EasyChair Preprint

№ 7360

Research on Kernel Approximation Algorithm Based on Random Fourier Feature Space

Kai Zhang, Changqian Men and Wenjian Wang

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

January 21, 2022

基于随机傅里叶特征空间的高斯核近似算法

张凯[1]　门昌骞[1]　王文剑[1,2]

[1]（山西大学 计算机与信息技术学院 山西太原 030006）

[2]（山西大学 计算智能与中文信息处理教育部重点实验室 山西太原 030006）

（wjwang@sxu.edu.cn）

Research on kernel approximation algorithm based on random Fourier feature space

Zhang Kai[1], Men Changqian[1], and Wang Wenjian[1,2]

[1]（*College of Computer and Information Technology, Shanxi University, Taiyuan 030006*）

[2]（*Key Laboratory of Computational Intelligence and Chinese Information Processing*（*Shanxi University*）, *Ministry of Education*, *Taiyuan* 030006）

Abstract Kernel method is a method to transform the linear non separable problem in low-dimensional space into the linear separable problem in high-dimensional space. It is widely used in a variety of learning models. The existing kernel selection methods may have low computational efficiency and high time cost in large-scale data. Aiming at above problem, this paper introduces the random Fourier feature to transform the original kernel feature space into another relatively low dimensional explicit random feature space. The theoretical analysis of the upper bound of the kernel approximation error and the upper bound of the error of training the learning model in the kernel approximation random feature space is given. The convergence consistency of kernel approximation and the relationship between error upper bound and kernel approximation parameters are obtained. Then, the optimal model parameters are selected based on random Fourier feature space, which can avoid the large-scale search for the optimal original Gaussian kernel model parameters, so as to greatly reduce the time cost required for the selection of the original Gaussian kernel model. Experiments show that the error upper bound proved in this paper is controlled by the kernel approximation parameters. The optimal model selected by the kernel approximation has good performance compared with the original Gaussian kernel function model, and the model selection time is greatly reduced compared with the grid search method.

Key words kernel method; random Fourier features; kernel approximation; Gaussian kernel; model selection

摘要 核方法是一种把低维空间的线性不可分问题转化为高维空间中线性可分问题的方法，其广泛应用于多种学习模型。其中现有的核模型选择方法在大规模数据中计算效率较低，时间成本很大。针对这一问题，本文引入随机傅里叶特征变换，将原始核特征空间转换为另一个相对低维的显式随机特征空间，并给出核近似误差上界理论分析以及在核近似的随机特征空间中训练学习模型的误差上界，得到核近似的收敛一致性和误差上界与核近似参数之间的关系。然后，基于随机傅里叶特征空间选择出最优模型参数，避免了对最优原始高斯核模型参数的大规模搜索，从而大幅降低原始高斯核模型选择所需的时间成本。实验表明本文给出的误

收稿日期：　修回日期：

基金内容：国家自然科学基金(U21A20513，62076154)；中央引导地方科技发展资金项目(YDZX20201400001224)；山西省国际科技合作计划项目(201903D421050)；山西省自然科学基金项目(201901D111030)

This work was supported by the National Natural Science Foundation of China (U21A20513,62076154), The central government guides local science and technology development fund projects (YDZX20201400001224), Key R&D program of Shanxi Province (International Cooperation, 201903D421050), National Natural Science Foundation of Shanxi Province (201901D111030).

通信作者：王文剑（wjwang@ sxu.edu.cn）

差上界确由核近似参数控制，核近似选择的最优模型相较于原始高斯核模型有较高的准确率并且模型选择时间相对网格搜索法大幅减小。
关键词 核方法；随机傅里叶特征；核近似；高斯核；模型选择

1 引言

核方法为解决机器学习中的许多问题提供了强大、灵活和坚实的理论基础。其核心思想是通过某种非线性映射将原始数据映射到合适的高维特征空间，再利用线性学习器在高维特征空间中分析和处理数据。核方法需要借助核函数实现低维空间向高维空间的转换，不同的核函数对应不同的特征空间。核方法的理论基础 Cover's theorem[1]，对于非线性可分的训练集，可以大概率通过将其映射到一个高维空间来转化成线性可分的训练集。

关于核方法对应的线性模型已经有大量的研究。经过核函数进行高维特征空间转化后，原始非线性问题可转化为高维线性可分问题，大量线性模型所对应的算法被提出，包括随机梯度下降算法[2]、割平面算法[3]、对偶坐标下降算法[4]等。利用具有线性时间复杂度的线性模型算法，普通 PC 机就可以有效地处理大规模问题[5] 。线性模型虽然在超高维文本和基因分类问题上能取得较好的学习效果，但是对于一般问题，预测精度很难超越非线性核模型，并且在计算机视觉、自然语言处理、系统设计和许多其他领域中的学习问题规模通常在数十万左右，甚至可能超过数百万，这就给运用非线性模型解决实际问题带来巨大的计算困难。然而，线性或亚线性时间复杂度的求解算法，计算开销小，易于扩展到大规模问题，这促进了应用线性学习算法求解非线性问题的研究工作。

非线性核模型的计算时间复杂度在理论上不低于$O(m^3)$，因此研究的重点在于设计快速核近似算法。利用核近似方法逼近原始核是利用线性算法来求解非线性问题的一种方法。随着核近似算法的深入研究，一些关于高斯核的核近似算法被提出，如 Nyström 方法[6]，Additive Chi Squared Kernel 方法等。Yang 等人[7]在低阶泰勒公式中近似高斯核，从而加速共轭梯度优化算法; Cao 等人[8]同样在低阶泰勒公式下近似高斯核，然后加速支持向量机(Support vector machine)预测。

随着大量研究的进行，出现了许多不同的核近似方法，其中 Rahimi 和 Recht[9]提出了一种随机傅里叶特征映射方法，该方法主要思想是通过近似平移不变核显式构造特征映射，保证任意两点对应随机特征映射的内积近似其对应的核函数值，从而为显式构造假设空间求解大规模核方法提供了一种有效途径。冯昌等人[10]在此基础上构造了一个循环随机假设空间，该方法运行效率较高，但是解释性不足。文献[11]定义了近似高斯核并建立其与高斯核的关系，实现了用线性 SVM 算法高效求解大规模非线性 SVM。另外，还有一些利用核近似来构建特征空间的算法，文献[12]显式描述核函数对应积分算子的特征值与特征函数，并构造其对应的再生核希尔伯特空间和特征映射。文献[13]在大规模样本进行子采样，然后应用随机傅里叶映射显式构造随机空间，从而达到高效求解的目的。

本文分析了核函数中基于随机傅里叶特征变换的核近似方法对支持向量机、核岭回归(Kernel Ridge Regression)模型这两种广泛使用的学习模型影响，并且给出这些算法的稳定性界。通过对核近似误差上界的分析，可以证明算法的准确率主要和傅里叶特征变换的空间维度以及核函数参数相关。

模型选择通常采用含交叉验证的网格搜索法找到最优参数组合，但这种方法需要多次训练模型，在大规模数据中多次训练模型所需的时间成本巨大。针对这一问题，本文在核近似的基础上提出一种在近似空间上的模型选择算法，该算法通过在近似线性空间上高效选择模型参数，将对应参数组合运用到原始核中，在保证参数有效性的前提下，大幅降低了模型选择的时间成本，并且得到的最优模型参数在原始核中仍有较好的效果，解决了传统模型选择带来的多次迭代以及巨大时间代价，达到模型选择优化的目的。

2 预备知识

本节简要介绍随机傅里叶特征变换的基本概念。

定理 1. Bochner 定理[15]。连续函数$f:\mathbb{R}^d \mapsto S$ 正定，当且仅当 f 为某一有限非负 Borel 测度ϵ 的傅里叶变换，则

$$f(x)=\int_{\mathbb{R}^d} e^{-ix^Tv}\,d\epsilon(v) \tag{1}$$

选择合适的核函数 $k(\cdot)$，由 Bochner 定理可以确定存在一个概率分布为 $p(\cdot)$的傅里叶变换与$k(\cdot)$对应，即有如下推论。

推论 1.任一正定平移不变核 $K(x,y)=k(x-y)$ 均满足

$$K(x,y)=\int_{\mathbb{R}^d} p(w)e^{-ix^T(x-y)}dw \tag{2}$$

其中，$p(w)$为w的概率密度函数。

现有核函数种类众多，不同的核函数在不同的样本中表现能力也各有差异，在众多核函数中，高斯核是一种通用核，是使用比较广泛的核函数，其定义形式为

$$K(x,y) =< \phi(x), \phi(y) >= exp(-\gamma||x-y||^2) \quad (3)$$

对于高斯核函数，通过$K(x,y)$的傅里叶变换计算$p(w)$，可得w~$N(0,2\gamma \boldsymbol{I})$，其中 $\boldsymbol{I}$ 表示单位矩阵。由上可得

$$K(x,y) = E_w[e^{-iw^T(x-y)}] = E_w[\cos(w^T(x-y))]$$
$$= E_{w,b}[\sqrt{2}\cos(w^T x + b)\sqrt{2}\cos(w^T y + b)] \quad (4)$$

其中令$Z_{w,b}(x) = \cos(w^T x + b)$，有

$$K(x,y)=E_w[< Z_{w,b}(x), Z_{w,b}(y) >] \quad (5)$$

则$< Z_{w,b}(x), Z_{w,b}(y) >$是高斯核函数的一个无偏估计。运用标准蒙特卡洛近似积分方法逼近高斯核，构造如下随机特征映射

$$\Psi_{RKS}: x \mapsto \sqrt{\frac{2}{D}}[cos(w_1^T x + b_1), cos(w_2^T x + b_2), \cdots, cos(w_D^T x + b_D)]^T \quad (6)$$

可以看出，随机映射由随机投影与余弦变换组成。其中$w_i \in \mathbb{R}^d$是一个高斯随机变量，每个元素均独立同分布采样于$N(0,2\gamma \boldsymbol{I})$，$b_i$为均匀随机向量，服从均匀分布$U[-\pi,\pi], i = 1,2,\cdots,D$。

3 随机特征空间的近似误差分析

3.1 核近似误差分析

本文利用傅里叶特征变换将原始特征空间转换为相对低维随机特征空间。下面分析傅里叶特征变换与原核函数的误差上界，命题 1 给出了随机傅里叶特征变换的近似一致收敛界[16]。

命题 1. 设 K 是一个连续可微核函数，其满足公式(5)的条件，并具有相关的测度 P。此外假设$\sigma_p^2 = \mathbb{E}_{w\sim p}[||w||^2] < \infty$，并且$X$是紧的，$d$ 表示其维度。设R表示包含X的欧氏球半径。然后对于公式(6)中定义的$\Psi \in R^D$和任意的 $0 < r \le 2R$和$\epsilon > 0$，以下成立

$$\mathbb{P}\left[\sup_{x,y \in X} |\Psi(x)\cdot\Psi(y) - K(x,y)| \ge \epsilon\right] \le 2N(2R,r)exp\left(-\frac{D\epsilon^2}{8}\right) + \frac{4r\sigma_p}{\epsilon} \quad (7)$$

根据命题 1 提出的随机傅里叶特征一致收敛界，核近似的一致收敛界与随机傅里叶特征变换的特征空间维度之间的关系表达如下。

推论 2.对于任意$\delta \in (0,1)$，下式(8) 以 $1-\delta$ 的概率成立

$$\mathbb{P}\left[\sup_{x,y\in X} |\Psi(x)\cdot\Psi(y) - K(x,y)| \ge \sqrt{\frac{1}{\frac{\delta}{(48R\sigma_p)^2} + \frac{D}{4(d+2)}}}\right] \le \delta \quad (8)$$

其中，d表示样本原始特征数，R是包含 x的欧几里得半径，$\sigma_p^2 = E_{w\sim p}[||w||^2] < \infty$，$D$ 是随机傅里叶变换的特征空间维度。

证明:

定义$F(x,y) = \Psi(x)\cdot\Psi(y) - K(x,y)$，$z(x,y) = \Psi(x)\cdot\Psi(y)$，$L_F$表示$F(x,y)$的利普希茨常数，对于所有的 i，当 $|F((x,y)_i)| < \frac{\epsilon}{2}$ 以及$L_F < \frac{\epsilon}{2r}$，对于所有$X \in R^d$，有$|F(x,y)| < \epsilon$。

根据马尔可夫不等式

$$\mathbb{P}\left[L_F \ge \frac{\epsilon}{2r}\right] \le \left(\frac{2r\sigma_p}{\epsilon}\right)^2 \quad (9)$$

在ϵ-net 联合霍夫丁不等式

$$\mathbb{p}\left[\cup_{i=1}^{W} |F((x,y)_i)| \ge \epsilon/8\right] \le 2Wexp(-D\epsilon^2/2) \quad (10)$$

其中$W = (6R/r)^d$，因为

$$N(R,r) \le \left(\frac{3R}{r}\right)^d \quad (11)$$

由(9)，(10)，(11)可得

$$\mathbb{P}\left[\sup_{x,y\in X} |F(x,y)| \le \epsilon\right] \ge 1 - 2\left(\frac{3R}{r}\right)^d exp(-D\epsilon^2/8) - \left(\frac{2r\sigma_p}{\epsilon}\right)^2 \quad (12)$$

上述不等式右式为 $1 - n_1 r^{-d} - n_2 r^2$。其中$n_1 = 2(3R)^d exp(-D\epsilon^2/8)$，$n_2 = \left(\frac{2\sigma_p}{\epsilon}\right)^2$。令$r = \left(\frac{n_1}{n_2}\right)^{\frac{1}{d+2}}$和$\frac{24R\sigma_p}{\epsilon} \ge 1$ 得

$$\mathbb{P}\left[\sup_{x,y\in X} |\Psi(x)\cdot\Psi(y) - K(x,y)| \ge \epsilon\right] \le \left(\frac{48R\sigma_p}{\epsilon}\right)^2 exp\left(-\frac{D\epsilon^2}{4(d+2)}\right) \quad (13)$$

令$\left(\frac{48R\sigma_p}{\epsilon}\right)^2 exp\left(-\frac{D\epsilon^2}{4(d+2)}\right) = \delta$。其中$\delta \in (0,1)$，由泰勒中值定理，将式(13)中部分变形，由于$\epsilon^2$趋近与 0，则

$$exp\left(-\frac{D\epsilon^2}{4(d+2)}\right) \approx 1 - \frac{D}{4(d+2)}\epsilon^2 \quad (14)$$

将式(14)带入式(13)得

$$\left(\frac{48R\sigma_p}{\epsilon}\right)^2 \left(1 - \frac{D}{4(d+2)}\epsilon^2\right) = \delta \quad (15)$$

将ϵ等式代换得

$$\epsilon = \sqrt{\frac{1}{\frac{\delta}{(48R\sigma_p)^2} + \frac{D}{4(d+2)}}}$$

证毕.

由式(8)可以得到，由于$\Psi(x)\cdot\Psi(y)=K^{'}(x,y)$，$\sqrt{1\Big/\left(\frac{\delta}{(48R\sigma_p)^2}+\frac{D}{4(d+2)}\right)}$ 是$|K^{'}(x,y)-K(x,y)|$ 的概率置信上界，并且上界由随机傅里叶变换的特征空间维度 D 和σ_p^2所影响，对于高斯核，$K(x,y)=exp(-\gamma||x-y||^2)$，有$\sigma_p^2=2d\gamma$。

近似特征空间维度 D 越大，γ越小，对应的σ_p^2也越小，近似的核函数与原始核函数误差上界越小。但是参数选择并不是仅要求误差最小，当维度 D 过大时，虽然两者的近似误差很小，但随机傅里叶特征空间中对于内核的计算效率会很低，失去了引入随机特征空间的加速效果和结构简明的优势。

当参数γ取值过小时，原高斯核空间数据分布间距趋于相等，分类效果极差。虽然核近似的核函数与原始核函数近似误差一直在减小，但得到的是两个极差的模型，失去了核近似的意义。为了得到一个效果较好，并且近似误差较小的核近似方法，所以γ会取在适中位置，而随机特征空间维度 D 要有合适的范围，在降低近似误差的同时，要求模型效果良好。

3. 2 核近似误差边界在模型中分析

本节通过核岭回归和支持向量机模型来进一步分析核近似误差在学习器中的影响因子和误差边界。

（1）核岭回归

KRR 解决的对偶优化问题[17]为

$$\max_{\alpha\in\mathbb{R}^m}\lambda\alpha^T\alpha+\alpha K\alpha-2\alpha^T y \tag{16}$$

其中$\lambda=m\lambda_0>0$ 是岭参数。该问题允许封闭形式的解$\alpha=(K+\lambda I)^{-1}y$，$h$ 表示使用精确核矩阵时核岭回归返回的假设。

命题 2.[18]设$h^{'}$表示使用近似核矩阵$K^{'}\in\mathbb{R}^{m\times m}$时核岭回归返回的假设。定义$r>0$，对于所有$x\in X$，$K(x,y)\le r$和$K^{'}(x,y)\le r$。以下的不等式适用于所有的$x\in X$

$$\left|h^{'}(x)-h(x)\right|\le\frac{rM}{\lambda_0^2 m}\left\|K^{'}-K\right\|_2 \tag{17}$$

根据式(8)和命题 2 中的式(17)，有以下推论成立。

推论 3.定义 $r>0$，对于所有$x\in X$，$K(x,y)\le r$和$K^{'}(x,y)\le r$，以下的不等式适用于所有的$x\in X$

$$\left|h^{'}(x)-h(x)\right|\le\frac{rM}{\lambda_0^2 m}\left\|K^{'}-K\right\|_2 \le\frac{rM}{\lambda_0^2 m}\sqrt{\frac{1}{\frac{\delta}{(48R\sigma_p)^2}+\frac{D}{4(d+2)}}} \tag{18}$$

根据以上推论可以得出，模型的近似误差上界仍由核函数近似误差所影响，其近似误差上界是由随机傅里叶变换的特征空间维度 D 和σ_p所影响。近似特征空间维度 D 越大，核近似对应的模型与原始核模型之间的误差上界越小；γ越小，σ_p^2越小，两者误差上界越小，但要在降低近似误差的同时，要求模型效果良好。

（2）支持向量机

下面分析支持向量机在 M 个点的样本 S 上训练时返回的假设 h 与使用核$K^{'}$在相同样本上训练时获得的假设$h^{'}$之间的差异。命题 3 度量了 SVM 应用核矩阵得到的假设与应用近似核矩阵得到的假设之间的误差界。

命题 3.设$h^{'}$表示使用近似核矩阵$K^{'}\in R^{M\times M}$时支持向量机返回的假设。定义$p>0$，对于所有$x\in X$，$K(x,y)\le p$和$K^{'}(x,y)\le p$。那么，以下不等式适用于所有的$x\in X$

$$\left|h^{'}(x)-h(x)\right|\le\sqrt{2}p^{\frac{3}{4}}C_0\left\|K^{'}-K\right\|_2^{\frac{1}{4}}+p^{\frac{1}{2}}C_0\left\|K^{'}-K\right\|_2^{\frac{1}{2}} \tag{19}$$

其中C_0为一常量，根据式(8)和命题 3 中的式(19)，有以下推论成立。

推论 4.定义$p>0$，对于所有$x\in X$，$K(x,y)\le p$和$K^{'}(x,y)\le p$，那么，以下不等式适用于所有的$x\in X$

$$\left|h^{'}(x)-h(x)\right|\le\sqrt{2}p^{\frac{3}{4}}C_0\left\|K^{'}-K\right\|_2^{\frac{1}{4}}+p^{\frac{1}{2}}C_0\left\|K^{'}-K\right\|_2^{\frac{1}{2}} \le\sqrt{2}p^{\frac{3}{4}}C_0\left(\sqrt{\frac{1}{\frac{\delta}{(48R\sigma_p)^2}+\frac{D}{4(d+2)}}}\right)^{\frac{1}{4}}+p^{\frac{1}{2}}C_0\left(\sqrt{\frac{1}{\frac{\delta}{(48R\sigma_p)^2}+\frac{D}{4(d+2)}}}\right)^{\frac{1}{2}} \tag{20}$$

通过式(18)和式(20)可以得出随机傅里叶特征变换后的核近似在核岭回归和支持向量机上的应用，模型中原始核矩阵和核近似矩阵对应的两种假设$h^{'}$和h之间存在误差上界，并且上界由具体的随机特征空间维度 D 以及σ_p^2所影响，维度 D 越大，对应模型近似误差上界越紧，即相应误差越小。在具体应用时应注意控制计算代价，随机特征空间 D 的取值以及σ_p^2所对应γ的取值应尽量控制在合理范围内。

在理论验证核近似收敛一致性的情况下，要保持近似核与原始核之间的误差在可接受范围内，同时通过核近似算法在近似空间上选择出一个性能较好的模型，并且利用核近似所得到的模型参数组合可以选出一个表现良好的原始核模型。

3. 3 算法设计

本文提出一种算法 RFFDG-RBF，目的在于选出一个核近似模型的最优参数组合。由于近似的核函数

与原始核函数有误差上界并且很小，所以在近似空间上选择出最优模型参数并运用到原始核空间，在保证性能的前提下，原始核搜索参数所需时间将大幅降低，所以既可以保证使用核近似参数的原始核准确度较高，又比传统网格搜索法选择模型速度快很多。选定合适大小的参数 D，输入需要搜索的σ_p^2值，其中 D 为随机特征空间维度，对于高斯核$K(x,y)=exp(-\gamma||x-y||^2)$，有$\sigma_p^2=2d\gamma$，所以需要输入$\gamma$的值。下面给出了 RFFDG-RBF 算法的主要步骤。

RFFDG-RBF 算法

输入 训练样本$N=\{(x_i,y_i)\}_{i=1}^n$、参数空间 R
输出 最优组合 (D_o,γ_o)
　　原始核准确率 rbf-bestacc
初始化 rff-btacc=0
for i=1, 2 ,3,…,D **do**
　　$w_i \in \mathbb{R}^d: w_i \sim \Lambda(0,2\gamma \boldsymbol{I})$
　　$b_i \in T[-\pi,\pi]$
end for
for each $(x,y) \in N$ **do**
　　计算(x,y)在随机特征空间中的映射$(\Psi(x),y)$
end for
在训练集$N=\{\Psi(x_i),y_i\}_{i=1}^n$上运用线性核模型训练
for each $(D,\gamma)_i \in (D,\gamma)_*$其中
$D \in (10^1,10^4)$:
　　$\gamma \in (10^{-2},10^1)$:
　　　　if result>rff-btacc
　　　　rff-btacc=result
　　　　(D_o,γ_o)=$(D,\gamma)_i$
Return (D_o,γ_o)
for 正则化系数$C \in (10^{-2},10^2)$ **do**
　　带入(D_o,γ_o)
　　计算 rbf-acc
　　输出 rbf-bestacc
end for

4. 1 随机特征空间维度 D 对近似误差的影响

本小节将验证核近似的误差上界受到近似特征空间维度 D 影响。随机傅里叶变换后的核近似与原始核的均方误差为：

$$MSE=\frac{1}{n}\sum_{i=1}^n\left(K'(x,x')-K(x,x')\right)^2$$

图 1 和图 2 分别为在回归和分类数据集上 MSE 随特征空间维度 D 变化的情况。由于本小节主要讨论 D 的变化与误差的关系，所以参数γ 统一设定为 1.0。

从图中可以看出，随着维度 D 增大，均方误差整体趋势在减小，当$D \in (200,800)$之间时，误差上界已经收敛在较小范围内。

4 实验与结果

本节首先对上文所提出的误差上界受参数影响的理论在回归和分类数据集上分别进行验证，然后利用 3.3 提出的算法分别在核岭回归模型和 SVM 模型找到最优参数组合，将最优参数组合运用到原始核上，从而达到通过近似核选出一个较优原始核模型的效果，并且参数选择的计算效率更高，最后将几种核方法进行对比。实验在 3.60GHz，Intel(R) Core(TM) i7-4700 CPU，16GB DDR3 内存的机器上运行，采用公开数据集[19]如表 1 所示，对于每个数据集根据其自身特征数，选择相应合适的核近似特征空间维度 D。

Table 1 Datasets used in the experiment

表 1 实验所实用数据集

Datasets	Dimension	Train	Test	Task
ijcnn1	22	49990	91701	classification
a8a	123	22696	9865	classification
letter	16	15000	5000	classification
w7a	300	24692	25057	classification
shuttle	9	43500	14500	classification
phishing	68	8844	2211	classification
mg	6	1039	346	regression
abalone	8	3133	1044	regression
housing	13	380	126	regression
space_ga	6	2263	754	regression
cpusmall	12	6144	2048	regression
mpg	7	294	98	regression

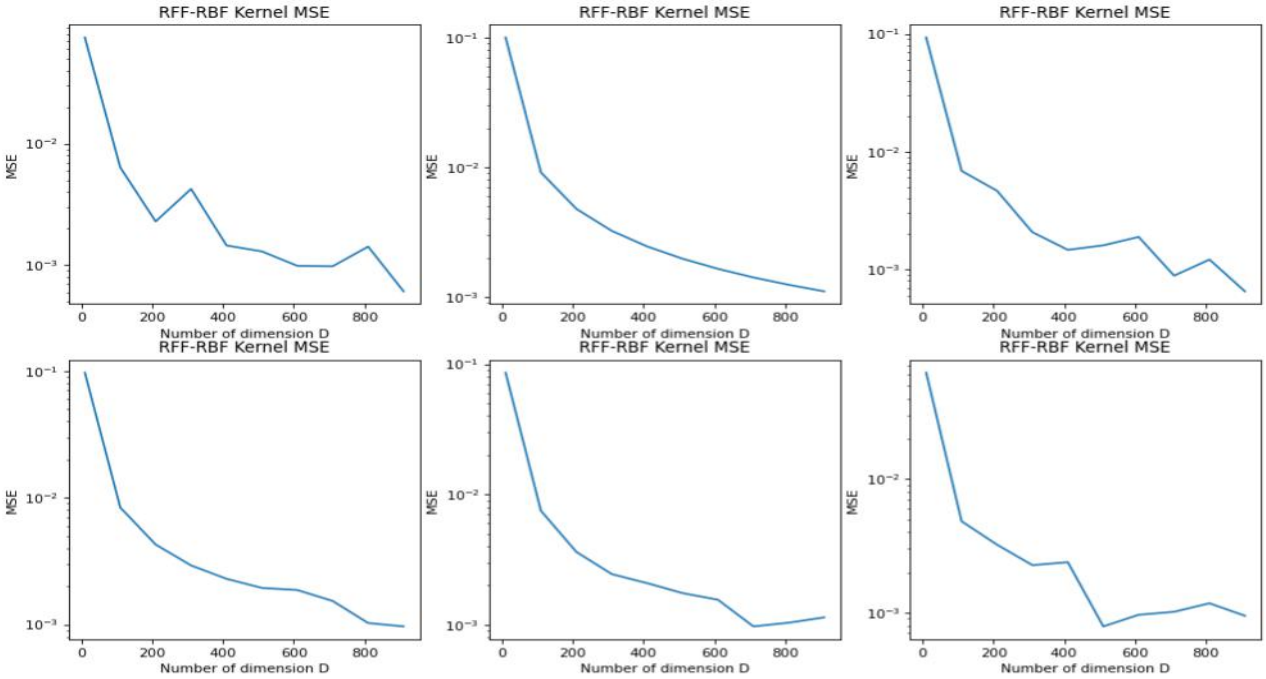


Fig. 1 Relationship between dimension D and MSE under each regression datasets

图 1 各回归数据集上维度 D 与 MSE 的关系

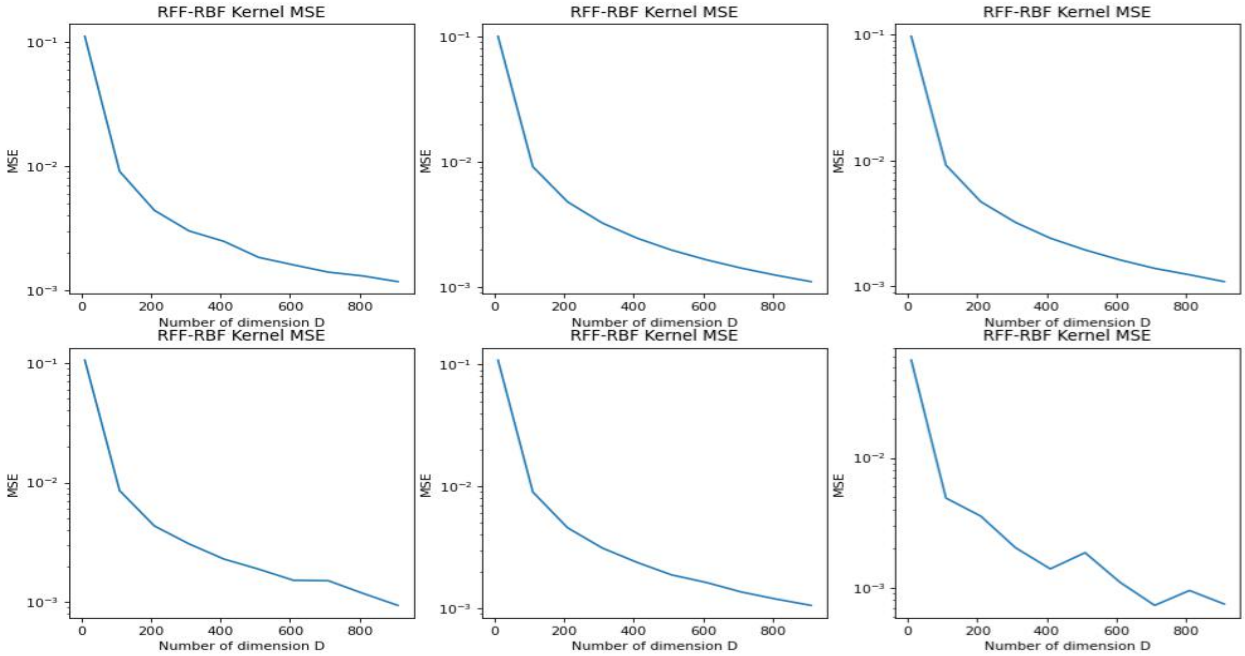


Fig. 2 Relationship between dimension D and MSE under each classification datasets

图 2 各分类数据集上维度 D 与 MSE 的关系

4.2 在核岭回归和支持向量机的模型选择上各种方法的对比

实验选取本文提出的RFFDG-RBF随机傅里叶核近似算法，RBF 为原始高斯核，Nyström 为 Nyström 近似方法。Nyström 近似方法是一种基于列的子集求解较大核矩阵的低秩逼近的有效方法。Nyström 近似的质量在很大程度上取决于所使用列的子集，这些列通常采用随机抽样方式选择。首先随机选择矩阵 K 的 c 列，并将 C 表示为由所选列形成的$l \times c$矩阵。用 W 表示由所选 c 列与相应的 c 行的相交组成的$c \times c$矩阵。Nyström 近似形式如下

$$\widetilde{K} = CW_k^{\dagger}C^T \approx K$$

其中W_k是 W的最优秩 k近似，$W_k^{\dagger}$是W_k的广义逆。

实验采用高斯核作为原始核函数实现核岭回归和 SVM 问题，在 6 个回归数据集和 6 个分类数据集中用 RFFDG-RBF 算法进行随机傅里叶特征变换，找出最优的参数组合(γ, D)，将最优参数组合运用到原始 RBF 核，利用核近似选择出的核参数γ，再对 SVM 中参数 C 以及核岭回归参数λ进行搜索。利用核近似选择出的参数γ，达到和原函数相近甚至相等的准确率。表 2 给出了各数据集上核近似选择最优模型对应的测试均方误差和训练时间，表 3 给出了各数据集上核近似选择最优模型对应的测试准确率和训练时间，所有随机算法在每个数据集上重复 10 次，并且记录平均值。

Table 2 The test MSE and training time corresponding to the optimal model of various methods

表 2 各种方法的最优模型对应的测试 MSE 以及训练时间(s)

Datasets	RFFDG-RBF		RBF		Nyström	
	Time	MSE	Time	MSE	Time	MSE
mg	**3.9255**	**0.12445**	4.3911	0.12447	16.4355	0.13899
abalone	**19.1972**	**2.08955**	50.1547	2.08954	39.6284	2.08956
housing	**0.8844**	3.95715	0.9331	3.79092	9.3699	5.26112
space_ga	**12.6758**	**0.10880**	25.6076	0.10880	29.3953	0.10881
cpusmall	**61.3288**	**2.96858**	201.8996	3.00552	93.4866	3.03640
mpg	0.8549	**2.70737**	0.6004	2.85580	7.8191	3.42480

Table 3 The test accuracy and training time corresponding to the optimal model of various methods

表 3 各种方法的最优模型对应的测试准确率以及训练时间(s)

Datasets	RFFDG-RBF		RBF		Nyström	
	Time	Acc/%	Time	Acc/%	Time	Acc/%
ijcnn1	**285.526**	96.68	1537.922	96.69	427.597	98.24
a8a	**547.090**	**85.26**	8506.023	85.26	632.908	82.90
letter	**209.395**	**96.74**	656.080	96.74	675.339	96.74
w7a	5144.605	**97.23**	9938.427	97.35	5106.867	97.05
shuttle	**219.801**	**99.86**	1490.305	99.86	862.808	99.77
phishing	**31.422**	**96.88**	326.978	97.06	58.124	95.86

实验表明通过近似的核函数选择模型相对原始 RBF 核的网格搜索法速度大幅加快。RFFDG-RBF 算法与原始 RBF 核以及 Nyström 近似方法在各数据集上选择出的最优模型对应的测试均方误差和训练时间的对比如表 2 和表 3 所示。表 2 为 3 种方法对应的最优模型和测试均方误差以及所需训练时间，表 3 为 3 种方法对应的最优模型和测试准确率以及所需训练时间。

从表 2 可以得出 RFFDG-RBF 算法最优模型的测试均方误差在 mg, abalone, space_ga, cpusmall, mpg 数据集上比 Nyström 方法最优模型的均方误差小，并且与原始 RBF 核对应的最优模型均方误差相近或相等。在 housing 数据集上均方误差较大是由于数据集本身样本量较小，再通过核近似会一定程度上增大均方误差的值，当在充足样本量的数据集中，误差就会相对较小。并且 RFFDG-RBF 核近似算法在 mg, abalone, space_ga, cpusmall, housing 数据集上的训练所需时间都比另外两种方法少，在样本规模较大的数据 abalone, space_ga, cpusmall 中的加速效果更加明显。通过表 2 实验结果说明 5 个数据集上的 RFFDG-RBF 算法都得到了与原始高斯核相近的均方误差，从而证明本文提出的方法可以在提升模型选择计算效率的基础上，并且运用核近似方法选择出表现较好的原始 RBF 核模型。

从表 3 可以得出 RFFDG-RBF 算法的最优模型的测试准确率在 a8a, letter, shuttle, w7a, phishing 数据集上比 Nyström 方法的准确率高，并且与原始高斯核准确度相近或相等。在 ijnn1 数据集上 RFFDG-RBF 算法比 Nyström 方法的准确率略低，但 RFFDG-RBF 算法的测试准确率仍然与原始 RBF 核准确率相近，并且计算效率大幅提升。另外，在 w7a 数据集上由于数据集本身的特征数量较大，导致 RFFDG-RBF 算法选择特征空间维度 D 也相应增大，但仍比 Nyström 方法准确度高，并且测试准确率与原始 RBF 核准确率相近。同时在 ijcnn1, a8a, letter, shuttle, phishing 数据集上 RFFDG-RBF 算法的模型选择计算效率相比原始 RBF 核以及 Nyström 近似方法的计算效率显著提高，在 w7a 数据集上 RFFDG-RBF 算法模型搜索时间仍然比原始 RBF 核模型搜索时间短。

结合表 2 和表 3，本文提出的算法在高斯核模型选择上大幅提高计算效率的同时，保证核近似在回归数据集上的均方误差与原始 RBF 核相近，在分类数据集上的准确率与原始 RBF 核相近，对原始高斯核模型的选择进行了优化。

4.3 各种方法在样本规模不同时的训练时间以及测试结果对比

随机傅里叶特征变换将数据映射到一个相对低维的特征空间中，其维度是 D。因为映射后的核近似算法在计算上可以避免其本身数据量 N，而原始高斯核计算仍需考虑数据量 N，所以本文提出的算法相对于原始 RBF 核，在规模越大的数据集上计算效率的提升幅度越大。本节实验分别在回归数据集 cpusmall 和分类数据集 shuttle 上进行实验，将 RFFDG-RBF 算法与原始 RBF 核以及 Nyström 方法的模型选择结果进行对比。随着样本规模的增大，得出其训练时间增长速率和最优模型选择的测试均方误差和测试准确率的变化规律。

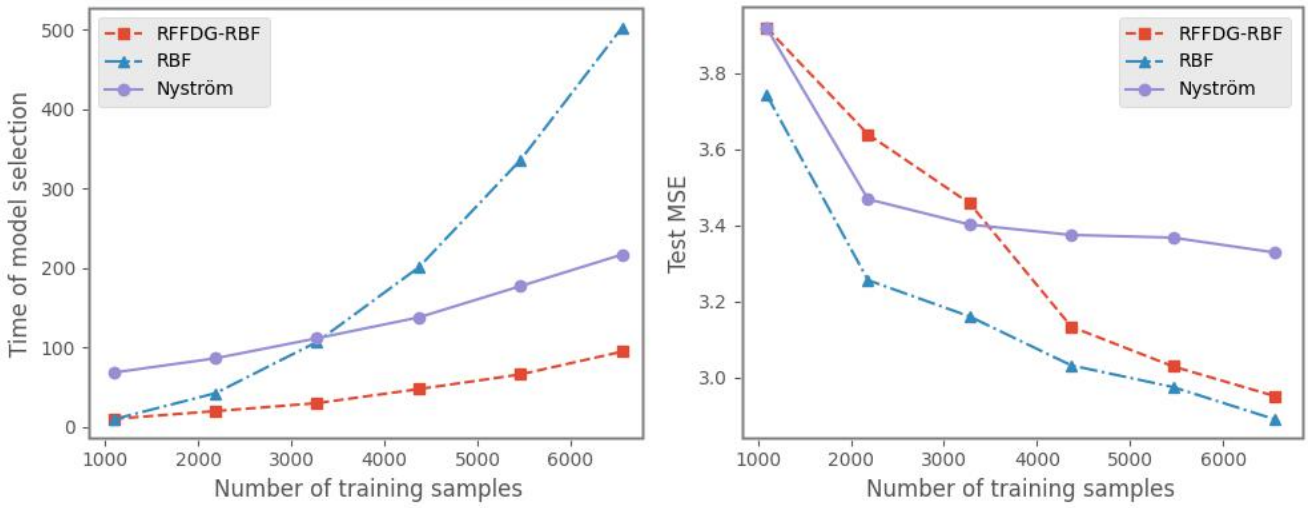


Fig. 3 The training time and test MSE of the three model selection algorithms increase with the sample size

图 3 样本规模增大三种模型选择算法的训练时间和测试 MSE

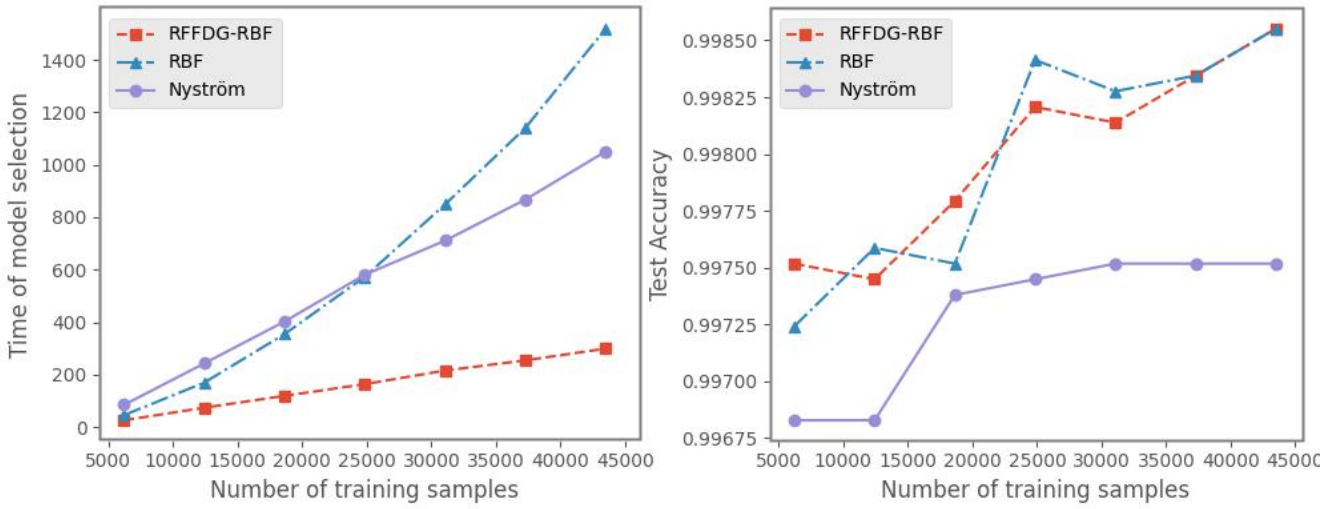


Fig. 4 The training time and test accuracy of the three model selection algorithms increase with the sample size

图 4 样本规模增大三种模型选择算法的训练时间和测试准确率

结合图 3 和图 4，随着样本规模的增大，RFFDG-RBF 算法在模型选择所需时间上相对于原始 RBF 核的模型选择时间大幅降低，同时比 Nyström 方法的模型选择所需的时间少，并且数据规模越大，三者训练所需时间差距越大。在 cpusmall 数据集上 RFFDG-RBF 算法比原始高斯核模型选择训练所需时间缩短 400 秒，仅为原始 RBF 核训练所需时间的五分之一，比 Nyström 方法训练所需时间缩短 100 秒，是 Nyström 方法所需时间的二分之一。在 shuttle 数据集上 RFFDG-RBF 算法比原始高斯核模型选择训练所需时间缩短 1200 秒，比 Nyström 方法所需时间缩短 800 秒。

在两个数据集上，原始 RBF 核计算所需时间增长速率都远大于 RFFDG-RBF 算法对应的所需时间增长速率，并且随着样本规模的增大，两者的时间成本差距越来越大，原始 RBF 核时间成本增长速率越来越快，而 RFFDG-RBF 算法计算时间成本增长速率则较为缓慢。实验表明 RFFDG-RBF 算法在不同规模数据集上的模型选择中都实现了计算效率的大幅提升，并且数据集规模越大，计算效率提升越明显。

最优模型测试结果方面，随着样本规模的增大，在 cpusmall 数据集上 RFFDG-RBF 算法对应的最优模型测试 MSE 逐渐减小，并且最后与原始 RBF 核对应的核岭回归最优模型的测试 MSE 很接近。在 shuttle 数据集上，RFFDG-RBF 算法的最优模型测试准确率逐渐增加，并且高于 Nyström 方法最优模型对应的测试准确率。在样本规模最大时 RFFDG-RBF 算法的最优模型测试准确率与原始 RBF 核支持向量机最优模型的测试准确率相近，仅相差 0.002% 。

上述实验表明本文提出的算法相对于原始 RBF 核模型，在规模越大的数据集上计算效率提升越大，提升计算效率的同时仍可以保证核近似选择的最优模型拥有良好的准确率。

5 结束语

本文首先提出基于随机傅里叶特征变换生成的核函数与原始核函数的近似误差存在上界，并且证明误差上界会受到随机傅里叶特征空间维度 D 以及核函数参数γ的影响，从理论上论证核近似的收敛一致性。然后，构建一个低维显式的特征空间，在这个特征空间中进行模型选择相对透明，增强了模型的可解释性。另一方面，选择出核近似相应的最优模型相对于原始核模型选择的网格搜索法，不仅实现了计算效率的提高，并且与原始高斯核的测试准确率相近。在规模越大的数据集上，计算效率的提升幅度越大，与其他核近似方法进行对比，仍有较好的准确率以及较少的训练时间。所以本文提出的算法提供了在模型选择中提升计算效率和保持准确率之间的优化。

未来，我们将进一步探索在其他核函数中运用核近似的合理性，并进行深入的理论依据研究。

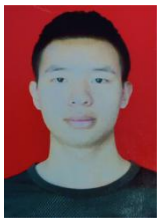
参 考 文 献

[1]Thomas H, Bernhard S, Alexander J. kernel methods in machine learning [C] //The Annals of Statistics, 2008: 1171–1220

[2]Zhang T. Solving large scale linear prediction problem sussing to chastic gradient descent algorithms [C] //Proc of the 25th Int Conf on Machine Learning, New York: ACM, 2004:919-926

[3]Joachims T. Training linear SVMs in linear time [C] //Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and

Data Mining, New York: ACM, 2006: 217-226
[4]Hsieh C J, Chang K W, Lin C J, et al. A dual coordinate descent method for large-scale linear SVM [C] //Proc of the 25th Int Conf on Machine Learning, New York: ACM, 2008:408-415
[5]Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification [J]. Journal of Machine Learning Research, 2008:9:1871-1874
[6]Assia O, Abdelaziz O, Mokhtar K. Low Complexity Method for DOA Estimation Based on Nyström Method [J]. International Journal on Communications Antenna and Propagation, 2017: 239-245
[7]Yang C J, Duraiswami R, Davis L. Efficient kernel machines using the improved fast Gauss transform [C] //Advances in Neural Information Processing Systems 16, Cambridge, MA: MIT Pres, 2004:1561-1568
[8]Cao H, Naito T, Ninomiya Y. Approximate RBF kernel SVM and Its applications in pedestrian classification [C] //Proc of the 1st Int Workshop on Machine Learning for Vision-based Motion Analysis. Berlin: Spinger, 2008:120-129
[9]Rahimi A, Recht B. Random features for large-scale kernel machines [C] //Proceedings of the 21st Annual Conference on Neural Information Processing Systems, Vancouver, Dec3-6.2007. Red Hook: Curan Associates, 2008:1177-1184.
[10]Feng C, Liao S Z. Large-Scale Kernel Methods via Random Hypothesis Spaces [C] //Journal of Frontiers of Computer Science and Technology, 2018: 1673-9418.
[11]Liu Y, Jiang L S, Liao S Z. Approximate Gaussian Kernel for Large-Scale SVM [C] //Journal of Computer Research and Development, 2014: 2171-2177.
[12]Liu Y, Liao S Z. Framework core selection method based on an explicit description of the integral operator space [J]. Scientia Sinica Informationis, 2014: 757-765
[13]Feng C, Li Z D, Liao S Z. Efficient Algorithm for Large-scale Support Vector Machine [M] //Computer Science, 2015:9-037
[14]Duan K, Keerthi S S, Poo A N. Evaluation of simple performance measures for tuning SVM hyperparameters [J]. Neurocomputing, 2003, 51:41-59
[15]Rudim W. Fourier analysis on groups [M]. San Francisco: John Wiley & Sons, 2011: 1-285
[16]Bharath K.S, Zoltán S. Optimal Rates for Random Fourier Features [J]. NIPS, 2015: 1144-1152
[17]Mehryar M, Afshin R, Ameet T. Foundations of Machine Learning [M]. The MIT, Press, 2018:022-812
[18]Cortes C, Mohri M, Talwalkar A. On the impact of kernel approximation on learning accuracy [C] //Proc of the 13th Int Conf on Artificial Intelligence and Statistics. Brookline, MA: Microtome Publishing, 2010:113-120
[19]Chang C C, Lin C J. LIBSVM : a library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.



张凯，男，1997 年，硕士研究生，主要研究方向为机器学习，数据挖掘.

门昌骞，男，1982 年，博士，讲师，CCF 会员，主要研究方向为支持向量机、机器学习理论、核方法.

王文剑，女，1968 年，教授，博士，CCFL 杰出会员，主要研究方向为机器学习、计算智能、图像处理.