Eigenvalues Perturbation of Integral Operator for Kernel Selection

Yong Liu Shali Jiang Shizhong Liao*

School of Computer Science and Technology
Tianjin University, Tianjin 300072, P. R. China
szliao@tju.edu.cn

ABSTRACT

Kernel selection is one of the key issues both in recent research and application of kernel methods. This is usually done by minimizing either an estimate of generalization error or some other related performance measure. It is well known that a kernel matrix can be interpreted as an empirical version of a continuous integral operator, and its eigenvalues converge to the eigenvalues of integral operator. In this paper, we introduce new kernel selection criteria based on the eigenvalues perturbation of the integral operator. This perturbation quantifies the difference between the eigenvalues of the kernel matrix and those of the integral operator. We establish the connection between eigenvalues perturbation and generalization error. By minimizing the derived generalization error bounds, we propose the kernel selection criteria. Therefore the kernel chosen by our proposed criteria can guarantee good generalization performance. To compute the values of our criteria, we present a method to obtain the eigenvalues of integral operator via the Fourier transform. Experiments on benchmark datasets demonstrate that our kernel selection criteria are sound and effective.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Parameter Learning*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier Design and Evaluation*; H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Theory, Experimentation

Keywords

Kernel Selection, Eigenvalues Perturbation, Integral Operator, Generalization Error.

*corresponding author

CIKM'13, October 27–November 01, 2013, San Francisco, CA, USA

http://dx.doi.org/10.1145/2505515.2505584.

1. INTRODUCTION

Kernel methods [33, 29, 10, 30, 31] have been widely used in pattern recognition and machine learning. Because the performance of kernel methods greatly depends on the choice of the kernel function, the kernel selection becomes an important topic in kernel methods. A related problem is the evaluation of the generalization ability of learning algorithms. In fact, it is common to select the optimal kernel function by choosing the one with the lowest generalization error.

Obviously, the generalization error is not directly computable, as the probability distribution generating the data is unknown, therefore it is necessary to resort to estimates of its value. The generalization error can be estimated either via theoretical bounds or testing on some unused data (hold-out testing or cross validation). To estimate the upper bounds of the generalization error, some complexity measures are introduced: such as VC dimension [33], Rademacher complexity [3], maximal discrepancy [2], regularized risk [29], radius-margin bound [33] and compression coefficient [24]. However, for most of these complexity measures, proposed to derive theoretical generalization error bounds, it is difficult to compute their values [25, 26], which make them hard to be used for kernel selection in practice. Minimizing the empirical estimate of the generalization error is an alternative to kernel selection. K-fold cross-validation (KCV) and leave-one-out cross-validation (LOO) [9, 23] are two popular empirical estimates. Although KCV and LOO are widely used in many fields, they have their dark sides: (a) the overall learning problem may over-fitting the cross-validation error [6, 7]; (b) high computational cost. For the sake of efficiency, some approximate KCV and LOO criteria are given: such as generalized cross-validation (GCV)[19], generalized comparative Kullback-Liebler distance (GCKL) [34], generalized approximate cross-validation (GACV) [35], span bound [8, 9] and influence function [15].

Based on the similarity, Cristianini et al. [14] present a kernel selection criterion called kernel target alignment (KTA). Nguyen and Ho [25, 26] point out several drawbacks of the KTA, and propose a surrogate measure (called FSM) to evaluate the goodness of a kernel function via the data distribution in the feature space. Similar to KTA, Cortes et al. [12] present a centered kernel target alignment criterion (CKTA) with a centered kernel matrix. Although KTA, CKTA and FSM are widely used, the connection between these criteria and generalization error for specific learning algorithms has not been established, so the kernels chosen by these criteria may not guarantee good generalization performance.

It is well known that the kernel matrix contains most of the information needed by the kernel methods, and its eigenvalues play an important role in kernel matrix. Because the kernel matrix can be interpreted as an empirical version of a continuous integral operator, and its eigenvalues converge to the eigenvalues of integral operator [5, 20, 28], therefore we aim at presenting new kernel selection criteria based on the eigenvalues perturbation of integral operator in this paper. This perturbation quantifies the difference between the eigenvalues of kernel matrix and those of integral operator. Different from most of the existing complexity measures, we can compute the value of eigenvalues perturbation for any given kernel function from empirical data, which makes it usable for kernel selection. We first use the eigenvalues perturbation to derive generalization error bounds for kernel ridge regression (KRR) and Support Vector Machine (SVM). Then, by minimizing the derived generalization error bounds, we propose two new kernel selection criteria: EPKRR (for KRR) and EPSVM (for SVM). To compute the values of our proposed criteria, we propose a method to compute the eigenvalues of integral operator based on the Fourier transform. For the popular Gaussian kernel and Laplacian kernel, the closed form of eigenvalues of their corresponding integral operators are given. Experimental results show that, for classification, the kernel chosen by EPSVM gives better results than those chosen by the popular classification criteria: CKTA, FSM and KCV, and for regression, EPKRR better than the popular regression criteria: KCV, LOO and GCV.

The rest of the paper is organized as follows. In Section 2, we introduce some elementary facts. In Section 3, we present the definition of eigenvalues perturbation, and use this definition to derive generalization error bounds for KRR and SVM. In Section 4, we propose the kernel selection criteria by minimizing the derived generalization error bounds, and present a method to compute the eigenvalues of integral operator. In Section 5 we empirically analyze the performance of our proposed kernel selection criteria compared with other popular criteria. We end in Section 6 with conclusion.

2. NOTATIONS AND PRELIMINARIES

Given a training set

$$S = \{z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^m$$

of size m drawn identically and independently distributed from a fixed, but unknown probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}$ for regression, and $\mathcal{Y} \subseteq \{+1, -1\}$ for classification.

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel, that is, K is symmetric and for any finite set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \subset \mathcal{X}$, the kernel matrix

$$\boldsymbol{K} = \left[\frac{1}{m} K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_{i,j=1}^m$$

is positive semidefinite. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined to be the completion of the linear span of the set of functions

$$\{K_{\boldsymbol{x}} = K(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}$$

with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying

$$\langle K_{\boldsymbol{x}}, K_{\boldsymbol{x}'} \rangle_K = K(\boldsymbol{x}, \boldsymbol{x}').$$

The kernel matrix $\boldsymbol{K}$ can be interpreted as an empirical version of the continuous integral operator $L_K : L_\rho^2(\mathcal{X}) \to L_\rho^2(\mathcal{X})$, which is defined by

$$(L_K f)(\boldsymbol{x}) = \int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{t}) f(\boldsymbol{t}) d\rho_{\mathcal{X}}(\boldsymbol{t}), \tag{1}$$

where $\rho_{\mathcal{X}}$ is the marginal distribution of ρ on $\mathcal{X}$ and $L_\rho^2(\mathcal{X})$ is the square-integrable space with respect to $\rho_{\mathcal{X}}$. This is a self-adjoint, compact operator that has eigenvalues

$$\lambda_1(L_K) \geq \lambda_2(L_K) \geq \ldots \geq \lambda_i(L_K) \geq \ldots \geq 0.$$

Since the kernel matrix $\boldsymbol{K}$ is positive semidefinite, its eigenvalues satisfy

$$\lambda_1(\boldsymbol{K}) \geq \lambda_2(\boldsymbol{K}) \geq \ldots \geq \lambda_m(\boldsymbol{K}) \geq 0.$$

The eigenvalue $\lambda_i(\boldsymbol{K})$ converges to the eigenvalue $\lambda_i(L_K)$ as the number of samples tends to infinity [5, 20, 28].

The learning algorithms we study here are the regularized algorithms [16]:

$$f_S := \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m \ell(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|_K^2 \right\}, \tag{2}$$

where $\ell(\cdot, \cdot)$ is a loss function, $\|f\|_K^2$ is the norm in RKHS and λ is the regularized parameter. Kernel Ridge Regression (KRR) [17] and Support Vector Machine (SVM) [31, 10, 30] are only different in the choice of loss function. For KRR

$$\ell(f(\boldsymbol{x}), y) = (y - f(\boldsymbol{x}))^2,$$

and for SVM

$$\ell(f(\boldsymbol{x}), y) = \max(0, 1 - yf(\boldsymbol{x})).$$

We will consider measuring the performance of the regularized algorithms. The main quantity we are interested in is the *risk* or *generalization error* which is a random variable depending on the training set S and is defined as

$$R(S) = \mathbb{E}_z[\ell(f_S(\boldsymbol{x}), y)],$$

where $\mathbb{E}_z[\cdot]$ is the expectation when $z = (\boldsymbol{x}, y)$ is sampled according to ρ. Unfortunately, $R(S)$ can't be computed since ρ is unknown. Thus, we estimate it using the empirical error $R_{\text{emp}}(S)$ defined as

$$R_{emp}(S) = \frac{1}{m} \sum_{i=1}^m \ell(f_S(\boldsymbol{x}_i), y_i).$$

We will bound the deviation between the empirical error and generalization error based on the eigenvalues perturbation which is defined in the next section.

3. GENERALIZATION ERROR BOUNDS WITH EIGENVALUES PERTURBATION

In this section, we first give the definition of eigenvalues perturbation and then use this definition to derive generalization error bounds for KRR and SVM.

3.1 Eigenvalues Perturbation

The way of defining of eigenvalues perturbation is to start from the goal: to get bounds on the generalization error and want these bounds to be tight when the kernel function satisfies the eigenvalues perturbation. In the following, we assume $\sup_{\boldsymbol{x} \in \mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}) = \kappa$ and $\forall y \in \mathcal{Y}, |y| \leq M$.

Because the kernel matrix contains most of the information needed by the regularized algorithms, and its eigenvalues converge to the eigenvalues of integral operator. Therefore we introduce the notion of eigenvalues perturbation, which quantifies the difference between the eigenvalues of kernel matrix and those of integral operator.

DEFINITION 1 (EIGENVALUES PERTURBATION). *The kernel function K is β eigenvalues perturbation if the following holds:*

$$\forall S \in \mathcal{Z}^m, \forall i \in \{1, \ldots, m\}, |\lambda_i(\boldsymbol{K}) - \lambda_i(L_K)| \leq \beta,$$

where $\boldsymbol{K}$ is the kernel matrix, $[\boldsymbol{K}]_{i,j} = \frac{1}{m}K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, L_K is the integral operator defined in (1), $\lambda_i(\boldsymbol{K})$ and $\lambda_i(L_K)$ are the eigenvalues of $\boldsymbol{K}$ and L_K, respectively.

The eigenvalues perturbation is defined on the kernel matrix and integral operator, therefore, if we obtain the eigenvalues of integral operator (the eigenvalues of integral operator induced by the popular radial kernels are given in Theorem 5), we can estimate its value for any given kernel function from empirical data, which makes it able to be used for kernel selection. In the next, we will show that the eigenvalues perturbation can yield upper bounds of generalization error for KRR and SVM.

3.2 Kernel Ridge Regression

KRR has successfully been applied to solve regression problems, which is a special case of the regularized algorithms when the loss function $\ell(f(\boldsymbol{x}), y) = (f(\boldsymbol{x}) - y)^2$. For KRR, the generalization error $R(S) = \mathbb{E}_z(f(\boldsymbol{x}) - y)^2$ and the empirical error $R_{\text{emp}}(S) = \frac{1}{m}\sum_{i=1}^m (f(\boldsymbol{x}_i) - y_i)^2$.

THEOREM 1. *If the kernel function K is β eigenvalues perturbation, then for the KRR, with probability $1-\delta$, we have*

$$R(S) \leq R_{\text{emp}}(S) + \sqrt{\frac{P^2 + 24Pm(C\beta + Q) + Pm(C\beta + Q)^2}{2m\delta}},$$

where $C = \frac{2\kappa M}{\lambda}$, $Q = \frac{2\kappa}{m-1}$ and $P = \frac{2\kappa^2 M^2}{\lambda^2} + 2M^2$.

The proof of this theorem is given in Appendix.A.

This theorem shows that small $R_{\text{emp}}(S)$ and β can guarantee good generalization performance for KRR.

Next, we also give a better exponential generalization error bound based on concentration inequalities.

THEOREM 2. *If the kernel function K is β spectral perturbation stability, then for the KRR, with probability $1-\delta$, we have*

$$R(S) \leq R_{emp}(S) + 4M(C\beta + Q) + 2M(C\beta + Q)^2 + \left(8mM(C\beta + Q) + 4m(C\beta + Q)^2 + P\right)\sqrt{\frac{\ln 1/\delta}{2m}},$$

where $C = \frac{2\kappa M}{\lambda}$, $Q = \frac{2\kappa}{m-1}$ and $P = \frac{2\kappa^2 M^2}{\lambda^2} + 2M^2$.

The proof of this theorem is given in Appendix.B.

3.3 Support Vector Machine

SVM has successfully been applied to solve classification problems, its loss function is the hinge loss $\ell(f(\boldsymbol{x}), y) = \max(0, 1 - yf(\boldsymbol{x}))$. For SVM, $R(S) = \mathbb{E}_z \max(0, 1 - yf(\boldsymbol{x}))$ and $R_{\text{emp}}(S) = \frac{1}{m}\sum_{i=1}^m \max(0, 1 - y_i f(\boldsymbol{x}_i))$.

THEOREM 3. *If the kernel function K is β eigenvalues perturbation, then for the SVM, with probability $1-\delta$,*

$$R(S) \leq R_{\text{emp}}(S) + \sqrt{\frac{Q^2 + 12Qm\beta^{\frac{1}{4}}(1 + (\beta/(2\kappa))^{\frac{1}{4}})}{2m\delta}},$$

where $C = 8^{\frac{1}{4}}\lambda\kappa^{\frac{3}{4}}$, $Q = 1 + M\lambda\kappa$.

The proof of this theorem is given in Appendix.C.

This theorem shows that, to guarantee good generalization performance for SVM, we should select the kernel function which has small $R_{\text{emp}}(S)$ and β.

In the next, we give a better exponential generalization error bound.

THEOREM 4. *If the kernel function K is β eigenvalues perturbation, then for the SVM, with probability $1-\delta$,*

$$R(S) \leq R_{emp}(S) + 2C\beta^{\frac{1}{4}}\left(1 + \left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right) + \left(4m\beta^{\frac{1}{4}}\left(1 + \left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right) + Q\right)\sqrt{\frac{\ln 1/\delta}{2m}},$$

where $C = 8^{\frac{1}{4}}\lambda\kappa^{\frac{3}{4}}$ and $Q = 1 + M\lambda\kappa$.

The proof of this theorem is given in Appendix.D.

The bounds we obtain for SVM are weaker than our bounds for KRR. This is due mainly to the different loss functions defining the optimization problems of these algorithms.

REMARK 1. *In this paper, although we only consider the KRR and SVM algorithms, the above results can be easy to be extended to other kernel-based methods, such as the kernel-based logistic regression, least squares Support Vector Machines (LSSVM) [32].*

4. KERNEL SELECTION CRITERIA

By the generalization error bounds in Theorem 1 and 2 for KRR, and Theorem 3 and 4 for SVM, to guarantee good generalization performance, we should select the kernel by minimizing $\sum_{i=1}^m (f_S(\boldsymbol{x}_i) - y_i)^2 + \delta\beta$ for KRR and minimizing $\sum_{i=1}^m \max(0, 1 - y_i f_S(\boldsymbol{x}_i)) + \delta\beta$ for SVM, where $\delta > 0$ is the regularization coefficient. However, by the definition of eigenvalues perturbation, we should try all the possibilities of the set S ($\forall S \in \mathcal{Z}^m$) to compute the β, which is infeasible in practice. We should estimate it from the available empirical data. Therefore, we consider using the following **eigenvalues perturbation** criteria:
For KRR,

$$EPKRR(K) = \frac{1}{m}\sum_{i=1}^m (y_i - f_S(\boldsymbol{x}))^2 + \delta\sum_{i=1}^m |\lambda_i(\boldsymbol{K}) - \lambda_i(L_K)|,$$

and for SVM,

$$EPSVM(K) = \frac{1}{m}\sum_{i=1}^m \max(0, 1 - y_i f_S(\boldsymbol{x}_i)) + \delta\sum_{i=1}^m |\lambda_i(\boldsymbol{K}) - \lambda_i(L_K)|.$$

In order to use these criteria for kernel selection, we should compute the eigenvalues of integral operator. In the following, we will present the method to compute the eigenvalues of integral operator induced by the popular radial kernels, such as Gaussian kernel and Laplacian kernel.

THEOREM 5. *Assuming the radial kernel $K(\boldsymbol{x}-\boldsymbol{x}')$ is defined on $[-M/2, M/2]^d$, $M > 0$, d is the dimension of the input data, then the eigenvalues of integral operator induced by the radial kernel $K(\boldsymbol{x}-\boldsymbol{x}')$ are*

$$F(\boldsymbol{n}) = \prod_{i=1}^{d} F(n_i),$$

where $\boldsymbol{n} = (n_1, \ldots, n_d)$, $n_i \in \mathbb{N} \cup \{0\}$,

$$F(n_i) = \int_{-M/2}^{M/2} K(t) \cos n_i w t \mathrm{d}t, w = 2\pi/M.$$

PROOF. We assume $d = 1$. A generalization to multidimensional kernels ($d > 1$) is straightforward. Since the radial kernel $K(x-x')$ is an even function defined on $[-M/2, M/2]$, therefore, by the Fourier transform, we have

$$K(x - x') = a_0 + \sum_{n=1}^{\infty} a_n \cos nw(x - x'),$$

where $a_0 = \frac{1}{M}F(0)$, $a_n = \frac{2}{M}F(n)$, $n = 1, 2, \ldots, \infty$. By the definition of integral operator (see (1)), we have

$$L_K(\sqrt{\frac{2}{M}} \sin nwx) = \int_{-M/2}^{M/2} K(x-t)\sqrt{\frac{2}{M}} \sin nwt \mathrm{d}t$$
$$= \int_{-M/2}^{M/2} \left(a_0 + \sum_{n=1}^{\infty} a_n \cos nw(x-t)\right) \sqrt{\frac{2}{M}} \sin nwt \mathrm{d}t.$$

Note that $\cos nw(x-t) = \cos nwx \cos nwt + \sin nwx \sin nwt$, and

$$\frac{1}{\sqrt{M}}, \sqrt{\frac{2}{M}} \sin(wx), \sqrt{\frac{2}{M}} \cos(wx), \ldots,$$
$$\sqrt{\frac{2}{M}} \sin(wnx), \sqrt{\frac{2}{M}} \cos(wnx), \ldots$$

is a standard orthogonal basis of the square integrable space $L^2_{[-M/2,M/2]}$. Thus, it is easy to verify that

$$\int_{-M/2}^{M/2} \left(a_0 + \sum_{n=1}^{\infty} a_n \cos nw(x-t)\right) \sqrt{\frac{2}{M}} \sin nwt \mathrm{d}t$$
$$= F(n)\sqrt{\frac{2}{M}} \sin nwx.$$

In the same way, we can obtain that

$$L_K(\sqrt{\frac{2}{M}} \cos nwx) = F(n)\sqrt{\frac{2}{M}} \cos nwx, n = 1, \ldots$$
$$L_K(\sqrt{\frac{1}{M}}) = F(0)\sqrt{\frac{1}{M}}.$$

Therefore, $F(n)$ is the eigenvalue of the integral operator L_K, its associate eigenfunction is $\sqrt{\frac{2}{M}} \sin nwx$ or $\sqrt{\frac{2}{M}} \cos nwx$, $n = 1, \ldots$. $F(0)$ is the eigenvalue of L_K, its associate eigenfunction is $\sqrt{\frac{1}{M}}$. □

Table 1: Radial kernels and eigenvalues of their corresponding integral operators.

Radial Kernel	$K(\boldsymbol{x} - \boldsymbol{x}')$	$F(\boldsymbol{n})$ $(M = \infty)$
Gaussian Kernel	$\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2}\right)$	$(2\pi)^{-\frac{d}{2}} \exp\left(-\frac{\|\boldsymbol{n}\|_2^2}{2}\right)$
Laplacian Kernel	$\exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|_1\right)$	$\prod_d \frac{1}{\pi(1+n_d^2)}$

By the above theorem, when the radial kernel $K(\boldsymbol{x}-\boldsymbol{x}')$ is defined on $\mathbb{R}^d$, that is $M = \infty$, it is easy to obtain the closed form of eigenvalues of the integral operators induced by the Gaussian kernel and Laplacian kernel. The close forms of Gaussian kernel and Laplacian kernel are given in Table 1.

REMARK 2. *Instead of choosing a single kernel, some researchers consider combining multiple kernels by some criteria, called multiple kernel learning (MKL), see, e.g., [22, 1, 27, 11, 21] and the references therein. Our criteria can be used for MKL. However, in this paper, we mainly want to verify the effectiveness of our eigenvalues perturbation criteria. How to utilize our proposed criteria for MKL is beyond the scope of this article.*

5. EXPERIMENTS

In this section, we will empirically analyze the performance of our proposed EPSVM and EPKRR criteria.

The evaluation is made on 15 publicly available data sets from UCI repository[1] and LIBSVM Data[2]: 8 data sets for classification seen in Table 2, and 7 data sets for regression seen in Table 3. All data sets are normalized to have zero-means and unit-variances on every attribute to avoid numerical problems caused by large value kernel matrices.

We use Gaussian kernels

$$K_{\text{Gauss}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{\tau}\right)$$

as our candidate kernels, $\tau \in \{2^i, i = -8, -9, \ldots, 5, 6\}$. For each data set, we have run all the methods 30 times with random partition of the datasets (50% of all the examples for training and the other 50% for testing).

5.1 Classification

We will compare our proposed EPSVM criterion with five popular classification criteria: centered kernel target alignment (CKTA) [12], feature space-based kernel matrix evaluation (FSM) [25], K-fold cross validation, $K = 3, 5, 10$. The learning algorithm we use here is the SVM.

In the first experiment, we set the regularization coefficient $\delta = 100$ (the parameter of the EPSVM criterion, we will explore the effect of this parameter in the next experiment). The average test errors with standard deviations are reported in Table 2. The elements in this table are obtained as follows. For each training set, each regularized parameter λ ($\lambda \in \{10^i, i = -3, -2, 1\}$), we choose the kernel by each kernel selection criterion on the training set, and evaluate the test error for the chosen parameters on testing set. Then we compute the means over all runs on the different partitions. The results in Table 2 can be summarized as follows: (a) EPSVM gives the best results on most data

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html
[2] http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Table 2: The test errors (%) with standard deviations ($\delta = 100$). For each training set, each regularized parameter λ, we choose the kernel by each kernel selection criterion on the training set, and evaluate the test error for the chosen parameters on the testing set. Then we compute the means over all runs on the different partitions.

$\lambda = 0.001$						
Method	EPSVM	CKTA	FSM	3-CV	5-CV	10-CV
australian	14.01 ± 2.66	14.97 ± 1.33	20.17 ± 1.65	14.03 ± 1.04	**13.97 ± 1.20**	14.03 ± 1.04
heart	**15.33 ± 3.33**	15.56 ± 2.92	25.33 ± 3.33	20.00 ± 3.47	21.19 ± 5.70	17.33 ± 2.65
ionosphere	11.23 ± 8.21	19.94 ± 1.43	29.26 ± 1.82	**10.29 ± 1.34**	10.40 ± 0.75	11.54 ± 1.69
breast-cancer	3.58 ± 0.38	**2.58 ± 0.56**	3.05 ± 0.71	3.87 ± 0.96	3.23 ± 0.41	2.93 ± 0.66
diabetes	25.16 ± 1.25	25.52 ± 1.22	31.30 ± 1.51	24.53 ± 1.24	24.27 ± 1.44	**24.27 ± 1.44**
german.numer	**26.40 ± 6.23**	27.68 ± 1.11	45.00 ± 1.12	29.84 ± 3.79	27.12 ± 2.63	28.12 ± 4.67
liver-disorders	38.93 ± 3.12	39.77 ± 2.62	40.93 ± 3.12	38.60 ± 3.69	37.09 ± 3.80	**35.93 ± 5.98**
sonar	**13.23 ± 5.04**	14.23 ± 3.93	21.15 ± 6.04	15.00 ± 3.76	14.62 ± 3.56	13.65 ± 3.75
$\lambda = 0.01$						
Method	EPSVM	CKTA	FSM	3-CV	5-CV	10-CV
australian	**12.38 ± 0.86**	13.65 ± 1.13	20.35 ± 2.08	13.80 ± 1.21	13.68 ± 1.13	13.80 ± 1.21
heart	19.56 ± 1.86	16.30 ± 1.74	24.00 ± 2.94	17.04 ± 2.46	16.15 ± 1.99	**16.13 ± 1.69**
ionosphere	14.74 ± 3.97	11.54 ± 2.12	31.77 ± 2.54	12.11 ± 3.17	12.00 ± 1.46	**10.51 ± 1.79**
breast-cancer	**2.28 ± 0.52**	2.99 ± 0.64	2.82 ± 0.87	3.52 ± 1.04	3.99 ± 0.71	3.99 ± 0.71
diabetes	24.22 ± 1.67	**23.28 ± 1.56**	29.48 ± 2.21	26.98 ± 3.53	28.75 ± 5.29	29.32 ± 4.39
german.numer	**25.84 ± 2.84**	26.44 ± 2.44	44.64 ± 0.80	29.56 ± 1.58	30.04 ± 1.65	29.36 ± 2.97
liver-disorders	39.42 ± 4.06	38.84 ± 2.95	39.42 ± 3.97	40.12 ± 1.70	**37.91 ± 5.11**	38.95 ± 5.71
sonar	**16.12 ± 2.39**	16.92 ± 2.60	18.08 ± 4.73	18.08 ± 2.58	16.92 ± 3.09	17.88 ± 1.61
$\lambda = 0.1$						
Method	EPSVM	CKTA	FSM	3-CV	5-CV	10-CV
australian	**13.51 ± 1.38**	17.91 ± 1.25	19.25 ± 1.08	16.64 ± 1.31	16.46 ± 4.15	17.10 ± 3.97
heart	**18.96 ± 3.08**	24.59 ± 11.04	24.74 ± 3.08	25.04 ± 10.65	20.30 ± 5.20	19.70 ± 4.37
ionosphere	22.40 ± 7.08	**20.00 ± 3.08**	29.94 ± 4.17	22.17 ± 5.13	22.40 ± 4.70	23.77 ± 6.98
breast-cancer	4.11 ± 0.36	**3.40 ± 0.26**	3.11 ± 0.39	5.81 ± 4.66	3.58 ± 0.91	3.64 ± 1.13
diabetes	30.47 ± 7.45	**25.57 ± 2.32**	28.85 ± 2.63	31.25 ± 3.01	33.85 ± 2.92	32.97 ± 2.26
german.numer	**29.52 ± 1.29**	29.84 ± 1.26	45.24 ± 1.17	29.84 ± 1.26	29.76 ± 1.16	29.76 ± 1.16
liver-disorders	40.58 ± 2.07	41.98 ± 1.99	42.67 ± 3.41	40.35 ± 4.64	40.70 ± 4.98	**40.12 ± 4.81**
sonar	**21.35 ± 7.94**	23.08 ± 7.54	21.92 ± 3.43	22.88 ± 9.06	22.50 ± 8.56	23.46 ± 9.82
$\lambda = 1$						
Method	EPSVM	CKTA	FSM	3-CV	5-CV	10-CV
australian	**16.41 ± 2.54**	44.70 ± 2.23	20.81 ± 0.72	22.84 ± 13.26	16.75 ± 1.89	28.06 ± 15.07
heart	**19.56 ± 3.42**	46.81 ± 6.78	24.59 ± 2.31	30.52 ± 11.24	36.89 ± 10.75	29.19 ± 10.42
ionosphere	28.77 ± 1.74	30.06 ± 3.01	30.40 ± 1.36	28.34 ± 10.91	27.89 ± 10.15	**26.74 ± 12.52**
breast-cancer	5.02 ± 2.55	5.87 ± 1.57	4.52 ± 1.38	6.69 ± 5.58	6.63 ± 5.62	**4.16 ± 0.67**
diabetes	35.68 ± 1.73	34.01 ± 3.00	**29.53 ± 2.21**	35.78 ± 1.85	35.73 ± 1.82	35.63 ± 1.72
german.numer	**30.48 ± 0.88**	30.48 ± 0.88	43.96 ± 1.30	30.48 ± 0.88	30.48 ± 0.88	30.48 ± 0.88
liver-disorders	**38.58 ± 3.52**	40.00 ± 4.47	39.56 ± 3.55	41.28 ± 1.79	40.58 ± 2.71	40.35 ± 3.17
sonar	**20.26 ± 4.89**	26.35 ± 11.35	21.54 ± 3.82	18.46 ± 2.67	20.58 ± 3.16	21.92 ± 7.30
$\lambda = 10$						
Method	EPSVM	CKTA	FSM	3-CV	5-CV	10-CV
australian	**16.75 ± 2.37**	44.81 ± 1.96	18.84 ± 1.88	21.86 ± 12.65	27.01 ± 16.24	17.45 ± 2.33
heart	**16.59 ± 3.34**	43.26 ± 2.00	26.07 ± 2.37	18.96 ± 3.65	18.52 ± 5.64	22.81 ± 11.86
ionosphere	16.11 ± 1.69	33.03 ± 4.75	29.94 ± 1.18	**14.40 ± 2.81**	14.63 ± 2.79	14.51 ± 2.73
breast-cancer	6.40 ± 4.74	12.43 ± 2.12	6.51 ± 1.49	4.57 ± 1.18	5.34 ± 1.98	**4.28 ± 1.22**
diabetes	35.42 ± 1.64	34.43 ± 2.49	**29.53 ± 0.84**	33.70 ± 1.61	34.11 ± 2.17	34.06 ± 3.47
german.numer	**28.34 ± 1.45**	29.88 ± 1.45	46.40 ± 1.11	28.80 ± 1.88	29.68 ± 1.88	29.64 ± 1.31
liver-disorders	44.88 ± 2.30	39.19 ± 4.94	**38.37 ± 4.85**	44.19 ± 4.48	42.44 ± 2.33	44.88 ± 2.30
sonar	**17.58 ± 3.84**	26.35 ± 8.26	19.81 ± 4.79	17.88 ± 3.16	17.50 ± 3.87	18.12 ± 5.28

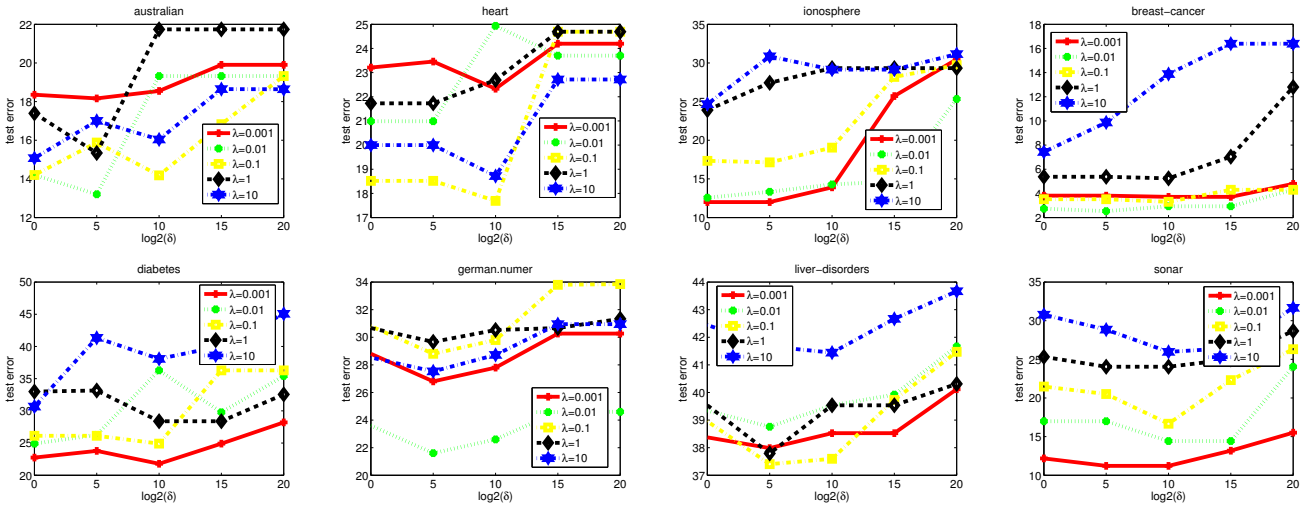

Figure 1: The average test errors using EPSVM with different δ. For each training set, each regularized parameter λ, each δ, we choose the kernel by each kernel selection criterion on the training set, and evaluate the test errors for the chosen parameters on testing set.

Table 3: The test mean square error (TMSE) with standard deviations ($\delta = 100$). For each training set, each λ, we choose the kernel by each kernel selection criterion on the training set, and evaluate the test error for the chosen parameters on the testing set.

Method	EPKRR	3-CV	5-CV	10-CV	LOO	GCV
			$\lambda = 0.0001$			
housing	23.90 ± 1.68	**22.29 ± 2.10**	22.50 ± 2.16	23.31 ± 1.38	23.31 ± 1.38	23.31 ± 1.38
mpg	10.26 ± 1.83	8.73 ± 0.68	**8.28 ± 0.97**	8.58 ± 0.94	8.81 ± 1.30	10.96 ± 1.08
pyrim	**4.40e-3 ± 3.47e-4**	8.48e-3 ± 3.37e-3	9.92e-3 ± 5.26e-3	1.24e-2 ± 6.54e-3	1.01e-2 ± 4.99e-3	2.72e-1 ± 1.68e-1
triazines	2.36e-2 ± 2.58e-2	**2.17e-2 ± 7.02e-3**	2.22e-2 ± 6.68e-3	2.22e-2 ± 6.80e-3	2.19e-2 ± 6.89e-3	2.09e-2 ± 5.79e-3
eunite2001	**418 ± 52.34**	422.72 ± 59.27	438.05 ± 46.20	447.78 ± 36.26	447.78 ± 36.26	447.78 ± 36.26
space-ga	1.58e-2 ± 3.22e-3	1.52e-2 ± 1.76e-3	1.48e-2 ± 1.05e-3	1.42e-2 ± 1.31e-3	**1.42e-2 ± 1.31e-3**	2.22e-2 ± 1.41e-3
cpusmall	37.61 ± 8.13	36.56 ± 17.90	35.54 ± 4.20	35.47 ± 15.30	**35.41 ± 15.10**	112.80 ± 9.85
			$\lambda = 0.001$			
housing	22.42 ± 21.92	23.29 ± 3.30	22.76 ± 2.59	20.34 ± 3.18	**19.06 ± 0.68**	28.24 ± 0.70
mpg	**9.85 ± 3.49**	10.20 ± 1.87	10.53 ± 1.23	9.52 ± 0.84	10.21 ± 0.91	13.91 ± 0.66
pyrim	6.33e-3 ± 1.41e-3	6.96e-3 ± 3.55e-3	6.34e-3 ± 2.65e-3	**6.09e-3 ± 2.03e-3**	7.43e-3 ± 2.01e-3	6.30e-2 ± 7.65e-2
triazines	**2.22e-2 ± 1.15e-3**	2.43e-2 ± 3.61e-3	2.36e-2 ± 2.93e-3	2.33e-2 ± 2.40e-3	2.51e-2 ± 4.90e-3	2.36e-2 ± 2.94e-3
eunite2001	**700.49 ± 118.49**	1.14e+3 ± 203.32	1.09e+3 ± 140.67	1.03e+3 ± 124.18	1.21e+3 ± 173.31	1.21e+3 ± 173.31
space-ga	1.75e-2 ± 3.90e-3	**1.74e-2 ± 2.91e-3**	1.71e-2 ± 2.79e-3	1.70e-2 ± 3.04e-3	1.68e-2 ± 2.97e-3	2.16e-2 ± 2.87e-3
cpusmall	42.09 ± 13.19	48.33 ± 13.01	41.72 ± 3.70	**38.68 ± 8.75**	38.68 ± 8.75	164.85 ± 24.84
			$\lambda = 0.01$			
housing	33.32 ± 6.40	35.56 ± 6.83	33.90 ± 7.14	33.37 ± 4.87	**32.81 ± 5.39**	47.00 ± 6.35
mpg	**15.48 ± 2.27**	16.17 ± 0.89	15.49 ± 1.64	17.81 ± 3.71	16.17 ± 2.36	21.80 ± 1.60
pyrim	1.23e-2 ± 4.00e-3	9.70e-3 ± 2.84e-3	1.03e-2 ± 4.63e-3	1.03e-2 ± 3.95e-3	**8.94e-3 ± 2.03e-3**	1.23e-2 ± 4.00e-3
triazines	2.78e-2 ± 3.52e-3	2.21e-2 ± 3.36e-3	2.35e-2 ± 4.51e-3	2.35e-2 ± 4.44e-3	**2.21e-2 ± 3.36e-3**	2.24e-2 ± 3.38e-3
eunite2001	2.17e+3 ± 99.41	**1.81e+3 ± 146.48**	1.83e+3 ± 132.19	1.83e+3 ± 154.29	1.83e+3 ± 151.97	2.17e+3 ± 99.41
space-ga	**2.07e-2 ± 7.94e-4**	2.16e-2 ± 2.36e-3	2.16e-2 ± 2.35e-3	2.32e-2 ± 3.11e-3	2.25e-2 ± 1.98e-3	2.24e-2 ± 1.77e-3
cpusmall	**77.63 ± 10.45**	79.51 ± 4.17	79.51 ± 4.17	79.51 ± 4.17	79.51 ± 4.17	204.56 ± 8.33
			$\lambda = 0.1$			
housing	75.36 ± 9.93	78.44 ± 23.11	88.96 ± 12.56	81.28 ± 19.66	81.28 ± 19.66	**69.15 ± 11.36**
mpg	52.95 ± 9.13	47.01 ± 18.01	48.45 ± 16.61	49.71 ± 15.68	55.47 ± 15.53	**46.72 ± 9.23**
pyrim	2.05e-2 ± 2.93e-3	1.97e-2 ± 4.05e-3	1.99e-2 ± 3.83e-3	1.98e-2 ± 3.48e-3	**1.95e-2 ± 3.62e-3**	2.05e-2 ± 2.93e-3
triazines	**3.27e-2 ± 3.32e-3**	4.35e-2 ± 1.02e-2	3.71e-2 ± 8.05e-3	3.63e-2 ± 5.99e-3	3.59e-2 ± 7.11e-3	3.27e-2 ± 3.32e-3
eunite2001	**6.88e+3 ± 648.63**	1.08e+4 ± 2.44e+3	9.85e+3 ± 639.14	1.044e+4 ± 2.83e+3	9.85e+3 ± 639.14	7.03e+3 ± 648.02
space-ga	4.15e-2 ± 3.65e-3	6.64e-2 ± 1.18e-2	7.07e-2 ± 5.10e-3	6.66e-2 ± 1.30e-2	7.07e-2 ± 5.10e-3	**3.59e-2 ± 5.75e-3**
cpusmall	**219.33 ± 11.14**	253.71 ± 14.46	269.47 ± 21.83	269.47 ± 21.83	269.47 ± 21.83	285.18 ± 18.65
			$\lambda = 1$			
housing	**218.29 ± 19.64**	282.97 ± 40.69	305.13 ± 49.95	282.92 ± 36.71	300.15 ± 22.14	256.51 ± 21.07
mpg	**190.34 ± 14.83**	272.45 ± 27.25	272.82 ± 44.78	272.45 ± 27.25	272.82 ± 44.78	249.64 ± 15.61
pyrim	**1.31e-1 ± 1.25e-2**	1.50e-1 ± 2.00e-2	1.57e-1 ± 1.71e-2	1.37e-1 ± 7.12e-3	1.52e-1 ± 2.12e-2	1.33e-1 ± 1.70e-2
triazines	1.61e-1 ± 4.63e-3	1.78e-1 ± 1.89e-2	1.67e-1 ± 1.60e-2	1.67e-1 ± 1.37e-2	1.67e-1 ± 1.37e-2	**1.59e-1 ± 4.63e-3**
eunite2001	**1.43e+4 ± 2.23e+3**	1.73e+4 ± 2.21e+3	1.73e+4 ± 2.21e+3	1.73e+4 ± 2.21e+3	1.73e+4 ± 2.21e+3	1.57e+4 ± 2.22e+3
space-ga	1.22e-1 ± 1.79e-3	2.03e-1 ± 1.87e-2	2.30e-1 ± 2.10e-2	2.39e-1 ± 2.03e-3	2.39e-1 ± 2.03e-3	**1.94e-1 ± 1.90e-3**
cpusmall	**2.14e+3 ± 38.95**	2.86e+3 ± 41.74	2.63e+3 ± 191.34	2.86e+3 ± 41.74	2.78e+3 ± 171.94	2.47e+3 ± 39.96

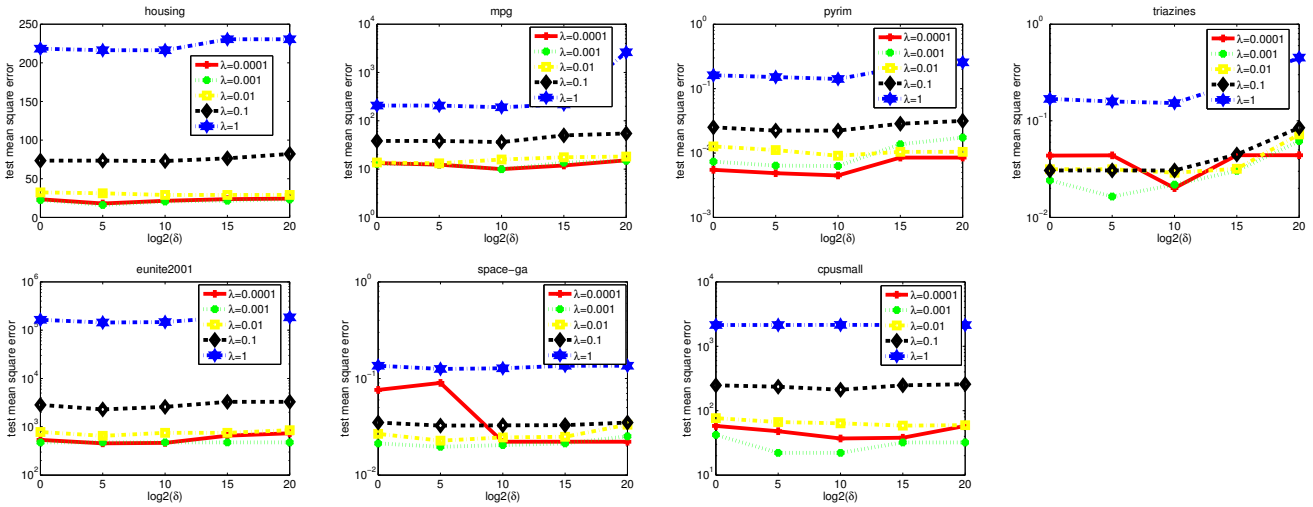

Figure 2: The average test mean square errors using EPKRR with different δ . For each training set, each regularized parameter λ, each δ, we choose the kernel by each kernel selection criterion on the training set, and evaluate the test errors for the chosen parameters on testing set.

sets. In particular, for each λ, EPSVM outperforms FSM on 6 (or more) out of 8 sets, and outperforms 3-CV, 5-CV and 10-CV on 5 (or more) out of 8 sets. Thus it implicates that choosing the kernel based on the eigenvalues perturbation can guarantee good generalization. (b) The test errors of EPSVM are more stable than these of other five criteria when changing the value of λ, that is the fluctuation of the test errors of the EPSVM are smaller than those of other methods when changing the value of λ. This property may bring some advantages in practical application.

In the next experiment, we will explore the effect of the regularization coefficient δ for EPSVM. The average test errors with different δ are given in Figure 1. We find that the optimal δ belong to $[2^5, 2^{10}]$ on most data sets. Therefore, the range of δ should be set between 2^5 and 2^{10}.

5.2 Regression

We will compare EPKRR criterion with five popular regression criteria: 3-CV, 5-CV, 10-CV, generalized cross-validation (GCV) [19] and leave-one-out (LOO) [3]. The learning algorithm we use here is the KRR.

The test mean square errors (TMSE) with standard deviations are reported in Table 3. In this experiment, the parameter $\delta = 100$. The results in Table 3 can be summarized as follows: (a) EPKRR criterion is much better than GCV on the nearly all data sets. In particular, for $\lambda \in \{0.0001, 0.001, 0.01\}$, EPKRR outperforms GCV on 6 out of 7 sets, and also gives the close result on the remaining 1 sets. (b) EPKRR criterion is comparable or better than 3-CV, 5-CV, 10-CV and LOO on most data sets. So it implicates that the EPKRR criterion is sound and effective.

Finally, we will explore the effect of the parameter δ for EPKRR. The TMSE with different δ are given in Figure 2. It is interesting to note that the TMSE does not quite depend on δ. Thus, we can set δ to be a constant for simplicity in practice (In Figure 2, we can find that the $\delta = 2^{10}$ is a reasonable choice).

[3]we only need to solve the KRR once to compute LOO [6].

6. CONCLUSION

In this paper, we introduce two new kernel selection criteria for KRR and SVM based on the eigenvalues perturbation of integral operator, which quantifies the difference between the eigenvalues of kernel matrix and these of integral operator. These criteria are theoretically justified and show good results in practice. We believe that our analysis opens new perspectives on the application of the integral operator to practical problems.

Future work will extend these criteria to other kernel based methods (such as kernel-based logistic regression, least squares Support Vector Machines), and use these criteria for multiple kernel learning.

Acknowledgments

The work is supported in part by the National Natural Science Foundation of China under grant No. 61170019, the Natural Science Foundation of Tianjin under grant No. 11JCYBJC00700, and Tianjin Key Laboratory of Cognitive Computing and Application.

Appendix.A Proof of Theorem 1

For each $i \in \{1, \ldots, m\}$, the removed training set is defined as follows:

$$S^i = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_m\}, \quad z = (\boldsymbol{x}, y).$$

In order to prove Theorem 1, we first prove the following theorem:

THEOREM 6. *If the kernel function K is β eigenvalues perturbation, then for the KRR, we have*

$$\forall S \in \mathcal{Z}^m, \forall i \in \{1, \ldots, m\}, |f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \leq C\beta + Q,$$

where $C = \frac{2kM}{\lambda}$ *and* $Q = \frac{2\kappa}{m-1}$.

PROOF. Denote vectors $\boldsymbol{k}$, $\boldsymbol{k}_i$, $\boldsymbol{y}$ and $\boldsymbol{y}_i$ as

$$\boldsymbol{k} = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_m))^{\mathrm{T}},$$
$$\boldsymbol{k}_i = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_{i-1}), K(\boldsymbol{x}, \boldsymbol{x}_{i+1}), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_m))^{\mathrm{T}},$$
$$\boldsymbol{y} = (y_1, \ldots, y_m)^{\mathrm{T}},$$
$$\boldsymbol{y}_i = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)^{\mathrm{T}},$$

respectively. Denote the $(m-1)\times(m-1)$ kernel matrix $\boldsymbol{K}_i$ with respect to the training set S^i as

$$[\boldsymbol{K}_i]_{j,k} = \frac{1}{m-1} K(\boldsymbol{x}_j, \boldsymbol{x}_k), \boldsymbol{x}_j, \boldsymbol{x}_k \in S^i.$$

According to [18], we know that the solutions of the KRR with respect to the training set S and S^i can be respectively written as

$$f_S(\boldsymbol{x}) = \frac{1}{m} \boldsymbol{k}^{\mathrm{T}} (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y},$$
$$f_{S^i}(\boldsymbol{x}) = \frac{1}{m-1} \boldsymbol{k}_i^{\mathrm{T}} (\boldsymbol{K}_i + \lambda \boldsymbol{I}_i)^{-1} \boldsymbol{y}_i,$$

where $\boldsymbol{I}$ and $\boldsymbol{I}_i$ are the $m \times m$ and $(m-1)\times(m-1)$ identity matrices, respectively. For each $i \in \{1, \ldots, m\}$, denote the $m \times m$ i-th removed kernel matrix as $\boldsymbol{K}^i$ with

$$\begin{cases} [\boldsymbol{K}^i]_{jk} = \dfrac{1}{m-1} K(\boldsymbol{x}_j, \boldsymbol{x}_k) & \text{if } j \text{ and } k \neq i, \\ [\boldsymbol{K}^i]_{jk} = 0 & \text{if } j \text{ or } k = i, \end{cases}$$

it is easy to verify that

$$\begin{aligned} f_{S^i}(\boldsymbol{x}) &= \frac{1}{m-1} \boldsymbol{k}_i^{\mathrm{T}} (\boldsymbol{K}_i + \lambda \boldsymbol{I}_i)^{-1} \boldsymbol{y}_i \\ &= \frac{1}{m-1} \boldsymbol{k}^{\mathrm{T}} \left((\boldsymbol{K}^i + \lambda \boldsymbol{I})^{-1} - \boldsymbol{A}_i \right) \boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{A}_i = \mathrm{diag}(0, \ldots, 0, \frac{1}{\lambda}, 0, \ldots, 0)$ is a diagonal matrix, with the i-th diagonal element $\frac{1}{\lambda}$, others 0. Therefore, we have

$$\begin{aligned} & f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x}) \\ &= \frac{\boldsymbol{k}^{\mathrm{T}}}{m} (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y} - \frac{\boldsymbol{k}^{\mathrm{T}}}{m-1} (\boldsymbol{K}^i + \lambda \boldsymbol{I})^{-1} \boldsymbol{y} \\ &\quad + \frac{1}{m-1} \boldsymbol{k}^{\mathrm{T}} \boldsymbol{A}_i \boldsymbol{y} \\ &= \frac{\boldsymbol{k}^{\mathrm{T}}}{m} \left((\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} - (\boldsymbol{K}^i + \lambda \boldsymbol{I})^{-1} \right) \boldsymbol{y} \\ &\quad - \frac{1}{m(m-1)} \boldsymbol{k}^{\mathrm{T}} (\boldsymbol{K}^i + \lambda I)^{-1} \boldsymbol{y} + \frac{1}{m-1} \boldsymbol{k}^{\mathrm{T}} \boldsymbol{A}_i \boldsymbol{y}. \end{aligned}$$

Since $\boldsymbol{M}'^{-1} - \boldsymbol{M}^{-1} = -\boldsymbol{M}'^{-1}(\boldsymbol{M}' - \boldsymbol{M})\boldsymbol{M}^{-1}$ is valid for any invertible matrices $\boldsymbol{M}$, $\boldsymbol{M}'$, so, we have

$$\begin{aligned} & (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} - (\boldsymbol{K}^i + \lambda \boldsymbol{I})^{-1} \\ &\qquad = -(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{K} - \boldsymbol{K}^i) (\boldsymbol{K}^i + \lambda \boldsymbol{I})^{-1}. \end{aligned}$$

Thus, we can obtain that

$$\begin{aligned} & \left\| \left((\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} - (\boldsymbol{K}^i + \lambda \boldsymbol{I})^{-1} \right) \boldsymbol{y} \right\| \\ &\leq \|(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\| \|\boldsymbol{K} - \boldsymbol{K}^i\| \|(\boldsymbol{K}^i + \lambda \boldsymbol{I})^{-1}\| \|\boldsymbol{y}\| \\ &\leq \frac{\|\boldsymbol{K} - \boldsymbol{K}^i\| \|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{K} + \lambda \boldsymbol{I}) \lambda_{\min}(\boldsymbol{K}^i + \lambda \boldsymbol{I})}, \end{aligned}$$

where $\lambda_{\min}(\boldsymbol{K} + \lambda \boldsymbol{I})$ is the smallest eigenvalue of $\boldsymbol{K} + \lambda \boldsymbol{I}$ and $\lambda_{\min}(\boldsymbol{K}^i + \lambda \boldsymbol{I})$ the smallest eigenvalue of $\boldsymbol{K}^i + \lambda \boldsymbol{I}$. Thus, we can obtain that

$$\begin{aligned} & |f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \\ &\leq \frac{\|\boldsymbol{k}\|}{m} \frac{\|\boldsymbol{K} - \boldsymbol{K}^i\| \|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{K} + \lambda \boldsymbol{I}) \lambda_{\min}(\boldsymbol{K}^i + \lambda \boldsymbol{I})} \\ &\quad + \frac{1}{m(m-1)} \frac{\|\boldsymbol{k}\| \|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{K}^i + \lambda \boldsymbol{I})} + \frac{|K(\boldsymbol{x}_i, \boldsymbol{x}_i) y_i|}{\lambda(m-1)}. \end{aligned}$$

Since the matrices $\boldsymbol{K}$ and $\boldsymbol{K}^i$ are positive semidefinite, so

$$\lambda_{\min}(\boldsymbol{K} + \lambda \boldsymbol{I}) \geq \lambda, \lambda_{\min}(\boldsymbol{K}^i + \lambda \boldsymbol{I}) \geq \lambda.$$

Note that $\|\boldsymbol{y}\| \leq \sqrt{m} M$ and $\|\boldsymbol{k}\| \leq \sqrt{m} \kappa$, so we have

$$|f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \leq \frac{\kappa M}{\lambda^2} \|\boldsymbol{K} - \boldsymbol{K}^i\| + \frac{\kappa M}{(m-1)\lambda} + \frac{\kappa M}{(m-1)\lambda}.$$

Denote the diagonal matrix $\boldsymbol{\Lambda}$ as

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1(L_K), \lambda_2(L_K), \ldots, \lambda_m(L_K)).$$

Therefore, we have

$$\|\boldsymbol{K} - \boldsymbol{K}^i\| \leq \|\boldsymbol{K} - \boldsymbol{\Lambda}\| + \|\boldsymbol{K}^i - \boldsymbol{\Lambda}\|.$$

By the definition of β-eigenvalues perturbation in Definition 1, it is easy to verify that

$$\|\boldsymbol{K} - \boldsymbol{\Lambda}\| \leq \beta, \|\boldsymbol{K}^i - \boldsymbol{\Lambda}\| \leq \beta.$$

Thus, we can obtain that

$$\|\boldsymbol{K} - \boldsymbol{K}^i\| \leq 2\beta. \tag{3}$$

Therefore,

$$|f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \leq \frac{2\kappa M \beta}{\lambda^2} + \frac{2\kappa M}{m-1}.$$

□

In order to prove Theorem 1, we need to introduce the pointwise hypothesis stability and a theorem given in [4].

DEFINITION 2 (POINTWISE HYPOTHESIS STABILITY). *An algorithm A has pointwise hypothesis stability γ with respect to the loss function ℓ if the following holds: $\forall S \in \mathcal{Z}^m, \forall i \in \{1, \ldots, m\}$,*

$$\mathbb{E}_S[|\ell(A_S, z_i) - \ell(A_{S^i}, z_i)|] \leq \gamma.$$

THEOREM 7 (THEOREM 11 IN [4]). *For any learning algorithm A with pointwise hypothesis stability γ with respect to a loss function ℓ such that $0 \leq \ell(A_S, z) \leq Q$, we have with probability $1 - \delta$,*

$$R(S) \leq R_{\mathrm{emp}}(S) + \sqrt{\frac{Q^2 + 12Qm\gamma}{2m\delta}}.$$

PROOF OF THEOREM 1. By Theorem 6, we have $\forall S \in \mathcal{Z}^m, \forall i \in \{1, \ldots, m\}$,

$$|f_S(\boldsymbol{x}) - f_{S^i}(\boldsymbol{x})| \leq C\beta + Q.$$

Therefore, $\forall z \in \mathcal{Z}$, we have

$$\begin{aligned} & |\ell(f_S, z) - \ell(f_{S^i}, z)| \\ &= |(y - f_S(\boldsymbol{x}))^2 - (y - f_{S^i}(\boldsymbol{x}))^2| \\ &= |f_S(\boldsymbol{x}_i) - f_{S^i}(\boldsymbol{x})| \cdot |2y - f_S(\boldsymbol{x}) + f_{S^i}(\boldsymbol{x})| \\ &\leq (C\beta + Q)(2M + C\beta + Q) \\ &= 2M(C\beta + Q) + (C\beta + Q)^2. \end{aligned} \tag{4}$$

By the definition of pointwise hypothesis stability (see Definition 2), it is easy to verity that the KRR with β eigenvalues perturbation is

$$2M(C\beta+Q)+(C\beta+Q)^2$$

pointwise hypothesis stability.

Since $|y|\leq M$ and $f_S(\boldsymbol{x})=\frac{1}{m}\boldsymbol{k}^{\mathrm{T}}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\boldsymbol{y}$, we have

$$|f(\boldsymbol{x})|\leq\frac{1}{m}\frac{\|\boldsymbol{k}\|\|\boldsymbol{y}\|}{\lambda_{\min}(\boldsymbol{K}+\lambda\boldsymbol{I})}\leq\frac{\kappa M}{\lambda}.$$

Note that

$$\begin{aligned}\ell(f_S,z)&=(f_S(\boldsymbol{x})-y)^2\\&\leq 2f_S^2(\boldsymbol{x})+2|y|^2\\&\leq\frac{2\kappa^2M^2}{\lambda^2}+2M^2.\end{aligned}\tag{5}$$

Thus, by using Theorem 7, this assertion can be proved. □

Appendix.B Proof of Theorem 2

We first give a definition of uniform stability and a theorem introduced in [4].

Definition 3 (Uniform Stability). *An algorithm A has uniform stability γ with respect to the loss function ℓ if the following holds: $\forall S\in\mathcal{Z}^m,\forall i\in\{1,\ldots,m\}$,*

$$\|\ell(f_S,\cdot)-\ell(f_{S^i},\cdot)\|_\infty\leq\gamma.$$

Theorem 8 (Theorem 12 in [4]). *Let A be an algorithm with uniform stability γ with respect to a loss function ℓ such that $0\leq\ell(f_S,z)\leq L$, for all $z\in\mathcal{Z}$ and all sets S. Then, for any $m\geq 1$, and any $\delta\in(0,1)$, the following bounds hold (separately) with probability at least $1-\delta$ over the random draw of the sample S,*

$$R(S)\leq R_{emp}(S)+2\gamma+(4m\gamma+L)\sqrt{\frac{\ln 1/\delta}{2m}}.$$

Proof of Theorem 2. By (4) and (5), we have

$|\ell(f_S,z)-\ell(f_{S^i},z)|\leq 2M(C\beta+Q)+(C\beta+Q)^2$ and

$$\ell(f_S,z)\leq\frac{2\kappa^2M^2}{\lambda^2}+2M^2.$$

By the definition of uniform stability (see in Definition 3), it is easy to verity that the KRR with β eigenvalues perturbation is

$$2M(C\beta+Q)+(C\beta+Q)^2$$

uniform stability. Thus, based on Theorem 8, we can prove this theorem. □

Appendix.C Proof of Theorem 3

Lemma 1 (Proposition 2 in [13]). *Let h' denote the hypothesis returned by SVMs when using the approximate kernel matrix $\boldsymbol{K}'$. Then, the following inequality holds for all $\boldsymbol{x}\in\mathcal{X}$:*

$$|h'(\boldsymbol{x})-h(\boldsymbol{x})|\leq\sqrt{2}\lambda\kappa^{\frac{3}{4}}\|\boldsymbol{K}'-\boldsymbol{K}\|_2^{\frac{1}{4}}\left[1+\left[\frac{\|\boldsymbol{K}'-\boldsymbol{K}\|_2}{4\kappa}\right]^{\frac{1}{4}}\right].$$

To prove Theorem 3, we first give the following theorem:

Theorem 9. *If the kernel function K is β eigenvalues perturbation, then for the SVM,*

$$|f_S(\boldsymbol{x})-f_{S^i}(\boldsymbol{x})|\leq\sqrt{2}\lambda\kappa^{\frac{3}{4}}(2\beta)^{\frac{1}{4}}\left[1+\left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right].$$

Proof. Note that $f_{S^i}(\boldsymbol{x})$ is the hypothesis returned by SVMs using the i-th removed kernel matrix $\boldsymbol{K}^i$,

$$\begin{cases}[\boldsymbol{K}^i]_{jk}=\dfrac{1}{m-1}K(\boldsymbol{x}_j,\boldsymbol{x}_k) & \text{if } j \text{ and } k\neq i,\\ [\boldsymbol{K}^i]_{jk}=0 & \text{if } j \text{ or } k=i.\end{cases}$$

By Lemma 1, it is easy to verify that for all $\boldsymbol{x}$,

$$|f_S(\boldsymbol{x})-f_{S^i}(\boldsymbol{x})|\leq\sqrt{2}\kappa^{\frac{3}{4}}\|\boldsymbol{K}^i-\boldsymbol{K}\|_2^{\frac{1}{4}}\left[1+\left[\frac{\|\boldsymbol{K}^i-\boldsymbol{K}\|_2}{4\kappa}\right]^{\frac{1}{4}}\right].$$

By (3), we know that

$$\|\boldsymbol{K}^i-\boldsymbol{K}\|\leq 2\beta.$$

Therefore, we can obtain that

$$|f_S(\boldsymbol{x})-f_{S^i}(\boldsymbol{x})|\leq\sqrt{2}\lambda\kappa^{\frac{3}{4}}(2\beta)^{\frac{1}{4}}\left[1+\left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right].$$

□

Proof of Theorem 3. Since the hinge loss ℓ is 1-Lipschitz, by Theorem 9, it is easy to verify that

$$\begin{aligned}&|\ell(f_S(\boldsymbol{x}),y)-\ell(f_{S^i}(\boldsymbol{x}),y)|\\&\leq\sqrt{2}\lambda\kappa^{\frac{3}{4}}(2\beta)^{\frac{1}{4}}\left[1+\left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right].\end{aligned}\tag{6}$$

Thus, it is easy to verity that the SVM with β eigenvalues perturbation is $\kappa^{\frac{3}{4}}(2\beta)^{\frac{1}{4}}\left[1+\left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right]$ pointwise hypothesis stability. Since $f_S(\boldsymbol{x})=\sum_{j=1}^m\alpha_jK(\boldsymbol{x}_j,\boldsymbol{x})$, $\alpha_i\leq\frac{\lambda}{m}$, therefore,

$$\begin{aligned}\ell(f_S(\boldsymbol{x}),y)&=|\max(0,1-yf_S(\boldsymbol{x}))|\\&\leq|1-yf_S(\boldsymbol{x})\leq 1+|y||f_S(\boldsymbol{x})|\\&\leq 1+M\lambda\kappa.\end{aligned}\tag{7}$$

Thus, due to Theorem 7, the assertion is proved. □

Appendix.D Proof of Theorem 4

Proof. By (6), we have

$$\begin{aligned}&|\ell(f_S(\boldsymbol{x}),y)-\ell(f_{S^i}(\boldsymbol{x}),y)|\\&\leq\sqrt{2}\lambda\kappa^{\frac{3}{4}}(2\beta)^{\frac{1}{4}}\left[1+\left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right].\end{aligned}$$

Thus, the SVM with β eigenvalues perturbation is

$$\kappa^{\frac{3}{4}}(2\beta)^{\frac{1}{4}}\left[1+\left[\frac{\beta}{2\kappa}\right]^{\frac{1}{4}}\right]$$

uniform stability. According to Eq.(7) and Theorem 8, we can prove this theorem. □

7. REFERENCES

[1] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21th International Conference on Machine Learning (ICML)*, pages 41–48, 2004.

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

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

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

[5] M. Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *The Journal of Machine Learning Research*, 7:2303–2328, 2006.

[6] G. C. Cawley and N. L. C. Talbot. Preventing over-fitting during model selection via bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research*, 8:841–861, 2007.

[7] G. C. Cawley and N. L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

[8] O. Chapelle and V. Vapnik. Model selection for Support Vector Machines. In *Advances in Neural Information Processing Systems 12 (NIPS)*, pages 230–236, 1999.

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

[10] C. Corte and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[11] C. Cortes, M. Mohri, and A. Rostamizadeh. L2 regularization for learning kernels. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 109–116, 2009.

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

[13] C. Cortes, M. Mohri, and A. Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 113–120, 2000.

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

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

[16] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and Support Vector Machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

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

[18] L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(3):331–368, 2008.

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

[20] L. Jia and S. Liao. Accurate probabilistic error bound for eigenvalues of kernel matrix. In *Proceedings of the 1st Asian Conference on Machine Learning*,2009.

[21] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

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

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

[24] U. V. Luxburg, O. Bousquet, and B. Schölkopf. A compression approach to Support Vector model selection. *Journal of Machine Learning Research*, 5:293–323, 2004.

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

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

[27] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[28] L. Rosasco, M. Belkin, and E. Vito. On learning with integral operators. *The Journal of Machine Learning Research*, 11:905–934, 2010.

[29] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT Press, Cambridge, MA, 2002.

[30] J. Shawe-Taylor and N. Cristianini. *Support Vector Machines*. Cambridge University Press, MA, 2000.

[31] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Verlag, 2008.

[32] J. A. K. Suykens and J. Vandewalle. Least squares Support Vector Machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[33] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 2000.

[34] G. Wahba. Support Vector Machine, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods-Support Vector Learning*, volume 6, pages 69–88, Cambridge, 1999. MIT Press.

[35] G. Wahba, Y. Lin, and H. Zhang. GACV for support vector machines. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge, 1999.