Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Eigenvalues Ratio for Kernel Selection of Kernel Methods

Yong Liu and **Shizhong Liao**
School of Computer Science and Technology,
Tianjin University, Tianjin 300072, P. R. China
{yongliu,szliao}@tju.edu.cn

Abstract

The selection of kernel function which determines the mapping between the input space and the feature space is of crucial importance to kernel methods. Existing kernel selection approaches commonly use some measures of generalization error, which are usually difficult to estimate and have slow convergence rates. In this paper, we propose a novel measure, called eigenvalues ratio (ER), of the tight bound of generalization error for kernel selection. ER is the ratio between the sum of the main eigenvalues and that of the tail eigenvalues of the kernel matrix. Different from most of existing measures, ER is defined on the kernel matrix, so it can be estimated easily from the available training data, which makes it usable for kernel selection. We establish tight ER-based generalization error bounds of order $O\left(\frac{1}{n}\right)$ for several kernel-based methods under certain general conditions, while for most of existing measures, the convergence rate is at most $O\left(\frac{1}{\sqrt{n}}\right)$. Finally, to guarantee good generalization performance, we propose a novel kernel selection criterion by minimizing the derived tight generalization error bounds. Theoretical analysis and experimental results demonstrate that our kernel selection criterion is a good choice for kernel selection.

Introduction

Kernel methods, such as SVM (Steinwart and Christmann 2008), least squares support vector machine (LSSVM) (Suykens and Vandewalle 1999) and kernel ridge regression (KRR) (Saunders, Gammerman, and Vovk 1998), have demonstrated great success in solving many machine learning and pattern recognition problems. These methods implicitly map data points from the input space to some feature space in which even relatively simple algorithms can deliver very impressive performance. The feature mapping is provided intrinsically via the choice of a kernel function. Therefore, for a kernel method to perform well, the kernel function plays a very crucial role.

Kernel selection is to select the optimal kernel by minimizing some kernel selection criterion that is usually defined via the estimate of the $generalization\ error$ (Bartlett, Boucheron, and Lugosi 2002). The estimate can be empirical or theoretical. The k-fold cross-validation (KCV) and leave-one-out cross-validation (LOO) are widely used empirical estimates of generalization error, but they require training the algorithm many times, which unavoidably incurs high computational burdens. For the sake of efficiency, some approximate KCV and LOO methods are given: such as generalized cross-validation (Golub, Heath, and Wahba 1979), span bound (Chapelle et al. 2002), influence function (Debruyne, Hubert, and Suykens 2008) and Bouligand influence function (Liu, Jiang, and Liao 2014). Kernel target alignment (KTA) (Cristianini et al. 2001) is another popular used empirical estimate, which is used to quantify the similarity between the kernel matrix and the label matrix. Several related criteria were also proposed, such as centered kernel target alignment (CKTA) (Cortes, Mohri, and Rostamizadeh 2010) and feature space-based kernel matrix evaluation (FSM) (Nguyen and Ho 2008). Although KTA, CKTA and FSM are widely used, the connection between these estimates and the generalization error of some special algorithms, such as SVM, LSSVM and KRR, has not established, hence the kernels chosen by these estimates may not guarantee good generalization performance for these algorithms (Liu, Jiang, and Liao 2013). Minimizing theoretical estimate bounds of generalization error is an alternative to kernel selection. The widely used theoretical estimates usually introduce some measures of the complexity of the hypothesis space, such as VC dimension (Vapnik 2000), radius-margin bound (Vapnik 2000), maximal discrepancy (Bartlett, Boucheron, and Lugosi 2002), Rademacher complexity (Bartlett and Mendelson 2002), compression coefficient (Luxburg, Bousquet, and Schölkopf 2004), eigenvalues perturbation (Liu, Jiang, and Liao 2013), spectral perturbation stability (Liu and Liao 2014a), kernel stability (Liu and Liao 2014b) and covering number (Ding and Liao 2014). Unfortunately, for most of these measures, it is difficult to estimate their specific values (Nguyen and Ho 2007), hence hard to use them for kernel selection in practice. Moreover, most of these measures usually have slow convergence rates of order $O\left(1/\sqrt{n}\right)$ at most.

In this paper, we propose a novel measure, called eigenvalues ratio (ER), for deriving a tight bound of generalization error for kernel selection. ER is the ratio between the sum of the main eigenvalues and that of the tail eigenvalues of the kernel matrix. Unlike most of existing measures, our measure is defined on the kernel matrix that can be estimated easily from the available training data. The tight ER-based generalization error bounds of SVM, KRR and LSSVM are established, which have convergence rates of the order $O\left(\frac{1}{n}\right)$ under some certain general conditions. While for the existing standard generalization error bounds, such as Rademacher complexity bound (Bartlett and Mendelson 2002) and radius-margin bound (Vapnik 2000), and the latest eigenvalues perturbation bound (Liu, Jiang, and Liao 2013), the convergence rates are $O(1/\sqrt{n})$ at most. Furthermore, we propose a new kernel selection criterion by minimizing the derived tight upper bounds to guarantee good generalization performance. Theoretical analysis and experimental results demonstrate the effectiveness of our criterion.

Related Work

One of the most useful data-dependent complexity measures used in the theoretical analysis is the notion of Rademacher complexity (Bartlett and Mendelson 2002; Koltchinskii and Panchenko 2002). Unfortunately, it provides global estimates of the complexity of the function class, that is, it does not reflect the fact that the algorithm will likely pick functions that have a small error. In recently years, several authors have considered the use of *local* Rademacher complexity to obtain better generalization error bounds. The local Rademacher complexity considers Rademacher averages of smaller subset of the hypothesis set, so it is always smaller than the corresponding global one.

Koltchinskii and Panchenkoy (2000) first considered using the local Rademacher complexity to obtain data dependent upper bounds using an iterative method. Bousquet, Koltchinskii and Panchenko (2002) proposed a more general result avoiding the iterative procedure. Lugosi and Wegkamp (2004) established the oracle inequalities using local Rademacher complexity and also demonstrated the advantages of local Rademacher complexity over those based on the complexity of the whole model class. Bartlett, Bousquet and Mendelson (2005) gave the optimal rates based on a local and empirical version of Rademacher, and presented some applications to classification and prediction with convex function classes, and with kernel classes in particular. Koltchinskii (2006) proposed new bounds on the error of learning algorithms in terms of the local Rademacher complexity, and applied these bounds to develop model selection techniques in abstract risk minimization problems. Srebro, Sridharan and Tewari (2010) established an excess risk bound for ERM with local Rademacher complexity. Mendelson (2003) presented sharp bounds on the localized Rademacher averages of the unit ball in a reproducing kernel Hilbert space in terms of the eigenvalues of the integral operator associated with the kernel function. Based on the connection between local Rademacher complexity and the tail eigenvalues of the integral operator, Kloft and Blanchard (2011) derived an upper bound of multiple kernel learning with the tail eigenvalues. Unfortunately, the eigenvalues of the integral operator are difficult to compute as the probability distribution is unknown, so Cortes, Kloft and Mohri (2013) used the tail eigenvalues of the kernel matrix, that is the empirical version of the tail eigenvalues of the integral operator, to design new algorithms for learning kernels. However, the theoretical error bound based on the tail eigenvalues of the kernel matrix was not established. Moreover, for different kinds of kernel functions or the same kind of kernel functions but with different parameters, the discrepancies of eigenvalues of different kernels are very large, hence the absolute value of the tail eigenvalues can't precisely reflect the goodness of different kernels. In this paper, we consider applying the relative value of eigenvalues of the kernel matrix, that is, the ratio between the sum of the main eigenvalues and that of the tail eigenvalues, for kernel selection. We establish the link between this relative value and the notion of local Rademacher complexity, and further present tight theoretical error bounds of KRR, LSSVM and SVM. To our knowledge, the generalization error bounds based on the eigenvalue analysis of the kernel matrix, of convergence rates of the order $O(\frac{1}{n})$ under certain general conditions, have never been given before.

The rest of the paper is organized as follows. Some preliminaries are introduced in Section 2. In Section 3, we propose the definition of ER and give some theoretical analysis. Section 4 shows how to use ER to derive tight generalization error bounds. In Section 5, we propose the kernel selection criterion by minimizing the derived error bounds. We empirically analyze the performance of our proposed criterion in Section 6. Finally, we conclude in Section 7.

Preliminaries

Let $S = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ be a sample set of size n drawn identically and independently from a fixed, but unknown probability distribution P on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ denotes the output domain, $\mathcal{Y} \subseteq \mathbb{R}$ in regression case and $\mathcal{Y} = \{+1, -1\}$ in classification case.

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function. Its corresponding kernel matrix is defined as $\mathbf{K} = \left[\frac{1}{n}K(\mathbf{x}_i, \mathbf{x}_j)\right]_{i,j=1}^n$. $\mathbf{K}$ is positive semidefinite, hence its eigenvalues satisfy $\lambda_1(\mathbf{K}) \geq \lambda_2(\mathbf{K}) \geq \cdots \geq \lambda_n(\mathbf{K}) \geq 0$. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with K is defined to be the completion of the linear span of the set of functions $\{K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}$ with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K(\mathbf{x}, \cdot), f \rangle_K = f(\mathbf{x}), \forall f \in \mathcal{H}_K$.

In this paper, we study the regularized algorithms:

$$f_S := \operatorname*{arg\,min}_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i) + \frac{\lambda}{n} \|f\|_K^2 \right\}, \quad (1)$$

where $\ell(\cdot, \cdot)$ is a loss function and λ is the regularization parameter. SVM, LSSVM and KRR are the special cases of the regularized algorithms. For SVM, $\ell(t, y) = \max(0, 1 - yt)$, for both KRR and LSSVM, $\ell(t, y) = (y - t)^2$.

The performance of the regularized algorithms is usually measured by the *generalization error*

$$R(S) := \int_{\mathcal{X} \times \mathcal{Y}} \ell(f_S(\mathbf{x}), y) \mathrm{d}P(\mathbf{x}, y).$$

Unfortunately, $R(S)$ can not be computed since P is unknown. Thus, we estimate it using the empirical error $R_{\text{emp}}(S) := \frac{1}{n} \sum_{i=1}^n \ell(f_S(\mathbf{x}_i), y_i)$.

In the following, we assume that $\|f\|_\infty \leq D, \forall f \in \mathcal{H}_K$, $|y| \leq M, \forall y \in \mathcal{Y}$, the trace of the kernel matrix $\mathrm{Tr}(\mathbf{K}) =: c < \infty$ and $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} K(\mathbf{x}, \mathbf{x}') =: \kappa < \infty$.

Eigenvalues Ratio

In this section, we first introduce the notion of eigenvalues ratio (ER) and then give some analysis of ER.

Definition 1 (Eigenvalues Ratio). *Assume K is a kernel function, $\mathbf{K} = \left[\frac{1}{n}K(\mathbf{x}_i, \mathbf{x}_j)\right]_{i,j=1}^n$ is its corresponding kernel matrix, and $\lambda_i(\mathbf{K})$ is the ith eigenvalue of $\mathbf{K}$, $\lambda_1(\mathbf{K}) \geq \lambda_2(\mathbf{K}) \geq \cdots \geq \lambda_n(\mathbf{K})$. Then the t-eigenvalues ratio of the kernel function K, $t \in \{1, 2, \cdots, n-1\}$, is defined as*

$$\frac{\sum_{i=1}^t \lambda_i(\mathbf{K})}{\sum_{i=t+1}^n \lambda_i(\mathbf{K})} =: \beta_t.$$

According to the above definition, ER is the ratio between the sum of the main eigenvalues and that of the tail eigenvalues of the kernel matrix. ER is closely related to the notion of local Rademacher complexity (Cortes, Kloft, and Mohri 2013). So, we can establish generalization error bound with ER.

Different from the existing notions of measure, see, e.g., (Vapnik 2000; Bartlett, Boucheron, and Lugosi 2002; Bartlett and Mendelson 2002; Koltchinskii 2006) and the references therein, ER is defined on the kernel matrix, hence we can estimate its value from the available training data, making this measure usable for kernel selection in practice.

Theorem 1. *If K satisfies the assumption of algebraically decreasing eigenvalues, that is $\exists \gamma > 1 : \lambda_i(\mathbf{K}) = O(\frac{1}{n}i^{-\gamma})$. Then the t-eigenvalues ratio of K satisfies*

$$\beta_t = \Omega\left(\frac{n(\gamma-1)}{t^{1-\gamma}}\right). \tag{2}$$

Proof. We sketch the proof. According to assumption of algebraically decreasing eigenvalues, we first show that $\sum_{i=t+1}^n \lambda_i(\mathbf{K}) = O(\frac{t^{(1-\gamma)}}{n(\gamma-1)})$. Then, from the definition of β_t, $\beta_t = \left(\mathrm{Tr}(\mathbf{K}) - \sum_{i=t+1}^n \lambda_i(\mathbf{K})\right) / \sum_{i=t+1}^n \lambda_i(\mathbf{K})$, it is easy to complete the proof. □

The assumption of algebraically decreasing eigenvalues of the kernel is a common assumption, for example, met for popular shift invariant kernel, finite rank kernels and convolution kernels (Williamson, Smola, and Scholkopf 2001).

One can see that ER is only related to kernel matrix and independent of specific learning task. Therefore, we can apply this measure for different learning tasks.

ER Based Generalization Error Bounds

In this section, we show how to apply ER to derive the tight generalization error bounds for SVM, KRR and LSSVM.

Support Vector Machine (SVM)

SVM has been successfully applied to solve classification problems, its loss function is the hinge loss.

Theorem 2. *If the t-eigenvalues ratio of the kernel function K is β_t. Then for SVM, with probability $1-\delta$, for any $k > 1$,*

$$R(S) \leq R_{\mathrm{emp}}(S) + C_1\sqrt{\frac{1}{n\beta_t}} + \frac{C_2}{n}\log\frac{3}{\delta} + C_3, \tag{3}$$

$C_1 = 12k\sqrt{2\kappa}$, $C_2 = k(96Ft + 96 + 5F) + 22$, $C_3 = (1+D)/(k-1)$, $F = 1 + D$.

Proof. We sketch the proof. We first show that the local Rademacher complexity of $\mathcal{H}_K$ can be bounded with ER: $\hat{R}_n(\mathcal{H}_K, r) \leq \sqrt{\frac{2}{n}(t \cdot r + \kappa/\beta_t)}$, where $\hat{R}_n(\mathcal{H}_K, r)$ is the local Rademacher complexity of $\mathcal{H}_K$. Then, by Theorem 4.1 of (Bartlett, Bousquet, and Mendelson 2005), we can obtain $R(S) \leq \frac{kR_{\mathrm{emp}}(S)}{k-1} + 6k\hat{r}^* + \frac{\log(\frac{3}{\delta})(22M+5k)}{n}$, where $\hat{r}^*$ is the fixed point of $2F\sqrt{\frac{2}{n}(t \cdot r + \kappa/\beta_t)} + 13\log(3/\delta)/n$. Finally, we estimate $\hat{r}^*$, and show that $\hat{r}^* \leq 16F^2t/n + 26\log(3/\delta)/n + 2F\sqrt{\kappa/(n\beta_t)}$, to complete the proof. □

The above theorem shows that the convergence rate of $R(S) - R_{\mathrm{emp}}(S)$ is $O\left(1/\sqrt{n\beta_t} + 1/n\right)$. Under the common assumption of algebraically decreasing eigenvalues, from Theorem 1, we have $\beta_t = \Omega\left(n(\gamma-1)/t^{1-\gamma}\right)$. Thus, the bound (3) converges at rate $O\left(\frac{1}{n}\sqrt{\frac{t^{1-\gamma}}{(\gamma-1)}} + \frac{1}{n}\right) = O\left(\frac{1}{n}\right)$.

Traditional Generalization Error Bounds Generalization error bounds based on Rademacher complexity are standard (Bartlett and Mendelson 2002). For any $\delta > 0$, $R(S) \leq R_{\mathrm{emp}}(S) + R_n(\mathcal{H}_K)/2 + \sqrt{\ln(1/\delta)/2n}$, where $R_n(\mathcal{H}_K)$ is the Rademacher complexity of $\mathcal{H}_K$. $R_n(\mathcal{H}_K)$ is in the order of $O(\frac{1}{\sqrt{n}})$ for various kernel classes used in practice, including the kernel class with bounded trace. Thus, in this case, this bound converges at rate $O(\frac{1}{\sqrt{n}})$.

For other measures, such as radius-margin bound (Vapnik 2000), $R(S) \leq R_{\mathrm{emp}}(S) + \sqrt{c(R^2\log^2 n/\rho^2 - \log\delta)/n}$. So, this bound converges at rate $O(\sqrt{\log^2 n/n})$.

For the latest eigenvalues perturbation bound (Liu, Jiang, and Liao 2013), $R(S) \leq R_{\mathrm{emp}}(S) + \sqrt{c\zeta/n}$, where ζ is the eigenvalues perturbation of kernel function, see definition 1 in (Liu, Jiang, and Liao 2013) for detail. This bound converges at rate $O(1/\sqrt{n})$.

The above theoretical analysis indicates that using the measure of eigenvalues ratio can obtain tight bound under some general conditions, which also demonstrates the effectiveness of the use of eigenvalues ratio to estimate the generalization error. Thus, to guarantee good generalization performance, it is reasonable to choose the kernel function with small $R_{\mathrm{emp}}(S)$ and $1/\beta_t$.

Kernel Ridge Regression (KRR)

KRR is a popular learning machine for solving regression problems, its loss function is the square loss.

Theorem 3. *If the t-eigenvalues ratio of K is β_t. Then for KRR, with probability at least $1-\delta$, $\forall k > 1$,*

$$R(S) \leq R_{\mathrm{emp}}(S) + C_4\sqrt{1/(n\beta_t)} + C_5/n + C_6, \tag{4}$$

Data sets	ER (ours)	EP	5-CV	LOO	CKTA	FSM
$\lambda = 0.01$						
australian	14.01±2.10	14.20±1.95	14.78±1.87	14.30±1.85	**13.98±2.10**	44.15±3.30
heart	17.08±3.28	18.56±3.59	16.91±3.64	**16.71±3.29**	16.75±3.50	44.65±4.35
ionosphere	8.98±2.39	9.65±2.58	**5.97±1.58**	6.32±2.33	32.89±8.71	35.68±3.13
breast-cancer	**3.59±0.97**	3.72±1.04	3.77±1.09	3.61±0.94	31.72±9.40	4.36±0.93
diabetes	22.30±2.30	22.81±2.67	22.43±2.69	**22.06±2.31**	35.09±2.87	35.09±2.87
german.numer	**23.76±2.34**	26.14±1.70	24.11±2.16	24.14±2.31	30.36±2.57	30.36±2.57
liver-disorders	**28.56±3.75**	30.80±3.40	29.17±4.24	29.04±4.37	36.83±6.53	40.83±4.72
a2a	17.74±0.95	19.22±1.07	17.80±1.03	**17.67±0.90**	25.21±1.40	25.21±1.40
$\lambda = 0.1$						
australian	**13.46±1.91**	13.59±2.03	14.59±2.07	14.41±1.73	19.89±6.18	44.15±3.30
heart	**16.87±3.39**	17.12±3.40	17.08±3.82	17.16±3.81	40.29±10.03	44.65±4.35
ionosphere	7.87±2.36	8.89±2.57	**5.30±1.29**	6.38±1.92	35.75±3.04	35.68±3.13
breast-cancer	3.53±0.97	4.23±1.09	3.79±0.91	**3.51±0.94**	32.70±6.57	3.84±0.81
diabetes	22.25±2.53	22.99±2.75	22.35±2.40	**22.16±2.35**	35.09±2.87	35.09±2.87
german.numer	**23.77±2.37**	23.89±2.13	24.08±2.15	23.99±2.31	30.36±2.57	30.36±2.57
liver-disorders	**29.01±4.19**	34.20±3.93	30.19±4.13	30.77±4.00	35.54±7.29	40.83±4.72
a2a	17.81±1.19	18.11±1.22	17.92±1.01	**17.68±0.97**	25.21±1.40	25.21±1.40
$\lambda = 1$						
australian	**14.12±1.62**	14.17±1.58	14.67±2.07	14.61±1.86	44.07±3.35	44.15±3.30
heart	**16.83±3.56**	17.12±3.30	17.08±3.01	17.12±3.72	44.65±4.35	44.65±4.35
ionosphere	7.81±2.60	11.02±3.13	**5.43±1.97**	6.57±2.08	35.75±3.04	35.68±3.13
breast-cancer	3.43±0.95	4.54±1.55	3.53±0.83	**3.25±0.94**	34.44±2.32	3.54±0.87
diabetes	22.39±2.53	24.65±3.29	22.48±2.22	**22.19±2.20**	35.09±2.87	35.09±2.87
german.numer	24.57±2.38	24.59±2.47	24.24±2.41	**24.13±2.37**	30.36±2.57	30.36±2.57
liver-disorders	**29.46±3.69**	41.96±4.45	30.38±3.99	30.87±3.42	36.19±6.09	40.83±4.72
a2a	18.17±1.23	18.66±1.31	18.01±1.00	**17.83±1.08**	25.21±1.40	25.21±1.40
$\lambda = 10$						
australian	**13.85±1.88**	14.19±1.71	14.52±1.73	14.35±2.06	44.07±3.35	44.07±3.35
heart	17.04±3.62	17.37±3.59	**16.79±3.44**	17.04±3.36	44.65±4.35	44.65±4.35
ionosphere	8.89±2.63	14.22±3.74	6.92±1.87	**6.76±1.88**	36.06±3.06	36.06±3.06
breast-cancer	3.24±0.89	5.71±1.59	3.37±0.93	**3.22±1.08**	35.04±2.45	3.61±0.88
diabetes	23.41±3.07	34.99±3.13	23.09±2.78	**22.54±2.51**	35.09±2.87	35.09±2.87
german.numer	27.13±2.65	29.64±3.30	26.94±2.80	**26.86±2.87**	30.36±2.57	30.36±2.57
liver-disorders	**34.97±4.47**	41.89±4.39	35.35±4.46	35.13±4.49	39.26±4.74	41.86±4.46
a2a	18.79±1.19	19.26±1.60	18.77±1.21	**18.71±1.21**	25.23±1.43	25.23±1.43

Table 1: Comparison of mean test errors between our eigenvalues ratio criterion (ER) and other ones including 5-CV, LOO, CKTA, FSM and EP. We bold the numbers of the best method, and underline the numbers of the other methods which are not significantly worse than the best one.

where $C_4 = 12hBk\sqrt{2\kappa}$, $C_5 = 96h^2Bkt + (96k + 22M + 5Bk)\log(3/\delta)$, $h = 2(M + D)$, $C_6 = (D + M)^2/(k - 1)$, $B = (D + M)^2$.

The idea of proof is same as that of Theorem 2.

The convergence rate of KRR is $O\left(1/\sqrt{n\beta_t} + 1/n\right)$. Under the assumption of algebraically decreasing eigenvalues, the converges rate can reach $O(1/n)$. This theorem also indicates that the kernel with small $R_{\mathrm{emp}}(S)$ and $1/\beta_t$ can guarantee good generalization performance.

Least Squares Support Vector Machine (LSSVM)

LSSVM is a popular classifier which has the same loss function as that of KRR. Thus, applying Theorem 3 with $M = 1$, we can obtain the following corollary:

Corollary 1. *If the t-eigenvalues ratio of K is β_t, then for LSSVM, with probability at least $1 - \delta$, for any $k > 1$,*

$$R(S) \leq R_{\mathrm{emp}}(S) + C_7\sqrt{1/(n\beta_t)} + C_8/n + C_9,$$

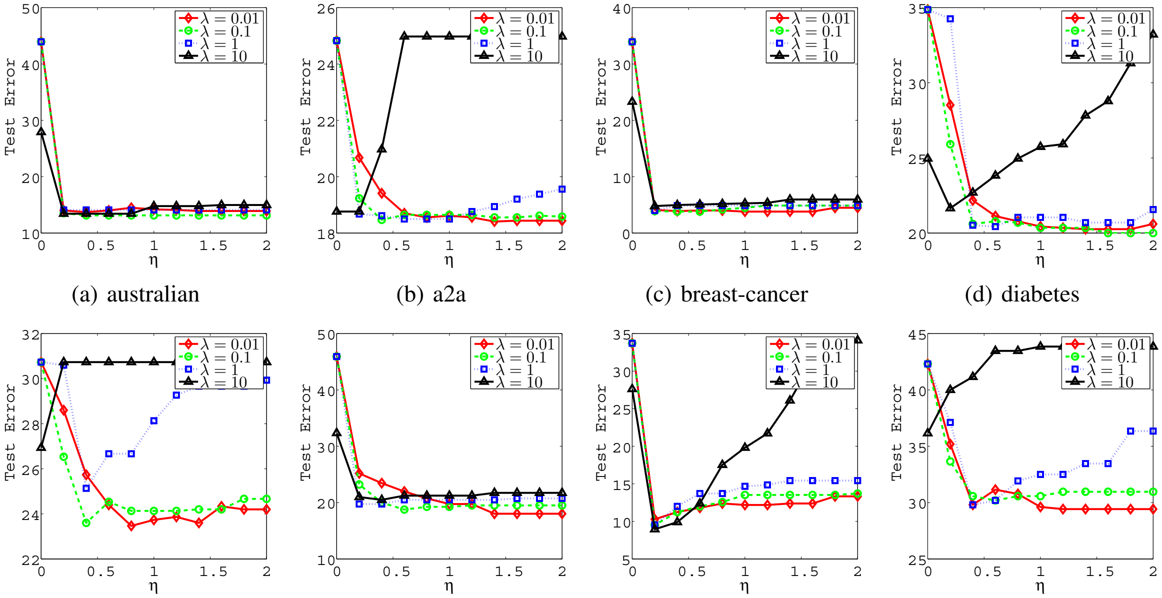
where $C_7 = 12hk\sqrt{2\kappa}$, $C_8 = 96Bh^2kt + (96k + 22 + 5Bk)\log(3/\delta)$, $h = 2(1 + D)$, $C_9 = (1 + D)^2/(k - 1)$, $B = (D + 1)^2$.

Kernel Selection with Eigenvalues Ratio

In this section, we will present a novel kernel selection criterion with ER to guarantee good generalization performance.

From the generalization error bounds derived in above section, to guarantee good generalization performance, we can choose the kernel function by minimizing $R_{\mathrm{emp}}(S)$ and $1/\beta_t$. Thus, we apply the following eigenvalues ratio criterion for kernel selection:

$$\operatorname*{arg\,min}_{K \in \mathcal{K}} R_{\mathrm{emp}}(S) + \eta \cdot n/\beta_t =: KS(K)$$



(a) australian (b) a2a (c) breast-cancer (d) diabetes

(e) german.numer (f) heart (g) ionosphere (h) liver-disorders

Figure 1: The test errors of our eigenvalue ratio criterion (ER) with different η. For each η, we choose the kernel by ER on the training set, and evaluate the test errors for the chosen parameters on test set.

where η is the trade-off parameter and $\mathcal{K}$ is a candidate set of kernel functions.

Time Complexity Note that $\sum_{i=t+1}^{n} \lambda_i(\mathbf{K}) = \mathrm{Tr}(\mathbf{K}) - \sum_{i=1}^{t} \lambda_i(\mathbf{K})$, where $\mathrm{Tr}(\mathbf{K})$ is the trace of matrix kernel $\mathbf{K}$. Thus, we only need $O(tn^2)$ to compute β_t. So the overall time complexity of eigenvalues ratio criterion is $O(tn^2 + J)$, where J is the time complexity of computing $R_{\mathrm{emp}}(S)$.

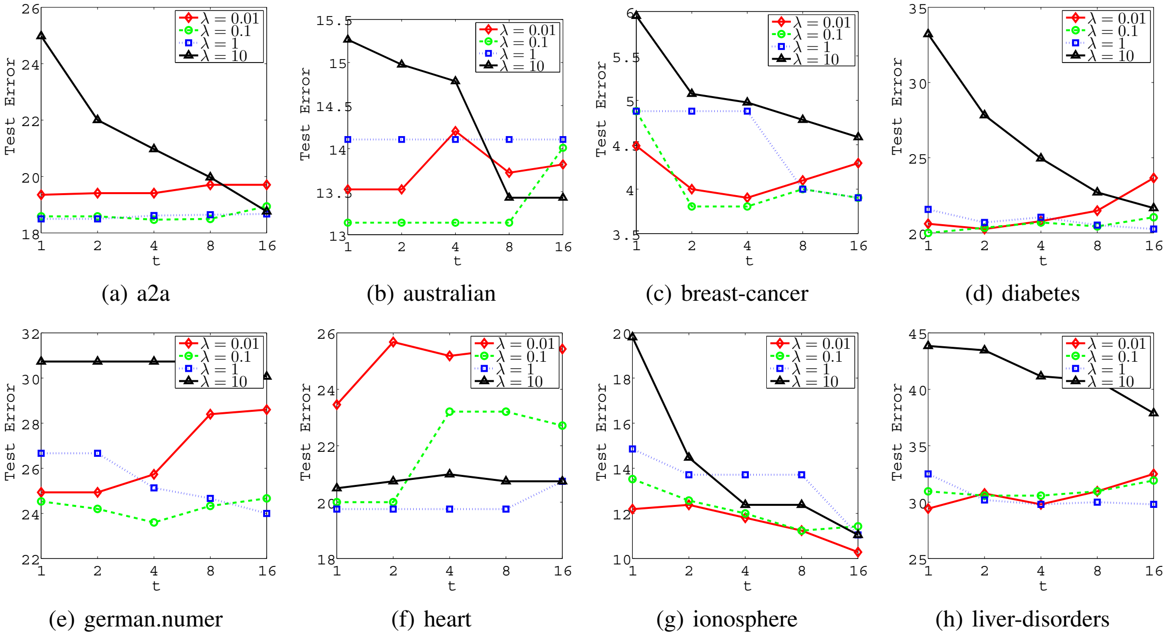
Remark 1. *Instead of choosing a single kernel, several authors consider combining multiple kernels by some criteria, called multiple kernel learning (MKL), see, e.g., (Lanckriet et al. 2004; Liu, Liao, and Hou 2011), etc. Our eigenvalues ratio criterion can be applied to MKL:* $\min_\mu KS(K_\mu)$ *s.t.* $\|\mu\|_p = 1, \mu \geq 0,\ K_\mu = \sum_{i=1}^{k} \mu_i K_i$. *The above optimization problem can be efficiently solved with gradient-based algorithms. However, in this paper, we mainly want to verify the effectiveness of ER criterion. Therefore, in our experiments, we focus on comparing our criterion with other popular kernel selection criteria.*

Experiments

In this section, we will empirically analyze the performance of our proposed eigenvalues ratio criterion (ER). The evaluation is made on 8 available public data sets from LIBSVM data seen in Table 1. For each data set, we run all methods 30 times with randomly selected 70% of all data for training and the other 30% for testing. The use of multiple training/test partitions allows an estimate of the statistical significance of differences in performance between methods. Let A_i and B_i be the test errors of methods A and B in partition i, and $d_i = B_i - A_i, i = 1, ..., 30$. Let $\bar{d}$ and S_d be the mean and standard error of d_i. Then under t-test, with confidence level 95%, we claim that A is significantly better than B (or equivalently B significantly worse than A) if the t-statistic $\frac{\bar{d}}{S_d/\sqrt{30}} > 1.699$. All statements of statistical significance in the remainder refer to a 95% level of significance.

In the first experiment, we compare our ER criterion with five popular kernel selection criteria: 5-fold cross-validation (5-CV), leave-one-out cross-validation criterion (LOO), centered kernel target alignment (CKTA) (Cortes, Mohri, and Rostamizadeh 2010), feature space-based kernel matrix evaluation (FSM) (Nguyen and Ho 2007) and the latest eigenvalues perturbation criterion (EP) (Liu, Jiang, and Liao 2013). We use the popular Gaussian kernels $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\tau\right)$ as our candidate kernels, $\tau \in \{2^i, i = -15, -14, \ldots, 15\}$. The learning machine we used is LSSVM.

For each kernel selection criterion and each training set, we chose the optimal kernel parameter τ for each fixed regularized parameter $\lambda \in \{0.01, 0.1, 1, 10\}$, and then evaluate the test error for the chosen parameters on the test set. The optimal values for the parameters $t \in \{1, 4, 16\}$ and $\eta \in \{0.2, 0.6, 1\}$ of ER, and the parameter $\delta \in \{2^i, i = 0, 5, 10, 15, 20\}$ of EP (following the same experimental setting of EP in (Liu, Jiang, and Liao 2013)) are determined by 3-fold cross-validation on the training set, we will explore the influence of parameters t and η in the next experiment. The average test errors are reported in Table 1. The results in Table 1 can be summarized as follows: (a) Our ER is significantly better than CKTA and FSM on nearly all data sets. This can possibly be explained by the fact that the connection between CKTA (or FSM) and generaliza-



(a) a2a

(b) australian

(c) breast-cancer

(d) diabetes

(e) german.numer

(f) heart

(g) ionosphere

(h) liver-disorders

Figure 2: The test errors of our eigenvalue ratio criterion (ER) with different t. For each t, we choose the kernel by ER on the training set, and evaluate the test errors for the chosen parameters on test set.

tion error of LSSVM has not been established, so the kernels chosen by this criterion can not guarantee good generalization performance; (b) ER is significantly better than EP on 5 (or more) of the 8 benchmarks without being significantly worse on any of the remaining data sets for each regularized parameter. The convergence rate of ER-based error bound is much faster than that of EP-based one, so the performance of ER being better than that of EP is conform to our theoretical analysis; (c) Our ER give comparable results to 5-CV and LOO. Specifically, for $\lambda = 0.1$, ER is significantly better than 5-CV and LOO on australian and german.numer, and is significantly worse on ionosphere. For other $\lambda \in \{0.01, 1, 10\}$, the results are similar with that of $\lambda = 0.1$. ER gives similar accuracies results with 5-CV and LOO, but ER only need to train once, which is more efficient than 5-CV and LOO, especially for LOO. The above results show that ER is a good choice for kernel selection.

In the next experiments, we will explore the influence of the parameters t and η for ER. The average test errors over different η are given in Figure 1 (in this experiment, we only report the results of $t = 4$, similar results can be found with other values, e.g. $t \in \{2, 8\}$). One can see that, for appropriate $\lambda \in \{0.01, 0.1, 1\}$, the test errors are stable with respect to $\eta \in [0.4, 1.5]$. However, for the largest λ ($\lambda = 10$), the test errors with respect to η is not very stable on some data sets, which is possibly because λ is unreasonably large. From Table 1, we can find that the optimal regularized parameter λ is almost in $\{0.1, 1\}$. In fact, we also consider using the large value of λ, such as $\lambda \in \{100, 1000\}$, but we find that the performance of this large λ are almost much worse than that of $\lambda \in \{0.01, 0.1, 1\}$. Thus, it is usually not necessary to set the value of λ too large. The test errors over different t are given in Figure 2 (Figure 1 shows that $\eta = 0.6$ is a good choice, so in this experiment, we set $\eta = 0.6$). From Figure 2, we can find that for all λ except the largest one, t is stable w.r.t $[2, 8]$. The robustness property of the parameters t and η implies that, for appropriate λ (not too large), we can randomly select $\eta \in [0.4, 1.5]$ and $t \in [2, 8]$, without sacrificing much accuracy. We believe that this robustness property can bring some advantages in practical application.

Conclusion

We introduced a novel measure of eigenvalues ratio (ER) which has two main advantages compared with most of existing measures of generalization error: 1) defined on the kernel matrix, hence can be estimated easily from available training data; 2) has a fast convergence rate of order $O(\frac{1}{n})$. To our knowledge, the theoretical error bounds via spectral analysis of the kernel matrix, of convergence rates of the order $O(\frac{1}{n})$, has never been given before. Furthermore, we proposed a kernel selection criterion by minimizing the derived tight generalization upper bound, which can guarantee good generalization performance. Our kernel selection criterion was theoretically justified and experimentally validated.

In future, we will consider applying the notion of ER for multiple kernel learning and for deriving tight generalization error bounds of other kernel-based methods.

Acknowledgments

The work is supported in part by the National Natural Science Foundation of China under grant No. 61170019.

References

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Bartlett, P. L.; Boucheron, S.; and Lugosi, G. 2002. Model selection and error estimation. *Machine Learning* 48:85–113.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local Rademacher complexities. *The Annals of Statistics* 33(4):1497–1537.

Bousquet, O.; Koltchinskii, V.; and Panchenko, D. 2002. Some local measures of complexity of convex hulls and generalization bounds. *Lecture Notes in Artificial Intelligence* 2575:59–73.

Chapelle, O.; Vapnik, V.; Bousquet, O.; and Mukherjee, S. 2002. Choosing multiple parameters for support vector machines. *Machine Learning* 46(1-3):131–159.

Cortes, C.; Kloft, M.; and Mohri, M. 2013. Learning kernels using local Rademacher complexity. In *Advances in Neural Information Processing Systems 25 (NIPS 2013)*. MIT Press. 2760–2768.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2010. Two-stage learning kernel algorithms. In *Proceedings of the 27th Conference on Machine Learning (ICML 2010)*, 239–246.

Cristianini, N.; Shawe-Taylor, J.; Elisseeff, A.; and Kandola, J. S. 2001. On kernel-target alignment. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, 367–373.

Debruyne, M.; Hubert, M.; and Suykens, J. A. 2008. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research* 9:2377–2400.

Ding, L., and Liao, S. 2014. Model selection with the covering number of the ball of RKHS. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM 2014)*, 1159–1168.

Golub, G. H.; Heath, M.; and Wahba, G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215–223.

Kloft, M., and Blanchard, G. 2011. The local Rademacher complexity of lp-norm multiple kernel learning. In *Advances in Neural Information Processing Systems 23 (NIPS 2011)*. MIT Press. 2438–2446.

Koltchinskii, V., and Panchenko, D. 2000. *Rademacher processes and bounding the risk of function learning*. Springer.

Koltchinskii, V., and Panchenko, D. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics* 30:1–50.

Koltchinskii, V. 2006. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* 34(6):2593–2656.

Lanckriet, G. R. G.; Cristianini, N.; Bartlett, P. L.; Ghaoui, L. E.; and Jordan, M. I. 2004. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* 5:27–72.

Liu, Y., and Liao, S. 2014a. Kernel selection with spectral perturbation stability of kernel matrix. *Science China Information Sciences* 57(11):1–10.

Liu, Y., and Liao, S. 2014b. Preventing over-fitting of cross-validation with kernel stability. In *Proceedings of the 7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML /PKDD 2014)*, 290–305. Springer.

Liu, Y.; Jiang, S.; and Liao, S. 2013. Eigenvalues perturbation of integral operator for kernel selection. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, 2189–2198.

Liu, Y.; Jiang, S.; and Liao, S. 2014. Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In *Proceedings of The 31st International Conference on Machine Learning (ICML 2014 (1))*, 324–332.

Liu, Y.; Liao, S.; and Hou, Y. 2011. Learning kernels with upper bounds of leave-one-out error. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011)*, 2205–2208.

Lugosi, G., and Wegkamp, M. 2004. Complexity regularization via localized random penalties. *The Annals of Statistics* 32:1679–1697.

Luxburg, U. V.; Bousquet, O.; and Schölkopf, B. 2004. A compression approach to support vector model selection. *Journal of Machine Learning Research* 5:293–323.

Mendelson, S. 2003. On the performance of kernel classes. *Journal of Machine Learning Research* 4:759–771.

Nguyen, C. H., and Ho, T. B. 2007. Kernel matrix evaluation. In *Proceedings of the 20th International Joint Conference on Artifficial Intelligence (IJCAI 2007)*, 987–992.

Nguyen, C. H., and Ho, T. B. 2008. An efficient kernel matrix evaluation measure. *Pattern Recognition* 41(11):3366–3372.

Saunders, C.; Gammerman, A.; and Vovk, V. 1998. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning (ICML 1998)*, 515–521.

Srebro, N.; Sridharan, K.; and Tewari, A. 2010. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 22*. MIT Press. 2199–2207.

Steinwart, I., and Christmann, A. 2008. *Support vector machines*. New York: Springer Verlag.

Suykens, J. A. K., and Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural Processing Letters* 9(3):293–300.

Vapnik, V. 2000. *The nature of statistical learning theory*. Springer Verlag.

Williamson, R.; Smola, A.; and Scholkopf, B. 2001. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory* 47(6):2156–2132.