软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2149−2159 [doi: 10.13328/j.cnki.jos.004650]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

支持向量学习的多参数同时调节*

丁立中, 贾 磊, 廖士中

(天津大学 计算机科学与技术学院,天津 300072)

通讯作者: 廖士中, E-mail: szliao@tju.edu.cn, http://cs.tju.edu.cn/faculty/szliao/

摘 要: 模型选择是支持向量学习的关键问题.已有模型选择方法采用嵌套的双层优化框架,内层执行支持向量学习,外层通过最小化泛化误差的估计进行模型选择.该框架过程复杂,计算效率低.简化传统的双层优化框架,提出一个支持向量学习的多参数同时调节方法,在同一优化过程中实现模型选择和学习器训练.首先,将支持向量学习中的参数和超参数合并为一个参数向量,利用序贯无约束极小化技术(sequential unconstrained minimization technique,简称 SUMT)分别改写支持向量分类和回归的有约束优化问题,得到多参数同时调节模型的多元无约束形式定义;然后,证明多参数同时调节模型目标函数的局部 Lipschitz 连续性及水平集有界性.在此基础上,应用变尺度方法(variable metric method,简称 VMM)设计并实现了多参数同时调节算法.进一步地,基于多参数同时调节模型的性质,证明了算法收敛性,对比分析了算法复杂性.最后,实验验证同时调节算法的收敛性,并实验对比同时调节算法的有效性.理论证明和实验分析表明,同时调节方法是一种坚实、高效的支持向量模型选择方法.

关键词: 核方法;支持向量学习;模型选择;参数调节;序贯无约束极小化技术

中图法分类号: TP181

中文引用格式: 丁立中,贾磊,廖士中.支持向量学习的多参数同时调节.软件学报,2014,25(9):2149−2159. http://www.jos.org.cn/1000-9825/4650.htm

英文引用格式: Ding LZ, Jia L, Liao SZ. Simultaneous multiple parameters tuning in support vector learning. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2149−2159 (in Chinese). http://www.jos.org.cn/1000-9825/4650.htm

Simultaneous Multiple Parameters Tuning in Support Vector Learning

DING Li-Zhong, JIA Lei, LIAO Shi-Zhong

(School of Computer Science and Technology, Tianjin University, Tianjin 300072, China)

Corresponding author: LIAO Shi-Zhong, E-mail: szliao@tju.edu.cn, http://cs.tju.edu.cn/faculty/szliao/

Abstract: Model selection is critical to support vector learning. Previous model selection methods mainly adopt a nested two-layer framework, where the inner layer trains the learner and the outer one conducts model selection by minimizing the estimate of the generalization error. Breaking from this framework, this paper proposes an approach of simultaneously tuning multiple parameters of support vector learning, which integrates model selection and learning into one optimization process. It first combines the parameters and hyperparameters involved in support vector learning into one parameter vector. Then, using sequential unconstrained minimization technique (SUMT), it reformulates the constrained optimization problems for support vector classification (SVC) and support vector regression (SVR) as unconstrained optimization problems to give the simultaneous tuning model of SVC and SVR. In addition, it proves the basic properties of the simultaneous tuning model of SVC and SVR, including the local Lipschitz continuity and the boundedness of their level sets. Further, it develops a simultaneous tuning algorithm to iteratively solve simultaneous tuning model. Finally, it proves the convergence of the developed algorithm based on the basic properties of the simultaneous tuning model and provides analysis on complexity of the algorithm as compared with related approaches. The empirical evaluation on benchmark datasets shows that the proposed simultaneous approach has lower running time complexity and exhibits similar predictive performance as existing approaches.

* 基金项目: 国家自然科学基金(61170019)
收稿时间: 2014-04-23; 定稿时间: 2014-05-14

Theoretical and experimental results demonstrate that the simultaneous tuning approach is a sound and efficient model selection approach for support vector learning.

Key words: kernel method; support vector learning; model selection; parameter tuning; SUMT (sequential unconstrained minimization technique)

支持向量学习(support vector learning,简称 SVL)是一类重要的机器学习方法[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.模型选择是支持向量学习的基本问题,对学习的泛化性有着重要影响,包括核函数及其参数的选择、正则化参数的选择以及回归问题中不敏感度参数的选择.典型的,核函数被确定为若干类型,如多项式核、Gaussian 核等.在这种情况下,核函数的选择等价于核参数的调节.本文统称核参数、正则化参数和不敏感度参数为超参数(hyperparameter).支持向量学习的模型选择等价于超参数的调节.已有模型选择方法可概括为一个内外双层的优化框架[3]:内层在超参数固定的情况下,通过凸二次优化训练学习器;外层基于内层的优化结果,通过最小化泛化误差来调节超参数.由于数据的潜在分布未知,泛化误差不可直接计算,可通过经验误差(如交叉验证误差)或理论误差界来估计.

k 折交叉验证可给出泛化误差较优的估计[4],交叉验证的极端形式——留一法(leave-one-out,简称 LOO)能够给出泛化误差几乎无偏的估计[5].然而,基于交叉验证的模型选择方法通常是格搜索整个超参数空间,对每一组候选的超参数向量都进行学习器训练,不可避免地带来了高的计算复杂性[6].为了提高交叉验证效率,Liu 等人利用 BIF(Bouligand influence function)给出了交叉验证的一种高效近似[7].另一方面,为了避免格搜索的低效性,进化计算[8]、基因算法[9]、粒子群算法[10]等被引入,以实现超参数的启发式搜索.最小化泛化误差的理论估计界是另一类模型选择方法.常见的误差界有支持向量张成(span)界[11]、半径间隔界[5]和特征值扰动界[12]等.整体上,无论经验方法还是理论误差界法,均是设计某种策略来约减超参数搜索空间,进而提高模型选择外层的效率.但搜索方向的确定具有较高的计算代价或有效性难以验证.另一方面,支持向量学习凸二次优化求解的复杂性为 $O(l^3)$,多核支持向量学习二阶锥规划求解的复杂性为 $O(Nl^{3.5})$[13],其中,l 为样本规模,N 为候选核矩阵的个数.对于大规模实际问题,若对每个搜索路径上的模型都进行一次学习器训练,计算代价太高.

本文简化传统的双层优化框架,提出了一种支持向量学习的多参数同时调节模型,将超参数的调节与学习器的训练在同一优化过程中实现.首先,分别重写支持向量分类(support vector classification,简称 SVC)和支持向量回归(support vector regression,简称 SVR)的凸二次优化形式,给出 SVC 和 SVR 的多参数优化形式.利用序贯无约束极小化技术(sequential unconstrained minimization technique,简称 SUMT)[14],将 SVC 和 SVR 的多参数优化形式改写为多元无约束优化问题,给出 SVC 和 SVR 多参数同时调节模型的形式定义.然后,证明了多参数同时调节模型目标函数的局部 Lipschitz 连续性及其水平集有界性.在此基础上,应用变尺度方法(variable metric method,简称 VMM)设计并实现了同时调节算法(simultaneous tuning algorithm,简称 STA),该算法较传统参数调节方法具有更低的计算复杂度.进一步证明了算法的收敛性.最后,通过标准数据集上的实验,验证了同时调节算法的收敛性,对比了同时调节算法与其他参数调节方法的有效性.

廖士中等人基于 SVC 的半径间隔界准则,提出了一种超参数和参数的同时调节方法[15].由于半径间隔界针对分类问题定义并不适用于回归,廖士中等人进一步从支持向量回归的优化问题出发,提出了 SVR 的多参数同时调节方法[16].整体而言,这两项工作仅从实验上初步验证了同时调节的可行性和正确性,没有给出理论证明.本文推导了一个新的 SVC 同时调节模型的形式定义,从理论上证明了 SVC 和 SVR 同时调节方法的正确性,细化了同时调节算法,提供了系统的比较实验,给出了一种完备的支持向量学习多参数同时调节方法.

1 支持向量学习

本节简述支持向量分类(support vector classification,简称 SVC)和支持向量回归(support vector regression,简称 SVR).令 $\mathcal{X}$ 表示输入空间,$\mathcal{Y}$ 表示输出域,通常有 $\mathcal{X}\subseteq\mathbb{R}^p$,二分类问题中 $\mathcal{Y}=\{-1,1\}$,回归问题中 $\mathcal{Y}\subseteq\mathbb{R}$.训练集可表示为 $\mathcal{S}=((\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_l,y_l))\in(\mathcal{X}\times\mathcal{Y})^l$,其中,$\boldsymbol{x}_i,y_i$ 为样例输入及对应的标签,l 为训练集规模.本文考虑的核 κ 是从 $\mathcal{X}\times\mathcal{X}$

到$\mathcal{R}$的函数,满足对于任意的有限样本$\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_l\}\subseteq\mathcal{X}$,矩阵 $\boldsymbol{K}=[\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{i,j=1}^{l}$ 是对称半正定的.

1.1 支持向量分类

SVC 的基本思想[1]是:将输入空间的点映射到由核函数κ隐式定义的特征空间,在特征空间中构造最优线性分类超平面.SVC 的分类函数可表示为 $f(\boldsymbol{x})=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i^* y_i\kappa(\boldsymbol{x}_i,\boldsymbol{x})+b\right)$,其中,Lagrange 乘子 α_i^* 是通过求解下述凸优化问题得到的:

$$\begin{cases}\min\ \dfrac{1}{2}\sum\limits_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum\limits_{i=1}^{l}\alpha_i\\ \text{s.t.}\ \sum\limits_{i=1}^{l}y_i\alpha_i=0,\alpha_i\geqslant 0,i=1,\ldots,l\end{cases}\tag{1}$$

SVC 的优化问题公式(1)被称作最大间隔分类器,这一分类器仅适用于特征空间中线性可分的数据.对于非线性可分的情况,需要引入软间隔 SVC[17].2-范数软间隔 SVC 的优化形式可表示为

$$\begin{cases}\min\ \dfrac{1}{2}\sum\limits_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\tilde{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum\limits_{i=1}^{l}\alpha_i\\ \text{s.t.}\ \sum\limits_{i=1}^{l}y_i\alpha_i=0,\alpha_i\geqslant 0,i=1,\ldots,l\end{cases}\tag{2}$$

公式(2)与公式(1)的不同在于核函数形式的不同,公式(2)中的核函数 $\tilde{\kappa}$ 称为修正核函数,其与核函数κ的关系为

$$\tilde{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\frac{1}{C}\delta_{ij}\tag{3}$$

其中,C 是正则化参数;δ_{ij} 为 Kronecker 函数,如果 $i=j$,则$\delta_{ij}=1$,否则$\delta_{ij}=0$.正则化参数 C 可看做是修正核 $\tilde{\kappa}$ 的参数.那么, $\tilde{\kappa}$ 的核参数向量可表示为 $\boldsymbol{\Theta}=(C,\theta_1,\ldots,\theta_d)^T\in\mathbb{R}_+^{d+1}$,其中,$\theta_1,\ldots,\theta_d$ 是核κ的参数,$\mathbb{R}_+$表示正实数.以 Gaussian 核 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i,\boldsymbol{x}_j\|^2/2\sigma^2)$ 为例, $\tilde{\kappa}$ 的核参数向量为$\boldsymbol{\Theta}=(C,\sigma)^T$.

1.2 支持向量回归

基于平方ε不敏感损失[1]的 SVR 优化问题可表示为

$$\begin{cases}\min\ \|\boldsymbol{w}\|^2+C\sum\limits_{i=1}^{l}(\xi_i^2+\hat{\xi}_i^2)\\ \text{s.t.}\ (\langle\boldsymbol{w}\cdot\boldsymbol{x}_i\rangle+b)-y_i\leqslant\varepsilon+\xi_i,i=1,\ldots,l\\ \qquad y_i-(\langle\boldsymbol{w}\cdot\boldsymbol{x}_i\rangle+b)\leqslant\varepsilon+\hat{\xi}_i,i=1,\ldots,l\\ \qquad \xi_i,\hat{\xi}_i\geqslant 0,i=1,\ldots,l\end{cases}\tag{4}$$

其中,ε为不敏感度参数,$\langle\boldsymbol{w}\cdot\boldsymbol{x}_i\rangle$表示 $\boldsymbol{w}$ 与 $\boldsymbol{x}_i$ 的点积,ξ_i 和 $\hat{\xi}_i$ 为松弛变量.

优化问题(4)的核化对偶形式为

$$\begin{cases}\min\ \dfrac{1}{2}\sum\limits_{i,j=1}^{l}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)\tilde{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum\limits_{i=1}^{l}y_i(\hat{\alpha}_i-\alpha_i)+\varepsilon\sum\limits_{i=1}^{l}(\hat{\alpha}_i+\alpha_i)\\ \text{s.t.}\ \sum\limits_{i=1}^{l}(\hat{\alpha}_i-\alpha_i)=0,\hat{\alpha}_i\geqslant 0,\alpha_i\geqslant 0,i=1,\ldots,l\end{cases}\tag{5}$$

其中, $\hat{\alpha}_i$,α_i 为 Lagrange 乘子, $\tilde{\kappa}$ 为修正核函数,同 SVC.

2 多参数同时调节模型

本节将 SVC 和 SVR 中的 Lagrange 乘子和超参数向量合并,重写对应优化问题,给出 SVC 和 SVR 的多参数表示形式;然后,利用序贯无约束极小化技术(SUMT)推导出 SVC 和 SVR 多参数同时调节模型的形式定义.

首先简述 SUMT.给定如下有约束优化问题:

$$\begin{cases}\min\ t(\boldsymbol{x})\\ \text{s.t. } g_i(\boldsymbol{x})=0, i=1,...,m\\ \quad h_j(\boldsymbol{x})\geqslant 0, j=1,...,p\end{cases} \tag{6}$$

公式(6)的解可通过求解一组无约束优化问题来逼近[14]:

$$\min\ J_k(\boldsymbol{x})=t(\boldsymbol{x})+\frac{1}{r_k}\sum_{i=1}^{m}g_i(\boldsymbol{x})^2+r_k\sum_{j=1}^{p}\frac{1}{h_j(\boldsymbol{x})} \tag{7}$$

其中,r_k称为障碍因子序列,满足$\{r_k|r_0>0, r_{k+1}=\beta r_k, 0<\beta<1\}$.当 $k\to\infty$,问题(7)的解将趋近于原问题(6)的解.

2.1 SVC多参数同时调节模型

本节推导 SVC 多参数同时调节模型的形式定义.令:

$$\begin{cases}\boldsymbol{\alpha}=(\alpha_1,...,\alpha_l)^T\\ \boldsymbol{X}=(\boldsymbol{\Theta}^T,\boldsymbol{\alpha}^T)^T\in\mathbb{R}^{d+l+1}\\ \boldsymbol{Y}=(y_1,...,y_l)^T\in\mathbb{R}^l\\ \tilde{\boldsymbol{Y}}=\left(\underbrace{0,...,0}_{d+1},\boldsymbol{Y}^T\right)^T\in\mathbb{R}^{d+l+1}\end{cases} \tag{8}$$

其中,$\boldsymbol{X}$将作为新的优化变量.

基于公式(8)中定义的新变量,重写 SVC 优化问题公式(2),可得:

$$\begin{cases}\min\ \frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\tilde{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i\\ \text{s.t. } \boldsymbol{X}^T\tilde{\boldsymbol{Y}}=0, \boldsymbol{X}\geqslant 0\end{cases} \tag{9}$$

利用 SUMT 将公式(9)表示为关于参数 r_k的无约束优化问题:

$$\min\ J_k^{SVC}(\boldsymbol{X})=\frac{1}{2}\sum_{i,j=1}^{l}y_iy_j\alpha_i\alpha_j\tilde{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{l}\alpha_i+\frac{1}{r_k}(\boldsymbol{X}^T\tilde{\boldsymbol{Y}})^2+r_k\sum_{i=1}^{d+l+1}\frac{1}{\boldsymbol{e}_i^T\boldsymbol{X}} \tag{10}$$

其中,$\boldsymbol{e}_i$表示第 i 个元素为 1 其余元素为 0 的单位列向量.公式(10)为 SVC 多参数同时调节模型的形式定义.

同时调节模型公式(10)是多变元无约束优化问题,求解得到的$\boldsymbol{X}$的最优解$\boldsymbol{X}^*$,同时包含了 Lagrange 乘子α^*、正则化参数 C^*和核函数参数 $\theta_1^*,...,\theta_d^*$的最优解.

2.2 SVR多参数同时调节模型

本节推导 SVR 多参数同时调节模型的形式定义.令:

$$\begin{cases}\boldsymbol{\alpha}_A=(\alpha_1,...,\alpha_l)^T\\ \boldsymbol{\alpha}_B=(\hat{\alpha}_1,...,\hat{\alpha}_l)^T\\ \boldsymbol{X}=(\boldsymbol{\Theta}^T,\varepsilon,\boldsymbol{\alpha}_A^T,\boldsymbol{\alpha}_B^T)^T\in\mathbb{R}^{d+2l+2}\\ \boldsymbol{Y}_1=(0,...,0)^T\in\mathbb{R}^{d+2}\\ \boldsymbol{Y}_2=-\boldsymbol{Y}_3=(1,...,1)^T\in\mathbb{R}^l\\ \tilde{\boldsymbol{Y}}=(\boldsymbol{Y}_1^T,\boldsymbol{Y}_2^T,\boldsymbol{Y}_3^T)^T\in\mathbb{R}^{d+2l+2}\end{cases} \tag{11}$$

利用公式(11)中定义的新变量,重写 SVR 优化问题公式(5),可得:

$$\begin{cases}\min\ \frac{1}{2}\sum_{i,j=1}^{l}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)\tilde{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{l}y_i(\hat{\alpha}_i-\alpha_i)+\varepsilon\sum_{i=1}^{l}(\hat{\alpha}_i+\alpha_i)\\ \text{s.t. } \boldsymbol{X}^T\tilde{\boldsymbol{Y}}=\boldsymbol{0}, \boldsymbol{X}\geqslant\boldsymbol{0}\end{cases} \tag{12}$$

利用 SUMT 重写公式(12),可得 SVR 多参数同时调节模型的形式定义:

$$\min\ J_k^{SVR}(\boldsymbol{X})=\frac{1}{2}\sum_{i,j=1}^{l}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)\tilde{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{i=1}^{l}y_i(\hat{\alpha}_i-\alpha_i)+\varepsilon\sum_{i=1}^{l}(\hat{\alpha}_i+\alpha_i)+\frac{1}{r_k}(\boldsymbol{X}^T\tilde{\boldsymbol{Y}})^2+r_k\sum_{i=1}^{d+2l+2}\frac{1}{\boldsymbol{e}_i^T\boldsymbol{X}} \tag{13}$$

为了表述清晰,将文中的重要符号记录于表 1.

Table 1 Table of notations

表 1 符号表

$\mathcal{S}$	训练集$((\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_l,y_l))\in(\mathcal{X}\times\mathcal{Y})^l$,其中,$\mathcal{X}$和$\mathcal{Y}$分别为输入输出空间
f	预测函数 $f:\mathcal{X}\to\mathcal{Y}$
$\boldsymbol{K}$	核矩阵 $\boldsymbol{K}=[\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{i,j=1}^{l}$,其中,$\kappa$为从$\mathcal{X}\times\mathcal{X}$到$\mathbb{R}$的核函数
$\tilde{\kappa}$	修正核函数
$\boldsymbol{\Theta}$	修正核的参数 $\boldsymbol{\Theta}=(C,\theta_1,\ldots,\theta_d)^T\in\mathbb{R}_+^{d+1}$,其中,$C$ 为正则化参数,$\theta_1,\ldots,\theta_d$ 为核κ的参数
ε	SVR 不敏感度参数
$\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_A$, $\boldsymbol{\alpha}_B$	Lagrange 乘子向量
r_k	障碍因子序列$\{r_k\mid r_0>0,r_{k+1}=\beta r_k,0<\beta<1\}$
$\boldsymbol{X}$	多参数同时调节模型的优化变量,包括 Lagrange 乘子向量和超参数
$\tilde{\boldsymbol{Y}}$	标签向量
J_k^{SVC}, J_k^{SVR}	SVC 与 SVR 多参数同时调节模型的目标函数
$k_{ij}^{\theta_t}$	核矩阵元素 $\boldsymbol{K}_{ij}$ 对核参数θ_t的偏导数
η, λ, ζ, τ, ς, ψ	超参数和 Lagrange 乘子取值的上下确界
$\Omega_1,\ldots,\Omega_6$	同时调节模型目标函数偏导的上下界

3 同时调节模型的基本性质

本节研究 SVC 和 SVR 多参数同时调节模型 J_k^{SVC} 和 J_k^{SVR} 的基本性质,包括 J_k^{SVC} 和 J_k^{SVR} 的局部 Lipschitz 连续性及其水平集的有界性.这些性质对于分析多参数同时调节模型求解算法的收敛性具有重要作用.

3.1 J_k^{SVC} 的基本性质

本节中,简记 J_k^{SVC} 为 J_k.J_k 的梯度可表示为 $\nabla J_k(\boldsymbol{X})=\left(\frac{\partial J_k}{\partial C},\frac{\partial J_k}{\partial\theta_1},\ldots,\frac{\partial J_k}{\partial\theta_d},\ldots,\frac{\partial J_k}{\partial\alpha_1},\ldots,\frac{\partial J_k}{\partial\alpha_l}\right)^T$.基于 SVC 多参数同时调节模型的形式定义公式(10),可求得$\nabla J_k(\boldsymbol{X})$中的各个分量:

$$\begin{cases}\dfrac{\partial J_k}{\partial C}=-\dfrac{1}{2C^2}\sum\limits_{i,j}\alpha_i\alpha_j y_i y_j-\dfrac{r^k}{C^2}\\ \dfrac{\partial J_k}{\partial\theta_t}=\dfrac{1}{2}\sum\limits_{i,j}\alpha_i\alpha_j y_i y_j k_{ij}^{\theta_t}-\dfrac{r^k}{\theta_t^2},t=1,\ldots,d\\ \dfrac{\partial J_k}{\partial\alpha_i}=\dfrac{1}{2}\sum\limits_{j}\alpha_j y_i y_j\left(\boldsymbol{K}_{ij}+\dfrac{\delta_{ij}}{C}\right)-1+\dfrac{2\alpha_i}{r^k}-\dfrac{r^k}{\alpha_i^2},i=1,\ldots,l\end{cases} \tag{14}$$

其中,$\boldsymbol{K}_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$表示核矩阵 $\boldsymbol{K}$ 的 ij 元素,$k_{ij}^{\theta_t}=\dfrac{\partial\boldsymbol{K}_{ij}}{\partial\theta_t}$ 表示核矩阵元素对核参数的导数.为了便于分析,对公式(14)中的参数取值范围做出假设.设 $\inf\{C\}=\eta>0$,$\sup\{C\}=\lambda>0$;$\inf\{\theta_t\}=\zeta>0$,$\sup\{\theta_t\}=\tau>0$.如果$\alpha_i=0$,则该α_i对 J_k 的函数值无影响.因此设 $\inf\{\alpha_i\}=\varsigma>0$,$\sup\{\alpha_i\}=\psi>0$.记优化变量 $\boldsymbol{X}$ 所属域为$\mathcal{D}$,满足 $\mathcal{D}\subseteq\mathbb{R}_+^{d+l+1}$.令 $p=\arg\max\limits_{\zeta\leqslant\theta_t\leqslant\tau}k_{ij}^{\theta_t}$,$\kappa_{\max}$ 为核矩阵 $\boldsymbol{K}$ 中的最大元素.利用上述参数的取值限定,公式(14)中各个偏导的上下界可表述为公式(15), 基于公式(15)的结果,可证明引理 1.

$$\left.\begin{aligned}
&\frac{\partial J_k}{\partial C}\geqslant -\frac{l^2\psi^2}{2\eta^2}-\frac{r_k}{\eta^2}\overset{def}{=}\Omega_1, && \frac{\partial J_k}{\partial C}\leqslant \frac{l^2\psi^2}{2\eta^2}-\frac{r_k}{\lambda^2}\overset{def}{=}\Omega_2,\\
&\frac{\partial J_k}{\partial \theta_t}\geqslant -\frac{l^2\psi^2}{2}k_{ij}^p-\frac{r_k}{\zeta^2}\overset{def}{=}\Omega_3, && \frac{\partial J_k}{\partial \theta_t}\leqslant \frac{l^2\psi^2}{2}k_{ij}^p-\frac{r_k}{\tau^2}\overset{def}{=}\Omega_4,\\
&\frac{\partial J_k}{\partial \alpha_i}\geqslant -\frac{l\psi}{2}\left(\kappa_{\max}+\frac{1}{\eta}\right)-1+\frac{2\varsigma}{r_k}-\frac{r_k}{\varsigma^2}\overset{def}{=}\Omega_5, && \frac{\partial J_k}{\partial \alpha_i}\leqslant \frac{l\psi}{2}\left(\kappa_{\max}+\frac{1}{\eta}\right)-1+\frac{2\psi}{r_k}-\frac{r_k}{\psi^2}\overset{def}{=}\Omega_6
\end{aligned}\right\} \tag{15}$$

引理 1. J_k是局部 Lipschitz 连续的.

证明:令 $M\overset{\text{def}}{=}\max\{|\Omega_1|,...,|\Omega_6|\}$,其中,$\Omega_1,\ldots,\Omega_6$的定义见公式(15).对于任意的 $\boldsymbol{X}\in\mathcal{D}\subseteq\mathbb{R}_+^{d+l+1}$,可得:

$$\|\nabla J_k(\boldsymbol{X})\|\leqslant M\sqrt{l+d+1}\overset{\text{def}}{=}L \tag{16}$$

因为 $J_k(\boldsymbol{X})$是 $\mathbb{R}_+^{l+d+1}$ 上的连续函数,利用 Lagrange 中值定理可得:对于任意的 $\boldsymbol{u},\boldsymbol{v}\in\mathcal{D}$,都存在$\rho\in(0,1)$,使得:

$$\nabla J_k(\boldsymbol{u}+\rho(\boldsymbol{v}-\boldsymbol{u}))=\frac{J_k(\boldsymbol{u})-J_k(\boldsymbol{v})}{\boldsymbol{u}-\boldsymbol{v}} \tag{17}$$

结合公式(16)与公式(17),有:

$$\left\|\frac{J_k(\boldsymbol{u})-J_k(\boldsymbol{v})}{\boldsymbol{u}-\boldsymbol{v}}\right\|=\|\nabla J_k(\boldsymbol{u}+\rho(\boldsymbol{v}-\boldsymbol{u}))\|\leqslant L.$$

那么,$|J_k(\boldsymbol{u})-J_k(\boldsymbol{u})|\leqslant L\|\boldsymbol{u}-\boldsymbol{v}\|$.因此,$J_k$是局部 Lipschitz 连续的. □

下一节将给出求解 J_k的同时调节算法,该算法是一个迭代算法.下面讨论中,设优化迭代的初始值为$(\boldsymbol{X}_0,r_0)$,其中,$\boldsymbol{X}_0=(\mathbf{1}^T,\mathbf{0}^T)^T$,也就是 $C=\theta_1=\ldots=\theta_t=1$ 且$\alpha_1=\ldots=\alpha_l=0$;$r_0$ 表示障碍因子 r_k的初始值.

为了方便描述,将 $J_k(\boldsymbol{X})$重记为 $J(\boldsymbol{X},r_k)$.下面引理表述了 $J(\boldsymbol{X},r_k)$的水平集有界性.

引理 2. 水平集$\mathcal{L}=\{\boldsymbol{X}|J(\boldsymbol{X},r_k)\leqslant J(\boldsymbol{X}_0,r_0)\}$是有界的.

证明:将$(\boldsymbol{X}_0,r_0)$带入公式(10),可得 $J(\boldsymbol{X}_0,r_0)=r_0(d+1)$.多参数同时调节模型是最小化过程,故 $r_0(d+1)$可看做$J(\boldsymbol{X},r_k)$的上界.基于各参数设定的取值范围可知:

$$J(\boldsymbol{X},r_k)\geqslant -\frac{l^2\psi^2}{2}\left(\kappa_{\max}+\frac{1}{\eta}\right)-l\psi+r_k\left(\frac{l}{\psi}+\frac{d}{\tau}+\frac{1}{\lambda}\right).$$

因此,$J(\boldsymbol{X},r_k)$有上界和下界.因 $J(\boldsymbol{X},r_k)$是关于 $\boldsymbol{X}$ 的连续函数,故 $J(\boldsymbol{X},r_k)$的上下界将约束 $\boldsymbol{X}$ 在一定的域内. □

3.2 J_k^{SVR} 的基本性质

假设对于 $i=1,\ldots,l$, $\inf\{\hat{\alpha}_i\}=\varsigma,\sup\{\hat{\alpha}_i\}=\psi$.

设 $\boldsymbol{X}_0=(\mathbf{1}^T,\mathbf{0}^T,\mathbf{0}^T)^T$,即,$C=\theta_1=\ldots=\theta_t=\varepsilon=1,\alpha_1=\ldots=\alpha_l=0,\hat{\alpha}_1=...=\hat{\alpha}_l=0$.

采用与 SVC 类似的证明方式,可得如下引理:

引理 3. J_k^{SVR} 是局部 Lipschitz 连续的.

引理 4. 水平集$\mathcal{L}=\{\boldsymbol{X}|J^{SVR}(\boldsymbol{X},r_k)\leqslant J^{SVR}(\boldsymbol{X}_0,r_0)\}$有界.

4 同时调节算法与分析

本节给出同时调节模型的求解算法、理论分析算法收敛性和复杂性,并与已有参数调节方法进行对比.

4.1 算 法

多参数同时调节模型的求解算法见算法 1.算法是基于 SVC 描述的,基本过程同样适用于 SVR.

算法 1. Simultaneous Tuning Algorithm.

Require: $\boldsymbol{X}_0\in\mathbb{R}^{l+d+1}$, r_0, $\varepsilon>0$, $\boldsymbol{H}_0=\boldsymbol{I}$, $k=0$;

1. **while** $\|\boldsymbol{g}_k=\nabla J(\boldsymbol{X}_k,r_k)\|>\varepsilon$ **do**
2. $\lambda_k=\arg\min_{\lambda} J(\boldsymbol{X}_k+\lambda\,\boldsymbol{p}_k,r_k)$;

3. $\boldsymbol{p}_k=-\boldsymbol{H}_k\boldsymbol{g}_k$;
4. $\boldsymbol{X}_{k+1}=\boldsymbol{X}_k+\lambda_k\boldsymbol{p}_k$;
5. **if** $k=l+d+1$ **then**
6. $\boldsymbol{X}_0=\boldsymbol{X}_k$, $k=0$;
7. **else**
8. $\Delta\boldsymbol{g}_k=\nabla J(\boldsymbol{X}_{k+1},r_{k+1})-\nabla J(\boldsymbol{X}_k,r_k)$;
9. $\Delta\boldsymbol{X}_k=\boldsymbol{X}_{k+1}-\boldsymbol{X}_k$;
10. $\boldsymbol{s}_k=\boldsymbol{H}_k\Delta\boldsymbol{g}_k$;
11. $\mu_k=1/(\boldsymbol{s}_k)^T\Delta\boldsymbol{g}_k$;
12. $\varphi_k=1/(\Delta\boldsymbol{X}_k)^T\Delta\boldsymbol{g}_k$;
13. $\boldsymbol{C}_k=\mu_k\boldsymbol{s}_k(\boldsymbol{s}_k)^T$;
14. $\boldsymbol{B}_k=\varphi_k\Delta\boldsymbol{X}_k(\Delta\boldsymbol{X}_k)^T$;
15. $\boldsymbol{H}_{k+1}=\boldsymbol{H}_k+\boldsymbol{B}_k-\boldsymbol{C}_k$;
16. $k=k+1$;
17. **end if**
18. **end while**
19. **return** $\boldsymbol{X}^*=\boldsymbol{X}_k$

同时调节算法应用了变尺度方法(variable metric method,简称 VMM)[18],采用梯度下降的方式最小化目标函数 J.在第 k 步迭代中,更新方向为 $\boldsymbol{p}_k$,计算 $\boldsymbol{p}_k$ 要用到 $\boldsymbol{H}_k$ 和 $\boldsymbol{g}_k$,$\boldsymbol{H}_k$ 为逆 Hessian 阵的近似形式,初始化为单位矩阵 $\boldsymbol{I}$;$\boldsymbol{g}_k$ 为当前目标函数梯度值.优化变元 $\boldsymbol{X}_k$ 的更新步长为 λ_k,λ_k 通过线性规划得到.更新完 $\boldsymbol{X}_k$ 后,需计算 $\boldsymbol{H}_{k+1}$ 的值,用于下次迭代过程.若算法在第 k 步终止,则返回最优参数向量 $\boldsymbol{X}^*=\boldsymbol{X}_k$.

4.2 收敛性

本节分析算法 1 的收敛性.Vlček 和 Lukšan 分析了一般 VMM 的收敛性[18]:如果无约束目标函数 $f(x)$是局部 Lipschitz 连续的且水平集$\{x|f(x)\leqslant f(x_0)\}$有界,那么 VMM 方法是收敛的.现给出如下定理:

定理 1. 算法 1 是收敛的.

证明:引理 1~引理 4 表明,SVC 和 SVR 多参数同时调节模型的目标函数满足局部 Lipschitz 连续性及水平集有界性.由文献[18]的引理 3.4 可知,算法 1 中,序列$\{\boldsymbol{g}_k\}$是有界的.另外,存在点 $\hat{\boldsymbol{X}}$ 和一个无限集合$\Xi\subset\{0,1,2,\ldots\}$满足 $\boldsymbol{X}_k\xrightarrow{\Xi}\hat{\boldsymbol{X}}$,$\Delta J_k=J(\boldsymbol{X}_{k+1},r_{k+1})-J(\boldsymbol{X}_k,r_k)\xrightarrow{\Xi}0$,使得 $0\in\partial J(\hat{\boldsymbol{X}},r_k)$.这意味着 $\hat{\boldsymbol{X}}$ 是目标函数的一个驻点.由文献[18]的引理 3.6 可知:如果迭代步数有限且最后的下降步出现在第 k 次迭代,那么点 $\boldsymbol{X}_{k+1}$ 为 J 的驻点. □

4.3 复杂性对比

传统支持向量学习的模型选择方法通过最小化泛化误差的经验或理论估计来进行模型选择,经验估计包括交叉验证误差或 LOO 误差,理论估计包括半径间隔界和 span 界[5,11]等.记泛化误差的某种估计为 $\Upsilon_{\boldsymbol{\alpha},\boldsymbol{\Theta}}$,支持向量学习的目标函数为 $G(\boldsymbol{\alpha},\boldsymbol{\Theta})$.传统模型选择方法采用的双层优化框架如算法 2 所示:内层固定超参数 $\boldsymbol{\Theta}_k$ 通过最小化 $G(\boldsymbol{\alpha},\boldsymbol{\Theta}_k)$计算参数 $\boldsymbol{\alpha}_k$;外层利用内层计算的结果 $\boldsymbol{\alpha}_k$,通过最小化 $\Upsilon_{\boldsymbol{\alpha}_k,\boldsymbol{\Theta}}$ 计算超参数 $\boldsymbol{\Theta}_{k+1}$.这类方法需要在内层进行多次学习器训练,以迭代地得到 $\Upsilon_{\boldsymbol{\alpha},\boldsymbol{\Theta}}$的最小值.利用二次规划求解 SVC 的复杂度为 $O(l^3)$,若优化 $\Upsilon_{\boldsymbol{\alpha},\boldsymbol{\Theta}}$所需的迭代步数为 S,则总的计算复杂度为 $O(Sl^3)$.

算法 2. Traditional Model Selection Framework.

1. Initialize $\boldsymbol{\alpha}$, $\boldsymbol{\Theta}^0$, $k=0$;
2. **repeat**
3. $\boldsymbol{\alpha}_k=\arg\min G(\boldsymbol{\alpha},\boldsymbol{\Theta}_k)$;
4. Calculate $\boldsymbol{\Theta}_{k+1}$ by minimizing $\Upsilon_{\boldsymbol{\alpha}_k,\boldsymbol{\Theta}}$;

5. **until** $Y_{\alpha,\Theta}$ is minimized

6. **return** α_k, Θ_k

多参数同时调节算法(算法 1)中,因优化变量 $\boldsymbol{X}$ 包括参数向量α和超参数向量Θ,可实现多参数在同一优化过程中同时调节.算法的主要计算代价源自 $\boldsymbol{H}_{k+1}$ 的计算.每次迭代的计算复杂度为 $O((l+d+1)^2)$.对于一般的核函数,如 Gaussian 核和多项式核,核参数个数远小于样本规模,即 $d<<l$,所以可知 $O((l+d+1)^2)\approx O(l^2)$.令 S'为迭代步数,则总的计算复杂度为 $O(S'l^2)$.

5 实验结果与分析

本节首先实验验证多参数同时调节算法的收敛性,然后实验对比多参数调节算法(simultaneous tuning algorithm,简称 STA)与其他经典的参数调节方法的有效性.对比方法包括 5 折交叉验证(5-fold cross validation,简称 5-fold CV)、半径间隔界(radius margin bound,简称 RMB)和 span 界.

5.1 数据及实验设置

实验数据选自 UCI 机器学习数据库\StatLog 数据库\Delve 数据库,详见表 2.每个数据集按照 7:3 随机分割为训练集和测试集.为了避免随机性的影响,所有实验重复 10 次.采用的核函数为 Gaussian 核.

Table 2 Datasets used in our experiments

表 2 实验数据集

数据集	特征数	样本数
Heart	13	270
Breast cancer	10	683
Diabetes	8	768
Titanic	103	1 313
Thyroid	21	7 200
A2a	123	32 561
W1a	300	49 749

5.2 收敛性

本节验证多参数同时调节算法的收敛性.研究目标函数值 J 随着迭代次数增加的变化规律,以及通过最小化 J 选择出的最优参数的测试误差随着迭代次数增加的变化规律.实验结果如图 1 和图 2 所示.可以发现:随着迭代次数的增加,目标函数值 J 呈现出收敛的趋势;最优参数的测试误差逐步减小并趋于稳定.

5.3 有效性

本节实验对比同时调节算法与其他参数调节方法的有效性,有效性包括泛化性和效率.泛化性由测试集上的平均测试误差来评估,效率由进行模型选择的平均计算时间来评估.应用不同的模型选择方法在训练集上进行模型选择得到最优超参数,然后在测试集上计算测试误差评估最优超参数的性能.每个数据集重复随机分割 10 次进行实验,得到的平均测试误差及标准差见表 3.利用 5%显著性 t 检验对实验结果进行分析.在前 6 个数据集上,4 个方法的测试误差没有显著不同;在 W1a 数据集上,同时调节算法和 span 界的测试误差显著低于 5 折交叉验证和 RMB.值得指出的是,span 界存在局部最小问题且难以实现[5],故不常采用.计算效率方面,所有方法的效率都明显高于 5 折交叉验证.同时调节算法的效率高于 RMB 和 span 界.另外,训练集规模越大,调节算法的效率优势越明显.因此,综合考虑效率和泛化性,同时调节算法的有效性优于其他参数调节方法.

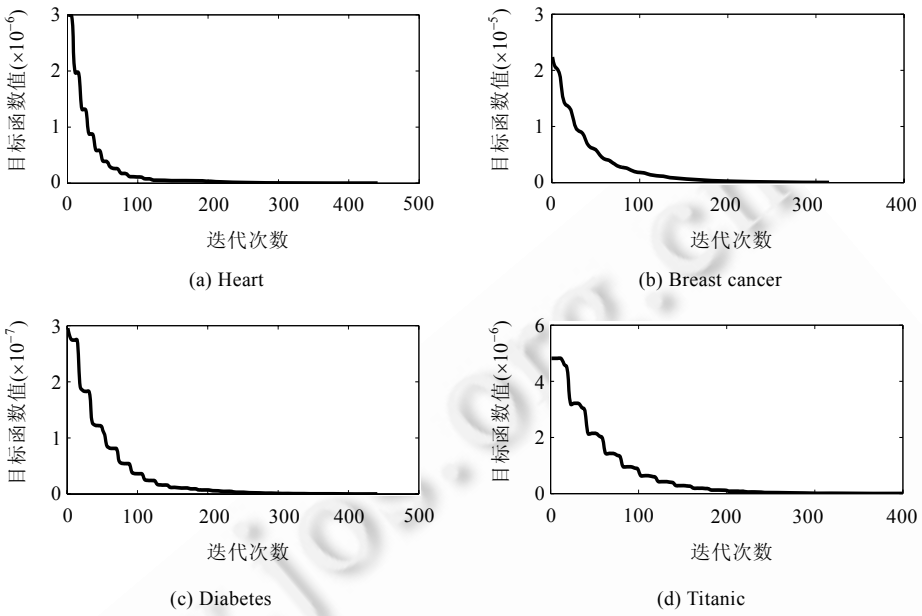

Fig.1 Evolution of the values of optimization objective J as iterations increase

图 1 目标函数 J 随着迭代次数增加的变化曲线

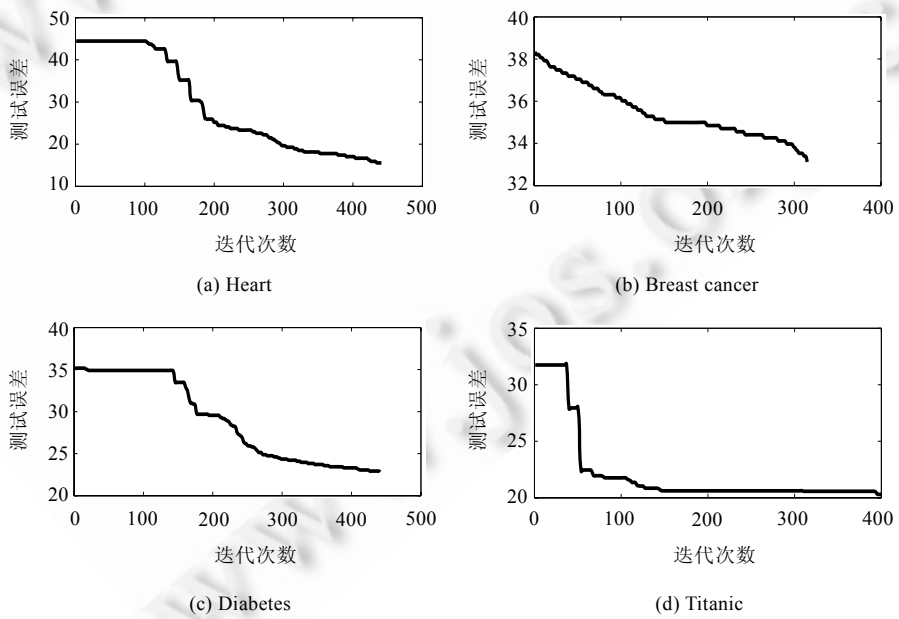

Fig.2 Evolution of the test errors as iterations increase

图 2 测试误差随着迭代次数增加的变化曲线

Table 3 Comparison of running time and test errors of different approaches

表 3 不同参数调节方法运行时间和测试误差的比较

数据集	度量指标	5-fold CV	RMB	Span	STA
Heart	测试误差(%)	15.95±3.26	15.92±3.18	16.13±3.11	16.02±2.35
	运行时间(s)	160.3	12.9	12.0	11.6
Breast cancer	测试误差(%)	26.04±4.74	26.84±4.71	25.59±4.18	26.64±4.15
	运行时间(s)	237.5	16.7	13.3	12.9
Diabetes	测试误差(%)	23.53±1.73	23.25±1.70	23.19±1.67	23.23±1.28
	运行时间(s)	330.0	28.1	26.5	23.6
Titanic	测试误差(%)	22.82±1.02	22.88±1.23	22.50±0.88	22.81±1.13
	运行时间(s)	965.0	19.1	16.5	11.9
Thyroid	测试误差(%)	4.80±2.19	4.63±2.03	4.56±1.97	4.63±1.85
	运行时间(s)	1802.5	20.8	51.8	13.6
A2a	测试误差(%)	18.24±2.72	16.03±3.11	15.66±3.43	15.44±3.23
	运行时间(s)	8151.6	126.8	115.7	56.3
W1a	测试误差(%)	2.97±0.39	2.98±0.28	2.34±0.58	2.17±0.43
	运行时间(s)	12455.0	174.7	400.3	59.8

6 结 语

本文提出一种支持向量学习的多参数同时调节方法,简化传统的双层迭代框架,为求解支持向量学习的模型选择问题提供了一个新的范型.给出 SVC 和 SVR 多参数同时调节模型的形式定义,理论分析模型的局部 Lipschitz 连续性及其水平集的有界性.设计并实现多参数同时调节算法,证明算法收敛性并对比分析算法复杂性.标准数据集上的实验结果表明,同时调节算法的有效性优于其他参数调节方法.理论证明和实验分析表明,多参数同时调节方法是坚实、高效的支持向量学习模型选择方法.

多参数同时调节模型适用于处理核参数个数较多的情况,因此,进一步工作考虑将这一模型扩展到更复杂情况,包括多核[13,19,20]和超核[21].

References:

[1] Vapnik V. The Nature of Statistical Learning Theory. 2nd ed., New York: Springer-Verlag, 2000.

[2] Ding SF, Huang HJ, Shi ZZ. Weighted smooth CHKS twin support vector machines. Ruan Jian Xue Bao/Journal of Software, 2013, 24(11):2548−2557 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4475.htm [doi: 10.3724/SP.J.1001.2013.04475]

[3] Guyon I, Saffari A, Dror G. Model selection: Beyond the Bayesian/frequentist divide. Journal of Machine Learning Research, 2010, 11:61−87.

[4] Duan K, Keerthi S, Poo A. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing, 2003, 51:41−59. [doi: 10.1016/S0925-2312(02)00601-X]

[5] Chapelle O, Vapnik V, Bousquet O. Choosing multiple parameters for support vector machines. Machine Learning, 2002,46(1-3): 131−159. [doi: 10.1023/A:1012450327387]

[6] Xu Z, Dai M, Meng D. Fast and efficient strategies for model selection of Gaussian support vector machine. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009,39(5):1292−1307. [doi: 10.1109/TSMCB.2009.2015672]

[7] Liu Y, Jiang SL, Liao SZ. Efficient approximation of cross-validation for kernel methods using Bouligand influence function. In: Xing EP, Jebara T, eds. Proc. of the 31st Int'l Conf. on Machine Learning. New York: ACM Press, 2014. 324−332.

[8] Friedrichs F, Igel C. Evolutionary tuning of multiple SVM parameters. Neurocomputing, 2005,64:107−117. [doi: 10.1016/j.neucom.2004.11.022]

[9] Huang C, Wang C. A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications, 2006,31(2):231−240. [doi: 10.1016/j.eswa.2005.09.024]

[10] Guo XC, Yang JH, Wu CG, Wang CY, Liang YC. A novel LS-SVMs hyper-parameter selection based on particle swarm optimization. Neurocomputing, 2008,71(16):3211−3215. [doi: 10.1016/j.neucom.2008.04.027]

[11] Vapnik V, Chapelle O. Bounds on error expectation for support vector machines. Neural Computation, 2000,12(9):2013−2036. [doi: 10.1162/089976600300015042]

[12] Liu Y, Jiang SL, Liao SZ. Eigenvalues perturbation of integral operator for kernel selection. In: He Q, Iyengar A, Nejdl W, Pei J, Rastogi R, eds. Proc. of the 22nd ACM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2013. 2189−2198. [doi: 10.1145/2505515.2505584]

[13] Jia L, Liao SZ, Ding LZ. Learning with uncertain kernel matrix set. Journal of Computer Science and Technology, 2010,25(4): 709−727. [doi: 10.1007/s11390-010-9359-4]

[14] McCormick G. The projective SUMT method for convex programming. Mathematics of Operations Research, 1989,14(2):203−223. [doi: 10.1287/moor.14.2.203]

[15] Liao SZ, Jia L. Simultaneous tuning of hyperparameter and parameter for support vector machines. In: Zhou ZH, Li H, Yang Q, eds. Proc. of the 11th Pacific-Asia Conf. on Knowledge Discovery and Data Mining. Berlin: Springer-Verlag, 2007. 162−172. [doi: 10.1007/978-3-540-71701-0_18]

[16] Liao SZ, Ding LZ, Jia L. Simultaneous tuning of multiple parameters for support vector regression. Journal of Nanjing University (Natural Sciences), 2009,45(5):585−592 (in Chinese with English abstract).

[17] Cortes C, Vapnik V. Support-Vector networks. Machine Learning, 1995,20(3):273−297. [doi: 10.1007/BF00994018]

[18] Vlček J, Lukšan L. Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. Journal of Optimization Theory and Applications, 2001,111(2):407−430. [doi: 10.1023/A:1011990503369]

[19] Lanckriet GRG, Cristianini N, Bartlett P. Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research, 2004,5:27−72.

[20] Sonnenburg S, Rätsch G, Schäfer C. Large scale multiple kernel learning. Journal of Machine Learning Research, 2006,7: 1531−1565.

[21] Ong C, Smola A, Williamson R. Learning the kernel with hyperkernels. Journal of Machine Learning Research, 2005,6:1043−1071.

附中文参考文献:

[2] 丁世飞,黄华娟,史忠植.加权光滑 CHKS 孪生支持向量机.软件学报,2013,24(11):2548−2557. http://www.jos.org.cn/1000-9825/4475.htm

[16] 廖士中,丁立中,贾磊.支持向量回归多参数的同时调节.南京大学学报(自然科学版),2009,45(5):585−592.

丁立中(1986－),男,内蒙古呼和浩特人,博士生,CCF 学生会员,主要研究领域为核方法模型选择.
E-mail: dinglizhong@tju.edu.cn

贾磊(1981－),男,博士,主要研究领域为机器学习,模型选择,核方法.
E-mail: ljia@tju.edu.cn

廖士中(1964－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为人工智能应用基础,理论计算机科学.
E-mail: szliao@tju.edu.cn