Multiple Empirical Kernel Learning with Discriminant Locality Preservation

Bolu Wang BLUECHASE@FOXMAIL.COM
East China University of Science and Technology, China.

Dongdong Li LDD@ECUST.EDU.CN
East China University of Science and Technology, China.

Zhe Wang WANGZHE@ECUST.EDU.CN
East China University of Science and Technology, China.

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

Multiple Kernel Learning (MKL) algorithm effectively combines different kernels to improve the performance of classification. Most MKL algorithms implicitly map samples into feature space by the form of inner-product. In contrast, Multiple Empirical Kernel Learning (MEKL) can explicitly map the input spaces into feature spaces so that the mapped feature vectors are explicitly represented, which is easy to process and analyze the adaptability of kernels for input space. Meanwhile, in order to pay attention to the structure and discriminant information of samples in empirical feature space, inspired by discriminant locality preserving projections, we introduce the discriminant locality preservation regularization into MEKL framework to propose the Multiple Empirical Kernel Learning with Discriminant Locality Preservation (MEKL-DLP). Experiments conducted on real-world datasets validate the effectiveness of the proposed MEKL-DLP compared with the classical kernel-based algorithms and state-of-art MKL algorithms.

Keywords: Multiple kernel learning; Empirical kernel mapping; Discriminant locality preserving projections; Machine learning.

1. Introduction

Kernel-based algorithms (Breneman (2006); Müller et al. (2001)) can effectively improve the classification performance of predictive machine learning algorithms and have been widely studied. These methods map the samples from input space X to feature space F via a mapping function $\Phi : X \to F$. In this way, the original linearly inseparable sample in X is mapped to a new linearly separable sample in F, which improves the accuracy of classification (Müller et al. (2001)). There are two kinds of mapping functions Φ which are Φ^i for implicit form and Φ^e for explicit form. The Implicit Kernel Mapping (IKM) implicitly maps samples to feature space via the inner-product $k(x_i, x_j) = \Phi^i(x_i) \cdot \Phi^i(x_j)$. But the necessity of inner-product in IKM restricts other methods unsatisfying the formulation to be kernelized (Lu et al. (2003); Ye (2005); Ye et al. (2004)). In contrast, the Empirical Kernel Mapping (EKM) (Xiong et al. (2005)) explicitly maps the samples into feature space by giving the explicit form of Φ^e. Due to the explicit representation of feature vectors

according to EKM, most algorithms can be kernelized directly. Thus, it is easier to process and analyze the adaptability of kernels for input space (Zhe et al. (2007)).

The choice of kernel function plays an important role in achieving superior performance in kernel-based algorithms. But it is difficult to select an appropriate kernel in a specific problem. Therefore, the Multiple Kernel Learning (MKL) (Gönen and Alpaydın (2011); Vedaldi et al. (2009); Bucak et al. (2014)) was proposed to address this issue. By combining multiple kernel functions in a certain way, the information brought by multiple kernel functions is considered simultaneously in training and testing to improve the classification performance. Lanckriet et al. (2002) integrated the linear combination of multiple kernels into the process of structural risk optimization and used a Quadratically Constrained Quadratic Program (QCQP) to solve the problem directly. To get more generalized performance, Kloft et al. (2011) extended MKL to arbitrary norms for robust kernel mixtures. Thus, they proposed a l_p-norm MKL where p was arbitrary. By putting priors on kernel combination parameters, Girolami and Rogers (2005) formulated a Bayesian hierarchical model and derived variational Bayes estimators for classification problems. Due to learning the optimal combinations of kernels in the process of optimization tasks is quite difficult to be solved, Hao and Hoi (2013) adopt boosting to solve a variant of MKL problem, which avoids solving the complicated optimization tasks. In order to reduce the time and space complexity, Aiolli and Donini (2015) proposed an efficient MKL method named EasyMKL which can easily cope with hundreds of thousands of kernels. Alternatively, Cortes et al. (2012) gave a centered-kernel alignment criterion. By maximizing the criterion between a nonnegative linear combination of kernels and the ideal kernel, a suitable combination weights of candidate kernels were acquired.

As mentioned above, most of the existing MKL methods use the IKM to map samples into feature space. In contrast, the MKL adopting the EKM to construct the feature spaces is denoted as Multiple Empirical Kernel Learning (MEKL). Conventional MEKL algorithms optimize the learning framework by minimizing the empirical risk and regularization risk (Zhe et al. (2007)), but the distribution information of samples is not considered. Since the samples after EKM are able to obtain its explicit representation in feature space, the distribution information of the samples is effectively integrated into MEKL. By introducing a locality preserving constrain regularization into MEKL, Fan et al. (2016) proposed the MEKL-LPC method which utilizes intra-class structure information to learn classifiers with robust performance. However, they ignored the inter-class structure information which was not able to effectively separate the different classes in projection space. Moreover, the intra-class graph of MEKL-LPC is constructed by all same class samples, but according to Yang and Chen (2014), the joint of locally constructed intra-class and globally constructed inter-class graphs is more discriminant. Thus, inspired by Discriminant Locality Preserving Projections (DLPP) (Yu et al. (2006)) which is a supervised linear dimensionality reduction method, we design a discriminant locality preservation regularization and introduce it into MEKL to propose a novel MEKL algorithm named Multiple Empirical Kernel Learning with Discriminant Locality Preservation (MEKL-DLP). MEKL-DLP increases the between-class distance and reduces the within-class distance locally, while guarantees lower generalization error. Therefore, structure and discriminant information in feature space is fully utilized by MEKL-DLP to achieve a favorable classification performance.

The proposed MEKL-DLP method first maps the input samples into multiple empirical feature space according to different EKMs. Then, by introducing the discriminant locality preservation regularization into the learning framework, the structure and discriminant information in each empirical space is fully considered to learning. Finally, different classifiers of each empirical feature space are combined to obtain the final classifier. In order to validate the effectiveness of MEKL-DLP, the experiments are conducted on a number of real-world data sets. The results demonstrate that our proposed MEKL-DLP method provides superior performance compared with state-of-art MKL algorithms and classical kernel-based algorithms. Moreover, we further use Bayesian analysis to prove the superiority of our method in the statistic.

The rest of this paper is organized as follows: Section 2 gives a brief introduction of EKM and DLPP. Section 3 describes the architecture of proposed MEKL-DLP and provides its pseudo-code. The experimental results and corresponding analysis of MIKL-DLP on real world datasets are reported in Section 4. Finally, the conclusions are presented in Section 5.

2. Relation Work

2.1. Empirical Kernel Mapping

Given the training samples $\{(x_i, \varphi_i)\}_{i=1}^N$, $\varphi_i \in \{+1, -1\}$ is the label of x_i, each x_i is mapped by a kernel mapping function Φ from input space to feature space. Traditionally, the mapping Φ is implicitly represented by a specified kernel function as the inner-product form between each pair samples in the feature space. But Xiong et al. (2005) give the mapping function an explicit form, which is named EKM and donated by Φ^e.

For the sample set $\{x_i\}_{i=1}^N$, the $K = [ker_{ij}]_{N\times N}$ donates the $N \times N$ kernel matrix where $ker_{ij} = \Phi(x_i) \cdot \Phi(x_j)$. Thus, K is a symmetrical positive semi-definite matrix. Suppose that the rank of K is r, then it can be decomposed as:

$$K = Q_{N\times r}\Lambda_{r\times r}Q_{N\times r}^T \tag{1}$$

where $\Lambda_{r\times r}$ is a $r \times r$ diagonal matrix with r positive eigen-values of kernel matrix K, and $Q_{N\times r}$ is the orthonormal eigen-vectors corresponding to the eigen-values. Thus, Φ^e can be represented as:

$$\Phi^e(x) = \Lambda_{r\times r}^{-\frac{1}{2}} Q_{N\times r}^T [ker(x, x_1), ..., ker(x, x_N)]^T \tag{2}$$

According to Eq.(2), different kernel function is corresponding to different EKM. The dimension of the mapped feature corresponds to the rank r of the kernel matrix K.

2.2. Discriminant Locality Preserving Projections

DLPP tries to find the subspace that best discriminates different classes by maximizing the between-class distance, while minimizing the within-class distance. Given the training samples $\{x_i\}_{i=1}^N, x \in R^d$ and each x_i belongs to exactly one of C classes $\{\varphi_1, \varphi_2 ... \varphi_C\}$. DLPP tries to maximize an objective function as follows (Yu et al. (2006)):

$$J = \frac{\sum_{p,q}^{C}(m_p - m_q)B_{pq}(m_p - m_q)^T}{\sum_{c=1}^{C}\sum_{i,j=1}^{n_c}(y_i^c - y_j^c)S_{ij}^c(y_i^c - y_j^c)^T} \tag{3}$$

where n_c is the number of samples in the cth class, y_i^c represents the ith projected vector in the cth class, m_p and m_q is the mean projected vector for the pth class and qth class, respectively, i.e., $m_p = \frac{1}{n_p}\sum_{k=1}^{n_p} y_k^p$ and $m_q = \frac{1}{n_q}\sum_{k=1}^{n_q} y_k^q$, where n_p and n_q is separately the number of samples in the pth class and qth class. S_{ij} and B_{pq} are the elements of *within-class* weight matrix S and *between-class* weight matrix B, respectively. They are defined as:

$$S_{ij}^c = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}), & x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ & \text{and } x_i, x_j \in \varphi_c \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$B_{pq} = \begin{cases} \exp(-\frac{\|f_p - f_q\|^2}{2\sigma^2}), & f_p \in N_k(f_q) \text{ or } f_q \in N_k(f_p) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where σ is an empirically determined parameter, $N_k(\cdot)$ denotes the k nearest neighbors, $f_p = (\frac{1}{n_p})\sum_{k=1}^{n_p} x_k^p$ is the mean vector of the pth class. Obviously, B and S are symmetric positive semi-definite matrices. Suppose that the mapping from x_i to y_i is W, i.e. $y_i = W^T x_i$, then the objective Eq.(3) can be rewritten as:

$$J(W) = \frac{W^T F H F^T W}{W^T X L X^T W} \tag{6}$$

where L and H are Laplacian matrices. $L = D - S, D = diag(D_1, ..., D_c)$,$D_i$ is a diagonal matrix and its elements are column (or row) sum of S^i. Similarly, $H = E - B$, E is a diagonal matrix and its elements are column (or row) sum of B. $F = [f_1, f_2, ..., f_c]$ is the mean vector matrix in the input space. The columns of transformation matrix $W = [w_1, w_2, ..., w_d]$ that maximizes the objective function Eq.(6) are given by maximum eigenvalues solutions to the generalized eigenvalues problem:

$$FHF^T w_i = \lambda_i X L X^T w_i, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \tag{7}$$

3. Multiple Empirical Kernel Learning with Discriminant Locality Preservation

The proposed MEKL-DLP algorithm adopts the idea of DLPP in empirical feature space and integrates the discriminant locality preservation regularization into empirical kernel leaning framework. Thus, in this section, we will give a description of the Discriminant Locality Preservation (DLP) constraint in each empirical feature space. Then, the architecture of MEKL-DLP is presented.

3.1. Discriminant Locality Preservation in Feature Spaces

Given the sample set $\{(x_i, \varphi_i)\}_{i=1}^N$, x_i can be mapped into different empirical feature spaces via different EKMs which is calculated according to Eq.(2) with different $ker(\cdot,\cdot)$. Supposing there are m EKMs $\left\{\Phi^{el}\right\}_{i=1}^m$, the mapped samples are donated as $\left\{\Phi_i^{el}\right\}_{i=1}^m$ in the lh feature space. Inspired by DLPP but different from its form of using generalized Rayleigh quotient

(Bathe and Wilson (1976)), DLP uses the form of maximum margin criterion (Li et al. (2004)). The DLP constraint in lth feature space is defined as:

$$\begin{aligned} R^l_{dlp} &= \sum_{c=1}^{C}\sum_{i,j=1}^{n_c}\left\|w_l^T(\Phi_i^{el})^c - w_l^T(\Phi_j^{el})^c\right\|_2^2 S_{ij}^c - \gamma\sum_{i,j=1}^{C}\left\|w_l^T(\Phi_{mi}^{el})^c - w_l^T(\Phi_{mj}^{el})^c\right\|_2^2 B_{ij}^c \\ &= w_l^T(\Phi^{el})^T L_l(\Phi^{el})w_l - \gamma w_l^T(\Phi_m^{el})^T H_l(\Phi_m^{el})w_l \\ &= w_l^T[(\Phi^{el})^T L_l(\Phi^{el}) - \gamma(\Phi_m^{el})^T H_l(\Phi_m^{el})]w_l \end{aligned} \tag{8}$$

where $L_l = D_l - S_l$ and $H_l = F_l - B_l$ are Laplacian matrices in lth feature space just like DLPP, S_{ij}^c and B_{ij} can be calculated according to the Eq.(4) and Eq.(5) substitute x_i with Φ_i^{el}, respectively. $\Phi^{el} = [[\Phi_l^{el}; 1]^T; ...; [\Phi^e_{N_{ms}} l; 1]]$ represents the mapped sample matrix, $\Phi_{mi}^{el} = \frac{1}{n_i}\sum_{k=1}^{n_i}\Phi_i^{el}$ is the mean vector of the ith class and w_l is the augmented weight vector in the lth feature space. $\gamma \geq 0$ is the regularization parameter to balance the relative merits of minimizing the *within-class* scatter to the maximization of the *between-class* scatter.

Minimizing the DLP constraint is an attempt to ensure that if Φ_i^{el} and Φ_j^{el} are 'close' in empirical feature space then they are close as well in output space and also ensure that if Φ_{mi}^{el} and Φ_{mj}^{el} are 'far' but they are far in output space.

3.2. Architecture of Proposed MEKL-DLP Method

After constructing m empirical feature spaces and integrating the discriminant locality preservation constraint into each empirical feature space respectively according to the previous sections, here we give the architecture of the proposed MEKL-DLP.

The proposed MEKL-DLP adopts the empirical risk term R_{emp} and the regularization term R_{reg} as done in traditional methods (Xiong et al. (2005); Leski (2003)), which guarantees the correctness of classifier. Moreover, in order to restrain the relationships between all kernels, we introduce the Inter-Function Similarity Loss term (Zhe et al. (2007)):

$$R_{IFSL} = \sum_{l=1}^{m}(f_l - \frac{1}{m}\sum_{j=1}^{m} f_j)^2 \tag{9}$$

where f_l is the output in the lth feature space. Since minimizing R_{IFSL} means minimizing variances of different outputs, information from different kernel spaces is effectively integrated. Finally, the objection function J of the proposed MEKL-DLP integrates the R_{emp}, R_{reg}, R_{dlp} and R_{IFSL} together, which is simply described as:

$$J = \sum_{l=1}^{m}[R_{emp} + cR_{reg} + \beta R_{dlp}] + \lambda R_{IFSL} \tag{10}$$

where m is the number of empirical feature space, c, β, λ are all nonnegative regularization parameter. Different regularization term has the different function in the process of classification. Specifically, the R_{emp} is empirical risk term which usually uses the mean-squared errors. The R_{reg} is the structural risk term which avoids the over-fitting phenomenon, thus the generalization performance of model is improved. The R_{dlp} is our proposed discriminant locality preservation term which increases the structure and discriminant ability of model.

The last term is R_{IFSL} which keeps the output of each kernel be maximally close to the weighted average outputs of all kernels. The R_{emp}, R_{reg} and R_{dlp} regularization terms are added in each empirical kernel space and the parameter c, β are used to adjust the influence of different regularization. The last regularization term is R_{IFSL} which aims to ensure the outputs of different kernel space are similarity and consistent.

In practice, given the training samples $\{(x_i, \varphi_i)\}_{i=1}^{N}, \varphi_i \in \{+1, -1\}$ and m kernels, we explicitly map the input training samples into m empirical feature spaces $\left\{\Phi^{el}(x_1), ..., \Phi^{el}(x_N)\right\}_{l=1}^{m}$. In the lth empirical feature space, let $Y_l = [\varphi_1(\Phi_1^{el}); ...; \varphi_i(\Phi_i^{el}); ...; \varphi_N(\Phi_N^{el})]$ where $\Phi_i^{el} = [\Phi^{el}(x_i); 1]$, and $w = [\hat{w}; w_0]$ where $\hat{w}$ and w_0 are the weight vector and bias, respectively. Then, the empirical risk term R_{emp}, regularization term R_{reg} and R_{IFSL} can be formulated as:

$$R_{emp} = (Y_l w_l - \mathbf{1}_{N\times 1} - \mathbf{b}_l)^T (Y_l w_l - \mathbf{1}_{N\times 1} - \mathbf{b}_l) \tag{11}$$

$$R_{reg} = \hat{w}_l^T \hat{w}_l \tag{12}$$

$$R_{IFSL} = \sum_{l=1}^{m} (Y_l w_l - \frac{1}{m} \sum_{j=1}^{m} Y_j w_j)^2 \tag{13}$$

where the N-dimensional vector $\mathbf{1}_{N\times 1}$ represents the vector with all elements set to 1, and $\mathbf{b}_l$ denotes a non-negative margin vector. After substituting Eq.(8), Eq.(11), Eq.(12) and Eq.(13) into Eq.(10), the final objective function can be expressed as follows:

$$\begin{aligned} \min_{w_1, \mathbf{b}_l \geq 0} J = & \sum_{l=1}^{m} \left\{ (Y_l w_l - \mathbf{1}_{N\times 1} - \mathbf{b}_l)^T (Y_l w_l - \mathbf{1}_{N\times 1} - \mathbf{b}_l) + c \hat{w}_l^T \hat{w}_l \right\} \\ & + \sum_{l=1}^{m} \beta w_l^T [(\Phi^{el})^T L_l (\Phi^{el}) - \gamma (\Phi_m^{el})^T H_l (\Phi_m^{el})] w_l \\ & + \lambda \sum_{l=1}^{m} (Y_l w_l - \frac{1}{m} \sum_{j=1}^{m} Y_j w_j)^T (Y_l w_l - \frac{1}{m} \sum_{j=1}^{m} Y_j w_j) \end{aligned} \tag{14}$$

Each w_l and b_l can be optimized separately by a heuristic gradient descent method (Leski (2003)). By setting gradient of J with respect to w_l and b_l to zero, we obtain:

$$\begin{aligned} w_l = & \left\{ [1 + \lambda(1 + (\frac{m-1}{m})^2)] Y_l^T Y_l + c\hat{I} + \beta[(\Phi^{el})^T L_l (\Phi^{el}) - \gamma (\Phi_m^{el})^T H_l (\Phi_m^{el})] \right\}^{-1} Y_l^T \\ & \times (\mathbf{b}_l + \mathbf{1}_{N\times 1} + \lambda \frac{1}{m} \sum_{j=1; j\neq l}^{m} Y_j w_j) \end{aligned} \tag{15}$$

$$\mathbf{b}_l = Y_l w_l - \mathbf{1}_{N\times 1} \tag{16}$$

where $\hat{I}$ is a diagonal matrix with full 1 except the last element set to 0. Then, in the lth feature space, we import the error vector

$$\mathbf{e}_l = Y_l w_l - \mathbf{1}_{N\times 1} - \mathbf{b}_l \tag{17}$$

By adopting a heuristic gradient descent method, we initialize $\mathbf{b}_l^1 \geq 0$ and update b_l^t at each iteration as follows to refuse decrease any elements of it.

$$\begin{cases} \mathbf{b}_l^1 \geq 0 \\ \mathbf{b}_l^{t+1} = \mathbf{b}_l^t + \rho(\mathbf{e}_l^t + \left\|\mathbf{e}_l^t\right\|) \end{cases} \tag{18}$$

where t is iteration index, and $\rho(\geq 0)$ is the learning rate. Then, the weight vector w_l^{t+1} can be obtained according to Eq.(15). Furthermore, the termination criterion to be $\| J^{t+1} - J^t \|_2 \leq \xi$ where the termination criterion parameter ξ is a small positive constant. The algorithm of our proposed MEKL-DLP is summarized in **Table 1**. Finally, with optimal

Table 1: Algorithm of MEKL-DLP

Input:Training samples $\{(x_i, \varphi_i)\}_{i=1}^N, \varphi \in \{1, -1\}$, and m candidate kernels $\{ker(x_i, x_j)\}_{l=1}^m$.
Output: The weight $w_l, l = 1, ..., m$.

1. Explicitly map $\{x_i\}_{i=1}^N$ into m feature spaces $\left\{\Phi^{el}(x_1), ..., \Phi^{el}(x_i), ..., \Phi^{el}(x_N)\right\}_{l=1}^m$ according to Eq.(2).
2. For each empirical feature spaces, let $\Phi_i^{el} = [\Phi^{el}(x_i); 1], Y_l = [\varphi_1(\Phi_1^{el})^T; ...; \varphi_N(\Phi_N^{el})^T]$, $l = 1, ..., m$
3. Initialize $c \geq 0, \beta \geq 0, \gamma \geq 0, \lambda \geq 0, \mathbf{b}_l^1 \geq 0, \xi \geq 0, t = 1, l = 1, ..., m$
4. Calculate S and B according to Eq.(4) and Eq.(5)
5. Do until the termination condition $\| J^{t+1} - J^t \|_2 \leq \xi$ is satisfied
   a) Calculate w_l^t according to Eq.(15) with $\mathbf{b}_l = \mathbf{b}_l^1$
   b) Calculate e_l^t according to Eq.(17)
   c) Calculate $\mathbf{b}_l^{t+1}$ according to Eq.(18)
6. Return $w_l, l = 1, ..., m$

$w_l, l = 1, ..., m$, the decision function for an input sample z with its corresponding mapped samples $\left\{\Phi^{el}(z)\right\}_{l=1}^m$, can be formulated as:

$$F(z) = \frac{1}{m} \sum_{l=1}^{m} w_l^T [\Phi^{el}(z); 1] \tag{19}$$

If $F(z) \geq 0$, then $z \in class + 1$ and if $F(z) \leq 0$, then $z \in class - 1$.

4. Experimental Studies

In this section, we compare MEKL-DLP with five state-of-art MKL methods and three representative single kernel learning methods to evaluate MEKL-DLPs performance on 10 real-world datasets which are obtained from UCI datasets. The experiment results prove that MEKL-DLP improves classification accuracy compared with these algorithms.

4.1. Datasets

In the experiment, 10 UCI data sets are used to evaluate the performance of MEKL-DLP by comparing the classification performance with other methods. These data sets can be

Table 2: Information for the adopted datasets

Dataset	Instances	Attributes	Classes
Iris	150	4	3
Liver Disorders	345	6	2
Ionosphere	351	34	2
House Vote	435	16	2
BCW	699	9	2
MM	961	6	2
Hill Valley	1212	100	2
CMC	1473	9	3
Semeion	1593	256	10
Segmentation	2310	18	7

accessed from UCI Machine Learning Repository (Asuncion and Newman (2007)), which are *Iris*, *Liver Disorders*, *Ionosphere*, *House Vote*, *Breast Cancer Wisconsin (BCW)*, *Mammographic Masses (MM)*, *Hill Valley*, *Contraceptive Method Choice (CMC)*, *Semeion* and *Segmentation*. **Table 2** shows the detail information about these data sets.

4.2. Experimental Settings

In the experiment, eight kernel-based methods form the comparison group to verify the effectiveness of our proposed MEKL-DLP method. The RBF kernels $ker(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ is chosen as the candidate kernel and the mean kernel bandwidth $\sigma^2 = \eta * \frac{1}{N^2} \sum_{i,j=1}^{N} \|x_i - x_j\|^2$ are used as the kernel parameter, where N is the number of training samples and η is a parameter to change the value of σ. Furthermore, the comparison group is categorized into two parts. The first part is the comparison between our proposed MEKL-DLP method and the single kernel learning methods such as KMHKS (Leski (2003)), SVM (Vapnik (1995)) and KFDA (Liu et al. (2004)). Because MEKL-DLP is a MKL method, the number of kernel is set to 3 and the parameter η is respectively set to $2^{-2}, 2^0, 2^2$, while the η is set to $2^0 = 1$ for KMHKS, SVM and KFDA. The second part is the comparison between our proposed MEKL-DLP method and the MKL methods such as MEKL-LPC(Fan et al. (2016)), MultiK-MHKS (Zhe et al. (2007)), SimpleMKL (Rakotomamonjy et al. (2008)), EasyMKL (Aiolli and Donini (2015)) and GLMKL (Xu et al. (2010)). For easy comparison, the number of RBF kernel is also set to 3 expect EasyMKL and the kernel bandwidth parameter η is selected from $\{2^{-4}, 2^{-2}, ..., 2^4\}$. Since EasyMKL is a scalable multiple kernel learning algorithm, the number of kernels is set to 50. For all MHKS-based algorithms such as MEKL-DLP, MEKL-LPC, MultiK-MHKS and KMHKS, the learning rate ρ and the initial value of b is set to 0.99 and 10^{-6}. Meanwhile, all the regularization parameter $c, \beta, \lambda, \gamma$ are chosen from $\{10^{-2}, 10^{-1}, ..., 10^2\}$. For MEKL-DLP, the local nearest neighbor parameters k is selected from 1,3,5,7,9. For all SVM-based algorithms, such as SimpleMKL and GLMKL, the parameter c is also chosen from $\{10^{-2}, 10^{-1}, ..., 10^2\}$. In addition, the 5-fold cross validation (Braga-Neto and Dougherty (2004)) approach is used for the parameter selection.

4.3. Performance Comparison with single kernel learning methods

To validate the effectiveness of MEKL-DLP for integrating multiple kernels, we compare the classification accuracy between our method and single kernel learning methods such as KMHKS, KFDA and SVM on 10 UCI data sets. The experiment results are listed in **Table 3** and the best results on each data set are highlighted in **BOLD**.

Table 3: Classification accuracy of MEKL-DLP and single kernel learning methods on UCI datasets

Data Set	MEKL-DLP	KMHKS	KFDA	SVM
Iris	$\mathbf{98.67 \pm 1.83}$	97.33 ± 2.79	94.67 ± 2.98	97.33 ± 2.79
Liver Disorders	$\mathbf{71.59 \pm 3.01}$	68.99 ± 4.76	61.16 ± 4.02	60.33 ± 4.59
Ionosphere	$\mathbf{94.87 \pm 3.29}$	92.01 ± 2.98	80.06 ± 6.04	93.72 ± 3.32
House Vote	$\mathbf{94.70 \pm 1.97}$	93.78 ± 1.08	87.60 ± 2.94	92.24 ± 3.44
BCW	96.84 ± 2.43	96.55 ± 2.01	90.12 ± 2.04	$\mathbf{97.11 \pm 1.81}$
MM	57.75 ± 3.67	$\mathbf{59.10 \pm 4.06}$	53.49 ± 2.00	55.29 ± 6.02
Hill Valley	$\mathbf{52.63 \pm 2.70}$	52.56 ± 1.99	48.76 ± 3.08	50.09 ± 0.19
CMC	56.22 ± 1.95	$\mathbf{54.44 \pm 2.05}$	48.66 ± 2.33	53.58 ± 3.18
Semeion	$\mathbf{95.63 \pm 1.42}$	93.00 ± 1.69	95.31 ± 0.94	94.84 ± 1.80
Segmentation	97.84 ± 1.05	94.03 ± 1.30	96.32 ± 1.46	94.33 ± 1.13
Average	$\mathbf{81.67 \pm 2.33}$	80.18 ± 2.47	75.61 ± 2.78	78.89 ± 2.83

According to the results, we can conclude that: 1) Compared with single kernel learning methods, MEKL-DLP achieves the best classification performance on 7 datasets and the highest average classification accuracy among the compared methods. It is evident that the proposed MEKL-DLP is significantly superior to the other single-kernel-based methods. 2) Compared with KMHKS which is the kernelization of MHKS, MEKL-DLP combines multiple kernel spaces information to achieve superior performance. 3) Compared with KFDA which also utilizes the with-class and between-class information, MEKL-DLP provides a better discriminant ability by introducing the class information of samples into the learning framework.

4.4. Performance Comparison with multiple kernel learning methods

To validate the performance of MEKL-DLP with other MKL methods, five state-of-art MKL methods such as MultiK-MHKS, MEKL-LPC, SimpleMKL, GLMKL and EasyMKL are compared and discussed in this section. The experiment results are listed in **Table 4** and the best results on each data set are highlighted in **BOLD**. Meanwhile, the average classification accuracies are given in the last row of **Table 4**.

According to the results, we can conclude that: 1) MEKL-DLP obtains the best classification accuracies on 7 datasets and the highest average classification accuracy among the compared methods, which demonstrates that the method achieves more robust classification performance by introducing the discriminant locality preservation into the learning framework. 2) Compared with the related methods such as MultiK-MHKS and MEKL-LPC, our proposed MEKL-DLP achieves the best performance on 9 datasets. It indicates that MEKL-DLP combines the local intra-class structure information and global inter-class discriminant information to further improve the generalization and robustness performance.

3) On *Liver Disorders*, *MM*, *Hill Valley* and *CMC* datasets which are difficult to classify, the performances of MEKL-DLP are significantly improved than SimpleMKL which is a state-of-art MKL method. It indicates that our proposed MEKL-DLP method is more advantageous on datasets which are difficult to classify.

Table 4: Classification accuracy of MEKL-DLP and MKL methods on UCI datasets

Data Set	MEKL-DLP	MultiK-MHKS	MEKL-LPC	SimpleMKL	GLMKL	EasyMKL
Iris	**98.67 ± 1.83**	98.00 ± 1.83	97.33 ± 2.79	96.00 ± 3.65	96.67 ± 2.36	98.00 ± 1.63
Liver Disorders	**73.04 ± 1.59**	70.43 ± 1.65	71.30 ± 3.14	72.17 ± 7.63	70.72 ± 3.30	66.67 ± 3.67
Ionosphere	**95.44 ± 2.75**	91.73 ± 5.93	92.28 ± 1.35	95.43 ± 3.41	94.30 ± 2.86	91.72 ± 5.61
House Vote	94.95 ± 1.00	94.52 ± 2.97	94.48 ± 1.51	94.22 ± 3.00	94.03 ± 0.89	**95.39 ± 2.64**
BCW	**97.42 ± 1.88**	96.69 ± 2.95	96.93 ± 1.41	97.28 ± 1.08	96.99 ± 1.94	94.98 ± 2.69
MM	**57.75 ± 3.67**	56.50 ± 4.38	55.15 ± 3.79	56.71 ± 4.06	53.69 ± 0.11	56.82 ± 4.39
Hill Valley	**53.13 ± 3.85**	52.64 ± 2.26	52.29 ± 3.92	51.73 ± 2.53	52.06 ± 3.34	53.05 ± 4.39
CMC	56.22 ± 1.95	55.89 ± 1.80	**56.49 ± 4.08**	54.58 ± 3.32	55.07 ± 2.20	50.92 ± 2.40
Semeion	**95.63 ± 1.42**	94.50 ± 1.26	95.49 ± 1.17	95.05 ± 0.34	94.94 ± 1.52	86.44 ± 1.96
Segmentation	97.84 ± 1.05	96.06 ± 1.41	97.49 ± 0.82	97.36 ± 0.64	97.45 ± 0.66	**98.05 ± 0.56**
Average	**82.01 ± 2.10**	80.7 ± 2.64	80.75 ± 2.42	81.05 ± 2.96	80.59 ± 1.92	79.21 ± 2.99

4.5. Bayesian analysis

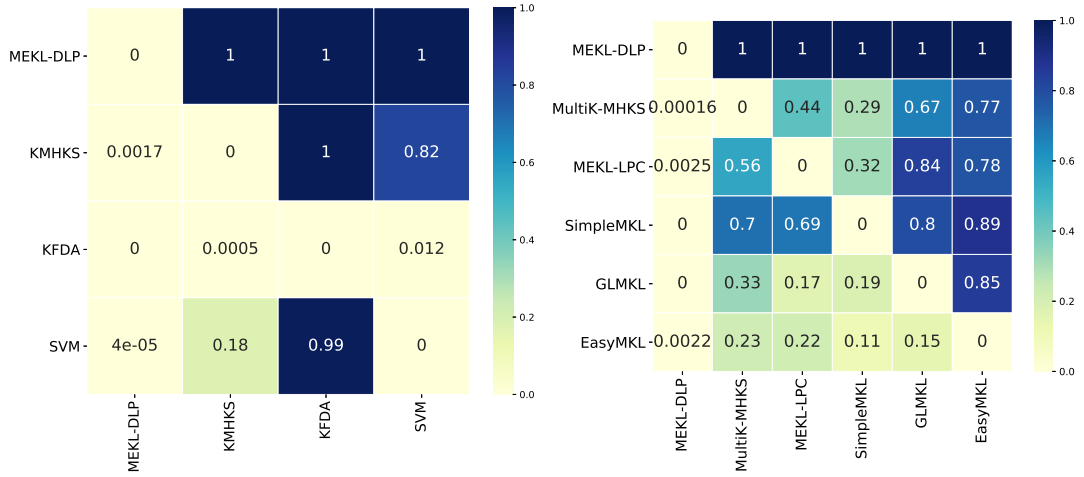
In our experiment, Bayesian analysis (Benavoli et al. (2016)) is considered to further compare the classification performance of different algorithms. Bayesian analysis takes both magnitude and uncertainty into account to estimate the performance of classifier. The assumption of Bayesian analysis is that the difference between two estimators in a certain metric is a normal distribution. Using a Bayesian signed rank test method, two probability matrixes are obtained and shown in **Fig.1**. **Fig.1-a** is the probability matrix of MEKL-DLP and single kernel learning methods and **Fig.1-b** is the probability matrix of MEKL-DLP and MKL methods. The value in row ith and column jth represents the probability that $method_{ith}$ exceeds $method_{jth}$. Actually, this probability matrix indicates the probability that the difference between the two methods is more than q. Commonly, it is non-equivalent that two classifiers whose mean difference is more than 1%. Thus, the parameter q is set to 1% in this analysis.

As can be seen from the first row of **Fig.1**, compared to the single kernel learning and multiple kernel learning methods, the classification performances of MEKL-DLP are both superior on all data sets with q=1%. Therefore, it is evident that the proposed MEKL-DLP is statistically superior to the other compared methods.

4.6. Analysis on β

In MEKL-DLP, the regularization parameter β controls the contribution of discriminant locality preservation regularization R_{dlp} to the decision hyperplane. Thus, experiments are designed on the 10 UCI data sets to track the influence of β. The rest parameters are fix as the best ones selected via 5-fold cross validation and the value of β is selected from $\{10^{-2}, 10^{-1}, ..., 10^{2}\}$. For each β, the average classification accuracies are shown in **Fig.2** and **Fig.3**.

It can be concluded that: 1) On most data sets, especially those with high accuracy, the classification accuracies are obviously fluctuate with the varying of β, which indicates



(a) (b)

Figure 1: Probability matrices obtained from a Bayesian signed rank test on the UCI datasets. The value in row ith and column jth represents the probability that $method_{ith}$ exceeds $method_{jth}$ with q for the corresponding metric. (a) Probability matrices of MEKL-DLP and single kernel learning methods. (b) Probability matrices of MEKL-DLP and MKL methods.

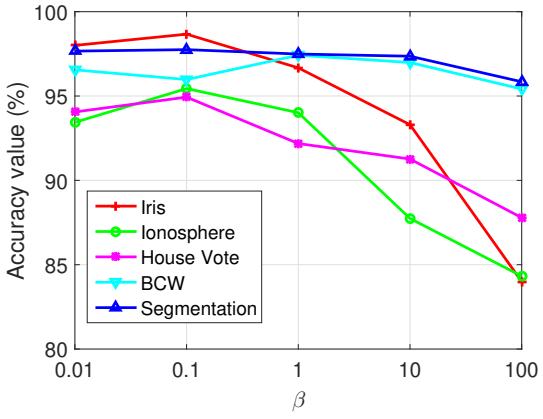


Figure 2: Accuracy values (%) of MEKL-DLP with the variation of parameter β on Iris, Ionosphere, House Vote, BCW, Segmentation.

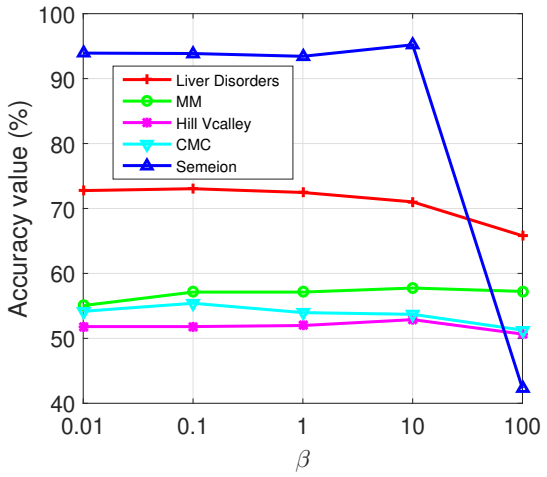


Figure 3: Accuracy values (%) of MEKL-DLP with the variation of parameter β on Liver Disorders, MM, Hill Vcalley, CMC, Semelon.

that β has the great influence on the classification performance of MEKL-DLP. 2) On several low accuracy data sets, such as *MM*, *CMC* and *Hill Vcalley*, the classification accuracy will rise first and then fall with the β increasing. Thus, our selection of value β in $\left\{10^{-2}, 10^{-1}, ..., 10^{2}\right\}$ is sufficient for MEKL-DLP to achieve favorable classification performance. 3) On all data sets, it is obvious that when β exceeds to 10, the classification performance decreases greatly which indicates that to attach more importance to R_{dlp} is not appropriate for improving the classification performance.

5. Conclusion

In this paper, we propose a novel multiple kernel learning method named MEKL-DLP. Inspired by DLPP, the proposed method integrates discriminate locality preservation into MEKL framework, which effectively uses the distribution and discriminant information of samples to improve the classification performance. Furthermore, to validate the effectiveness of MEKL-DLP, experiments are designed to compare MEKL-DLP with five stats-of-art multiple kernel learning algorithms and three representative single kernel learning algorithms on 10 real-world UCI data sets. Experimental results indicate that MEKL-DLP outperforms the compared algorithms. Meanwhile, Bayesian analysis is used to further demonstrate the superiority of our algorithm. In addition, we discuss the influence of discriminate locality preservation regularization parameter β to the classification performance. The results express that β greatly impacts the classification accuracy and it should not be set to a large value. In General, by introducing the discriminate locality preservation into MEKL, the proposed MEKL-DLP results in more robust classification performance.

References

Fabio Aiolli and Michele Donini. Easymkl: a scalable multiple kernel learning algorithm. *Neurocomputing*, 169:215–224, 2015.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

Klaus-Jürgen Bathe and Edward L Wilson. Numerical methods in finite element analysis. 1976.

Alessio Benavoli, Giorgio Corani, Janez Demsar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18, 2016.

Ulisses M. Braga-Neto and Edward R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004.

James Breneman. Kernel methods for pattern analysis. john shawe-taylor and nello cristianini. *Journal of the American Statistical Association*, 101(December):1730–1730, 2006.

Serhat S Bucak, Rong Jin, and Anil K Jain. Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1354–1369, 2014.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(2):795–828, 2012.

Qi Fan, Daqi Gao, and Zhe Wang. Multiple empirical kernel learning with locality preserving constraint. *Knowledge-Based Systems*, 105:107–118, 2016.

Mark Girolami and Simon Rogers. Hierarchic bayesian models for kernel learning. In *International Conference on Machine Learning*, 2005.

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

Xia Hao and Steven C. H. Hoi. Mkboost: A framework of multiple kernel boosting. *IEEE Transactions on Knowledge & Data Engineering*, 25(7):1574–1586, 2013.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12(Mar):953–997, 2011.

Gert R. G Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semi-definite programming. In *Nineteenth International Conference on Machine Learning*, 2002.

Jacek Leski. Ho-kashyap classifier with generalization control. *Pattern Recognition Letters*, 24(14):2281–2290, 2003.

Haifeng Li, Tao Jiang, and Keshu Zhang. Efficient and robust feature extraction by maximum margin criterion. In *Advances in neural information processing systems*, pages 97–104, 2004.

Qingshan Liu, Hanqing Lu, and Songde Ma. Improving kernel fisher discriminant analysis for face recognition. *IEEE transactions on circuits and systems for video technology*, 14 (1):42–49, 2004.

Juwei Lu, Plataniotis K N, and Venetsanopoulos A N. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Trans Neural Netw*, 14(1):117–126, 2003.

Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2):181, 2001.

Alain Rakotomamonjy, Francis R Bach, Stéphane Canu, and Yves Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9(3):2491–2521, 2008.

Vladimir N. Vapnik. *The nature of statistical learning theory*. 1995.

Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *2009 IEEE 12th international conference on computer vision*, pages 606–613. IEEE, 2009.

Huilin Xiong, MNS Swamy, and M Omair Ahmad. Optimizing the kernel in the empirical feature space. *IEEE transactions on neural networks*, 16(2):460–474, 2005.

Zenglin Xu, Rong Jin, Haiqin Yang, Irwin King, and Michael R Lyu. Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1175–1182. Citeseer, 2010.

Bo Yang and Songcan Chen. A comparative study: globality versus locality for graph construction in discriminant analysis. *Journal of Applied Mathematics*, 2014, 2014.

Jieping Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6(1):483–502, 2005.

Jieping Ye, Tao Li, Tao Xiong, and Ravi Janardan. Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(4):181–190, 2004.

Weiwei Yu, Xiaolong Teng, and Chongqing Liu. *Face recognition using discriminant locality preserving projections*. 2006.

Wang Zhe, Chen Songcan, and Sun Tingkai. Multik-mhks: a novel multiple kernel learning algorithm. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 30(2):348–353, 2007.