# 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, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[3]Tianjin University, China

## Introduction

- Kernel function
  - operate in a high-dimensional, implicit feature space
  - get cheaper computation by using kernel trick
- Kernel methods is a class of algorithms by using kernel functions(SVM, KRR, etc)
- Performance of kernel methods depends on kernel selection
- Exist Kernel Selection Methods
  - cross-validation(CV) and approximate CV:GCV, GACV, ELOO, etc
  - Kernel target aligment(KTA) and improved KTA: CKTA, FSM, etc
  - Minimizing theoretical estimate bounds of generalizaiton:VC dimension, Rademacher complexiy, ER, etc
- Drawbacks of current kernel selection methods:
  - No theoretical guarantee
  - High computational complexity
- Our proposed kernel selection criteria: Spectral Measure
  - sound theoretical foundation
  - high computational efficiency
- 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 |

## Notations and Preliminaries

- The regularized algorithms to study

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

  - where $\ell(\cdot,\cdot)$ is a loss function and $\lambda$ is the regularization parameter.
  - The performance of the regularized algorithms for classification is usually measured by the ***generalization error*** or ***risk***
  - $R(S) = \mathbf{Pr}_{(\mathbf{x},y)\sim D}[yf_S(\mathbf{x}) < 0]$. Unfortunately, $R(S)$ can not be computed since the probability distribution $D$ is unknown, hence we should estimate it from empirical data.
- Kernel matrix

$$\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^{n}$$

- Normal kernel matrix

$$\mathbf{N} = \mathbf{K}/|\mathbf{K}|_1$$

  - where $|\mathbf{K}|_1 = \sum_{i,j=1}^{n} K(\mathbf{x}_i, \mathbf{x}_j)$.
  - $(\lambda_i, \mathbf{v}_i)$ is the *spectral decomposition* of $\mathbf{N}$
  - $\lambda_i$ is the eigenvalue and $\mathbf{v}_i$ is the eigenvector, $i = 1, \ldots, n$.

## Definition (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, $\varphi(\lambda_i) \leq \lambda_i$ for all $i \in \{1, \ldots, n\}$. Then the spectral measure of $K$ with respect to $\varphi$ 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 $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$.

## Theorem

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

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

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

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* $b = \max\{1, \frac{1}{2\lambda} - 1\}$.

## Spectral measure criterion(SM)

- Weighted spetral measure criterion(SM):

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

  - 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 $O(n^2)$.

## Experiments

Table: 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 | 16.88±1.41 | 18.86±1.49 | 24.72±1.67 | 16.97±1.52 |
| a2a | **17.78±1.28** | 17.96±1.25 | 17.94±1.27 | 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 | 13.71±2.10 | 13.84±2.18 | 13.82±2.04 | 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 | 29.13±4.46 |
| colic | **15.62±3.00** | 16.47±2.78 | 15.73±2.97 | 19.27±2.58 | 36.42±3.28 | 17.35±3.09 |
| 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 | 21.82±5.68 | **20.95±4.82** | 26.41±7.13 | 43.00±9.22 | 22.50±5.08 |
| german.numer | 24.09±2.15 | 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 | 15.98±3.47 |
| hepatitis | **15.57±4.68** | 17.09±5.74 | 16.63±4.64 | 15.74±5.00 | 21.22±5.41 | 18.91±6.20 |
| ionosphere | 4.88±2.10 | 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** | 14.47±8.08 | 14.82±8.34 | 15.41±8.80 | 34.59±8.70 | 18.82±8.81 |
| pima | 23.80±2.14 | 22.78±2.36 | **22.51±2.41** | 24.38±2.28 | 34.47±2.42 | 22.78±2.07 |
| 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 | 21.93±4.45 | **21.87±4.13** | 22.13±4.19 | **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** | 2.43±1.07 | 2.73±1.11 | 2.82±1.20 | 37.49±3.83 | 4.75±1.66 |

## Conclusions

- A novel notion kernel selection crierion based on spectral analysis
  - sound theoretical foundation
  - high computation efficiency

## Main References

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

[2] Yong Liu, Shali Jiang, and Shizhong Liao.
Eigenvalues perturbation of integral operator for kernel selection.
Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)


lijian9026@iie.ac.cn