Elastic Multiple Kernel Learning

WU Zheng-Peng1, ZHANG Xue-Gong1

Abstract Multiple kernel learning (MKL) was proposed to deal with kernel fusion. MKL learns a linear combination of several kernels and solves the supporting vector machine (SVM) associated with the combined kernel simultaneously. Current framework of MKL encourages sparsity of the kernel combination coefficients. When a significant portion of the kernels are informative, forcing sparsity tends to select only a few kernels and may ignore useful information. In this paper, we propose elastic multiple kernel learning (EMKL) to achieve adaptive kernel fusion. EMKL makes use of a mixing regularization function to compromise sparsity and non-sparsity. Both MKL and SVM could be regarded as special cases of EMKL. Based on gradient descent algorithm for MKL problem, we propose a fast algorithm to solve EMKL problem. Results on the simulation datasets demonstrate that the performance of EMKL compares favorably to both MKL and SVM. We further apply EMKL to gene set analysis and get promising results. Finally, we study the theoretical advantage of EMKL comparing to other non-sparse MKL.

Key words Support vector machine (SVM), multiple kernel learning (MKL), elastic multiple kernel learning (EMKL), regularization


Since Support vector machine (SVM) was proposed in 1999[3], kernel methods have obtained a great development and been applied in various learning problems. Kernel methods implicitly embed the data into a Hilbert space through kernel function $K(x, z)$ which defines the inner product of the Hilbert space. Performance of kernel methods strongly depends on the kernel function. In order to deal with kernel fusion problem and give more flexibility to kernel function, Lanckriet et al. proposed multiple kernel learning (MKL) which considers a group of kernels simultaneously[2]. MKL pursues the optimal linear combination of a group of kernels:

$$K(x, z) = \sum_{k=1}^{m} \mu_k K_k(x, z)$$  \hspace{1cm} (1)

Given some constraints on combination coefficients $\mu_k$ (e.g., if sum is constant) or on combined kernel $K$ (e.g., positive definite), MKL solves $\mu_k$ and the SVM associated with combined kernel simultaneously. The kernel $K_k$ can be obtained from different kernel functions (e.g., linear, polynomial, and Gaussian function) using the same features, or obtained from the same kernel function using several data sources. The former case could be used for kernel selection and the latter could be used for data fusion. These two cases can be treated in a unified framework of kernel learning, so we do not distinguish them in present paper.

Bach et al.[3] proposed the equivalence between MKL and support kernel machines (SKM). From the formulation of SKM, it can be concluded that the regularization function encourages sparse kernel combinations. When informative kernels take a significant portion of all, forcing sparsity may lead to discarding classification information. This is the starting point of our study.

In this paper, we propose elastic multiple kernel learning (EMKL) to deal with adaptive multiple kernel learning problem. EMKL makes use of a mixing norm regularization function to compromise sparsity and non-sparsity. In order to solve EMKL efficiently, we develop a rapid iterative gradient descent algorithm which significantly reduces the computation cost comparing with the general optimization software SeDuMi. A series of simulation studies demonstrate that EMKL outperforms MKL and SVM in classification accuracies and achieves more efficient kernel learning than MKL. The application of EMKL to the gene set analysis also gets promising results. Finally, we analyze the theoretical advantage of EMKL comparing to other non-sparse MKL.

1 Methods

1.1 Notations and concepts

In this paper, we consider binary classification problems. Given data set $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X}$ for some feature space $\mathcal{X}$ and $y_i \in \{+1, -1\}$ is the label of $x_i$. Kernel methods embed the data into a Hilbert space $\mathcal{H}$ through mapping $\phi: \mathcal{X} \rightarrow \mathcal{H}$. Kernel function is defined as the inner product in $\mathcal{H}$: $K(x, z) = \langle \phi(x), \phi(z) \rangle$. Aim of kernel methods is to find a function: $f(x) = \langle w, \phi(x) \rangle$ which minimizes the following regularization problem:

$$\min_{f} \sum_{i=1}^{n} L(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2$$  \hspace{1cm} (2)

where $L$ is loss function, $\lambda$ is regularization parameter, and $\|f\|_{\mathcal{H}}$ is the norm of $f$ in $\mathcal{H}$. Under modest assumptions about the loss function and the Hilbert space, from classical representer theorem[4–5], the solution of (2) has a linear form:

$$f(z) = \sum_{i=1}^{n} \alpha_i \phi(x_i), \phi(z) \rangle = \sum_{i=1}^{n} \alpha_i K(x_i, z)$$  \hspace{1cm} (3)

Kernel methods only deal with the kernel function $K(x, z)$, rather than the explicit form of $\phi$. This is so-called kernelization technique. In the following sections, we will consider Euclidean norm and inner product, due to kernelization technique, all conclusions will be valid for general mappings or kernels.

1.2 Multiple kernel learning

MKL considers a group of mappings $\phi_k : \mathcal{X} \rightarrow \mathcal{H}_k$. Giving weights to these mappings and stacking them together, we have: $\phi(z) = (\sqrt{\mu_1} \phi_1(z)^T, \cdots, \sqrt{\mu_m} \phi_m(z)^T)^T$ (all vectors appearing in this paper will be column vectors by default). Then, the kernel function has the following form:

$$K(x, z) = \langle \phi(x), \phi(z) \rangle = \sum_{k=1}^{m} \mu_k \langle \phi_k(x), \phi_k(z) \rangle = \sum_{k=1}^{m} \mu_k K_k(x, z)$$  \hspace{1cm} (4)

Manuscript received May 5, 2010; accepted January 22, 2011.

Supported by National Natural Science Foundation of China (61021063)

1 Ministry of Education Key Laboratory of Bioinformatics and Bioinformatics Division, Tsinghua National Laboratory of Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100080, P. R. China
Assuming $X = R^{d_1} \times \cdots \times R^{d_m}$ and $Y = R^{d_k}$, we define $\phi_k(x_i) = x_{i,k}$, where $x_{i,k}$ is the component of $x_i$ belonging to $R^{d_k}$. (all through this paper, we will always use $n$ and $m$ to denote the number of samples and kernels respectively, and use $i$ and $k$ to index samples and kernels, respectively). Thus, the kernels $\{K_1, \cdots , K_m\}$ can be viewed as defined on $R^{d_1} \times \cdots \times R^{d_m}$, respectively. The kernelization technique still applies. Above formulation is equivalent to split the feature space into several subspaces, and we will take this view on MKL in this paper.

Recalling the formulation of support vector machines, it seeks the minimizer of the following regularization problem:

$$\min_{w} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

w.r.t. $w \in R^{d_1} \times \cdots \times R^{d_m}$

s.t. $y_i \langle w, x_{i,k} \rangle + b \geq 1 - \xi_i$

Its decision function has the form:

$$f(z) = \langle w, z \rangle + b = \sum_{i=1}^{n} \alpha_i y_i \langle x_i, z \rangle + b$$

According to the block-based formulation proposed in [3], MKL seeks the optimizer of following problem:

$$\min_{w} \frac{1}{2} \left( \sum_{k=1}^{m} \|w_k\|^2 \right)^2 + C \sum_{i=1}^{n} \xi_i$$

w.r.t. $w = (w_1^T, \cdots , w_m^T)^T \in R^{d_1} \times \cdots \times R^{d_m}$

$b \in R^d \times \cdots \times R^d$

s.t. $y_i \left( \sum_{k=1}^{m} (w_k, x_{i,k}) + b \right) \geq 1 - \xi_i$

Its decision function has the form:

$$f(z) = \langle w, z \rangle + b = \sum_{i=1}^{n} \left( \sum_{k=1}^{m} \mu_k \langle x_{i,k}, z_k \rangle \right) + b$$

Here, $\mu_k$ is kernel combination coefficient and can be recovered from the primal-dual solution of (7). MKL uses $\ell_1$-norm at block level. Therefore, it encourages the sparsity of $w$ at block level.

### 1.3 Elastic multiple kernel learning

Using $\ell_1$-norm to encourage sparsity was first proposed in multiple linear regression[6]. In a multiple linear regression problem given observations $Y = (y_1, \cdots , y_n)^T$ and predictors $X = (x_{11}, \cdots , x_{np})$ where $x_j = (x_{1j}, \cdots , x_{nj})^T$, the underlying model is:

$$Y = Xw + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I_n)$ and $w = (w_1, \cdots , w_p)^T$. Lasso involves the following regularization:

$$\min_{\beta} \frac{1}{2n} \|Y - Xw\|^2 + \lambda_n \|w\|_1$$

where $\|\|$ means $\ell_2$-norm and $\|\|_1$ means $\ell_1$-norm. Zou et al.[7] proposed elastic-net which employs a mixing norm regularization to adaptively adjust sparsity:

$$\min_{\beta} \frac{1}{2n} \|Y - Xw\|^2 + \lambda_n \|w\|_1 + \frac{\mu_n}{2} \|w\|^2$$

Elastic-net controls the compromise between $\ell_1$-norm term and $\ell_2$-norm term through the regularization parameters $(\lambda_n, \mu_n)$. It achieves more flexible variable selection compared to Lasso[7].

In present paper, we introduce a mixing norm regularized multiple kernel learning which has the following form:

$$\min_{\beta} \frac{1}{2} (1 - \lambda) \left( \sum_{k=1}^{m} \|w_k\|^2 \right)^2 + \frac{1}{2} \lambda \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

w.r.t. $w = (w_1^T, \cdots , w_m^T)^T \in R^{d_1} \times \cdots \times R^{d_m}$

$b \in R^d \times \cdots \times R^d$

s.t. $y_i \left( \sum_{k=1}^{m} (w_k, x_{i,k}) + b \right) \geq 1 - \xi_i$

where $\|w\|^2 = \sum_{k=1}^{m} \|w_k\|^2$ and $\lambda$ is the tuning parameter between 0 and 1. Due to the usage of the mixing norm regularization, we named it as elastic multiple kernel learning. Different from elastic-net regression, EMKL treats the hinge loss function instead of square loss and deals with the blocks instead of single variables. Furthermore, we use $(\sum_{k=1}^{m} \|w_k\|)^2$ other than $\sum_{k=1}^{m} \|w_k\|$. As we will show in following sections, this square term of sum of $\ell_1$-norms has an intuitive geometric interpretation and makes optimization easier in multiple kernel learning framework.

In EMKL framework, the compromise between sparsity and non-sparsity is controlled by the regularization parameter $\lambda$. When $\lambda$ is relatively large, the $\ell_2$-norm term becomes main part of optimization. This will reduce the sparsity of weight vector at the level of blocks. When $\lambda$ approaches 1, EMKL degenerates into SVM whose kernel is the sum of all sub-kernels. This case is also known as canonical MKL. For simplicity, we will not distinguish canonical MKL and SVM in following sections. When $\lambda$ is relatively small, the $\ell_1$-norm term becomes main part of optimization. This situation will encourage sparsity of the weight vector at the level of blocks. When $\lambda$ approaches 0, EMKL degenerates into MKL. We noticed that Bach et al.[3] used a similar formulation to solve MKL by letting $\lambda$ approach 0. For specific classification problem, we use cross-validation to tune the regularization parameter $\lambda$.

### 1.4 Geometric interpretation of mixing norm regularization

We rewrite the regularization function of EMKL as following (ignore the constant):

$$(1 - \lambda) \left( \sum_{k=1}^{m} \|w_k\|^2 \right)^2 + \lambda \|w\|^2 = \left( \begin{array}{c} 1 \cdots 1 - \lambda \end{array} \right) \left( \begin{array}{c} \|w_1\| \vdots \vdots \|w_m\| \end{array} \right) \left( \begin{array}{c} 1 - \lambda \cdots 1 \end{array} \right)$$

So the regularization function is a quadratic form of $(\|w_1\|, \cdots , \|w_m\|)$. With different regularization parameter, this quadratic form has different characteristics. In order to illustrate this, we consider a simple case where the three dimensional feature space is separated into two blocks, $w_1 = \{\omega_1\}$ and $w_2 = \{\omega_2, \omega_3\}$. 
Fig. 1 Contours and contour sections of EMKL regularization function with different parameters

Figs. 1 (a) ∼ (c) are contours of regularization function in different parameters. Fig. 1 (a) corresponds to \( \lambda = 1 \). Regularization function becomes \( \omega_1^2 + \left( \sqrt{\omega_2^2 + \omega_3^2} \right)^2 = 1 \). This is also the case of SVM. Fig. 1 (b) corresponds to \( \lambda = 0.5 \), where the regularization function is \( \omega_1^2 + \left( \sqrt{\omega_2^2 + \omega_3^2} \right)^2 + \omega_1 \sqrt{\omega_2^2 + \omega_3^2} = 1 \). It is a typical case of EMKL. Fig. 1 (c) corresponds to \( \lambda = 0 \). Regularization function is \( \left( \omega_1 + \sqrt{\omega_2^2 + \omega_3^2} \right)^2 = 1 \). This is just the case of MKL. Figs. 1 (d) ∼ (f) show the horizontal sections of the contours when \( \omega_1 = 0 \). Within the block \( w_2 = \{ \omega_2, \omega_3 \} \), three cases treat all directions equally and do not encourage sparsity. Figs. 1 (g) ∼ (i) show the longitudinal sections of the contours. At block level, SVM treats all directions equally, MKL treats coordinate directions differently from other directions, and EMKL is intermediate between these two extreme cases.

Through above illustrations, we can see the regularization function of MKL has the sharpest contour and tends to produce sparse solution. On the other hand, the regularization function of SVM has a spherical shape contour and tends to produce non-sparse solution. EMKL builds a bridge between SVM and MKL.

2 Algorithm

In this section we propose a rapid algorithm for solving EMKL problem, based on gradient descent algorithm for MKL. Due to a variational formulation of \( \ell_1 \)-norm block regularization[8-9], objective function of MKL has the following equivalent formulation:

\[
\min_{\sum_{k=1}^{m} v_k = 1, v_k \geq 0} \frac{1}{2} \sum_{k=1}^{m} \frac{1}{v_k} \|w_k\|^2 + C \sum_{i=1}^{n} \xi_i \quad (14)
\]

Rakotomamonjy et al. [4] proposed an algorithm for solving MKL through above variational formulation. We also adopt the variational formulation and transform EMKL problem into the following form:

\[
\min_{\sum_{k=1}^{m} v_k = 1, v_k \geq 0} \frac{1}{2} (1 - \lambda) \sum_{k=1}^{m} \frac{1}{v_k} \|w_k\|^2 + \frac{1}{2} \lambda \|w\|^2 + C \sum_{i=1}^{n} \xi_i
\]

w.r.t. \( w = (w_1^T, \ldots, w_m^T)^T \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_m} \)

\( \xi_i \in \mathbb{R}_+, b \in \mathbb{R} \)

s.t. \( y_i \left( \sum_{k=1}^{m} w_k x_{i,k} + b \right) \geq 1 - \xi_i \quad (15) \)

Consider (15) as a constrained optimization problem:

\[
\min_{\sum_{k=1}^{m} v_k = 1, v_k \geq 0} \theta(v) \quad (16)
\]

where \( \theta(v) \) is the optimal value of following optimization
problem:

\[
\min \frac{1}{2} (1 - \lambda) \sum_{k=1}^{m} \frac{1}{v_k} \|w_k\|^2 + \frac{1}{2} \lambda \|w\|^2 + C \sum_{i=1}^{n} \xi_i
\]

w.r.t. 

\[
w = (w_1^T, \cdots, w_m^T)^T \in \mathbb{R}^{d_1 \times \cdots \times d_m}
\]

\[
\xi_i \in \mathbb{R}_+, \quad b \in \mathbb{R}
\]

s.t. 

\[
y_i \left( \sum_{k=1}^{m} w_k(x_{i,k}) + b \right) \geq 1 - \xi_i
\]

Then we have following statement:

**Proposition 1.** The dual problem of (17) is

\[
\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \alpha^T \left( \sum_{k=1}^{m} \frac{v_k}{1 - \lambda + \lambda v_k} K_k \right) \alpha
\]

s.t. \[\sum_{i=1}^{n} \alpha_i y_i = 0, \quad C \geq \alpha \geq 0\] (18)

It is straightforward to find the optimal value \(\theta(v)\) using any SVM algorithm. Existence of derivatives of \(\theta(v)\) refers to [4,10]. The derivatives of \(\theta(v)\) can be calculated as following:

\[
\frac{\partial \theta}{\partial v_k} \bigg|_{v=v^*} = - \frac{1}{2} \frac{\alpha^T}{(1 - \lambda + \lambda v_k)^2} K_k \hat{\alpha}
\]

where \(\hat{\alpha}\) is optimal solution at \(v = \hat{v}\). Using the gradients of \(\theta(v)\), we update \(v\) through gradient descent method. The outline of this algorithm is shown in Algorithm 1.

We take into account the equality constraint and positivity constraint on \(v\) when we compute descent gradient \(\nabla \theta_i\). The optimal step \(\lambda_i\) is calculated using Armijo’s rule. The stopping criterion can be based on duality gap or maximum number of iterations. The details of calculation and convergence issues refer to [4,10]. Through conic duality, EMKL can also be solved using general-purpose optimization software such as SeDuMi[11].

Using simulation data, we compare the computation time of the proposed algorithm with the SeDuMi. We choose 100 samples in each experiment and let the dimension of each kernel to be 10. Then, we increase the number of kernels to investigate the computation time of these two methods. When the number of kernels is not very big (say less than 100), the time cost of SeDuMi is 2–3 times greater than the gradient descent algorithm. Fig. 2 illustrates the results which compare the average computation time of 50 repetitions.

**Algorithm 1.** Iterative gradient descent algorithm for elastic multiple kernel learning:

Set \(t = 1\) and \(v_k^0 = \frac{1}{m}\) for \(k = 1, \cdots, m\);

While stopping criterion not meet do

Solve SVM problem with \(K = \sum_{k=1}^{m} \frac{v_k}{1 - \lambda + \lambda v_k} K_k\), get solution \(\hat{\alpha}\);

Compute gradient \(\frac{\partial \theta}{\partial v} \bigg|_{v=v^*} = - \frac{1}{2} \frac{\alpha^T}{(1 - \lambda + \lambda v_k)^2} K_k \hat{\alpha}\);

Compute descent gradient \(\nabla \theta_i\) and optimal step size \(\lambda_i\) based on \(\frac{\partial \theta}{\partial v} \bigg|_{v=v^*}\);

Update \(v\) with \(v^{t+1} = v^t + \lambda_i \nabla \theta_i\);

\(t = t + 1\).
3 Experimental results

Through simulation study, we investigate the performances of SVM, MKL and EMKL. With each experiment, simulated data set consists of an independent training set and an independent testing set. Methods were trained and tuned on training set. Test errors were computed on testing set using the trained model and tuned parameters.

For each experiment, we generate 20 kernel matrices. Each kernel matrix corresponds to a 10-dimensional feature space. In order to control the sparsity of informative kernels, we set \( n_I \) kernels of the 20 kernel matrices to be informative and others non-informative. The informative kernels are generated from two Gaussian distributions with the following configurations:

\[
\Theta_+ = (\theta_1, \ldots, \theta_5, 0, \ldots, 0), \quad \Sigma_+ = I
\]  

(20)

\[
\Theta_- = (-\theta_1, \ldots, -\theta_5, 0, \ldots, 0), \quad \Sigma_- = I
\]  

(21)

where \( \Theta_+ \) and \( \Theta_- \) are mean vectors of two classes, \( \Sigma_+ \) and \( \Sigma_- \) are covariance matrices of two classes. The non-informative kernels are generated from the Gaussian distribution with following configuration:

\[
\Theta = (0, \ldots, 0), \quad \Sigma = I
\]  

(22)

We adopt linear kernels and use 10-fold cross-validation to tune parameters. There are two tuning parameters in EMKL, so we use 10-fold cross-validation on a grid. Every time we generated 20 samples for training (10 positive and 10 negative) and independently generated 200 samples for testing (100 positive and 100 negative). Each experiment has 100 repeats.

Table 1 summarizes the results of all simulation experiments. Columns 4 ~ 6 show classification error rates of all methods. We can see when \( n_I \) is small, MKL outperforms SVM. Along with the increase of \( n_I \), the performances of SVM gradually exceed MKL. EMKL achieves best performances among all these experiments. Column 7 shows the average of regularization parameters \( \lambda \) of EMKL which are obtained on the training sets using cross-validation. \( \lambda \) increases along with the decrease of sparsity. This also proves our analysis on the mixing regularization previously. Columns 8 and 9 show the ratio between the number of selected informative kernels and the number of informative kernels of MKL and EMKL respectively. We define a kernel is selected if its combination coefficient exceeds 0.001. As shown in the table, along the decrease of sparsity MKL tends to lose more informative kernels but EMKL keeps more informative kernels in the combined kernel.

Besides above simulation studies, we applied EMKL to pathway analysis of microarray data. By measuring thousands of genes expression simultaneously, microarray technology supplies a powerful tool to detect disease-related genes. Realizing the complexity of interactions among the disease-related genes, people hope to find gene groups rather than single genes. A pathway contains a group of genes which are closely related to specific biological functions or processes. The aim of pathway analysis is to find disease-related pathways.

Microarray data was taken from [12] which focuses on breast cancer research. Hess et al. [12] investigated the gene expression profiles of 130 patients with breast cancer using oligonucleotide microarrays. Taking the estrogen receptor status as classification labels, our aim is to identify breast cancer related pathways which will be useful for further research.

### Table 1 Simulation study configurations and results (Columns 4 ~ 6 summarize the average test error rates and their standard errors (shown in parenthesis) of SVM, MKL and EMKL, respectively. Column 7 shows the learnt regularization parameter \( \lambda \) of EMKL. Columns 8 ~ 9 summarize the ratio between the number of selected informative kernels and the number of informative kernels of MKL and EMKL respectively. Columns 7 ~ 9 are all averaged results of all experiments.)

<table>
<thead>
<tr>
<th>( n_I )</th>
<th>( \theta )</th>
<th>SVM</th>
<th>MKL</th>
<th>EMKL</th>
<th>( \lambda )</th>
<th>( R_{MKL} )</th>
<th>( R_{EMKL} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.7</td>
<td>28.20 (7.21)</td>
<td>22.60 (9.98)</td>
<td>22.29 (9.01)</td>
<td>0.49</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.6</td>
<td>20.34 (5.93)</td>
<td>18.21 (7.48)</td>
<td>16.46 (6.59)</td>
<td>0.57</td>
<td>0.91</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.5</td>
<td>21.80 (5.82)</td>
<td>21.53 (7.66)</td>
<td>19.31 (6.95)</td>
<td>0.57</td>
<td>0.84</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.4</td>
<td>23.43 (5.54)</td>
<td>25.34 (7.34)</td>
<td>22.49 (6.99)</td>
<td>0.63</td>
<td>0.73</td>
</tr>
</tbody>
</table>

### Table 2 Pathway analysis results (Stars indicate the pathways found only by EMKL.)

**MKL**

- PPAR signaling pathway
- ErbB signaling pathway
- Calcium signaling pathway
- Phosphatidylinositol signaling system
- VEGF signaling pathway

**EMKL**

- PPAR signaling pathway
- ErbB signaling pathway
- Calcium signaling pathway
- Phosphatidylinositol signaling system
- VEGF signaling pathway
  * Aminoacyl-tRNA biosynthesis
  * Biosynthesis of unsaturated fatty acids
  * Neuroactive ligand-receptor interaction
  * Apoptosis
  * TGF-beta signaling pathway
  * ECM-receptor interaction
  * Toll-like receptor signaling pathway
  * Taste transduction
  * Regulation of actin cytoskeleton
  * Glioma
We apply MKL and EMKL to the above problem. Taking the genes within a pathway as a block, we get 200 blocks from more than 15 000 genes according to the pathway annotation taken from http://www.genome.jp/kegg. Each block makes a kernel. The dataset was randomly partitioned into 80% training and 20% testing set. Within each training set, parameters were tuned using 5-fold cross-validation. Final results are the averages over 20 random partitions.

EMKL shows higher accuracy than MKL (86.33% vs. 85.6%). More importantly, EMKL identified more disease-related pathways than MKL. Table 2 summarizes the results. The shared pathways identified by MKL and EMKL are mostly important pathways which are closely related with breast cancer, such as ErbB signaling pathway[13] and VEGF signaling pathway[14]. Among the pathways identified only by EMKL, many of them have been reported to be important in breast cancer. Several examples are as follows. As reported by [15], estrogen receptor (ER) of a breast tumor is a potential mediator of the transition of TGF-beta from tumor suppressor to tumor promoter. Hancock et al.[16] found that the ECM protein tenascin-C is frequently up-regulated in breast cancer. Ilvesaro et al.[17] showed that Toll-like receptor 9 (TLR9) expression is increased in breast cancer. Jing et al.[18] studied the role of FIP3, which is a binding protein of a master regulator of endocytic membrane traffic and cytoskeletal dynamics, in breast carcinoma cell motility.

In summary, both EMKL and MKL could be used in microarray-based pathway analysis to identify disease-related pathways. EMKL achieves higher accuracy and identifies more meaningful disease-related pathways.

4 Discussions and conclusions

Motivated by the fact that MKL may discard useful classification information, we propose EMKL in present paper. Contrary to MKL which encourages the sparsity of kernel combination coefficients, EMKL introduces the mixing norm regularization to compromise sparsity and non-sparsity, so it has more flexibility than MKL and SVM.

There exist other works concerning the remedy for shortcomings of MKL. Kloft et al.[19] proposed a non-sparse multiple kernel learning which alternates the constraint on $v_k$ in (14) but keeps the other components unchanged:

$$\min_{\sum_{k=1}^{m} v_k^2 = 1, v_k \geq 0} \frac{1}{2} \sum_{k=1}^{m} \left\| w_k \right\|^2 + C \sum_{i=1}^{n} \xi_i$$

For above formulation of non-sparse MKL, we obtained following proposition:

Proposition 2. The problem (23) is equivalent to

$$\min \frac{1}{2} \left( \sum_{k=1}^{m} \left\| w_k \right\|^2 \right)^{\frac{3}{4}} + C \sum_{i=1}^{n} \xi_i$$

It can be seen that original MKL regularizes $\xi_i$ norm on $w_k$ while the modified version in (23) regularizes $\ell_{4/3}$ norm which will lead to less sparse solutions. Although this formulation seeks non-sparse kernel fusion, it just uses the fixed sparsity and will not adaptively adjust it. So there is significant distinction between above non-sparse strategy and EMKL strategy. EMKL has more flexibility.

For the computational issue, in learning process MKL and EMKL are more time-consuming than SVM because multiple kernel methodologies have more parameters to learn. For the classification process, single kernel and multiple kernels methodology have no difference because they both use one kernel matrix (single kernel or fused kernel based on multiple ones). We propose a rapid gradient descent algorithm for EMKL based on gradient descent algorithm of MKL. The computation cost of this algorithm is linear about the number of kernels thus works well in large scale multiple kernels learning. Although the gradient descent algorithm greatly improves the speed of optimization, more efficient algorithms, such as regularization-path algorithm developed for Lasso[20], are in demand.

Through a series of experiments, we demonstrate efficiency of EMKL framework in classification and kernel learning. The application of EMKL on the gene set analysis demonstrates its usefulness on real world learning problem.

Current formulation of EMKL mainly focuses on binary classification problems. In the future, we want to address regression and multi-class variants of EMKL. Finally, theoretical analysis about the statistical consistency of EMKL will be valuable.

Appendix

1) Proof of Proposition 1

Proof. Lagrangian of problem (17) is:

$$\mathcal{L} = \frac{1}{2} (1 - \lambda) \sum_{k=1}^{m} \frac{1}{v_k} ||w_k||^2 + \frac{1}{2} \lambda ||w||^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left( y_i \sum_{k=1}^{m} w_k(x_i,k) + y_i b - 1 + \xi_i \right) - \sum_{i=1}^{n} \beta_i \xi_i$$

(A1)

where $\alpha_i, \beta_i \geq 0$ are Lagrangian multipliers. Calculating the derivatives of the Lagrangian with respect to primal variables, we have:

$$\frac{\partial \mathcal{L}}{\partial w_k} = (1 - \lambda) \frac{1}{v_k} w_k + \lambda w_k - \sum_{i=1}^{n} \alpha_i y_i x_{i,k}$$

(A2)

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \beta_i$$

Letting these derivatives to be zero and inserting them into Lagrangian, we get the dual problem, thus leading to the proposition.\hfill \square

2) Proof of Proposition 2

Proof.

$$\left( \sum_{k=1}^{m} \frac{1}{v_k} ||w_k||^2 \right)^{\frac{3}{4}} = \left( \sum_{k=1}^{m} \left( \frac{||w_k||^2}{v_k^2} \right)^{\frac{3}{4}} \right)^{\frac{4}{3}}$$

$$\left( \sum_{k=1}^{m} \left( \frac{||w_k||^2}{v_k^2} \right)^{\frac{3}{4}} \right)^{\frac{4}{3}} \geq \sum_{k=1}^{m} ||w_k||^2$$

(A3)

The first equality is simple algebraic operation and second equality is because the constraint on $v_k$: $\sum_{k=1}^{m} v_k^2 = 1$. The inequality comes from Hölder inequality.

Hölder inequality. Given $\frac{1}{p} + \frac{1}{q} = 1$ with $p, q > 1$, the
Hölder’s inequality states that
\[
\left( \sum_{k=1}^{m} |a_k|^p \right)^{\frac{1}{p}} \left( \sum_{k=1}^{m} |b_k|^q \right)^{\frac{1}{q}} \geq \sum_{k=1}^{m} |a_k b_k| \quad (A4)
\]
with equality when \( |b_k| = c |a_k|^{p-1} \).
In our proof, we just let
\[
p = \frac{3}{2}, \quad q = 3, \quad a_k = \left| \frac{w_k}{v_k^\frac{3}{2}} \right|, \quad b_k = v_k^\frac{2}{3}
\]
So we have the following conclusion:
\[
\min \sum_{k=1}^{n} \frac{1}{2} \sum_{k=1}^{m} \|w_k\|^2 + C \sum_{i=1}^{n} \xi_i = \frac{1}{2} \left( \sum_{k=1}^{m} \|w_k\|^2 \right)^{\frac{3}{2}} + C \sum_{i=1}^{n} \xi_i \quad (A5)
\]
\[\square\]

References


WU Zheng-Peng: Ph. D. candidate in the Department of Automation, Tsinghua University. He received his B.S. degree in automatic control from Tsinghua University in 2004. His research interest covers pattern recognition, statistical learning theory, and statistics and computational genomics. E-mail: wuzhengpeng99@mails.tsinghua.edu.cn

ZHANG Xue-Gong: Professor in the Department of Automation, Tsinghua University. He received his B.S. degree and Ph. D. degree from Tsinghua University in 1989 and 1994, respectively. His research interest covers pattern recognition, machine learning, and bioinformatics and system biology. Corresponding author of this paper. E-mail: zhangxg@tsinghua.edu.cn