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

*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 \( x_i \) and \( x_j \), where \( x_i \in \mathbb{R}^d \) and \( d \) is the dimension of the feature space, the Mahalanobis distance between the two samples \( x_i \) and \( x_j \), is defined as

\[
d_M(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j)
\]

where \( M \in \mathbb{R}^{d \times d} \) is a positive semi-definite (PSD) matrix. The parameter \( M \) 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 \( y = [y_1, ..., y_k, ..., y_c] \) be the label vector of \( x \). For the traditional single-label classification task, \( x \) can only belong to one class. If \( x \) 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 \( x \) can belong to multiple classes. For label distribution learning, there are two constraints for \( y \), 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 aim 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.

![Feature space](image)

![Decision space](image)

Figure 2: The consistency of the sample relations in feature and decision spaces. The sample relation matrix \( E \) is calculated via \( g(y_i, y_j) \). Then we can use \( E \) to guide metric learning (i.e., learning of \( M, b \)) in the feature space to get a more discriminative sample relation \( A \).
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, \( f(x_i, x_j, M, b) \) is used to measure the sample relations in feature space while \( g(y_i, y_j) \) is used to measure the sample relations in decision space. \( g(y_i, y_j) \) is specially designed for different tasks. Let \( A \in \mathbb{R}^{n \times n} \) and \( E \in \mathbb{R}^{n \times m} \) 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.,

\[
\begin{bmatrix}
a_{11} & \ldots & a_{1n} \\
\vdots & \ddots & \vdots \\
a_{n1} & \ldots & a_{nn}
\end{bmatrix}
= 
\begin{bmatrix}
e_{11} & \ldots & e_{1n} \\
\vdots & \ddots & \vdots \\
e_{n1} & \ldots & e_{nn}
\end{bmatrix}
\]

where \( a_{ij} \) and \( e_{ij} \) represent the sample relation of \( x_i \) and \( x_j \) in the feature space and decision space, respectively. Here, to measure the sample relations in feature space while \( x \) is actually a linear combination of the scaled sample moments of the difference between two samples in one pair.

\[
\begin{align*}
f(x_i, x_j, M, b) & = (x_i - x_j)^T M (x_i - x_j) + b \\
& = \langle M, T_{ij} \rangle + b
\end{align*}
\]

(3)

where \( \langle \cdot, \cdot \rangle \) is defined as the Frobenius inner product of two matrices, \( b \) is the bias item, and \( T_{ij} = (x_i - x_j) (x_i - x_j)^T \).

Then we rewrite (2) to

\[
g(y_i, y_j) = \langle M, T_{ij} \rangle + b
\]

(4)

Once the relation function \( g(y_i, y_j) \) 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 \( T_{ij} \) of sample pair \( (x_i, x_j) \) 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 \( z_i \) denote the sample pair \((x_{i1}, x_{i2})\). Then the 2-degree polynomial kernel is defined as

\[
k(z_i, z_j) = \langle T_i, T_j \rangle
\]

\[
= tr \left( (x_{i1} - x_{i2}) (x_{i1} - x_{i2})^T (x_{j1} - x_{j2}) (x_{j1} - x_{j2})^T \right)
\]

\[
= \left( (x_{i1} - x_{i2}) (x_{i1} - x_{i2})^T \right)^2
\]

(5)

With the sample pair kernel, given a sample pair \( z = (x_1, x_2) \), the regression function can be rewritten as

\[
f(z) = \sum_{i=1}^{n} \beta_i \langle T_i, T_i \rangle + b = \langle M, T \rangle + b
\]

(6)

where \( T = (x_1 - x_2) (x_1 - x_2)^T \) and \( T_i = (x_{i1} - x_{i2}) (x_{i1} - x_{i2})^T \). Here \( M = \sum_{i=1}^{n} \beta_i T_i \) 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 \( M \) for Mahalanobis distance with a kernel regression formulation, then our Relation Alignment Metric Learning (RAML) can be formulated as

\[
\begin{align*}
\min_{M} & \quad \lambda r(M) + \sum_{i=1}^{n} l(M, z_i, g(z_i)) \\
\text{s.t.} & \quad \langle M, T_i \rangle + b = g(z_i)
\end{align*}
\]

(7)

where \( r(M) \) is the regularization item imposed on \( M \), \( \lambda \) is a positive constant and \( l(M, z_i, g(z_i)) \) is the regression loss function. The parameter \( M \) 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 \( M \) 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:

\[
\begin{align*}
\min_{M, \xi, \xi^*} & \quad \lambda r(M) + \rho(\xi, \xi^*) \\
\text{s.t.} & \quad \langle M, T_i \rangle + b = g(z_i) - \xi_i \leq \varepsilon + \xi_i^* \\
& \quad \xi_i, \xi_i^* \geq 0
\end{align*}
\]

(8)

where \( \xi_i \) and \( \xi_i^* \) are slack variables, and \( \rho(\xi, \xi^*) \) is the margin loss item. If we adopt the Frobenius norm to regularize \( M \) and \( \varepsilon \)-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 Huber 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 \( r(M) \) and \( \varepsilon \)-sensitive loss function for \( \rho(\xi, \xi^*) \), the metric learning problem in (8) can be formulated as:

\[
\begin{align*}
\min_{M, \xi, \xi^*} & \quad \frac{1}{2} \| M \|^2_F + \lambda \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\
\text{s.t.} & \quad \langle M, T_i \rangle + b = g(z_i) - \xi_i \leq \varepsilon + \xi_i^* \\
& \quad \xi_i, \xi_i^* \geq 0
\end{align*}
\]

(9)

where \( \| M \|^2_F \) is the Frobenius norm of \( M \) and \( \lambda \) is a trade-off constant. By using the Lagrange multipliers, we have

\[
L = \left\{ \frac{1}{2} \| M \|^2_F + \lambda \sum_{i=1}^{n} (\xi_i + \xi_i^*) - \sum_{i=1}^{n} a_i (\varepsilon + \xi_i^* - g(z_i) - \langle M, T_i \rangle - b) - \sum_{i=1}^{n} \eta_i \xi_i \right. \\
\left. - \sum_{i=1}^{n} \eta_i^* \xi_i^* \right\}
\]

(10)
Algorithm 1 The algorithms of our proposed RAML-SVR and RAML-KRR

Input:
Training data $X \in \mathbb{R}^{d \times m}$, where $d$ and $m$ are the numbers of feature dimension and samples, respectively.
1: Generate sample pairs $(x_i, x_j), i, j = 1, 2, ..., n$.
2: Compute sample relation $g(x_i, x_j), i, j = 1, 2, ..., n$.
3: RAML-SVR: Solve (9) by SVR solvers
RAML-KRR: Solve (17) by (20)
4: RAML-SVR: $M = \sum_{i=1}^{n} \beta_i T_i.$
RAML-KRR: $M = \sum_{i=1}^{n} \beta_i T_i.$

Output:
Distance metric matrix $M$

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.,

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

$$\frac{\partial L}{\partial M} = -\sum_{i=1}^{n} (a_i - a_i^*) T_i = 0$$ (12)

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

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

$$\max \left\{ -\frac{1}{2} \sum_{i,j=1}^{n} (a_i - a_i^*) (a_j - a_j^*) \langle T_i, T_j \rangle \right\}$$

$$-\sum_{i=1}^{n} (a_i + a_i^*) + \sum_{i=1}^{n} g(z_i) (a_i - a_i^*)$$ (14)

s.t. $\sum_{i=1}^{n} g(z_i) (a_i - a_i^*) = 0, a_i, a_i^* \in [0, \lambda]$

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

$$M = \sum_{i=1}^{n} (a_i - a_i^*) T_i$$ (15)

Then, the corresponding regression function can be rewritten as

$$f(z) = \sum_{i=1}^{n} (a_i - a_i^*) \langle T_i, T \rangle + b$$ (16)

For the metric learning task, $M$ is required to be positive semi-definite. Whereas, the solution for (14) cannot ensure that $M$ is a PSD matrix. We compute the singular value decomposition of $M = U \Lambda V$ and only keep the positive part of $\Lambda$ to form a new matrix $\Lambda_+$. Finally, we obtain the PSD matrix $M = U \Lambda_+ V$. 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 $M = \sum_{i=1}^{n} \beta_i T_i$, it is equivalent to regularize $M$ by regularizing $\beta$. The metric learning problem in (7) can be formulated as:

$$\min_{\beta} \sum_{j=1}^{n} (f(z_j) - g(z_j))^2 + r(\beta)$$ (17)

where $f(z_j) = \sum_{j=1}^{n} \beta_j \langle T_j, T_i \rangle + b$, and $r(\beta)$ is the regularization item for $\beta$. When $\ell_2$-norm regularization is imposed on $\beta$, the objective function becomes

$$L(\beta) = \sum_{j=1}^{n} (\langle T_j, M \rangle - g(z_j))^2 + \lambda \|\beta\|_2^2$$

$$= \sum_{j=1}^{n} (\sum_{i=1}^{n} \beta_i \langle T_j, T_i \rangle - g(z_j))^2 + \lambda \|\beta\|_2^2$$

$$= \sum_{j=1}^{n} (\sum_{i=1}^{n} \beta_i K_{ij} - g(z_j))^2 + \lambda \|\beta\|_2^2$$ (18)

By setting the partial derivatives of $L$ with respect to the variable $\beta$ to 0, we have

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

Then we can get

$$\beta = (K^T (K K^T + \lambda I))^{-1} g(z)$$ (20)

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

6 Discussion

6.1 Generalization Error Analysis

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

$$\min_{M \succeq 0} L(M) = \mathbb{E}_{(x, y), (\bar{x}, \bar{y}) \sim P} I(M; (x, y), (\bar{x}, \bar{y})),$$ (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., $\|x\|_2 \leq R$ for all $x \in \mathcal{X}$. With probability at least $1-\delta$ over the sampling of the dataset $\mathcal{D}$, the solution $M$ to the optimization problem (7) satisfies

$$L(M) \leq \inf_{M \succeq 0} \mathbb{E} \left\{ \left( \frac{L^2 + (v^2 + \| M\|_F^2) R^4}{n \log \frac{1}{\delta}} \right)^{1/2} \right\}$$ (22)

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

We can see (22) shows that the optimal solution $M$ to (7) minimizes (21) with an additive approximation error. Additionally, the generalization error bound is independent of the feature dimension.
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 $y_i$ and $y_j$ denote the label vector of $x_i$ and $x_j$. The sample relation function is defined as:

$$g(y_i, y_j) = \| y_i - y_j \|_1$$  \hspace{1cm} (23)

where $\| a \|_1$ is the $l_1$-norm of $a$. For single label classification, when $g(y_i, y_j)$ 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 $g(y_i, y_j)$.

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], DoubtSVM (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 on 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\(^1\), 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 ”\(\downarrow\)" after the measures indicates “the smaller the better” and ”\(\uparrow\)" 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.

#### 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 are 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(Intersection) [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 ”\(\downarrow\)" after the measures indicates “the smaller the better”. ”\(\uparrow\)" 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.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>AAKNN</th>
<th>RAML-SVR</th>
<th>RAML-KRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chebyshev(\downarrow)</td>
<td>0.3261±0.0120</td>
<td>0.3102±0.0123</td>
<td>0.3139±0.0157</td>
</tr>
<tr>
<td>Clark(\downarrow)</td>
<td>1.8448±0.0233</td>
<td>1.6986±0.0386</td>
<td>1.6865±0.0468</td>
</tr>
<tr>
<td>Canberra(\downarrow)</td>
<td>4.3412±0.0650</td>
<td>3.8576±0.1316</td>
<td>3.8419±0.1316</td>
</tr>
<tr>
<td>Cosine(\uparrow)</td>
<td>0.6905±0.0167</td>
<td>0.7051±0.0167</td>
<td>0.7057±0.0126</td>
</tr>
<tr>
<td>Intersection(\uparrow)</td>
<td>0.5506±0.0120</td>
<td>0.5739±0.0261</td>
<td>0.5743±0.0172</td>
</tr>
</tbody>
</table>

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


