

随机傅里叶特征空间中高斯核支持向量机模型选择

冯 昌 廖士中

(天津大学计算机科学与技术学院 天津 300350)
(changfeng@tju.edu.cn)

Model Selection for Gaussian Kernel Support Vector Machines in Random Fourier Feature Space

Feng Chang and Liao Shizhong

(School of Computer Science and Technology, Tianjin University, Tianjin 300350)

Abstract Model selection is very critical to support vector machines (SVMs). Standard SVMs typically suffer from cubic time complexity in data size since they solve the convex quadratic programming problems. However, it usually needs to train hundreds/thousands of SVMs for model selection, which is prohibitively time-consuming for medium-scale datasets and very difficult to scale up to large-scale problems. In this paper, by using random Fourier features to approximate Gaussian kernel, a novel and efficient approach to model selection of kernel SVMs is proposed. Firstly, the random Fourier feature mapping is used to embed the infinite-dimensional implicit feature space into an explicit random feature space. An error bound between the accurate model obtained by training kernel SVM and the approximate one returned by the linear SVM in the random feature space is derived. Then, in the random feature space, a model selection approach to kernel SVM is presented. Under the guarantee of the model error upper bound, by applying the linear SVMs in the random feature space to approximate the corresponding kernel SVMs, the approximate model selection criterion can be efficiently calculated and used to assess the relative goodness of the corresponding kernel SVMs. Finally, comparison experiments on benchmark datasets for cross validation model selection show the proposed approach can significantly improve the efficiency of model selection for kernel SVMs while guaranteeing test accuracy.

Key words model selection; support vector machines (SVMs); random Fourier features; Gaussian kernel; cross validation

摘 要 模型选择是支持向量机(support vector machines, SVMs)学习的关键问题. 标准支持向量机学习本质上是求解一个凸二次优化问题, 求解的时间复杂度为数据规模的立方级, 而经典的模型选择方法往往需要多次训练支持向量机, 这种模型选择方法对于中等规模的支持向量机学习计算代价已较高, 更难以扩展到大规模支持向量机学习. 基于高斯核函数的随机傅里叶特征近似, 提出一种新的、高效的核支持向量机模型选择方法. 首先, 利用随机傅里叶特征映射, 将无限维隐式特征空间嵌入到一个相对低维的显式随机特征空间, 并推导在 2 个不同的特征空间中分别训练支持向量机所得到的模型的误差

收稿日期: 2015-06-12; 修回日期: 2015-10-30

基金项目: 国家自然科学基金项目(61170019)

This work was supported by the National Natural Science Foundation of China (61170019).

通信作者: 廖士中(szliao@tju.edu.cn)

上界;然后,以模型误差上界为理论保证,提出随机特征空间中核支持向量机的模型选择方法,应用随机特征空间中的线性支持向量机来逼近核支持向量机,计算模型选择准则的近似值,从而评价所对应的核支持向量机的相对优劣;最后,在标杆数据集上验证所提出方法的可行性和高效性.实验结果表明,所提出的模型选择方法与标准交叉验证方法的测试精度基本相当,但可显著地提高核支持向量机模型选择效率.

关键词 模型选择;支持向量机;随机傅里叶特征;高斯核;交叉验证

中图法分类号 TP181; TP301

支持向量机(support vector machines, SVMs)是在统计学习理论上发展起来的一类重要的学习方法,是目前流行的数据挖掘方法之一^[1-2].该方法在核诱导的特征空间中训练线性学习器,并应用泛化性理论来避免过拟合现象.

SVM的泛化性能主要依赖于核函数及模型参数的选择,不合适的选择将会导致过拟合(over-fitting)或欠拟合(under-fitting),因此模型选择是SVM学习的重要问题.本文主要考虑高斯核SVM的模型选择问题,即核参数 γ 和特征空间中线性学习器正则化系数 C 的选择,其中 (γ, C) 合称为超参数^[3].

已有的模型选择方法可概括为一个内外双层的优化框架^[4]:内层在超参数固定的情况下在隐式特征空间中训练学习器,本质上是求解一个标准的凸二次优化问题;基于内层的学习结果,外层通过最小化泛化误差来调节超参数. k -折交叉验证(k -fold cross validation, k -CV)是泛化误差的有效估计^[5],被广泛应用在不同问题的学习过程中.对于参数空间中的每一组参数 (γ, C) , k -CV需要训练学习器 k 次,因此计算代价较高.最小化泛化误差的理论上是另外一类模型选择方法^[6-7],通常可以应用梯度下降算法实现.常见理论上界有半径间隔界(radius margin bound)^[6]、span界^[7]等.一般来说,在优化框架外层,基于理论上界的模型选择方法仅仅只需在梯度下降方向评估模型,不再需要搜索整个参数空间,从而提高模型选择效率,但是该类方法仍需多次训练SVM.

在核诱导的特征空间中训练线性学习器本质上是求解一个凸二次优化问题,求解时间复杂度为 $O(l^3)$,其中 l 为训练样本的个数.基于分解法实现的高效SVM求解器,最小序贯优化算法(sequential minimal optimization, SMO),求解时间复杂度为 $O(l^2)$ ^[8].若利用10折交叉验证模型选择,在 $10 \times 10 = 100$ 个 (γ, C) 参数组合上进行SVM学习,需要训练1000个SVM;若应用理论上界方法进行模型

选择,仍需要训练多个SVM.对于一般规模的数据,模型选择过程已十分耗时,更难扩展到大规模问题.

线性SVM研究引起了广泛关注.大量计算高效的线性SVM算法被提出,包括随机梯度下降算法^[9]、割平面方法^[10]、对偶坐标下降算法^[11]等.利用具有线性时间复杂度的线性SVM算法,日常PC机就可以有效地处理大规模问题^[12].但是对于非线性问题,线性SVM的预测准确率远不及核SVM. Rahimi和Recht^[13]提出利用随机傅里叶特征近似高斯核函数,将核诱导的特征空间嵌入到一个显式的随机特征空间,从而可以应用高效的线性SVM一致逼近核SVM.该方法不但具有线性SVM的求解效率,而且能保持核SVM的预测精度.刘勇等人^[14]基于高斯核的显式描述给出了另外一种显式特征映射方法,从而也能够应用线性算法求解非线性问题.该类方法将核SVM线性化,可很大程度地提升核SVM模型选择优化框架的内层效率.

对于模型选择而言,没有必要对每个模型都计算其精确结果,所需的仅是一个能够评估模型相对优劣且可快速计算的近似解.基于这一思想,核矩阵近似已经成功应用在支持向量机和最小二乘支持向量机的高效模型选择过程^[15-17].本文基于高斯核函数的随机傅里叶特征近似,应用线性SVM近似核SVM,提出一种新的高斯核支持向量机模型选择方法.首先,证明随机特征空间中近似模型与其对应精确模型的误差上界.然后,在该误差上界的保证下,提出一种随机特征空间中核SVM模型选择的新方法.在模型选择优化框架内层,不再求解凸二次优化问题,而是将核诱导的特征空间嵌入到一个相对低维的显式随机特征空间,训练线性SVM,计算模型选择准则的近似值,评价对应核SVM的相对优劣.以 k -折交叉验证模型选择实验为例,验证所提出的方法的有效性及其高效性.实验结果表明,在模型选择标准数据集上,所提出的模型选择方法与标准交叉验证方法测试精度相当,但可显著地提高核SVM模型选择效率.

1 预备知识

首先简要回顾支持向量机,然后介绍高斯核函数的随机傅里叶特征近似.

1.1 支持向量机

令 $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_l, y_l)\} \in (\mathcal{X} \times \mathcal{Y})^l$ 表示样本集合,其中 $\mathcal{X}$ 为输入空间, $\mathcal{Y}$ 为输出域. 考虑二分类问题,即 $\mathcal{Y} = \{-1, +1\}$, 输入空间 $\mathcal{X} \in \mathbb{R}^d$.

软间隔 SVM 原优化问题描述如下:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi_i, \\ \text{s. t.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \\ & i = 1, 2, \dots, l, \end{aligned} \quad (1)$$

其中,常量 $C > 0$,称为正则化参数. SVM 的对偶问题为一个标准的凸二次优化问题:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i,j=1}^l y_i y_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^l \alpha_i, \\ \text{s. t.} \quad & \sum_{i=1}^l y_i \alpha_i = 0, \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \dots, l. \end{aligned} \quad (2)$$

令核函数 $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. 若优化问题式(2)的解为 $\alpha_i^*, i = 1, 2, \dots, l$, 对于任意输入 $\mathbf{x} \in \mathcal{X}$, 由核技巧与表示定理^[18]可知,决策函数

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i k(\mathbf{x}, \mathbf{x}_i) + b\right). \quad (3)$$

通过上述优化过程可知,支持向量机的性能主要依赖于拉格朗日乘子、核函数、超参数的选择. 然而,当核函数与超参数给定时,拉格朗日乘子可以通过优化过程求得. 因此核函数与超参数的选择对 SVM 的性能起着决定性作用. 不同的核函数表现能力不尽相同,其中高斯核是一种通用核,是使用最为广泛的支持向量机核函数,定义形式为

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2). \quad (4)$$

1.2 随机傅里叶特征

定义 1. 傅里叶变换与傅里叶逆变换, $f(t)$ 表示变量 t 的函数,如果 $f(t)$ 满足 Dirichlet 条件:具有有限个间断点、有限个极值点, $f(t)$ 绝对可积,则

$$\mathcal{F}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \quad (5)$$

称为 $f(t)$ 的傅里叶变换,且有

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathcal{F}(\omega) e^{i\omega t} d\omega \quad (6)$$

称为 $\mathcal{F}(\omega)$ 的傅里叶逆变换,其中 $i = \sqrt{-1}$.

定理 1. Bochner 定理^[19]. 连续函数 $f: \mathbb{R}^d \mapsto \mathbb{C}$ 正定,当且仅当 f 为某一有限非负 Borel 测度 μ 的傅里叶变换,

$$f(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-i\mathbf{x}^T \mathbf{v}} d\mu(\mathbf{v}). \quad (7)$$

只要核函数 $\kappa(\cdot)$ 选择合适, Bochner 定理保证存在一个概率分布 $p(\cdot)$ 的傅里叶变换与 $\kappa(\cdot)$ 对应,即有如下推论.

推论 1. 任一正定平移不变核 $k(\mathbf{x}, \mathbf{y}) = \kappa(\mathbf{x} - \mathbf{y})$ 均满足式(8):

$$k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} p(\mathbf{w}) e^{-i\mathbf{x}^T (\mathbf{x} - \mathbf{y})} d\mathbf{w}, \quad (8)$$

其中, $p(\mathbf{w})$ 为 $\mathbf{w}$ 的概率密度函数.

对于高斯核函数,通过 $k(\mathbf{x}, \mathbf{y})$ 的傅里叶逆变换计算 $p(\mathbf{w})$,可得 $\mathbf{w} \sim \mathcal{N}(0, 2\gamma \mathbf{I})$, 其中 $\mathbf{I}$ 表示单位矩阵. 易知

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= E_w[e^{-i\mathbf{w}^T (\mathbf{x} - \mathbf{y})}] = E_w[\cos(\mathbf{w}^T (\mathbf{x} - \mathbf{y}))] = \\ &= E_{w,b}[\sqrt{2} \cos(\mathbf{w}^T \mathbf{x} + b) \sqrt{2} \cos(\mathbf{w}^T \mathbf{y} + b)]. \end{aligned} \quad (9)$$

若令 $Z_{w,b}(\mathbf{x}) = \sqrt{2} \cos(\mathbf{w}^T \mathbf{x} + b)$, 有

$$k(\mathbf{x}, \mathbf{y}) = E[\langle Z_{w,b}(\mathbf{x}), Z_{w,b}(\mathbf{y}) \rangle], \quad (10)$$

则 $\langle Z_{w,b}(\mathbf{x}), Z_{w,b}(\mathbf{y}) \rangle$ 是高斯核函数的一个无偏估计. 利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

$$\begin{aligned} \psi: \mathbf{x} \mapsto & \sqrt{\frac{2}{D}} (\cos(\mathbf{w}_1^T \mathbf{x} + b_1), \cos(\mathbf{w}_2^T \mathbf{x} + b_2), \dots, \\ & \cos(\mathbf{w}_D^T \mathbf{x} + b_D))^T, \end{aligned} \quad (11)$$

其中 $\mathbf{w}_i \in \mathbb{R}^d$ 是一个高斯随机向量,服从正态分布 $\mathcal{N}(0, 2\gamma \mathbf{I})$; 随机变量 b_i 服从均匀分布 $\mathcal{U}(-\pi, \pi)$, $i = 1, 2, \dots, D$. 同样有 $k(\mathbf{x}, \mathbf{y}) = E[\langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle]$.

2 随机特征空间中模型近似误差

算法 1. 随机特征空间中线性 SVM.

输入: 训练样本 $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^l$ 、随机特征空间维度 D 、高斯核参数 γ 、正则化参数 C ;

输出: 随机特征空间中决策函数 f .

① for $i = 1, 2, \dots, D$ do

② 采样 $\mathbf{w}_i \in \mathbb{R}^d; \mathbf{w}_i \sim \mathcal{N}(0, 2\gamma \mathbf{I})$;

③ 采样 $b_i \in \mathcal{U}[-\pi, \pi]$;

④ endfor

⑤ for each $(\mathbf{x}, y) \in \mathcal{S}$ do

⑥ 根据式(1)计算 $\phi(\mathbf{x})$;

⑦ endfor

⑧ 在训练集 $\mathcal{T} = \{(\phi(\mathbf{x}_i), y_i)\}_{i=1}^l$ 上利用正则化参数 C 训练线性支持向量机。

随机傅里叶特征方法通过近似高斯核函数,将原始数据映射到一个相对低维的有限维特征空间,并在新的特征空间中训练线性学习器,算法描述见算法 1。本节将分析在随机特征空间中训练线性 SVM 得到模型与在原始数据空间上直接训练核 SVM 得到模型的误差上界。

在误差分析中,令 $\mathbf{K} = \frac{1}{l} [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^l \in \mathbb{R}^{l \times l}$ 表示标准化的核矩阵。根据式定义的随机特征映射,定义如下近似核矩阵 $\mathbf{Q}$,

$$\mathbf{Q} = \frac{1}{l} \mathbf{H}^T \mathbf{H}, \tag{12}$$

其中 $\mathbf{H} = [\phi(\mathbf{x}_1)^T, \phi(\mathbf{x}_2)^T, \dots, \phi(\mathbf{x}_l)^T]$ 。

命题 1 表明在 Frobenius 范数定义下,随机特征空间中潜在的近似核矩阵 $\mathbf{Q}$ 以 $O(1/D)$ 的收敛速度逼近 $\mathbf{K}^{[20]}$ 。

命题 1. 对于任意 $\delta \in (0, 1)$, 下面的式(13)以 $1-\delta$ 的概率成立,

$$\|\mathbf{K} - \mathbf{Q}\|_F \leq \frac{2\ln(2/\delta)}{D} + \sqrt{\frac{2\ln(2/\delta)}{D}}. \tag{13}$$

假设在训练数据 $\mathcal{S}$ 上利用核矩阵训练核 SVM 得到的假设记为 f , 在随机特征中训练线性 SVM 得到的假设记为 f' , 潜在的近似核矩阵为 $\mathbf{Q}$ (注: 训练线性 SVM 不需要计算核矩阵)。对于任意 $\mathbf{x} \in \mathcal{X}$, 比较 $f(\mathbf{x})$ 和 $f'(\mathbf{x})$ 之间的差异。

命题 2 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^[21]。

命题 2. 记 h 为利用核矩阵 $\mathbf{K} \in \mathbb{R}^{l \times l}$ 训练核支持向量机得到的假设, h' 为利用近似核矩阵 $\mathbf{K}'$ 训练支持向量机得到的假设。进一步, 对于所有 $\mathbf{x} \in \mathcal{X}$, 定义 $\tau > 0$, 使得 $k(\mathbf{x}, \mathbf{x}) \leq \tau$ 及 $k'(\mathbf{x}, \mathbf{x}) \leq \tau$ 。那么, 对于所有 $\mathbf{x} \in \mathcal{X}$ 如下不等式成立:

$$|h'(\mathbf{x}) - h(\mathbf{x})| \leq \frac{\sqrt{2}C_0\tau^{\frac{3}{4}}}{l} \|\mathbf{K}' - \mathbf{K}\|_2^{\frac{1}{4}} \left(1 + \left(\frac{\|\mathbf{K}' - \mathbf{K}\|_2}{4\kappa}\right)^{\frac{1}{4}}\right), \tag{14}$$

其中 C_0 为某一常量。

定理 2. 模型近似误差界。对于任意 $\delta \in (0, 1)$, 所有 $\mathbf{x} \in \mathcal{X}$, 下列不等式以 $1-\delta$ 概率成立,

$$|f'(\mathbf{x}) - f(\mathbf{x})| \leq \frac{\sqrt{2}C_0}{l} \epsilon(\delta, D)^{\frac{1}{4}} \left(1 + \left(\frac{\epsilon(\delta, D)}{4}\right)^{\frac{1}{4}}\right), \tag{15}$$

其中, $\epsilon(\delta, D) = \frac{2\ln(2/\delta)}{D} + \sqrt{\frac{2\ln(2/\delta)}{D}}$, C_0 为某一常量。

证明. 根据命题 1 和命题 2 的结果, 对于高斯核 SVM 有 $\tau=1$, 并且 $\|\cdot\|_2 \leq \|\cdot\|_F$, 本定理即可得证。证毕。

定理 2 给出了分别训练基于随机特征的线性 SVM 及对应高斯核 SVM 得到决策函数之间的误差上界。随着随机特征空间维度 D 的增加, 该误差界将变得更小。该定理保证了, 在随机特征空间中训练线性 SVM 得到的近似模型能够被用来评价在核诱导的特征空间中对应精确模型的相对优劣。

3 高斯核 SVM 模型选择

模型选择优化框架外层通过使得泛化误差最小化搜索参数空间, 内层则训练多个核 SVM。本节通过应用随机特征方法提升模型选择优化框架内层效率, 即在优化框架内层利用基于随机特征的线性 SVM 近似核 SVM。

对于模型选择而言, 选择准则没有必要是泛化误差的无偏估计, 近似准则值只要能够区分不同模型的相对优劣选择出最优的模型即可。由定理 2 可知, 在随机特征空间中利用基于随机特征的线性 SVM 近似核 SVM, 得到的近似模型相对于精确模型有界。因此, 能够利用此近似方法计算模型选择准则的近似值, 从而高效地评价原核诱导的特征空间中对应模型的相对优劣。具体描述见算法 2。

算法 2. 基于随机傅里叶特征的近似模型选择算法 AMS-RFF。

输入: 训练样本 $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^l$ 、参数空间 $\mathcal{P}$;
输出: $(\gamma, C)_{\text{opt}}$ 。

- ① 初始化: $(\gamma, C) = (\gamma^0, C^0)$;
- ② repeat
- ③ 利用算法 1 在训练集 $\mathcal{S}$ 上训练学习器, 得到线性决策函数 f ;
- ④ 利用 f 计算模型选择准则 T 的值;
- ⑤ 若 T 的值减小则更新 $(\gamma, C)_{\text{opt}}$;
- ⑥ 在 $\mathcal{P}$ 中更新 (γ, C) ;
- ⑦ until 准则 T 的值最小化;
- ⑧ 输出 $(\gamma, C)_{\text{opt}}$ 。

记 l 为训练样本规模, d 为训练样本维度, D 为随机傅里叶特征空间维度。训练高斯核 SVM 应用高效的 SMO 算法实现^[8, 22], 其计算时间复杂度为

$O(l^2)$. 线性支持向量机求解的时间复杂度为 $O(lD)^{[12]}$. 随机傅里叶特征映射的时间复杂度为 $O(lDd)$.

对于半径间隔界和 Span 界方法,需要在模型选择内层的每一次迭代完整地求解核 SVM. 记 S 为模型选择需要迭代的次数. 标准核 SVM 模型选择时间复杂度为 $O(Sl^2)$, 而算法 2 的时间复杂度为 $O(SdlD)$.

对于 k -折交叉验证方法,记 S_γ, S_C 分别表示参数 γ, C 格搜索的长度. 标准核 SVM 模型选择时间复杂度为 $O(kS_\gamma S_C l^2)$. 算法 1 中,随机特征映射时间复杂度为 $O(S_\gamma l d D)$, 训练线性 SVM 的时间复杂度为 $O(kS_\gamma S_C l D)$, 因此算法 2 总的时间复杂度为 $O((d+kS_C)S_\gamma l D)$.

4 实验结果与分析

交叉验证是一种广泛应用且有效的模型选择方法. 以 k -折交叉验证($k=5, 10$)为例,验证提出算法 AMS-RFF 的可行性及计算效率. 实验中数据集为标准模型选择数据集^[23],如表 1 所示:

Table 1 Specification of Benchmark Datasets
表 1 标准数据集说明

Dataset	Dimension	# Training	# Test	# Repetition
Thyroid	5	140	75	100
Titanic	3	150	2 051	100
Heart	13	170	100	100
Breast	9	200	77	100
Banana	2	400	4 900	100
Ringnorm	20	400	7 000	100
Twonorm	20	400	7 000	100
Waveform	21	400	4 600	100
Diabetes	8	468	300	100
Flare solar	9	666	400	100
German	20	700	300	100
Splice	60	1 000	2 175	20
Image	18	1 300	1 010	20

为了消除数据划分带来的影响,每一个数据集根据训练集与测试集大小随机划分 100 次(对于 Image 和 Splice 各进行 20 次),得到 100 组训练集和测试集. 实验中,首先在训练集上进行模型选择得到最优参数组合,然后利用最优参数组合训练核 SVM,最后在测试集上评价最优参数下的核 SVM 模型. 应用统计分析方法评价模型选择算法.

使用 LIBSVM^[22] 作为高斯核 SVM 的 SMO 算法实现,使用 LIBLINEAR^[12] 训练线性 SVM. 实验环境为 Core i5 个人 PC 机,主频 3.2 GHz,内存 4 GB,算法实现采用 C++ 语言.

核 SVM k -折交叉验证模型选择过程记为 k -CV,与之对应,在模型选择优化框架内层训练基于随机特征的线性 SVM,进行 k -折交叉验证过程记为 Approximate k -CV(简称为 Ak -CV),其中 $k=5, 10$. 交叉验证在 13×11 大小的格搜索空间上进行, $\gamma \in \{2^{-15}, 2^{-13}, \dots, 2^0\}, C \in \{2^{-5}, 2^{-3}, \dots, 2^{15}\}$. 算法 1 中随机特征空间维度统一设置为 $D=30$.

实验结果如表 2 所示,其中,Time 表示交叉验证所花费的时间,Acc 表示利用交叉验证所选择出来的参数 $(\delta, C)_{\text{opt}}$ 重新训练核 SVM 得到的测试准确率. 利用 z 统计量估计不同方法测试准确率的统计显著性^[24]. 记 p, q 分别表示 2 种不同方法的测试准确率均值, e_p, e_q 为对应标准差,那么 z 统计量的计算方法为 $z = (p - q) / \sqrt{e_p^2 + e_q^2}$,其中 $z=1.64$ 对应 95% 的置信度. 仅仅只在数据集 Splice 上, k -CV 显著地优于 Ak -CV. 对于 5-CV,测试准确率差异为 4.1% ($z=3.05$),对于 10-CV 测试准确率差异为 3.4% ($z=1.84$). 另外,根据 Wilcoxon 符号化排序测试(Wilcoxon signed rank test)^[25]可知, k -CV 与 Ak -CV 并没有相互以 95% 的置信度优于另外一种方法($k=5, 10$). 该组结果表明基于随机傅里叶特征的高斯核 SVM 模型选择是可行的.

再看模型选择效率,对于所有 13 个数据集,仅仅在 3 个数据集上 Ak -CV ($k=5, 10$)没有带来效率上的提升,这 3 个数据集(Thyroid, Titanic, Heart)的数据量均小于 200. 对于数据规模较大的数据集, Ak -CV 能带来效率上 10 倍甚至 20 倍的提升. 随着数据量的增加,效率提升更加明显. 这是因为所提出的方法与标准交叉验证方法在小规模数据上效率差异不大,随着数据规模的增加,所提出的方法效率提高越明显.

在数据集 Splice 和 Image 上进行另外一组实验,验证标准 5-折交叉验证(5-CV)与所提出的方法(A5-CV)性能随着训练样本规模的变化情况,实验结果如图 1 和图 2 所示. 随着训练样本规模的变化,2 种模型选择方法选择的模型测试准确率相差不大(如图 1 所示). 在模型选择的效率上, A5-CV 的时间几乎与训练样本规模呈线性关系,而 5-CV 的时间与样本规模呈平方关系,这正好与第 3 节模型选择时间复杂度分析结果一致. 因而,所提出的方法更加适用于大规模核方法模型选择问题.

Table 2 Comparison of Computation Time and Test Accuracy (Acc) of Model Selection among k -CV and Approximate k -CV ($k=5,10$)

表 2 k -折交叉验证、近似 k -折交叉验证模型选择的计算时间与测试准确率比较 ($k=5,10$)

Dataset	5-CV		A5-CV		10-CV		A10-CV	
	Computation Time/s	Acc	Computation Time/s	Acc	Computation Time/s	Acc	Computation Time/s	Acc
Thyroid	3.278	0.951±0.022	6.704	0.949±0.026	8.340	0.949±0.022	18.522	0.949±0.025
Titanic	4.938	0.774±0.009	4.396	0.746±0.044	16.593	0.774±0.008	13.643	0.751±0.040
Heart	3.960	0.834±0.035	3.544	0.812±0.075	9.525	0.838±0.034	9.889	0.818±0.062
Breast	6.719	0.733±0.049	4.329	0.713±0.050	15.073	0.732±0.049	10.370	0.720±0.053
Banana	16.186	0.892±0.008	5.572	0.885±0.013	37.218	0.892±0.007	12.325	0.884±0.013
Ringnorm	9.838	0.982±0.004	3.902	0.968±0.017	20.051	0.982±0.004	7.050	0.967±0.024
Twonorm	8.841	0.975±0.003	3.965	0.970±0.005	17.666	0.974±0.004	7.123	0.968±0.020
Waveform	10.696	0.895±0.008	4.382	0.884±0.015	22.316	0.896±0.008	9.358	0.882±0.024
Diabetes	24.462	0.765±0.020	4.836	0.756±0.028	59.911	0.762±0.017	8.893	0.758±0.023
Flare Solar	24.895	0.672±0.020	6.733	0.646±0.050	58.838	0.669±0.022	11.292	0.642±0.050
German	40.972	0.760±0.023	5.539	0.749±0.030	96.627	0.758±0.023	8.485	0.749±0.033
Splice	83.548	0.886±0.006	5.849	0.845±0.012	187.652	0.886±0.007	7.380	0.852±0.017
Image	66.890	0.968±0.005	8.043	0.944±0.020	152.404	0.967±0.006	13.610	0.946±0.021

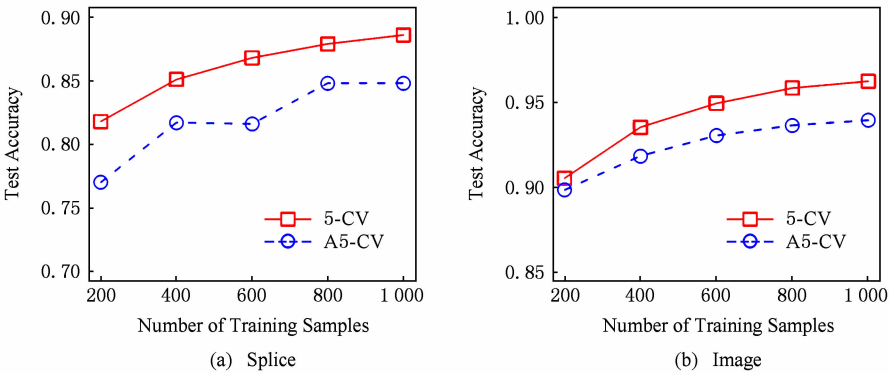


Fig. 1 The test accuracy varies with respect to the number of training samples.

图 1 测试准确率随着训练样本数的变化情况

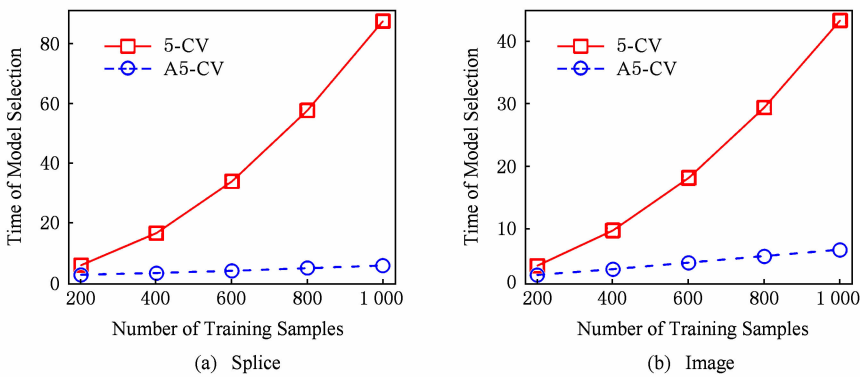


Fig. 2 The time of model selection varies with respect to the number of training samples.

图 2 模型选择时间随着训练样本数的变化情况

5 结束语

核 SVM 优化问题求解的复杂度为 $O(l^3)$, 基于分解策略的高效 SMO 算法求解的复杂度为 $O(l^2)$, 其中 l 为训练样本的个数. 模型选择过程往往需要多次训练核 SVM, 并且精确评估每一个模型, 计算代价过高. 针对这一问题, 提出基于随机傅里叶特征的高斯核模型选择方法. 通过随机傅里叶特征映射将数据映射到显式的随机特征空间, 训练具有 $O(l)$ 复杂度的线性 SVM, 进而可高效计算模型选择准则的近似值, 大幅度提升模型选择效率. 推导了随机特征空间中模型与对应核诱导的特征空间中模型的误差上界, 从理论上保证了可利用随机特征方法评价原特征空间中模型的相对优劣. 所提出的方法适用于大规模核 SVM 的模型选择.

进一步工作考虑随机特征空间维度 D 的选择问题.

参 考 文 献

- [1] Vapnik V N. Statistical Learning Theory [M]. New York: John Wiley & Sons, 1998
- [2] Schölkopf B, Smola A J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond [M]. Cambridge, MA: MIT Press, 2002
- [3] Chapelle O, Vapnik V. Model selection for support vector machines [C] //Advances in Neural Information Processing Systems 12. Cambridge, MA: MIT Press, 2000: 230-236
- [4] Guyon I, Saffari A, Dror G, et al. Model selection: Beyond the Bayesian/frequentist divide [J]. Journal of Machine Learning Research, 2010, 11: 61-87
- [5] Duan K, Keerthi S S, Poo A N. Evaluation of simple performance measures for tuning SVM hyperparameters [J]. Neurocomputing, 2003, 51: 41-59
- [6] Chapelle O, Vapnik V N, Bousquet O, et al. Choosing multiple parameters for support vector machines [J]. Machine Learning, 2002, 46(1/2/3): 131-159
- [7] Vapnik V N, Chapelle O. Bounds on error expectation for support vector machines [J]. Neural Computation, 2000, 12(9): 2013-2036
- [8] Platt J C. Fast Training of support vector machines using sequential minimal optimization [C] //Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999: 185-208
- [9] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms [C] //Proc of the 21st Int Conf on Machine Learning. New York: ACM, 2004: 919-926

- [10] 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
- [11] 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
- [12] 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
- [13] Rahimi A, Recht B. Random features for large-scale kernel machines [C] //Advances in Neural Information Processing Systems 20. Cambridge, MA: MIT Press, 2007: 1177-1184
- [14] Liu Yong, Jiang Shali, Liao Shizhong. Approximate Gaussian kernel for large-scale SVM [J]. Journal of Computer Research and Development, 2014, 51(10): 2171-2177 (in Chinese)
(刘勇, 江沙里, 廖士中. 基于近似高斯核显式描述的大规模 SVM 求解[J]. 计算机研究与发展, 2014, 51(10): 2171-2177)
- [15] Ding L, Liao S. Approximate model selection for large scale LSSVM [C] //Proc of the 3rd Asian Conf on Machine Learning. Brookline, MA: Microtome Publishing, 2011: 165-180
- [16] Ding L, Liao S. Nyström approximate model selection for LSSVM [C] //Proc of the 16th Pacific Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2012: 282-293
- [17] Ding Lizhong, Liao Shizhong. Approximate model selection on regularization path for support vector machines [J]. Journal of Computer Research and Development, 2012, 49(6): 1248-1255 (in Chinese)
(丁立中, 廖士中. 基于正则化路径的支持向量机近似模型选择[J]. 计算机研究与发展, 2012, 49(6): 1248-1255)
- [18] Kimeldorf G S, Wahba G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines [J]. The Annals of Mathematical Statistics, 1970, 41(2): 495-502
- [19] Rudin W. Fourier Analysis on Groups [M]. New York: John Wiley & Sons, 1962
- [20] Chitta R, Jin R, Jain A K. Efficient kernel clustering using random Fourier features [C] //Proc of the 12th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2012: 161-170
- [21] 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

[22] Chang C C, Lin C J. LIBSVM: A library for support vector machines [J]. ACM Trans on Intelligent Systems and Technology, 2011, 2: 27:1-27:27

[23] Ratsch G, Onoda T, Müller K R. Soft margins for AdaBoost [J]. Machine Learning, 2001, 42(3): 287-320

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

[25] Demšar J. Statistical comparisons of classifiers over multiple data sets [J]. Journal of Machine Learning Research, 2006, 7: 1-30



Feng Chang, born in 1989, PhD candidate. Student member of China Computer Federation. His main research interests include kernel methods, model selection and randomized computation.



Liao Shizhong, born in 1964. PhD, professor, and PhD supervisor. Member of China Computer Federation. His main research interests include artificial intelligent and theoretical computer science.