0333	0.7971±0.0365	0.8253±0.0232	0.8271±0.0353	**0.8429±0.0228**	0.7071±0.0267	**0.8450±0.0305**	0.7936±0.0341
MSRA25	1799/256/12	**0.9989±0.0168**	0.9978±0.0046	0.9923±0.0397	0.9950±0.0121	0.9956±0.0140	0.9817±0.0211	**0.9989±0.0023**	0.9980±0.0142
news20	3970/8014/4	0.8678±0.0200	0.8816±0.0145	0.8734±0.0290	0.8594±0.0159	0.8647±0.0143	0.8166±0.0222	**0.9025±0.0132**	**0.9217±0.0145**
TDT2_20	1938/3677/20	0.9587±0.0358	0.9531±0.0306	0.9352±0.0197	0.9499±0.0175	0.9437±0.0275	0.6333±0.0176	**0.9679±0.0244**	**0.9762±0.0164**
uspst	2007/256/10	0.8979±0.0261	0.9084±0.0243	0.9096±0.0217	0.9125±0.0172	0.8858±0.0168	0.8030±0.0330	**0.9525±0.0147**	**0.9447±0.0157**

Table 2: Classification accuracy on image datasets

6.2 Sample Relation Function

The motivation of RAML is keeping relation consistency in different spaces, including feature space and label space. As the sample relations in the decision space are used to guide the metric learning in feature space, it is important to choose proper sample relation functions for different kinds of decision spaces. We consider three learning tasks, *i.e.*, single label learning, multi-label learning and label distribution learning. Let $\mathbf{y}_i$ and $\mathbf{y}_j$ denote the label vector of $\mathbf{x}_i$ and $\mathbf{x}_j$. The sample relation function is defined as:

$$g(\mathbf{y}_i, \mathbf{y}_j) = \|\mathbf{y}_i - \mathbf{y}_j\|_1 \tag{23}$$

where $\|\mathbf{a}\|_1$ is the l_1-norm of $\mathbf{a}$. For single label classification, when $g(\mathbf{y}_i, \mathbf{y}_j)$ is defined as (23), RAML degenerates to a sample pair classification problem. For multi-label learning, (23) reflects the difference with respect to positive classes of two samples. For label distribution learning, there are many metrics to evaluate the difference between two distributions. Here, we experimentally find that (23) reflects sample difference in the decision space and achieves superior performance. Therefore, we choose (23) for all the three learning tasks. The choice of optimal relation functions for different tasks are still an open problem, which will be investigated in our future work. If we want to learn a similarity metric in feature space, the inner product of two vectors, or other kernel functions can be used for $g(\mathbf{y}_i, \mathbf{y}_j)$.

6.3 Sample Pair Selection

Relation alignment learning aims to preserve the consistency of the sample relations between the feature space and the decision space. However, we do not need to use the relations of all sample pairs. For support vector regression, the support vectors are mainly lying on the decision boundary. Therefore, sample pairs are only generated in the k nearest neighbors (default value of k is 3 for our method RAML), which is similar to most existing metric learning algorithms. Besides, using only part of sample pairs can greatly reduce computational complexity and storage burden.

7 Experiments

In this section, we conduct experiments to validate the performance of the proposed metric learning methods. We consider three applications, including single-label classification, multi-label classification and label distribution learning. The following part will be organized as the corresponding three parts.

7.1 Single-Label Classification

Experiment setup. We first verify the effectiveness of RAML on seven UCI datasets [1] and eight image datasets [2]. The detailed information of these datasets is listed in Table **??** and Table **??**, where "S/F/C" represents the number of samples, features and classes. We compare RAML with the state-of-the-art methods, *i.e.*, ITML [Davis *et al.*, 2007], LMNN [Weinberger and Saul, 2009], DML [Ying and Li, 2012], DoubletSVM (DSVM) [Wang *et al.*, 2015], GMML [Zadeh *et al.*, 2016] on each dataset. For fair comparison, the parameters of all compared methods are set as the default setting of the original references. For DSVM, we set $k = 1$, and the penalty factor $C < 10,000$. For GMML, the weight t is set within [0,1] and chosen by greedy search. Ten-fold cross validation is introduced to evaluate the metric learning performance, *i.e.*, 90% for training and 10% for testing. The average accuracy of 10-fold cross validation is reported.

Experimental analysis. Table **??** and Table **??** list the classification accuracy of different metric learning methods on UCI datasets and image datasets, respectively, where the best results are marked in bold face. RAML-SVR and RAML-KRR indicate support vector regression metric learning and ridge regression metric learning, respectively. RAML achieves superior results in terms of the evaluation criteria on most dataset. The performance of RAML-SVR is similar to RAML-KRR. For RAML-KRR, when the number of samples increase significantly, the efficiency will be reduced because

[1] http://archive.ics.uci.edu/ml/index.php

[2] http://www.escience.cn/people/fpnie/index.html

its time complexity is $o(n^3)$, where n is the number of samples.

7.2 Multi-Label Classification

Datasets. In this section, we evaluate the proposed method using three datasets [3], *i.e.*, emotion [Trohidis *et al.*, 2008], flags, and corel800 dataset [Hoi *et al.*, 2006]. The emotion dataset [Trohidis *et al.*, 2008] consists of 100 songs from each of the following 7 different genres, *Classical, Reggae, Rock, Pop, Hip-Hop, Techno and Jazz*. The collection was created from 233 musical albums choosing three songs from each album. The flag dataset contains 194 instances, 19 features and 7 labels (red, green, blue, yellow, white, black, orange). The corel 800 dataset [Hoi *et al.*, 2006] contains 800 grayscale images of 10 individuals with 80 images per class.

Evaluation metrics. We employ five multi-label classification measures as evaluation metrics including Hamming loss, ranking loss, one error, coverage and average precision. Hamming loss measures accuracy in a multi-label classification task. Ranking loss has the property that the minimization of the loss functions will lead to the maximization of the ranking measures. MLKNN is the multi-label version of KNN [Zhang and Zhou, 2007] and it is based on statistical information derived from the label sets of an unseen instance's neighboring instances. As no specific metric learning algorithms are developed for MLKNN, here we use MLKNN as the baseline. If the performance of RAML is superior to MLKNN, the effectiveness of RAML is verified.

Experimental analysis. Experimental results of RAML and MLKNN are reported in Table 3, where the best result on each evaluation criterion is shown in bold face. The "↓" after the measures indicates "the smaller the better" and "↑" after the measures indicates "the larger the better". As shown in Table 3, both RAML-SVR and RAML-KRR achieve superior results in terms of the five evaluation measures. Compared with MLKNN, RAML can learn a discriminative distance metric, making the sample relation in the feature space more consistent with that in the decision space.

	Data	emotion	flags	corel800
MLKNN	Hamming Loss↓	0.2137	0.3099	0.0137
	Ranking Loss↓	0.1729	0.2228	0.1888
	One Error↓	0.3317	0.2154	0.6825
	Coverage↓	1.9158	3.8154	88.5100
	Average Precision↑	0.7808	0.8084	0.3276
RAML-SVR	Hamming Loss↓	**0.2054**	**0.2967**	**0.0135**
	Ranking Loss↓	**0.1577**	**0.2179**	**0.1882**
	One Error↓	**0.2376**	**0.2000**	**0.6425**
	Coverage↓	**1.8960**	**3.8115**	**88.2350**
	Average Precision↑	**0.8101**	**0.8128**	**0.3386**
RAML-KRR	Hamming Loss↓	**0.2046**	**0.2967**	**0.0134**
	Ranking Loss↓	**0.1382**	**0.2113**	**0.1888**
	One Error↓	**0.2574**	**0.2000**	**0.6550**
	Coverage↓	**1.7327**	**3.7692**	**88.5100**
	Average Precision↑	**0.8225**	**0.8112**	**0.3388**

Table 3: The performance of RAML-SVR, RAML-KRR and MLKNN in terms of five evaluation measures.

[3] http://mulan.sourceforge.net/datasets-mlc.html

7.3 Label Distribution Learning

Datasets. The dataset employed in this experiment includes 2,000 natural scene images [Zhang and Zhou, 2007]. There are nine possible labels associated with these images, *i.e.*, *plant, sky, cloud, snow, building, desert, mountain, water and sun*. The image features are extracted using the method in [Boutell *et al.*, 2004]. Each image is represented by a feature vector of 294 dimensions. The output of each instance is a distribution rather than discrete labels. AAKNN is the extended version of KNN in label distribution learning. Here AAKNN is used as the baseline without metric learning in the label distribution task.

Evaluation metrics. Different from both the single label output and the label set output of multi-label learning, the output of label distribution learning algorithm is a label distribution. The evaluation measures for label distribution learning is the average distance or similarity between the predicted and real label distributions. On a particular dataset, each of the measures may reflect some aspects of an algorithm. It is hard to say which evaluation metric is the best. Therefore, we use several measures to evaluate the proposed algorithm, and compare RAML with the classical AAKNN method. Finally we employ five measures: Chebyshev distance (Cheb), Clark distance (Clark), Canberra metric (Canber), cosine coefficient (Cosine), and intersection similarity(Intersec) [Cha, 2007]. The first three are distance measures and the last two are similarity measures.

Experimental analysis. Table 4 shows RAML and AAKNN in terms of five measures. We show the best result with respect to each measure in bold face. The "↓" after the measures indicates "the smaller the better". "↑" after the measures indicates "the larger the better". One can easily conclude that our RAML methods perform better than AAKNN in terms of five different measures. It owes to more discriminative metric learned by the proposed RAML formulation.

Criterion	AAKNN	RAML-SVR	RAML-KRR
Chebyshev↓	0.3261±0.0120	**0.3102±0.0123**	**0.3139±0.0157**
Clark↓	1.8448±0.0233	**1.6986±0.0386**	**1.6865±0.0468**
Canberra↓	4.3412±0.0650	**3.8576±0.1316**	**3.8419±0.1316**
Cosine↑	0.6905±0.0167	**0.7051±0.0167**	**0.7057±0.0126**
Intersection↑	0.5506±0.0120	**0.5739±0.0261**	**0.5743±0.0172**

Table 4: The performance of RAML-SVR, RAML-KRR and AAKNN in terms of five measures on Nature Scene dataset.

8 Conclusions

In this paper, we proposed a relation alignment metric learning (RAML) formulation to learn distance metrics for various kinds of learning tasks. Different from all existing methods relying on similar and dissimilar relations between sample pairs, we formulated metric learning problem as a kernel regression model via relation alignment learning. Based on our RAML formulation, two metric learning methods are instantiated with support vector regression and kernel ridge regression. Experimental result show RAML is very competitive with state-of-the-art metric learning methods on single-label classification, moreover it can improve the performance of multi-label learning and label distribution learning tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 61502332 and 61732011, Natural Science Foundation of Tianjin Under Grants 17JCZDJC30800.

References

[Bhadra *et al.*, 2017] Sahely Bhadra, Samuel Kaski, and Juho Rousu. Multi-view kernel completion. *Machine Learning*, 106(5):713–739, 2017.

[Bhatia *et al.*, 2015] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In *NIPS*, pages 730–738, 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.

[Cha, 2007] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

[Cortes *et al.*, 2012] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(Mar):795–828, 2012.

[Davis *et al.*, 2007] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[Drucker *et al.*, 1997] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. Support vector regression machines. *NIPS*, 28(7):779–784, 1997.

[Geng, 2016] Xin Geng. Label distribution learning. *IEEE TKDE*, 28(7):1734–1748, 2016.

[Gu *et al.*, 2016] Yuwen Gu, Hui Zou, et al. High-dimensional generalizations of asymmetric least squares regression and their applications. *The Annals of Statistics*, 44(6):2661–2694, 2016.

[Guillaumin *et al.*, 2009] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505. IEEE, 2009.

[Hoi *et al.*, 2006] Steven CH Hoi, Wei Liu, Michael R Lyu, and Wei-Ying Ma. Learning distance metrics with contextual constraints for image retrieval. In *CVPR*, volume 2, pages 2072–2078, 2006.

[Law *et al.*, 2013] Marc T Law, Nicolas Thome, and Matthieu Cord. Quadruplet-wise image similarity learning. In *ICCV*, pages 249–256, 2013.

[Li *et al.*, 2012] Zhen Li, Liangliang Cao, Shiyu Chang, John R Smith, and Thomas S Huang. Beyond mahalanobis distance: Learning second-order discriminant function for people verification. In *CVPRW*, pages 45–50. IEEE, 2012.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[McFee and Lanckriet, 2011] Brian McFee and Gert Lanckriet. Learning multi-modal similarity. *Journal of machine learning research*, 12(Feb):491–523, 2011.

[Mensink *et al.*, 2012] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. *ECCV*, pages 488–501, 2012.

[Mu *et al.*, 2010] Yadong Mu, Jialie Shen, and Shuicheng Yan. Weakly-supervised hashing in kernel space. In *CVPR*, pages 3344–3351. IEEE, 2010.

[Oh Song *et al.*, 2016] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016.

[Shaw *et al.*, 2011] Blake Shaw, Bert Huang, and Tony Jebara. Learning a distance metric from a network. In *NIPS*, pages 1899–1907, 2011.

[Trohidis *et al.*, 2008] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.

[Wang *et al.*, 2015] Faqiang Wang, Wangmeng Zuo, Lei Zhang, Deyu Meng, and David Zhang. A kernel classification framework for metric learning. *IEEE TNNLS*, 26(9):1950–1962, 2015.

[Wang *et al.*, 2016] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE TPAMI*, 38(10):2010–2023, 2016.

[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(Feb):207–244, 2009.

[Yang *et al.*, 2006] Liu Yang, Rong Jin, Rahul Sukthankar, and Yi Liu. An efficient algorithm for local distance metric learning. In *AAAI*, volume 2, pages 543–548, 2006.

[Ying and Li, 2012] Yiming Ying and Peng Li. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13(Jan):1–26, 2012.

[Zadeh *et al.*, 2016] Pourya Zadeh, Reshad Hosseini, and Suvrit Sra. Geometric mean metric learning. In *ICML*, pages 2464–2471, 2016.

[Zhang and Wu, 2015] Min-Ling Zhang and Lei Wu. Lift: Multi-label learning with label-specific features. *IEEE TPAMI*, 37(1):107–120, 2015.

[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.