Efficient Kernel Selection via Spectral Analysis

Jian Li[1,2], **Yong Liu**[1,3*], **Hailun Lin**[1], **Yinliang Yue**[1], **Weiping Wang**[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Tianjin University
{lijian9026,liuyong,yueyinliang,linhailun,wangweiping}@iie.ac.cn

Abstract

Kernel selection is a fundamental problem of kernel methods. Existing measures for kernel selection either provide less theoretical guarantee or have high computational complexity. In this paper, we propose a novel kernel selection criterion based on a newly defined spectral measure of a kernel matrix, with sound theoretical foundation and high computational efficiency. We first show that the spectral measure can be used to derive generalization bounds for some kernel-based algorithms. By minimizing the derived generalization bounds, we propose the kernel selection criterion with spectral measure. Moreover, we demonstrate that the popular minimum graph cut and maximum mean discrepancy are two special cases of the proposed criterion. Experimental results on lots of data sets show that our proposed criterion can not only give the comparable results as the state-of-the-art criterion, but also significantly improve the efficiency.

1 Introduction

Kernel methods, such as support vector machine (SVM) and least square support vector machine (LSSVM), have been widely used in data mining, pattern recognition and machine learning. The performance of these algorithms greatly depends on the choice of kernel function. Therefore, kernel selection is foundational to kernel methods and is also a challenging problem in kernel methods.

The standard technique for kernel selection is cross-validation (CV). It consists in using a subset of the data for training, and then testing on the remaining data in order to estimate the generalization error. However, CV requires training the learning algorithm several times, which is computationally intensive. For the sake of efficiency, some approximate CV criteria are introduced: such as generalized cross-validation (GCV) [Golub *et al.*, 1979], generalized approximate cross-validation (GACV) [Wahba *et al.*, 1999], span bound [Chapelle *et al.*, 2002], efficient leave-one-out cross-validation (ELOO) [Cawley, 2006], influence function [Debruyne *et al.*, 2008], Bouligand Influence Function [Liu *et al.*, 2014; Liu and Liao, 2017], et al.

Kernel target alignment (KTA) [Cristianini *et al.*, 2001] is another widely used criterion that can be effectively calculated in $O(n^2)$ time complexity, where n is the size of data set. Based on the centered kernel matrix, an improved centered KTA (called CKTA) is proposed [Cortes *et al.*, 2010], which gives better performance than KTA. FSM is another criterion that can be effectively calculated [Nguyen and Ho, 2008], which evaluates the goodness of a kernel function via the data distribution in the feature space. Due to their simplicity and efficiency, KTA, CKTA and FSM have been widely used. However, the connection between these measures and the generalization error of some special algorithms, such as SVM and LSSVM, has not established. Therefore, the kernels chosen by these criteria can not guarantee good performance for these specific algorithms.

Minimizing theoretical estimate bounds of generalization error is an alternative to kernel selection. To this end, some measures of complexity are introduced: such as VC dimension [Vapnik, 2000], Rademacher complexity [Bartlett and Mendelson, 2002], local Rademacher complexity [Bartlett *et al.*, 2005], covering number [Zhang, 2002], uniform stability [Bousquet and Elisseeff, 2002], compression coefficient [Luxburg *et al.*, 2004], eigenvalues perturbation [Liu *et al.*, 2013], kernel stability [Liu and Liao, 2014], eigenvalues ratio [Liu and Liao, 2015], principal eigenvalue proportion [Liu *et al.*, 2017d; 2017c], et al.

However, the focus of these measures lies on deriving theoretical generalization bounds, which are usually difficult to be used for kernel selection in practice. To address this problem, Cortes et al. [2013] used the tail eigenvalues of kernel matrix to design new algorithms for learning kernels. However, the theoretical error bound based on the tail eigenvalues of kernel matrix is lacking. Liu et al. [2015] investigate it further, and introduce a notion of eigenvalue ratio for kernel selection, which can be used to derive generalization bounds. However, the time complexity of this measure is high. Moreover, this measure brings two extra parameters that should be tuned, making this measure hard to use.

In this paper, we propose a novel measure, called spectral measure, with sound theoretical foundation and high computational efficiency. The spectral measure is defined based on the spectral decomposition of kernel matrix, and can be

[*] Corresponding author

effectively calculated in $O(n^2)$. We show that the generalization error of SVM and LSSVM can be bounded with spectral measure. Furthermore, we propose a new kernel selection criterion with spectral measure by minimizing the derived upper bounds to guarantee good generalization performance, and prove that the minimum graph cut and maximum mean discrepancy are its two special cases. Theoretical analyses and experimental results show that our criterion is sound and effective. To our knowledge, the kernel selection criterion with both theoretical guarantee and $\mathcal{O}(n^2)$ time complexity for kernel methods has never been given before.

2 Notations and Preliminaries

We consider supervised learning where a learning algorithm receives a sample of n labeled points $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, where $\mathcal{X}$ denotes the input space and $\mathcal{Y} = \{-1, +1\}$ denotes the output space. We assume S is drawn identically and independently from a fixed, but unknown probability distribution D on $\mathcal{X} \times \mathcal{Y}$.

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function. 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 $\langle K(\mathbf{x}, \cdot), f \rangle_K = f(\mathbf{x}), \forall f \in \mathcal{H}_K$. We use $\| \cdot \|_K$ to denote the norm in $\mathcal{H}_K$. In this paper, we study the regularized algorithms:

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

where $\ell(\cdot, \cdot)$ is a loss function and λ is the regularization parameter. Choosing $\ell(\cdot, \cdot)$ to be the square loss $\ell(t, y) = (t - y)^2$ gives rise to LSSVM while choosing it to be the hinge loss $\ell(t, y) = \max\{0, 1 - yt\}$ produces SVM.

The performance of the regularized algorithms for classification is usually measured by the *generalization error* or *risk* $R(S) = \Pr_{(\mathbf{x},y)\sim D}[y f_S(\mathbf{x}) < 0]$. Unfortunately, $R(S)$ can not be computed since the probability distribution D is unknown, hence we should estimate it from empirical data. In this paper, we will introduce a novel measure to estimate it.

$\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^n$ is the kernel matrix and $\mathbf{N}$ is the normal kernel matrix denoted as $\mathbf{N} = \mathbf{K}/|\mathbf{K}|_1$, where $|\mathbf{K}|_1 = \sum_{i,j=1}^n K(\mathbf{x}_i, \mathbf{x}_j)$. Let $(\lambda_i, \mathbf{v}_i)$ be the *spectral decomposition* of $\mathbf{N}$, where λ_i is the eigenvalue and $\mathbf{v}_i$ is the eigenvector, $i = 1, \ldots, n$.

In the following, we assume that $|\mathbf{K}|_1 =: C$, and $0 \le K(\mathbf{x}, \mathbf{x}') \le \kappa, \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Let $\mathbb{R}_+ = \{t | t \in \mathbb{R}, t \ge 0\}$.

3 Spectral Measure

In this section, we will first introduce the definition of spectral measure, and then use it to derive generalization bounds for LSSVM and SVM.

3.1 Definition of Spectral Measure

It is well known that the kernel matrix contains most of the information needed by kernel methods, and its spectral decomposition plays an important role in kernel matrix. Therefore, we aim at proposing a novel measure based on the spectral decomposition of kernel matrix for kernel selection.

Definition 1 (Spectral Measure (SM)). *Let $(\lambda_i, \mathbf{v}_i)$ be the spectral decomposition of the normal kernel matrix $\mathbf{N}$, $i = 1, \ldots, n$. Assume that $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a function, $\varphi(\lambda_i) \le \lambda_i$ for all $i \in \{1, \ldots, n\}$. Then the spectral measure of K with respect to φ is defined as*

$$\mathrm{SM}(K, \varphi) := \frac{1}{n} \sum_{i=1}^n \varphi(\lambda_i) \langle \mathbf{y}, \mathbf{v}_i \rangle^2,$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$.

The condition of $\varphi(\lambda_i) \le \lambda_i$ looks strange at first glance. Then, we will give two forms of φ to show that it is a very important element to remove the noise. Two special cases of φ are given as follows:

- Hinge form: $h \ge 0$,
$$\varphi(\lambda_i) = \begin{cases} 0 & \text{if } \lambda_i \le h, \\ \lambda_i & \text{otherwise.} \end{cases}$$
- High degree form:
$$\varphi(\lambda_i) = \lambda_i^r, r \ge 1.$$

It is easy to verity that the Hinge form satisfies the assumption of $\varphi(\lambda_i) \le \lambda_i$. From the definition of $\mathbf{N}$, one can see that $0 \le \lambda_i \le 1$, so the high degree form also satisfies the assumption of $\varphi(\lambda_i) \le \lambda_i$.

From the definition of the above two forms of φ, one can see that $\varphi(\lambda_i) = 0$ (Hinge form) or $\varphi(\lambda_i) \to 0$ quickly (High degree form) when λ_i is small. Note that the small value of eigenvalue usually expresses the noise [Steinwart and Christmann, 2008], thus we can use them to remove the noise.

When we use the High degree form,

$$\begin{aligned} \mathrm{SM}(K, \varphi) :&= \frac{1}{n} \sum_{i=1}^n \lambda_i^r \langle \mathbf{y}, \mathbf{v}_i \rangle^2 \\ &= \mathbf{y}^{\mathrm{T}} \left(\sum_{i=1}^n \lambda_i^r \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \right) \mathbf{y} = \mathbf{y}^{\mathrm{T}} \mathbf{N}^r \mathbf{y} \end{aligned}$$

The above equations shows that the $\mathrm{SM}(K, t^r)$ can be effectively calculated in $O(n^2)$.

Remark 1. *The time cost of high degree form is $O(n^2)$, which is much faster than that of Hinge form. However, the accuracy of these two forms is similar, so in this paper, we only consider the use of high degree form for kernel selection.*

3.2 SM-based Generalization Error Bounds

In this subsection, we will derive the generalization bounds for LSSVM and SVM with spectral measure.

Theorem 1. *Consider the LSSVM, and assume that $\|f\|_K \le 1, \forall f \in \mathcal{H}_K$. Then, with probability $1 - \delta$ over the random choice of sample S with size $n \ge 5$, we have*

$$\begin{aligned} R(S) \le 1 - c_0 \cdot \mathrm{SM}(K, \varphi) + \\ \inf_{\theta \in (0,1]} \left[\theta + \frac{7\mu + 3\sqrt{3\mu} + 6}{3n} + \sqrt{\frac{3\mu}{n}} \right], \end{aligned}$$

where $\mu = \frac{8}{\theta^2} \ln n \ln(2n) + \ln \frac{2n}{\delta}$, $c_0 = \frac{C\lambda}{C+\lambda}$.

Table 1: Time complexity and theoretical guarantee of Cross-Validation, KTA, CKTA, FSM and ER

Criteria	Time complexity	Theory
Cross-Validation	$O(n^3)$ at least	Yes
KTA, CKTA, FSM	$O(n^2)$	No
ER	$O(n^3)$	Yes
SM (Ours)	$O(n^2)$	Yes

Proof. The proof is given in Appendix A. □

The assumption of $\forall f \in \mathcal{H}_K, \|f\|_K \leq 1$ is a common assumption which is used in [Bartlett *et al.*, 2005; Kloft and Blanchard, 2011; Cortes *et al.*, 2013].

The above theorem shows that the generalization error of LSSVM can be bounded with spectral measure $\mathrm{SM}(K, \varphi)$. Therefore, we can choose the kernel function by maximizing $\mathrm{SM}(K, \varphi)$ to guarantee good generalization performance.

Theorem 2. *Consider the SVM, and assume that $\|f\|_K \leq 1, \forall f \in \mathcal{H}_K$. Then, for SVM, with probability $1 - \delta$ over the random choice of sample S with size $n \geq 5$, we have*

$$R(S) \leq 1 - C \cdot \mathrm{SM}(K, \varphi) + \inf_{\theta \in (0,1]} \left[\theta + \frac{7\mu + 3\sqrt{3\mu} + 3/(2\lambda) + 3b}{3n} + \sqrt{\frac{3\mu}{n}} \right],$$

where $\mu = \frac{8}{\theta^2} \ln n \ln(2n) + \ln \frac{2n}{\delta}$, and $b = \max\{1, \frac{1}{2\lambda} - 1\}$.

Proof. The proof is given in Appendix B. □

The above theoretical analysis demonstrates the theoretical guarantee of spectral measure for kernel selection.

4 Spectral Kernel Selection

In this section, we will present a novel kernel selection criterion with spectral measure, and show the minimum graph cut and maximum mean discrepancy are its two special cases.

According to Theorem 1 and Theorem 2, to guarantee good generalization performance, we can select the kernel function by maximizing the spectral measure $\mathrm{SM}(K, \varphi)$. Note that we can set $\varphi(t) = t^r$ to ignore the noise, and we know that $\mathrm{SM}(K, t^r)$ can be calculated effectively. Thus, we consider the use of the following kernel selection criterion:

$$\operatorname*{arg\,max}_{K \in \mathcal{K}} \mathrm{SM}(K, t^r) = \frac{1}{n} \mathbf{y}^{\mathrm{T}} \mathbf{N}^r \mathbf{y}$$

where $\mathcal{K}$ is a candidate set of kernel functions. Note that we can give different weights to positive and negative classes according to their sample sizes to avoid the imbalance problem of positive and negative classes. Thus, we finally consider the following weighted **spectral measure criterion** (SM):

$$\operatorname*{arg\,max}_{K \in \mathcal{K}} \overline{\mathrm{SM}}(K, t^r) = \frac{1}{n} \bar{\mathbf{y}}^{\mathrm{T}} \mathbf{N}^r \bar{\mathbf{y}}, \qquad (1)$$

where $\bar{y}_+ = \frac{n}{n_+}$ and $\bar{y}_- = -\frac{n}{n_-}$, n_+ and n_- are respective the sizes of positive and negative classes. One can see that the time complexity of SM criterion is $O(n^2)$. The time complexity of existing popular kernel selection criteria, and whether they have theoretical guarantee are reported in Table 1. We can find that our SM is the only criterion with both theoretical guarantee and $O(n^2)$ time complexity.

Connections to Graph Cut

In graph theory, graph cut is used to measure the degree of dissimilarity of different segmentation [Shi and Malik, 2000]. Note that $K(\mathbf{x}_i, \mathbf{x}_j)$ can be considered as the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, thus we can use kernel matrix $\mathbf{K}$ as similar matrix to construct a graph. In this case, the normalized graph cut (Ncut) [Shi and Malik, 2000] can be written as

$$\mathrm{Ncut}(K) = \frac{\mathbf{y}^{\mathrm{T}} \mathbf{L} \mathbf{y}}{\mathbf{y}^{\mathrm{T}} \mathbf{D} \mathbf{y}},$$

where $\mathbf{L} = \mathbf{D} - \mathbf{K}$, $\mathbf{K}$ is the kernel matrix, $\mathbf{D}$ is the diagonal matrix with the diagonal element $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{K}_{ij}$.

If the positive and negative classes are balanced, that is $n_+ = n_-$, then $\mathbf{y}^{\mathrm{T}} \mathbf{D} \mathbf{y} = \sum_{i=1}^{n} \mathbf{D}_{ii} = \sum_{i,j=1}^{n} \mathbf{K}_{ij} = |\mathbf{K}|_1$. Thus, we have

$$\mathrm{Ncut}(K) = 1 - \frac{\mathbf{y}^{\mathrm{T}} \mathbf{K} \mathbf{y}}{\mathbf{y}^{\mathrm{T}} \mathbf{D} \mathbf{y}} = 1 - \frac{\bar{\mathbf{y}}^{\mathrm{T}} \mathbf{K} \bar{\mathbf{y}}}{2|\mathbf{K}|_1} = 1 - \frac{n}{2} \cdot \overline{\mathrm{SM}}(K, t),$$

The above equation shows that minimizing $\mathrm{Ncut}(K)$ is equal to maximizing $\overline{\mathrm{SM}}(K, \varphi)$ when setting $\varphi(t) = t$ for kernel selection.

Connections to Mean Discrepancy

Mean discrepancy (MD) [Gretton *et al.*, 2007] is proposed to test whether two distributions are different on the basis of samples drawn from each of them. The estimate of MD in RKHS $\mathcal{H}_K$ [Gretton *et al.*, 2007] can be written as

$$\mathrm{MD}(K) = \Big\| \frac{1}{n_+} \sum_{i}^{n_+} \phi(\mathbf{x}_i) - \frac{1}{n_-} \sum_{j}^{n_-} \phi(\mathbf{x}_j) \Big\|_K,$$

where $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_K = \mathbf{K}_{ij}$.

If $\mathbf{K}$ has been normalized, that is $\mathbf{K} = \mathbf{N}$, it is easy to verity that

$$\begin{aligned} \overline{\mathrm{SM}}(K, t) &= \frac{1}{n} \bar{\mathbf{y}}^{\mathrm{T}} \mathbf{N} \bar{\mathbf{y}} = \frac{1}{n} \bar{\mathbf{y}}^{\mathrm{T}} \mathbf{K} \bar{\mathbf{y}} \\ &= \frac{1}{n_+^2} \sum_{i,j}^{n_+} \mathbf{K}_{ij} + \frac{1}{n_-^2} \sum_{i,j}^{n_-} \mathbf{K}_{ij} - \frac{2}{n_- n_+} \sum_{i}^{n_+} \sum_{j}^{n_-} \mathbf{K}_{ij} \\ &= \Big\| \frac{1}{n_+} \sum_{i}^{n_+} \phi(\mathbf{x}_i) - \frac{1}{n_-} \sum_{j}^{n_-} \phi(\mathbf{x}_j) \Big\|_K. \end{aligned}$$

The above equation will show that $\mathrm{MD}(K)$ is a special case of $\overline{\mathrm{SM}}(K, \varphi)$, which demonstrates the effectiveness of our proposed kernel selection criterion again.

Remark 2. *The measure of graph cut and mean discrepancy are two special case of our SM when $\varphi(t) = t$, but in our analysis of the above subsection, we know that $\varphi(t) = t^r$, $r \geq 1$ can used to ignore the noise.*

Remark 3. *Instead of choosing a single kernel, several authors consider the use of multiple kernels by some criteria, called multiple kernel learning (MKL), see [Lanckriet* et al.*, 2004; Kloft* et al.*, 2011; Liu* et al.*, 2017a; 2017b] and references therein. Our spectral measure criterion can be applied to MKL: $\min_{\boldsymbol{\mu}} \overline{\mathrm{MS}}(K_{\boldsymbol{\mu}}, t^r)$ s.t.$\|\boldsymbol{\mu}\|_p = 1, \boldsymbol{\mu} \geq 0$, where $K_{\boldsymbol{\mu}} = \sum_{i=1}^{k} \mu_i K_i$. The above optimization problem can be*

Table 2: Comparison of test errors (%) among our spectral measure criterion (SM) and other five popular ones including 5-fold cross-validation (CV), efficient leave-one-out cross-validation (ELOO), centered kernel target alignment (CKTA), feature space-based kernel matrix evaluation (FSM) and eigenvalue ratio (ER). 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.

	SM	CV	ELOO	CKTA	FSM	ER
a1a	**16.84±1.39**	17.02±1.57	$\underline{16.88\pm1.41}$	18.86±1.49	24.72±1.67	$\underline{16.97\pm1.52}$
a2a	**17.78±1.28**	$\underline{17.96\pm1.25}$	$\underline{17.94\pm1.27}$	18.52±1.26	25.62±1.47	18.99±1.37
anneal	**2.69±3.28**	3.81±4.11	**2.69±3.28**	4.75±4.78	5.13±4.18	5.50±4.95
australian	$\underline{13.71\pm2.10}$	$\underline{13.84\pm2.18}$	$\underline{13.82\pm2.04}$	13.91±1.89	44.71±2.47	**13.53±2.06**
autos	**11.81±11.67**	**11.81±11.67**	12.75±11.06	13.71±12.03	12.71±8.06	12.14±11.51
breast-w	**3.27±1.01**	3.56±1.16	3.59±1.08	3.51±1.05	3.50±1.05	4.26±1.40
breast-cancer	**3.18±1.15**	3.63±1.16	3.50±1.23	3.63±1.16	3.60±1.14	4.04±1.12
bupa	30.29±3.48	**29.10±4.04**	30.31±4.27	35.81±3.45	39.77±3.68	$\underline{29.13\pm4.46}$
colic	**15.62±3.00**	16.47±2.78	$\underline{15.73\pm2.97}$	19.27±2.58	36.42±3.28	$\underline{17.35\pm3.09}$
diabetes	24.22±2.41	24.69±2.71	**23.51±2.75**	24.85±2.46	35.30±3.00	23.90±2.48
glass	22.09±5.07	$\underline{21.82\pm5.68}$	**20.95±4.82**	26.41±7.13	43.00±9.22	22.50±5.08
german.numer	$\underline{24.09\pm2.15}$	25.28±2.38	**23.81±2.26**	26.02±2.16	29.89±2.41	25.33±2.14
heart	16.53±3.27	16.69±3.36	**15.95±3.29**	18.67±3.78	44.37±5.50	$\underline{15.98\pm3.47}$
hepatitis	**15.57±4.68**	17.09±5.74	16.63±4.64	$\underline{15.74\pm5.00}$	21.22±5.41	18.91±6.20
ionosphere	$\underline{4.88\pm2.10}$	5.28±2.11	6.42±2.17	11.70±3.43	35.77±4.00	**4.86±1.99**
labor	**13.65±8.10**	$\underline{14.47\pm8.08}$	14.82±8.34	15.41±8.80	34.59±8.70	18.82±8.81
pima	23.80±2.14	$\underline{22.78\pm2.36}$	**22.51±2.41**	24.38±2.28	34.47±2.42	$\underline{22.78\pm2.07}$
segment	**0.01±0.00**	0.06±0.24	0.20±0.04	0.32±0.03	0.21±0.01	0.24±0.04
liver-disorders	31.94±3.21	**29.00±4.11**	30.02±4.76	36.27±3.93	40.90±4.10	29.69±4.97
sonar	15.06±4.80	14.26±4.93	**13.68±4.43**	15.00±5.51	49.32±6.93	18.84±5.75
vehicle	**3.02±1.79**	3.33±1.77	**3.02±1.79**	3.77±1.51	53.32±3.38	5.52±2.44
vote	**4.31±1.71**	4.78±1.74	4.82±1.73	5.25±1.72	6.37±3.96	7.80±2.33
wpbc	23.10±4.58	22.83±4.32	$\underline{21.93\pm4.45}$	**21.87±4.13**	$\underline{22.13\pm4.19}$	**21.87±4.13**
tic-tac-toe	10.10±1.93	10.28±1.66	**9.78±1.66**	33.62±5.31	34.44±2.04	14.62±2.05
wdbc	**2.29±1.15**	$\underline{2.43\pm1.07}$	2.73±1.11	2.82±1.20	37.49±3.83	4.75±1.66

efficiently solved with gradient-based algorithms. However, in this paper, we mainly want to verify the effectiveness of our spectral measure criterion. Therefore, in our experiments, we focus on comparing our criterion with other popular kernel selection criteria.

5 Experiments

In this section, we will empirically analyze the performance of our proposed spectral measure criterion $\overline{\mathrm{SM}}(K, t^r)$ (SM).

The evaluation is made on 25 publicly available data sets from UCI, StatLib and Weka Collections seen in Table 2. For each data set, we run all methods 50 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 $d_i = B_i - A_i$, $i = 1, \ldots, 50$. 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 $\frac{\bar{d}}{S_d/\sqrt{50}} > 1.676$. All statements of statistical significance in the remainder refer to a 95% level of significance. Experiments are conducted on a Dell PC with 3.1-GHz 4-core CPU and 4-GB memory. We use the popular Gaussian kernels $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\tau}\right)$ as our candidate kernels, $\tau \in \{2^i, i = -15, -14, \ldots, 15\}$. In this paper, we mainly focus on choosing the kernel selection, so we set the regularization parameter $\lambda = 1$ for all methods. The learning machine we used is LSSVM.

In the first experiment, we compare the performance of our proposed SM criterion with five popular kernel selection criteria: 5-fold cross-validation (CV), efficient leave-one-out cross-validation (ELOO) [Cawley, 2006], centered kernel target alignment (CKTA) [Cortes *et al.*, 2010], feature space-based kernel matrix evaluation (FSM) [Nguyen and Ho, 2007] and the latest eigenvalue ratio (ER) [Liu and Liao, 2015]. For each data set, we choose the kernel parameter τ by each kernel selection criterion on the training set and then evaluate the test error for the chosen parameters on the test set. In this experiment, we set $r = 3$. We will explore the influence of r later. Table 2 reports the average test errors that can be summarized as follows: (a) SM is significantly better than FSM and CKTA on nearly all data sets. This can be explained by the fact that FSM and CKTA don't have theoretical guarantee for special learning algorithm, so the kernels chosen by these criteria can not guarantee good generalization performance; (b) SM is better than CV and ER on most data sets. In particular, SM is significantly better than ER (or CV) on 15 (or 11) of 25 data sets, but only 4 (or 5) sets significantly worse on the remaining data sets; (c) SM gives compara-

Table 3: Comparison of run time (second) among our proposed SM and other five ones including CV, ELOO, CKTA, FSM and ER.

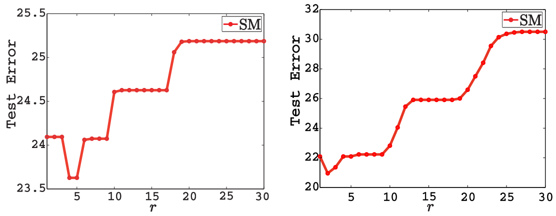
	SM	CV	ELOO	CKTA	FSM	ER
a1a	**63.77**	704.50	217.03	87.51	79.71	422.71
a2a	**124.53**	1995.76	810.87	172.28	156.68	1558.70
anneal	**0.28**	4.60	0.84	0.43	0.55	1.36
australian	**8.30**	111.31	30.33	11.34	9.51	52.53
autos	**0.10**	2.18	0.23	0.17	0.29	0.37
breast-cancer	**8.05**	104.55	27.13	11.10	9.36	47.68
breast-w	**8.59**	105.39	29.77	11.80	10.11	49.86
colic	**1.55**	25.72	6.52	2.15	2.17	11.60
glass	**0.33**	5.74	1.15	0.53	0.66	2.00
heart	**1.02**	14.71	3.66	1.12	1.20	6.33
hepatitis	**0.38**	6.55	1.42	0.58	0.70	2.44
ionosphere	**1.59**	24.60	6.08	2.03	2.06	10.60
labor	**0.16**	2.88	0.42	0.28	0.42	0.67
pima	**10.99**	137.19	36.60	15.11	12.74	62.45
segment	**7.50**	91.73	23.83	10.49	8.93	42.63
diabetes	**10.70**	134.71	36.26	14.90	12.51	62.46
german.numer	**21.80**	249.98	72.63	29.59	26.27	127.51
liver-disorders	**1.37**	21.04	5.46	1.91	1.94	9.27
sonar	**0.62**	10.53	2.34	0.87	0.98	4.06
vehicle	**2.03**	35.64	9.35	2.86	2.85	15.95
vote	**2.03**	34.43	9.33	2.79	2.81	15.92
wpbc	**0.52**	9.06	2.07	0.76	0.92	3.33
bupa	**1.36**	21.09	5.34	1.85	1.88	9.19
tic-tac-toe	**18.93**	224.13	63.10	26.13	23.04	107.12
wdbc	**5.75**	61.32	19.18	7.86	6.80	33.65

ble results with ELOO. In particular, SM is significantly better than ELOO on 9 data sets (autos, breast-w, breast-cancer, hepatitis, ionosphere, labor, segment, vote and wdbc) and is significantly worse on 8 data sets (diabetes, glass, heart, pima, liver-disorders, sonar, wpbc and tic-tac-toe). The run time of SM, CV, ELOO, CKTA, FSM and ER are reported in Table 3. It turns out that SM is much faster than CV, ELOO and ER, and gives comparable results with CKTA and FSM. The above results manifest that SM can guarantee generalization performance and has high computational efficiency as well.

In the following experiment, we explore the influence of the parameter r of SM. Figure 1 plots the average test errors with different r (due to space limit, we randomly select 2 data sets). For each fixed r, we choose the kernel parameter τ by SM on the training set, and evaluate the test errors for the chosen parameters on test set. We find that the optimal r belong to $[2, 5]$ on most data sets. And we can also find out that we can randomly set $r = 2$, $r = 3$ or $r = 4$ in practice without sacrificing accuracy.

6 Conclusion

In this paper, we define a new spectral measure of a kernel matrix, which can be effectively calculated in $O(n^2)$ time complexity. With the spectral measure, we can bound the generalization errors of LSSVM and SVM, and propose the kernel selection criterion by minimizing the derived generalization error bounds, of which the minimum graph cut and maximum mean discrepancy are two special cases. The high computational efficiency and theoretical guarantee of the spectral measure may bring a new perspective on designing kernel selection criterion using spectral analysis.



(a) glass (b) german-numer

Figure 1: The test errors of SM criterion with different r. For each r, we choose the kernel by SM on the training set, and evaluate the test errors for the chosen parameters on test set. (Due to space limited, we randomly select 2 data sets)

Appendix A: Proof of the Theorem 1

To prove Theorem 1, we first prove the following Lemma:

Lemma 1. *If f_S is the solution of LSSVM, then*

$$\frac{1}{n}\sum_{i=1}^{n} y_i f_S(\mathbf{x}_i) \geq c_0 \cdot \mathrm{SM}(K, \varphi). \tag{2}$$

Proof. Note that $(f_S(\mathbf{x}_1), \ldots, f_S(\mathbf{x}_n))^{\mathrm{T}} = \mathbf{K}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha} = [\mathbf{K} + \lambda\mathbf{I}]^{-1}\mathbf{y}$. Thus

$$\sum_{i=1}^{n} y_i f_S(\mathbf{x}_i) = \mathbf{y}^{\mathrm{T}}\mathbf{K}\boldsymbol{\alpha} = \mathbf{y}^{\mathrm{T}}\mathbf{K}[\mathbf{K} + \lambda\mathbf{I}]^{-1}\mathbf{y}. \tag{3}$$

Let $(\beta_i, \mathbf{u}_i)$ be the spectral decomposition of $\mathbf{K}$, then we have

$$\mathbf{y}^{\mathrm{T}}\mathbf{K}[\mathbf{K} + \lambda\mathbf{I}]^{-1}\mathbf{y} = \sum_{i=1}^{n} \frac{\beta_i}{\lambda + \beta_i}\langle \mathbf{y}, \mathbf{u}_i\rangle^2. \tag{4}$$

Since $\mathrm{Tr}(\mathbf{K}) \leq |\mathbf{K}|_1 = C$, we know that $\frac{\beta_j}{C} \leq \frac{\mathrm{Tr}(\mathbf{K})}{C} \leq 1$. According to Equation (4), we have $\mathbf{y}^{\mathrm{T}}\mathbf{K}[\mathbf{K} + \lambda\mathbf{I}]^{-1}\mathbf{y} = \sum_{i=1}^{n} \frac{\beta_i/C}{\beta_i/C + \lambda/C}\langle \mathbf{y}, \mathbf{u}_i\rangle^2 \geq \sum_{i=1}^{n} \frac{\beta_i/C}{1+\lambda/C}\langle \mathbf{y}, \mathbf{u}_i\rangle^2$. Note that $\lambda_i = \beta_i/|\mathbf{K}|_1 = \beta_i/C$, $\mathbf{v}_i = \mathbf{u}_i$, where $(\lambda_i, \mathbf{v}_i)$ is the spectral decomposition of $\mathbf{N}$, thus

$$\mathbf{y}^{\mathrm{T}}\mathbf{K}[\mathbf{K} + \lambda\mathbf{I}]^{-1}\mathbf{y} \geq \sum_{i=1}^{n} \frac{\lambda_i}{1 + \lambda/C}\langle \mathbf{y}, \mathbf{v}_i\rangle^2. \tag{5}$$

From the assumption $\varphi(\lambda_i) \leq \lambda_i$, Equations (3), (4) and (5), we have $\sum_{i=1}^{n} y_i f_S(\mathbf{x}_i) \geq \sum_{i=1}^{n} \frac{\varphi(\lambda_i)}{1+\lambda/C}\langle \mathbf{y}, \mathbf{v}_i\rangle^2$, which completes the proof.

□

Theorem 3 (Theorem 8 [Gao and Zhou, 2013])**.** *For any $\delta > 0$, with probability at least $1 - \delta$ over the random choice of sample S with size $n \geq 5$, every $f \in \mathcal{H}_K$ satisfies the following bound:*

$$\Pr_D[yf(\mathbf{x}) < 0] \leq \frac{2}{n} + \inf_{\theta \in (0,1]}\Bigg[\Pr_S[yf(\mathbf{x}) \leq \theta] + \frac{7\mu + 3\sqrt{3\mu}}{3n} + \sqrt{\frac{3\mu}{n}\Pr_S[yf(\mathbf{x}) \leq \theta]}\Bigg],$$

where $\mu = \frac{8}{\theta^2}\ln n \ln(2n) + \ln\frac{2n}{\delta}$.

Proof of the Theorem 1. Note that $\Pr_S[yf_S(\mathbf{x}) \leq \theta] = \frac{1}{n}\sum_{i=1}^{n} 1_{[y_i f_S(\mathbf{x}_i)-\theta]}$, where $1_{[t]} = 1$, if $t < 0$, otherwise $1_{[t]} = 0$. Thus, it is easy to verify that

$$\Pr_S[yf_S(\mathbf{x}) \leq \theta] \leq \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, 1 - y_i f_S(\mathbf{x}_i) + \theta\right\}. \quad (6)$$

From the assumption that $\mathcal{H}_K$ ranges in [-1,1], we have $1 - y_i f_S(\mathbf{x}_i) + \theta > 0$. Thus, by Equation (6), we have

$$\Pr_S[yf_S(\mathbf{x}) \leq \theta] \leq \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f_S(\mathbf{x}_i) + \theta). \quad (7)$$

According to Equation (7) and Equation (2), we have

$$\Pr_S[yf_S(\mathbf{x}) \leq \theta] \leq 1 + \theta - \frac{1}{n}\sum_{i=1}^{n} y_i f_S(\mathbf{x}_i) \leq$$
$$1 + \theta - c_0 \cdot \mathrm{SM}(K, \varphi). \quad (8)$$

Substituting Equation (8) to Theorem 3, we have

$$\Pr_D[yf_S(\mathbf{x}) < 0] \leq \frac{2}{n} + \inf_{\theta\in(0,1]} \Big[[1 + \theta - c_0 \cdot \mathrm{SM}(K, \varphi)] +$$
$$\frac{7\mu + 3\sqrt{3\mu}}{3n} + \sqrt{\frac{3\mu}{n}}\Big].$$

This completes the proof for Theorem 1.

□

Appendix B: Proof of the Theorem 2

Similar with the proof of the Theorem 1, we first prove the following lemma:

Lemma 2. *If f_S be the solution of SVM, then*

$$\frac{1}{n}\sum_{i=1}^{n} y_i f_S(\mathbf{x}_i) \geq C\left(\mathrm{SM}(K, \varphi) + \frac{d}{n}\right),$$

where $d = \min\{-1, 1 - \frac{1}{2\lambda}\}$.

Proof. Note that $f_S(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x})$, where $\boldsymbol{\alpha}$ is the solution of the dual form of SVM. Thus, we have

$$\frac{1}{n}\sum_{i=1}^{n} y_i f_S(\mathbf{x}_i) = \frac{1}{n}\mathbf{y}^{\mathrm{T}}\mathbf{K}(\mathbf{y} \otimes \boldsymbol{\alpha}), \quad (9)$$

where $\otimes$ is the entrywise matrix product (also known as the Hadamard product). Note that

$$\begin{aligned}
&\mathbf{y}^{\mathrm{T}}\mathbf{K}(\mathbf{y} \otimes \boldsymbol{\alpha}) - \mathbf{y}^{\mathrm{T}}\mathbf{K}\mathbf{y} \\
&= \Big[\sum_{y_i=y_j} \alpha_i \mathbf{K}_{ij} - \sum_{y_i\neq y_j} \alpha_i \mathbf{K}_{ij}\Big] - \Big[\sum_{y_i=y_j} \mathbf{K}_{ij} - \sum_{y_i\neq y_j} \mathbf{K}_{ij}\Big] \\
&= \Big[\sum_{y_i=y_j} \alpha_i \mathbf{K}_{ij} + \sum_{y_i\neq y_j} \mathbf{K}_{ij}\Big] - \Big[\sum_{y_i\neq y_j} \alpha_i \mathbf{K}_{ij} + \sum_{y_i=y_j} \mathbf{K}_{ij}\Big] \\
&\geq \sum_{i,j} \min\{\alpha_i, 1\} \cdot \mathbf{K}_{ij} - \sum_{i,j} \max\{\alpha_i, 1\} \cdot \mathbf{K}_{ij} \\
&= \sum_{i,j} (\min\{\alpha_i, 1\} - \max\{\alpha_i, 1\}) \cdot \mathbf{K}_{ij} \\
&= \sum_{i,j} c_i \mathbf{K}_{ij}, \qquad (10)
\end{aligned}$$

where $c_i = \min\{\alpha_i, 1\} - \max\{\alpha_i, 1\}$. Note that $0 \leq \alpha_i \leq \frac{1}{2\lambda}$, so we can obtain that

$$c_i = \min\{\alpha_i - 1, 1 - \alpha_i\} \geq \min\{-1, 1 - \frac{1}{2\lambda}\} =: d$$

From Equation (10), we have

$$\mathbf{y}^{\mathrm{T}}\mathbf{K}(\mathbf{y} \otimes \boldsymbol{\alpha}) - \mathbf{y}^{\mathrm{T}}\mathbf{K}\mathbf{y} \geq d\sum_{i,j} \mathbf{K}_{ij} = d \cdot C.$$

Thus, we can obtain that $\mathbf{y}^{\mathrm{T}}\mathbf{K}(\mathbf{y} \otimes \boldsymbol{\alpha}) \geq dC + \mathbf{y}^{\mathrm{T}}\mathbf{K}\mathbf{y} = dC + \sum_i \beta_i \langle \mathbf{y}_i, \mathbf{u}_i\rangle^2 = dC + C\sum_i \lambda_i \langle \mathbf{y}_i, \mathbf{v}_i\rangle^2 \geq dC + C\sum_i \varphi(\lambda_i) \cdot \langle \mathbf{y}_i, \mathbf{v}_i\rangle^2$, where $(\beta_i, \mathbf{u}_i)$ is the spectral decomposition of $\mathbf{K}$. This completes the proof. □

Proof of Theorem 2. According to Lemma 2 and Equation (7), we can obtain that

$$\Pr_S[yf_S(\mathbf{x}) \leq \theta] \leq 1 + \theta - C \cdot \mathrm{SM}(K, \varphi) - \frac{Cd}{n}. \quad (11)$$

Substituting Equation (11) to Theorem 3, it is easy to complete the proof. □

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No.61673293, No.61602467, No.61303056), and Excellent Talent Introduction, Institute of Information Engineering, CAS (Y7Z0111107), and Youth Innovation Promotion Association, CAS (No.2016146).

References

[Bartlett and Mendelson, 2002] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[Bartlett *et al.*, 2005] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[Bousquet and Elisseeff, 2002] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[Cawley, 2006] Gavin C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proceeding of the International Joint Conference on Neural Networks (IJCNN 2006)*, pages 1661–1668, 2006.

[Chapelle *et al.*, 2002] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.

[Cortes *et al.*, 2010] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Two-stage learning kernel algorithms. In *Proceedings of the 27th Conference on Machine Learning (ICML 2010)*, pages 239–246, 2010.

[Cortes *et al.*, 2013] Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local Rademacher complexity. In *Advances in Neural Information Processing Systems 25 (NIPS 2013)*, pages 2760–2768. MIT Press, 2013.

[Cristianini *et al.*, 2001] Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz S. Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pages 367–373, 2001.

[Debruyne *et al.*, 2008] Michiel Debruyne, Mia Hubert, and Johan A.K. Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9:2377–2400, 2008.

[Gao and Zhou, 2013] Wei Gao and Zhi-Hua Zhou. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

[Golub *et al.*, 1979] Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[Gretton *et al.*, 2007] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the twosample problem. In *Advances in Neural Information Processing Systems 15 (NIPS 2007)*, pages 513–520, 2007.

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

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

[Lanckriet *et al.*, 2004] Gert R. G. Lanckriet, Nello Cristianini, Peter L. Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[Liu and Liao, 2014] Yong Liu and Shizhong Liao. Preventing over-fitting of cross-validation with kernel stability. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML)*, pages 290–305, 2014.

[Liu and Liao, 2015] Yong Liu and Shizhong Liao. Eigenvalues ratio for kernel selection of kernel methods. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2814–2820, 2015.

[Liu and Liao, 2017] Yong Liu and Shizhong Liao. Granularity selection for cross-validation of svm. *Information Sciences*, 378:475–483, 2017.

[Liu *et al.*, 2013] Yong Liu, Shali Jiang, and Shizhong Liao. Eigenvalues perturbation of integral operator for kernel selection. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, pages 2189–2198, 2013.

[Liu *et al.*, 2014] Yong Liu, Shali Jiang, and Shizhong Liao. 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))*, pages 324–332, 2014.

[Liu *et al.*, 2017a] Xinwang Liu, Miaomiao Li, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel k-means with incomplete kernels. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 2259–2265, 2017.

[Liu *et al.*, 2017b] Xinwang Liu, Sihang Zhou, Yueqing Wang, Miaomiao Li, Yong Dou, En Zhu, and Jianping Yin:. Optimal neighborhood kernel clustering with multiple kernels. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 2266–2272, 2017.

[Liu *et al.*, 2017c] Yong Liu, Shizhong Liao, Hailun Lin, Yinliang Yue, and Weiping Wang. Generalization analysis for ranking using integral operator. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 2273–2279, 2017.

[Liu *et al.*, 2017d] Yong Liu, Shizhong Liao, Hailun Lin, Yinliang Yue, and Weiping Wang. Infinite kernel learning: generalization bounds and algorithms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 2280–2286, 2017.

[Luxburg *et al.*, 2004] Ulrike Von Luxburg, Olivier Bousquet, and Bernhard Schölkopf. A compression approach to support vector model selection. *Journal of Machine Learning Research*, 5:293–323, 2004.

[Nguyen and Ho, 2007] Canh Hao Nguyen and Tu Bao Ho. Kernel matrix evaluation. In *Proceedings of the 20th International Joint Conference on Artifficial Intelligence (I-JCAI 2007)*, pages 987–992, 2007.

[Nguyen and Ho, 2008] Canh Hao Nguyen and Tu Bao Ho. An efficient kernel matrix evaluation measure. *Pattern Recognition*, 41(11):3366–3372, 2008.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[Steinwart and Christmann, 2008] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Verlag, New York, 2008.

[Vapnik, 2000] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

[Wahba *et al.*, 1999] Grace Wahba, Yi Lin, and Hao Zhang. GACV for support vector machines. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge,, 1999.

[Zhang, 2002] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.