



Manifold regularized multiple kernel learning with Hellinger distance

Tao Yang¹ · Dongmei Fu¹ · Xiaogang Li² · Kamil Říha³

Received: 10 August 2017 / Revised: 29 November 2017 / Accepted: 7 February 2018 / Published online: 16 March 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

The aim of this paper is to solve the problem of unsupervised manifold regularization being used under supervised classification circumstance. This paper not only considers that the manifold information of data can provide useful information but also proposes a supervised method to compute the Laplacian graph by using the label information and the Hellinger distance for a comprehensive evaluation of the similarity of data samples. Meanwhile, multi-source or complex data is increasing nowadays. It is desirable to learn from several kernels that are adaptable and flexible to deal with this type of data. Therefore, our classifier is based on multiple kernel learning, and the proposed approach to supervised classification is a multiple kernel model with manifold regularization to incorporate intrinsic geometrical information. Finally, a classifier that minimizes the testing error and considers the geometrical structure of data is put forward. The results of experiments with other methods show the effectiveness of the proposed model and computing the inner potential geometrical information is useful for classification.

Keywords Multiple kernel learning · Manifold regularization · Hellinger distance

1 Introduction

For classification tasks, there are mainly three types: supervised, unsupervised and semi-supervised. Supervised learning has both input and output data, of which the output data are labels of corresponding input data; while, unsupervised learning has only input data. Intuitively, semi-supervised learning has full input data but has partial output data. In the field of supervised classification, many types of classifiers have been proposed. These classifiers are the Bayesian methods, least square linear or non-linear models [1], neural networks [2], decision trees [3], hierarchical methods [4] and so on [5–7]. Most of them are based solely on the values or relations of the training data

and are solved by learning parameters or a linear system, and a method to estimate the potential geometry within data has not yet been fully developed. In this study, we consider the geometry aspect of data as important information, introducing it into a learning classifier to obtain a more appropriate classifier. As the manifold theory shows, any data assume an intrinsic geometrical structure and, particularly, when $\{(x_i, y_i)\}_{i=1}^n \in X \times R$, we assume that the marginal probability distribution P_X is supported on the manifold $\mathcal{M} \subset X$. Although we may not know the distribution information, the manifold structure can be learned to estimate P_X . In order to consider the inner geometrical structure in the classifier, we adopt the manifold regularization technique, one of the important works is proposed by [8], in that work a geometrical framework for solving semi-supervised problem was proposed. Manifold regularization is a geometrically motivated penalty; which considers the classifier f restricted to \mathcal{M} and forces f to be smooth along \mathcal{M} . The computation of manifold regularization is based on a graph. The graph is an important tool to represent the similarity of pair-wise data samples and facilitate to extract the geometrical structure within data. When we build a graph, the traditional methods used are unsupervised methods and the relation between vertices is kNN or the Euclidean distance. In supervised classification

✉ Tao Yang
yangtao@ustb.edu.cn

¹ School of Automation and Electrical Engineering, University of Science & Technology Beijing, No. 30 Haidian District Xueyuan Road, Beijing 100083, China

² Institute for Advanced Materials and Technology, University of Science & Technology Beijing, No. 30 Haidian District Xueyuan Road, Beijing 100083, China

³ Department of Telecommunications, Brno University of Technology, Antonínská 548/1, 601 90 Brno, Czech Republic

scenario, the label information is given to the classifier; therefore, we propose a supervised graph method that uses labels. To evaluate the similarity of pair-wise vertices in a graph, we use the Hellinger distance instead of the Euclidean distance as the Hellinger distance can represent distance from four aspects, such as similarity, density, dimension, and orientation [9] and thus provide a comprehensive consideration of data relations. After obtaining the graph, we need to choose the classifier f . Nowadays many datasets are from multiple sources and, have multiple attributes that differ in features and correlate with each other. Therefore, two points should be considered while selecting a classifier: One is that the classifier should be suitable for manifold regularization, and the other is that it should be flexible in dealing with a complicated dataset. Therefore, multiple kernel learning (MKL) is a good option. Unlike a support vector machine (SVM) or a logistic model, MKL searches for a linear combination of the base kernel functions by optimizing the generalized performance evaluation. Given a series of base kernel functions, MKL displays good adaptability and interpretation [10, 11], and thus far, many algorithms for MKL have been developed [12–16]. Moreover we know that regularization techniques are often used in kernel-based methods [5].

In this study, under the assumption that the inner geometrical structure of data improves the performance of a learning machine, we present a classifier based on MKL with the manifold regularization, expecting the regularizer can extract the inner geometrical structure. In the supervised scenario, we propose a supervised graph construction method. This classifier makes use of an intrinsic geometrical structure obtained by using a graph and has the ability to deal with a multiple feature dataset via advanced MKL. The related work is referred to [17], in which a manifold regularization is also added to multiple kernels classifier. The differences between previous study and this study are that previous study adopted an approximated formulation while this study is accurate, previous study adopted original manifold regularization while this study applied an improved manifold regularization with Hellinger distance, especially fit for the supervised classification. The main contributions of this paper can be outlined as follows:

- (1) We propose the supervised manifold regularization using the Hellinger distance, where the Hellinger distance can describe data relationships in details, to address the problem of the supervised classification scenario. The proposed supervised manifold regularization is applied to MKL to construct a reasonable classifier, which can deal with large ranges of datasets.

- (2) We conduct experiments on several public datasets; the results show that the proposed classifier can achieve competent performances as compared to other methods, leading to results related to the effectiveness and efficiency of the proposed method.

The rest of this paper is organized as follows. In Sect. 2, we present the manifold regularization MKL, and introduce the supervised graph method using the Hellinger distance. In Sect. 3, the classifier is solved, along with a preliminary analysis on complexity; some concerns are also presented in this section. Experimental results are described in Sect. 4. Finally, we present the conclusion and future research directions in Sect. 5.

2 Methodology

2.1 Classifier formulation

Given a supervised dataset $\{(x_i, y_i)\}_{i=1}^n \in X \times R$, $x_i \in R^d$, $y_i = \pm 1$, our task is to find a classifier function f with the ability to make generalizations. Assume that (x, y) are drawn from a probability distribution P , and the marginal distribution P_X is supported on a compact manifold \mathcal{M} ; thus, the conditional distribution $P_{Y|X}$ varies smoothly along the geodesics in the geometry of \mathcal{M} , which implies that similar $P(y_i|x_i)$ and $P(y_j|x_j)$ correspond to a close relation between x_i and x_j . Therefore, an additional regularization term is needed to control the classification function f along \mathcal{M} . This additional regularization is called manifold regularization and we use $\|f\|_f^2$ to note it. In practice, P_X is usually unknown; therefore, $\|f\|_f^2$ is estimated by using a graph [18]. Now, as we use multiple kernel machine to develop f , we first introduce the reproducing kernel Hilbert space (RKHS) [19]. For a Mercer kernel $K(x, y) : X \times X \rightarrow R$, there is an RKHS \mathcal{H} of functions $f : X \rightarrow R$ with the corresponding norm $\|\cdot\|_{\mathcal{H}}$. In \mathcal{H} , we have the following property,

$$f(x) = (f(\cdot), K(\cdot, x)) \quad (1)$$

where $f \in \mathcal{H}$ and (\cdot, \cdot) denotes the inner product defined in \mathcal{H} . In order to introduce MKL, we use the idea proposed by [13] in SimpleMKL; i.e., any function f in the multiple kernel space \mathcal{H} is a sum of functions f_m , $m = 1, \dots, M$, each belonging to \mathcal{H}_m . The space \mathcal{H}_m is also an RKHS endowed with an inner product $(\cdot, \cdot)_m$ and a positive definite kernel K_m ; finally, a classical result on RKHS [19] reveals that \mathcal{H} is an RKHS, with the following form:

$$K(x, y) = \sum_{i=1}^M d_m K_m(x, y) \tag{2}$$

$$\sum_{i=1}^M d_m = 1; d_m \geq 0, \forall m$$

where we consider the ℓ_1 norm constraint to the weights $\{d_m\}_{m=1}^M$. For MKL only, given the data $\{(x_i, y_i)\}_{i=1}^n \in X \times R$, the following form is considered in the classification task:

$$f^* = \arg \min_{f \in \mathcal{H}} C \sum_{i=1}^n V(x_i, y_i, f) + \gamma \|f\|_{\mathcal{H}}^2 \tag{3}$$

where V denotes a loss function, such as $(y_i - f(x_i))^2$ for a least-square model or the hinge loss function for a SVM; C represents the penalty parameter; and, the RKHS norm $\|\cdot\|_{\mathcal{H}}^2$ forces smoothness conditions on the possible solutions to obtain a better generalization and avoid over-fitting. According to the Representer The-orem [20, 21] and multiple kernels, the optimal function of problem (3) is given by:

$$f^*(x) = \sum_{i=1}^n \alpha_i K(x_i, x) = \sum_{i=1}^n \alpha_i d_m K_m(x_i, x) \tag{4}$$

where $\sum_{i=1}^M d_m = 1; d_m \geq 0, \forall m$ and $\alpha_i \in R, \forall i$.

As mentioned earlier, the intrinsic geometrical information is introduced by a regularization form, which is as follows:

$$f^* = \arg \min_{f \in \mathcal{H}} C \sum_{i=1}^n V(x_i, y_i, f) + \gamma_h \|f\|_{\mathcal{H}}^2 + \gamma_l \|f\|_l^2 \tag{5}$$

where γ_h controls the function complexity in the space \mathcal{H} and γ_l controls the manifold penalty.

We also noted that in (5), the optimal f^* lies in the linear space $\mathcal{S} = span\{K(\cdot, x) | x \in \mathcal{M} \subset X\}$, and for any $f \in \mathcal{H}$, $f = f_{\mathcal{S}} + f_{\mathcal{S}}^{\perp}$, where $f_{\mathcal{S}}$ denotes the projection of f to \mathcal{S} and $f_{\mathcal{S}}^{\perp}$ represents its orthogonal complement.

Lemma 1. [8] All functions $f_{\mathcal{S}}^{\perp}$ in \mathcal{H} vanish on \mathcal{M} .

By using lemma 1, we have $f(x_i) = f_{\mathcal{S}}(x_i)$; on the other hand, $\|f\|_{\mathcal{H}}^2 = \|f_{\mathcal{S}}\|_{\mathcal{H}}^2 + \|f_{\mathcal{S}}^{\perp}\|_{\mathcal{H}}^2$. Therefore, $\|f\|_{\mathcal{H}}^2 = \|f_{\mathcal{S}}\|_{\mathcal{H}}^2$ and the minimizer f^* is in \mathcal{S} . And f^* admits the Representer Theorem in (4). This is helpful in reducing the problem and optimizing over the results over a finite dimensional space.

2.2 Supervised graph

Initially, $\|f\|_l^2$ is estimated using a graph. Define a graph $G = (V, E)$, where V denotes vertices on all samples, i.e., $v_i = x_i$; E represents the edges linking the adjacency pair, i.e., $e_{ij} : v_i \sim v_j$; and ‘ \sim ’ implies that two vertices are adjacent. Two vertices are considered adjacent on the basis of kNN or the Euclidean distance if $\|x_i - x_j\| \leq \epsilon$. Further, the edges are weighted by a weight matrix W , whose elements are as follows:

$$w_{ij} = \begin{cases} 1 & v_i \sim v_j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

or $w_{ij} = \begin{cases} \exp(-\sigma^2 \|x_i - x_j\|^2) & v_i \sim v_j \\ 0 & \text{otherwise} \end{cases}$

The natural way to compute manifold regularization is: $\|f\|_l^2 = \sum_{ij} (f(x_i) - f(x_j))^2 w_{ij}$; this implies that the outputs of f maintain adjacent relations. By using a Laplacian graph, we obtain the following computations:

$$\|f\|_l^2 = \mathbf{f}^T L \mathbf{f} \tag{7}$$

$$L = D - W$$

where D denotes a diagonal matrix with the entries $D_{ii} = \sum_j w_{ij}$, and $\mathbf{f}^T = (f(x_1), \dots, f(x_n))$. L represents a Laplacian graph. This method is unsupervised, and in a supervised setting, it is insufficient from two aspects. One is the loss of useful information, i.e., labels, and the other is the fact that the weights are biased if the adjacent vertices are decided solely on the basis of distance, particularly a scenario where two points are close but they bear different labels.

We propose a graph that uses label information for showing more relations between samples. As we have already seen, the edges on the graph are based on the similarity of vertices; therefore, the more we consider the similarity, the more accurate would be the sample relations. Here, we adopt the Hellinger distance to depict a similarity. As we mentioned above, Hellinger distance can represent data from four aspects: similarity, density, dimension, and orientation. This distance describes data based on the probability where the data is generated from, different from Euclidean distance or Manhattan Distance or Chebyshev distance or Minkowski distance, all of which are based on the coordinate or computation of absolute value, and Mahalanobis distance which emphasizes the distance of covariance. As for manifold assumption, the probability-based measure is proper; therefore, the Hellinger distance is adopted. We notice that the Bhattacharyya distance is also the probability-based distance but it fits for discreet random variables. Kullback–Leibler divergence measures two probabilities but it is not symmetric and mainly for entropy variations.

After the kNN method, we can obtain the neighbors of each sample, for instance, for a point x , $N(x)$ is to be a neighbor around x . Define the local sample covariance matrix Δ_x :

$$\Delta_x = \sum_{x' \in N(x)} \frac{(x' - \mu_x)(x' - \mu_x)^T}{(|N(x)| - 1)} \tag{8}$$

where μ_x is defined as the neighborhood mean and, $|N(x)|$ denotes the cardinality, which is a quantity that captures

the local geometry feature. We know that initially, the Hellinger distance quantifies the similarity between two probability distributions [22]; therefore, we define two Gaussian distributions, namely $p(x_i) = N(0, \Delta_{x_i})$, $p(x_j) = N(0, \Delta_{x_j})$, with zero mean and covariance $\Delta_{x_i}, \Delta_{x_j}$. We can get the Hellinger distance of x_i, x_j using the following:

$$H_{ij}^2 = \frac{1}{2} \int (p^{1/2}(x_i) - p^{1/2}(x_j))^2 dx \tag{9}$$

After obtaining the Hellinger distance matrix H , we need to construct the graph G , which would consider the label information. Therefore, G is split into two graphs G_w and G_b ; G_w denotes a within-class graph, G_b represents a between-class graph. Let $\ell(y_i)$ be the class label of y_i , and N be the notation of neighbors. Therefore, for each point y_i , we define the following:

$$\begin{cases} N_{w1} = \{y_j | \ell(y_j) = \ell(y_i), H(\Delta_{x_i}, \Delta_{x_j}) \leq \lambda_1\} \\ N_{w2} = \{y_j | \ell(y_j) = \ell(y_i), H(\Delta_{x_i}, \Delta_{x_j}) > \lambda_1\} \\ N_{b1} = \{y_j | \ell(y_j) \neq \ell(y_i), H(\Delta_{x_i}, \Delta_{x_j}) \leq \lambda_2\} \\ N_{b2} = \{y_j | \ell(y_j) \neq \ell(y_i), H(\Delta_{x_i}, \Delta_{x_j}) > \lambda_2\} \end{cases} \tag{10}$$

It is noted that N_{w1} denotes within-class neighbors with a low Hellinger distance and, N_{w2} denotes within-class neighbors with a high Hellinger distance; thus, N_{b1}, N_{b2} represent between-class neighbors with low and high Hellinger distances. Here, we treat the Hellinger distance differently, i.e., $\lambda_1 \leq \lambda_2$ implying that in a same label occasion we would tighten the bound; otherwise, we would relax the bound. This is done to avoid a bias. Thus, we can adapt the set of neighbors according to the class information and the local geometry similarity between samples.

Let W_w and W_b be the weight matrices of G_w and G_b , defined them as follows:

$$\begin{aligned} G_w : & \begin{cases} \text{if } y_i \in N_{w1}(y_j), w_{w,ij} = 1 \\ \text{if } y_i \in N_{w2}(y_j), w_{w,ij} = \exp(-\|x_i - x_j\|^2) \end{cases} \\ G_b : & \begin{cases} \text{if } y_i \in N_{b1}(y_j), w_{b,ij} = \exp(-r_{ij}\|x_i - x_j\|^2) \\ \text{if } y_i \in N_{b2}(y_j), w_{b,ij} = 0 \end{cases} \end{aligned} \tag{11}$$

where $r_{ij} = \max\{Radius\{N_{b1}(y_i), N_{b1}(y_j)\}\}$ for the weight between different classes; here, a relatively high similarity needs to be treated as non-zero. Finally, it is easy to show that the global weight matrix W can be written as $W = W_w + W_b$. In terms of manifold regularization, the Laplacian graph is computed in the same way, i.e., $L = D - W$. To use a different notation, here we denote the Laplacian graph as L_H . In the experiment, we use both unsupervised L and supervised L_H for the manifold

regularized MKL classification, to determine their performances. Then, we use only L for the following formulas to simplify the notation. As shown in (5), by taking the Representer Theorem and (7), we transform (5) into the following:

$$\min_{f \in \mathcal{H}} C \sum_{i=1}^n V(x_i, y_i, f) + \gamma_h \alpha^T K \alpha + \gamma_l \alpha^T K L K \alpha \tag{12}$$

where $K = \sum_{m=1}^M d_m K_m$ and K_m denotes the Gram-matrix with entry $K_m(x_i, x_j)$. Here, (12) extends the model in [8] by using multiple kernels; therefore, (12) has the advantages of both multiple kernels and a manifold regularization.

3 Algorithm and discussion

3.1 Optimization analysis

Using the hinge loss function, we can rewrite (12) as follows:

$$\begin{aligned} & \min_{\alpha, \xi} C \sum_{i=1}^n \xi_i + \gamma_h \alpha^T K \alpha + \gamma_l \alpha^T K L K \alpha \\ \text{s.t. } & y_i \left(\sum_{j=1}^n \alpha_j K(x_j, x_i) + b \right) \geq 1 - \xi_i; \xi_i \geq 0, \forall i \\ & K = \sum_{m=1}^M d_m K_m, \sum_{m=1}^M d_m = 1; d_m \geq 0, \forall m \end{aligned} \tag{13}$$

Here, $\alpha^T K \alpha = \|\alpha\|_K^2$ and $\alpha^T K L K \alpha = \|K \alpha\|_L^2$. Thus, by adding Lagrangian multipliers to (13), we obtain the following:

$$\begin{aligned} & \min_{d, \alpha, \xi, b} J(d) \\ \text{s.t. } & \sum_{m=1}^M d_m = 1; d_m \geq 0, \forall m \\ J(d) = & \max_{\beta, \lambda} C \xi + \gamma_h \|\alpha\|_K^2 + \gamma_l \|K \alpha\|_L^2 + \\ & \beta^T (\mathbf{1} - \xi - Y(K \alpha + b)) - \lambda^T \xi \end{aligned} \tag{14}$$

where β, λ denote the multiplier vectors, $\mathbf{1}$ represents a column vector with 1, ξ denotes a column vector with the entry ξ_i , and Y refers to a diagonal matrix with the entry $Y_{ii} = y_i$.

First, we consider the problem with a temporary fixed d , then take the derivative of $J(d)$ w.r.t. primal variables:

$$\begin{aligned} \partial J / \partial \alpha &= 0 \rightarrow (2\gamma_h K + 2\gamma_l K L K) \alpha = K Y \beta \\ \partial J / \partial \xi_i &= 0 \rightarrow 0 \leq \beta_i \leq C \\ \partial J / \partial b &= 0 \rightarrow \sum_i \beta_i y_i = 0 \end{aligned} \tag{15}$$

By inserting the derivative results back into (14), we obtain the following:

$$\begin{aligned}
 J(d) = \min_d \max_{\beta} \sum_{i=1}^n \beta_i - \frac{1}{2} \beta^T Y K Q Y \beta \\
 \text{s.t. } \sum_{i=1}^n \beta_i y_i = 0; 0 \leq \beta_i \leq C
 \end{aligned} \tag{16}$$

where $Q = (2\gamma_h I + 2\gamma_l LK)^{-1}$ and I denotes an identity matrix. (16) can be solved by using a classical SVM solver with fixed d . After getting an intermediate optimal β^* , we go back to update each d_m ; this two-layer problem continues until the stop criterion is met. To update each d_m , we develop the gradient $\partial J / \partial d_m := g_m$ to update d as follows:

$$g_m = -\beta^{*T} Y K_m Q / 2 - \gamma_l K Q L K_m Y \beta^* \quad \forall m \tag{17}$$

It is certain that the minus sign on the right side of (17) denotes the descent direction. Meanwhile, we pay attention to the constraints on d in order to estimate a feasible descent direction. Therefore, we use the Reduced Gradient Method [23]. Suppose that d_v denotes the maximum element in d and v represents the index, then according to the constraint $\sum d_m = 1$, the reduced gradient is as follows:

$$g_m^{reduced} = \begin{cases} g_m - g_v & m \neq v \\ 0 & m = v \end{cases} \tag{18}$$

Therefore, the feasible descent direction can be computed as follows:

$$p = \begin{cases} p_m = \begin{cases} -g_m^{reduced} & g_m^{reduced} \leq 0 \\ -d_m g_m^{reduced} & g_m^{reduced} > 0 \end{cases} \\ p_v = -\sum_{m \neq v} p_m \end{cases} \tag{19}$$

After obtaining direction p , we perform a one-dimensional search, and hence, the constraints on d have to be considered here. Suppose that the update of $d^{t+1} = d^t + \lambda p$ and, the step length λ is in the range $[0, \lambda_{max}]$. Then, we need to pay attention to the maximum admissible step λ_{max} to avoid the violation of the feasible field:

$$\lambda_{max} = \begin{cases} \min\{-d_m/p_m\} & p_m < 0, d_m > 0 \\ \lambda_0 & \{p_m < 0, d_m > 0\} = \emptyset \end{cases} = \phi \tag{20}$$

where λ_0 denotes a pre-defined maximum step.

For (16), we consider the KKT conditions; i.e., we assume that $J(d)$ denotes the objective and $\sum_m d_m = 1$ represents the constraints. Then, the first-order optimality conditions can be expressed as follows:

$$\begin{aligned}
 g_m + \delta - \eta_m = 0 \quad \forall m \\
 \eta_m \cdot d_m = 0 \quad \forall m
 \end{aligned} \tag{21}$$

where δ and η_m denote the Lagrange multipliers for the equality and inequality constraints, respectively. The KKT conditions tell us that for active $d_m > 0$, we have $g_m = \delta$, and for inactive $d_m = 0$, we have $g_m \geq \delta$. Therefore, we define a tolerance ε away from δ , and the necessary optimality conditions could be the following:

$$\begin{aligned}
 \left| \min_{d_m > 0} g_m - \max_{d_m > 0} g_m \right| \leq \varepsilon \\
 g_\mu \geq \max_{d_m > 0} g_m, d_\mu = 0
 \end{aligned} \tag{22}$$

Note that for vanishing d_m , we put the gradient outside the tolerance tube. Therefore, the stop criterion is decided by (22), and the other way is the norm of the feasible descent direction; i.e., when $\|p\|_2 \leq \varepsilon$, the iteration can be terminated.

3.2 Bound

As for the problem of error bounds, we often take two aspects into consideration: one is the sample-based classification accuracy, and the other is the measurement of model complexity. According to [24], the Rademacher complexity is suitable for the case of MKL.

Theorem 1 *Let P be the probability distribution on $X \times \{\pm 1\}$, let \mathcal{H} be $\{\pm 1\}$ -valued functions defined on X , and $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^n$ be the training samples drawn from P . With a probability of at least $1 - \delta$, every function f in \mathcal{H} satisfies the following:*

$$P(Y \neq f(x)) \leq P_n(Y \neq f(x)) + \frac{R_n(\mathcal{H})}{2} + \sqrt{\frac{\ln(1/\delta)}{2n}} \tag{23}$$

where P_n denotes the average output of an indicator function, and $R_n(\mathcal{H})$ represents the Rademacher complexity of \mathcal{H} [25]. Let $\|f\|_{\mathcal{H}}^2 \leq a^2$ and the manifold regularization $\|f\|_l^2 \leq b^2$. Thus, by using the Representer Theorem, we obtain $\alpha^T K \alpha \leq a^2$ and $\alpha^T K L K \alpha \leq b^2$. We use set \mathcal{A} to denote them. Here, K denotes the Gram matrix over \mathcal{T} and the multiple kernel weight d is still under ℓ_1 norm constraints. For \mathcal{T} , the Rademacher complexity can be expressed as follows:

$$R_n(\mathcal{H}) = \frac{1}{n} E_\sigma \left[\sup_{f \in (\mathcal{H})} \sum_{i=1}^n \sigma_i f(x_i) \right] = \frac{1}{n} E_\sigma \left[\sup_{d, \alpha \in \mathcal{A}} \sigma^T K \alpha \right]$$

Here, we take $\lambda, \mu \geq 0$ as the Lagrangian multipliers to the inequalities in \mathcal{A} and obtain the relation $\sigma = (\lambda I + \mu K L) \alpha$. Thus, we obtain the following:

$$R_n(\mathcal{H}) = \frac{1}{n} E_\sigma \left[\sup_d \sigma^T K (\lambda I + \mu K L)^{-1} \sigma \right] \tag{24}$$

Note that $\mu_\sigma^m = \sigma^T K_m (\lambda I + \mu K L)^{-1} \sigma$ and the dual norm $\sup_{\|u\|_1 \leq 1} u \cdot v = \|v\|_\infty$. Therefore, for any integer $r \geq 1$, we have the following:

$$R_n(\mathcal{H}) = \frac{1}{n} E_\sigma \left[\sup_d \sum_{m=1}^M d_m \mu_\sigma^m \right] \leq \frac{1}{n} E_\sigma [\|u_\sigma\|_r] \tag{25}$$

The algorithm computes the inverse of a Gram matrix which has $O(n^3)$ complexity; this is not practical for large data sets. For high-dimensional data, we may use the Laplacian Eigenmap (LE) method for the dimensionality reduction. In this process, we would obtain the Laplacian matrix L ; this matrix is just usable in manifold regularization again, which should help to improve the time efficiency. Further, with respect to the algorithm itself, our future work would be to design an efficient algorithm to implement the update of multiple kernel weights d and the computation of an inverse matrix.

4 Algorithm and discussion

In this section, we will present experimental results on a number of benchmark datasets, which are the UCI dataset [26], USPS dataset [27], Spoken letter dataset and a synthetic dataset.

4.1 Experimental setting

In practice, we fix the value of the hyper-parameter C ($C = 100$) and the Hellinger distance parameter λ ($\lambda_1 = 0.4, \lambda_2 = 0.6$). Further, the value of γ_h is taken from $[10^{-5}, \dots, 10^{-2}]$ and γ_l from $[10^{-6}, \dots, 10^{-2}]$ on the basis of a five-fold cross validation. For multiple kernels, the candidate kernels are divided into two parts: One part is composed of ten Gaussian kernels with 10 different bandwidths $[2^{-4}, \dots, 2^5]$ and each kernel is set on all variables and on each variable. The other part is composed of three polynomial kernels of degree $[1-3]$ on all variables and on each variable. All kernel matrices are a normalized to unit trace. Our algorithm programming refers to SimpleMKL. The neighboring graph vertices are found using the kNN method, where the parameter k is chosen from $[6-8]$ with cross-validation along with γ_h and γ_l ; a large value of k is not suitable and a small value may correspond to a sparse graph. In an unsupervised Laplacian graph, the weights on the edges are determined by a binary function, and in a supervised Laplacian graph, the weights are decided by using (11). We will use a SVM, the method of [8], i.e., LapSVM, and generic MKL for the sake of comparison. In the case of the SVM, we select two kernels: one is the Gaussian kernel, and the other is the polynomial kernel. The parameters are tuned by using LIBSVM [28], in LapSVM, kernel function is Gaussian, the parameter k in kNN is 7, and in generic MKL, multiple kernels are set in the same way as described above.

4.2 Results

We selected 12 groups of the UCI dataset, and the “g50c” was generated from two unit-covariance normal distributions with equal probabilities. For each group, the experiments were run 30 times with 50% training data and 50% test data each time; the data were randomly chosen. The training data were normalized to zero mean and unit variance, and the test data were normalized using the mean and variance of the training data. The results are shown in Table 1.

From Table 1, we can see that MKL with manifold regularization could have competent or better accuracy rates. This proves that the potential geometrical information obtained using regularization favors a supervised classification. In the process, we have found that the kernel numbers in the first two methods are more than in the case of generic MKL. We believe that it is implicitly implied that the kernels preserve more information and may be fit for highly relevant data. In the case of the SVM, we elaborate on parameter tuning and obtain good results, but the SVM lacks a mixture of kernels and does not provide much information on sample geometry. Therefore, we expect MKL with the regularization technique to become a useful tool in the case of relatively complicated and multi-source data.

We reported our results of the handwritten digits in the USPS dataset for 3 situations, namely 2 versus 6, 4 versus 9, 5 versus 8 of classification accuracy. Each gray-scale image is scaled 16×16 , and we vectorize it into a 256-dimensional sample with normalization. In the first three groups, we select 100 images and set the number of training samples varying between 5 and 20; the rest are considered test data. Further, we run each test 30 times and calculate the average of the results. Lastly, we calculated the average of the final results. When the number of training samples was 5, we assumed $tr = 5$ and set the kNN neighbor $k = 2$. Similarly, for $tr = 10$, we set $k = 5$, and for $tr = 15, 20, 25$, we set $k = 7$. In the case of the SVM, we used a Gaussian kernel with parameter tuning. Tables 2 presents the classification accuracy rates.

From Table 2, we find that manifold regularized MKL classifiers associated with either an unsupervised Laplacian graph or a supervised one remain on the top accuracy level in most of the cases. The results presented in Table 2 imply that the unsupervised Laplacian graph is accurate. This could be attributed to the facts that pairwise digits are almost separable in space and, the weights in the graph are sufficient to define the relations between data.

Isolet database is the letters of English alphabet spoken in isolation from UCI machine learning repository. There are 26 classes containing the utterances of 150 subjects

Table 1 The classification accuracy (mean \pm std. %) in UCI dataset

Dataset	#n;#d	MKL with L_H	MKL with L	MKL	LapSVM	SVM1	SVM2
Breast	277; 9	73.3 \pm 1.0	73.6 \pm 1.2	71.8 \pm 1.1	70.8 \pm 0.4	72.2 \pm 2.2	72.3 \pm 2.5
German	500; 24	75.6 \pm 1.2	74.6 \pm 1.6	73.4 \pm 1.3	72.8 \pm 0.1	74.3 \pm 0.8	71.9 \pm 2.2
Heart	300; 13	82.4 \pm 1.3	82.5 \pm 3.0	83.7 \pm 1.9	76.8 \pm 3.3	82.1 \pm 1.9	78.5 \pm 2.4
Iono	351; 33	92.8 \pm 1.5	91.7 \pm 0.5	91.3 \pm 1.4	65.4 \pm 0.7	92.3 \pm 0.7	89.9 \pm 0.9
Liver	345; 6	70.1 \pm 1.8	66.9 \pm 2.0	63.1 \pm 2.7	65.3 \pm 0.4	67.2 \pm 1.2	68.3 \pm 2.2
Pima	200; 8	74.1 \pm 1.8	73.4 \pm 3.6	70.9 \pm 2.4	72.2 \pm 1.6	71.8 \pm 2.6	74.8 \pm 0.6
Sonar	208; 60	73.4 \pm 1.7	78.9 \pm 2.8	79.2 \pm 2.3	63.9 \pm 2.7	79.5 \pm 2.1	79.9 \pm 1.0
Spambase	200; 57	94.7 \pm 1.9	94.3 \pm 2.5	93.7 \pm 1.2	88.1 \pm 2.1	87.6 \pm 1.3	91.8 \pm 2.6
Vote	435; 16	95.6 \pm 1.1	95.9 \pm 1.0	94.5 \pm 1.3	93.4 \pm 1.3	93.5 \pm 1.7	94.3 \pm 1.1
Wine	178; 13	98.0 \pm 0.7	97.1 \pm 2.4	96.8 \pm 2.4	94.7 \pm 1.2	97.7 \pm 0.8	97.1 \pm 2.8
Wpbc	198; 33	77.4 \pm 2.1	76.4 \pm 0.8	75.6 \pm 2.1	76.5 \pm 0.1	76.5 \pm 0.6	74.9 \pm 2.8
Wdbc	569; 14	97.4 \pm 0.8	96.8 \pm 1.1	94.7 \pm 0.9	92.5 \pm 1.2	95.1 \pm 0.8	83.4 \pm 2.1
g50c	550; 50	95.0 \pm 1.3	91.6 \pm 1.1	90.5 \pm 1.0	90.7 \pm 1.9	94.7 \pm 1.2	94.0 \pm 1.1

Bold values indicate the best accuracy

#n number of data samples, #d dimension of data samples, SVM1 is used with a Gaussian kernel, SVM2 is used with a polynomial kernel

Table 2 Classification accuracy (mean \pm std. %) of 2 versus 6, 4 versus 9 and 5 versus 8 data

		MKL with L_H	MKL with L	MKL	LapSVM	SVM
2 versus 6	$tr = 5$	77.6 \pm 2.8	75.9 \pm 5.9	74.6 \pm 6.2	76.8 \pm 10.4	74.7 \pm 10.8
	10	83.3 \pm 2.6	84.6 \pm 2.3	81.5 \pm 4.2	85.2 \pm 9.2	82.6 \pm 10.6
	15	90 \pm 2.5	87.1 \pm 2.9	84.8 \pm 2.8	87.6 \pm 5.8	86.5 \pm 3.7
	20	90.5 \pm 2.3	90.6 \pm 2.7	89.2 \pm 2.7	88.3 \pm 5.3	88.6 \pm 3.5
	25	92.5 \pm 2.3	81.8 \pm 3.1	90.2 \pm 2.3	90.7 \pm 4.4	90.6 \pm 3.4
4 versus 9	$tr = 5$	81.1 \pm 7.8	81.7 \pm 10.1	78.1 \pm 11.0	67.5 \pm 12.2	69.6 \pm 11.3
	10	86.5 \pm 6.5	84.9 \pm 8.4	83.5 \pm 8.4	69.6 \pm 10.6	76.3 \pm 6.2
	15	90.3 \pm 5.3	89.8 \pm 6.7	87.8 \pm 5.5	72.7 \pm 7.6	81.6 \pm 3.4
	20	92.5 \pm 5.5	92.5 \pm 7.5	90.7 \pm 4.7	79.0 \pm 4.9	84.7 \pm 3.8
	25	93.8 \pm 2.9	93.6 \pm 4.9	91.7 \pm 3.9	81.1 \pm 4.9	86.4 \pm 4.0
5 versus 8	$tr = 5$	74.1 \pm 10.0	78.1 \pm 12.7	72.3 \pm 10.2	72.1 \pm 8.2	71.7 \pm 10.5
	10	79.8 \pm 2.1	79.5 \pm 4.2	73.9 \pm 4.1	76.2 \pm 7.7	78.6 \pm 9.1
	15	85.7 \pm 1.8	85.2 \pm 2.3	84.2 \pm 2.5	83.2 \pm 5.5	84.1 \pm 5.6
	20	85.8 \pm 1.5	86.3 \pm 1.6	85.1 \pm 1.7	85.5 \pm 5.1	85.6 \pm 4.2
	25	89.1 \pm 0.8	88.9 \pm 1.7	88.3 \pm 1.6	87.4 \pm 4.3	88.7 \pm 4.3

Bold values indicate the best accuracy

who spoke each alphabet twice. In experiment, we selected 20 samples from each class. The number of training data is from 30 to 70% of all samples. For computation, we adopted Laplacian Eigenmap to reduce the dimension of Isolet from original 618–3 and the loss of information is inevitable, therefore we omitted the method of SVM because of poor performance. We reported the classification accuracy on all the data in Fig. 1.

Figure 1 shows that the proposed model obtained top classification accuracy. MKL with original manifold regularization also performed good classification ability.

Through the above experimental results, the proposed model benefits from the application of manifold

regularization, which introduces the inner geometrical structures within data into the kernel-based classifier. While, the traditional SVM and MKL model do not consider that structure information, losing possible useful information in classifying data. The more detailed formula differences can be seen from formula (3) and (12).

4.3 Time complexity

Assuming there are n data samples with dimension d . We use M different kernels in classifier. First, as for the graph computation, the distances of all samples take time complexity $O(nd)$ and Laplacian graph L has the time

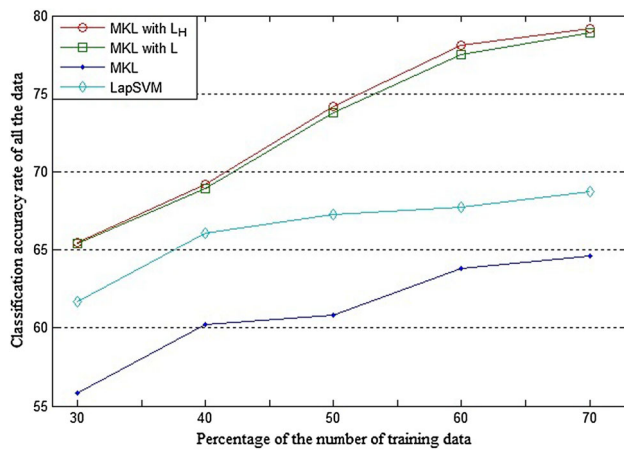


Fig. 1 Isolet experiment: classification accuracy rate of all the data samples

complexity $O(n^2)$. The hellinger distance of all samples has approximate time complexity $O(n^2)$. The combined kernel takes time complexity $O(Mnd)$. In formula (16), there is a dense inverse matrix Q for $O(2n^3)$ and calculating the intermediate β would have $O(n^3)$. The update of d has $O(Mn^3)$ time complexity.

The entire time complexity is $O(((M+1)dn + 2n^2 + (3+M)n^3) \times T)$, where T is the number of iterations in optimization process.

5 Conclusion and future

5.1 Conclusion

In this study, we added manifold regularization to a multiple kernel machine for supervised classification; the regularization was computed using a Laplacian graph, which can exploit the potential geometrical distribution of data samples. Further, in the case of a supervised classification, we proposed a supervised Laplacian graph to take the label information into account in order to obtain a good representation of the true data structure, relying on the measures obtained using the Hellinger distance, which help define data relations from the aspects of similarity, density, dimension and orientation. The base classifier that we selected was based on multiple kernels, which showed good adaptability and interpretability. The proposed model is expected to have minimum error as well as maintain the geometrical properties, represented by the assumed manifold. As shown in the experiments, the proposed classifier could achieve competent results. Further, we found that manifold regularization could be applied to supervised and semi-supervised data, particularly to datasets that show an obvious manifold when combined with MKL.

5.2 Future work and application

There are several research directions worthy of being explored further: Choices of parameters in the model and sensitivity analysis; selection of Gaussian widths in MKL; and development of an efficient algorithm to solve the entire optimal problem.

In addition to above works relevant to the model itself, the proposed classifier can be possibly develop to other kernel-based methods, such as signal processing frameworks based on kernel learning [29–31]. Moreover, the manifold regularization techniques are useful in analysing data especially for those like covariance features, shape variations and other data with specific structures.

Acknowledgements The authors acknowledge the China Postdoctoral Science Foundation (No. 2017M620615) and Fundamental Research Funds for the Central Universities (Grant: FRF-TP-16-082A1) and National Natural Science Foundation of China (No. 61272358).

References

- Murty, M.N., Devi, V.S.: Introduction to pattern recognition and machine learning. *J. Cell. Physiol.* **200**(1), 71–81 (2015)
- Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2014)
- Sheu, J., Chen, Y., Chu, K., et al.: An intelligent three-phase spam filtering method based on decision tree data mining. *Secur. Commun. Netw.* **9**(17), 4013–4026 (2016)
- Campello, B., Moulavi, D., Zimek, A., Sander, J.: A framework for semisupervised and unsupervised optimal extraction of clusters from hierarchies. *Data Min. Knowl. Discov.* **27**(3), 344–371 (2013)
- Sørensen, A.P.W.: Geometric classification of simple graph algebras. *Ergod. Theory Dyn. Syst.* **33**(4), 1199–1220 (2013)
- Criminisi, A., Shotton, J.: Semi-supervised classification forests. *Adv. Comput. Vis. Pattern Recognit.* **161**, 544–563 (2013)
- Wang, B., Tu, Z., Tsotsos, J.K.: Dynamic label propagation for semi-supervised multi-class multi-label classification. *Pattern Recognit.* **52**, 75–85 (2016)
- Belkin, M., Niyogi, P., Sindhvani, V.: Manifold regularization: a geometric framework for learning from labels and unlabeled examples. *J. Mach. Learn. Res.* **7**(1), 2399–2434 (2006)
- Xing, X., Yu, Y., Jiang, H., et al.: A multi-manifold semi-supervised Gaussian mixture model for pattern classification. *Pattern Recognit. Lett.* **34**(16), 2118–2125 (2013)
- Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.: Learning the kernel matrix with sem-definite programming. *J. Mach. Learn. Res.* **5**(1), 323–330 (2004)
- Liang, Z., Zhang, L., Liu, J.: A novel multiple kernel learning method based on the Kullback-Leibler divergence. *Neural Process. Lett.* **42**(3), 1–18 (2015)
- Bucak, S., Jin, R., Jain, A.K.: Multiple kernel learning for visual object recognition: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(7), 1354–1369 (2014)
- Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. *J. Mach. Learn. Res.* **9**(3), 2491–2521 (2008)
- Althloothi, S., Mahoor, M.H., Zhang, X.: Human activity recognition using multi-features and multiple kernel learning. *Pattern Recognit.* **47**(5), 1800–1812 (2014)

15. Nazarpour, A., Adibi, P.: Two-stage multiple kernel learning for supervised dimensionality reduction. *Pattern Recognit.* **48**(5), 1854–1862 (2015)
16. Aiolli, F., Donini, M.: EasyMKL: a scalable multiple kernel learning algorithm. *Neurocomputing* **169**, 215–224 (2015)
17. Yang, T., Fu, D.: Semi-supervised classification with Laplacian multiple kernel learning. *Neurocomputing* **140**, 19–26 (2014)
18. Cao, Y., Chen, D.R.: Generalization errors of Laplacian regularized least squares regression. *Sci. China Math.* **55**(9), 1859–1868 (2012)
19. Arqub, O.A., Al-Smadi, M., Momani, S., et al.: Numerical solutions of fuzzy differential equations using reproducing kernel Hilbert space method. *Soft. Comput.* **20**(8), 3283–3302 (2016)
20. Mcfee, B., Lanckriet, G.: Learning multi-modal similarity. *J. Mach. Learn. Res.* **12**(8), 491–523 (2010)
21. Dinuzzo, F., Neve, M., Necolao, G.D.: On the representer theorem and equivalent degrees of freedom of SVR. *J. Mach. Learn. Res.* **8**(8), 2467–2495 (2007)
22. Ladd, A.M., Kaviraki, L.E.: Measure theoretic analysis of probabilistic path planning. *IEEE Trans. Robot. Autom.* **20**(2), 229–242 (2004)
23. Chrétien, B., Escande, A., Kheddar, A.: GPU robot motion planning using semi-infinite nonlinear programming. *IEEE Trans. Parallel Distrib. Syst.* **27**(10), 1–1 (2016)
24. Micchelli, C.A., Pontil, M., Wu, Q., et al.: Error bounds for learning the kernel. *Anal. Appl.* **14**(06), 849–868 (2016)
25. Ying, Y., Campbell, C.: Rademacher chaos complexities for learning the kernel problem. *Neural Comput.* **22**(11), 2858–2886 (2014)
26. Ashok, P., Nawaz, G.M.K.: Outlier detection method on UCI repository dataset by entropy based rough K-means. *Def. Sci. J.* **66**(2), 113–119 (2016)
27. Johnson, D., Xiong, C., Corso, J.: Semi-supervised nonlinear distance metric learning via forests of max-margin cluster hierarchies. *IEEE Trans. Knowl. Data Eng.* **28**(4), 1035–1046 (2016)
28. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *Trans. Intell. Syst. Technol.* **2**, 27–27 (2011)
29. Ding, G., Wu, Q., Yao, Y.D., et al.: Kernel-based learning for statistical signal processing in cognitive radio networks: theoretical foundations, example applications, and future directions. *IEEE Signal Process. Mag.* **30**(4), 126–136 (2013)
30. Harchaoui, Z., Bach, F., Cappe, O., et al.: Kernel-based methods for hypothesis testing: a unified view. *IEEE Signal Process. Mag.* **30**(4), 87–97 (2013)
31. Bazerque, J.A., Giannakis, G.B.: Nonparametric basis pursuit via sparse kernel-based learning: a unifying view with advances in blind methods. *IEEE Signal Process. Mag.* **30**(4), 112–125 (2013)



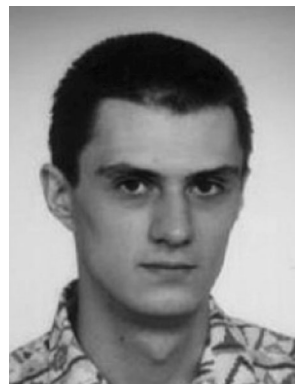
Tao Yang is post-doctoral in School of Automation and Electrical Engineering, University of Science & Technology Beijing, China. He received his Bachelor degree and Doctor degree in Automation Science and Engineering from USTB in 2010 and 2016. His researches focus on manifold learning, machine learning.



interests include automation control theory, image processing and data mining.



Xiaogang Li vice president, professor and Ph.D. supervisor of New Materials Technology Research Institute, University of Science and Technology Beijing. Prof. Li is the 973 chief scientist of the national marine corrosion and receives the government special allowance. Prof. Li is also Director of the National Association of Corrosion and Protection of Corrosion and Protection.



Kamil Říha Ph.D., associate professor at The Faculty of Electrical Engineering and Communication, Brno University of Technology, with the main workload comprised by research & development and pedagogic activities. In the long-term, he deals with digital image processing, mainly with extraction of spatial information from an image stereo pair. He actively participated in many research & development and pedagogic project solutions, mainly dealing with digital image processing, mainly for medical utilization.