计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20150110
52(8):1722-1734,2015

高斯核函数选择的广义核极化准则

田 萌[1,2] 王文剑[1]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](山东理工大学理学院 山东淄博 255049)
(wjwang@sxu.edu.cn)

Generalized Kernel Polarization Criterion for Optimizing Gaussian Kernel

Tian Meng[1,2] and Wang Wenjian[1]
[1](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)
[2](*School of Science*, *Shandong University of Technology*, *Zibo*, *Shandong* 255049)

Abstract The choice of kernel function is a basic and challenging problem in researches on kernel methods. Gaussian kernel is a popular and widely used one in various kernel methods, and many universal kernel selection methods have been derived for Gaussian kernel. However, these methods may have some disadvantages, such as heavy computational complexity, the difficulty of algorithm implement, and the requirement of the classes generated from underlying multivariate normal distributions. To remedy these problems, generalized kernel polarization criterion has been proposed to tune the parameter of Gaussian kernel for classification tasks. By taking the within-class local structure into account and centering the kernel matrix, the criterion does better in maximizing the class separability in the feature space. And the final optimized kernel parameter leads to a substantial improvement in the performance. Furthermore, the criterion function can be proved to have a determined approximate global minimum point. This good characteristic, coupled with its independence of the actual learning machine, makes the optimal parameter easier to find by many algorithms. Besides this, local kernel polarization criterion function, a special case of generalized kernel polarization criterion function, can also be proved to have a determined approximate global minimum point. The extensions of generalized kernel polarization criterion and local kernel polarization criterion to the multiclass domain have been proposed. Experimental results show the effectiveness and efficiency of our proposed criteria.

Key words kernel method; kernel selection; classification; kernel polarization criterion; generalized kernel polarization criterion

摘 要 核函数及其参数的选择是核方法研究中的一个基本却很困难的问题,高斯核是目前各类核方法中最常使用的一种核函数.关于高斯核参数的优化已有很多研究,然而这些方法大多存在时间复杂度高,或是算法实现困难,或是样本数据需服从多元正态分布的前提假设等不足.提出的广义核极化准则可用来解决分类问题中的高斯核参数优化,该准则通过保持类内局部结构信息及中心化核矩阵以更准确地刻画特征空间中类别间的分离度,进而获得更好的高斯核参数来提高分类性能.给出了广义核极化

收稿日期:2015-02-09;**修回日期**:2015-05-20
基金项目:国家自然科学基金项目(61273291);山西省回国留学人员科研资助项目(2012-008)
通信作者:王文剑(wjwang@sxu.edu.cn)

准则对应目标函数的近似最优解的存在唯一性证明,且由于该准则独立于学习算法,因此可用许多成熟的优化算法来寻找最优参数.此外,还补充了已有文献提出的局部核极化准则对应目标函数近似最优解的存在唯一性证明,并且指出该准则是所提出的广义核极化准则的一个特例.针对多分类问题,分别给出广义核极化准则及局部核极化准则的多分类拓展形式.在标准数据集上的实验结果表明所提准则的有效性.

关键词　核方法;核选择;分类;核极化准则;广义核极化准则

中图法分类号　TP181

核方法[1]是由统计学习理论发展起来并已经得到广泛应用的模式识别方法.一般说来,核方法通过核函数所隐式定义的非线性映射将输入空间映射到一个高维(甚至无穷维)的特征空间,通过在这个特征空间实施线性算法来间接实现原始输入空间的非线性算法.虽然核方法本质上是一种基于非线性映射的非线性方法,但其算法的实现却仅需使用线性的手段,这也就是人们常说的"核技巧",即通过核函数来刻画特征空间向量间的内积来回避非线性映射的显式表达.此外,多数核方法需要解决凸优化问题,这样就避免了局部极小化问题的困扰.同神经网络与决策树这些传统非线性学习算法相比,核方法有着坚实的数学基础、较低的运算代价且能从容处理高维数据,这些优良性质使得核方法发展成为当前机器学习及模式识别领域的主流学习算法.目前,常见的核方法有支持向量机[2]、核 Fisher 判别分析、高斯过程、核主成分分析、核聚类等,其中支持向量机是最早被提出也是目前应用最成功的核方法.

虽然核方法在很多领域得到了广泛的应用,但其性能表现却直接依赖于核函数的选择.这是由于核函数及其参数直接决定非线性映射所对应的特征空间.当选用不合适的核函数或核参数时,人们可能得到一个比在原始空间更差的识别结果[3].正因如此,核函数选择研究一直以来都是核方法研究中的基本问题和难点问题.

核函数选择具体可分为核函数类型的确定与选定核函数中核参数的优化.常用的核函数有线性核函数、多项式核函数、高斯核函数以及 Sigmoid 核函数等[2].实际问题中人们往往根据具体问题先确定核函数类型再优化核参数.一般来说,不存在万能核函数,核函数的表现总是相对确定数据集而言的.近些年来,研究者们依据核函数性质或针对具体问题设计出一些新的核函数[4-6].但纵观核方法的研究及应用,高斯核函数凭借其稳定优越的性能得到最广泛的重视与应用.

高斯核参数的优化方法主要可分为数据独立的优化方法与数据依赖的优化方法.经典的数据独立的优化方法是交叉验证法和留一法,其中留一法可视为一种特殊的交叉验证法.这类方法通过遍历寻参空间的每一组参数组合来寻找性能最优的参数组合,多数情况下该类方法是有效的但受限于最初确定的参数搜寻空间,且伴随着较大的时间复杂度.为解决这些问题,研究者开始尝试数据依赖的优化算法,如支持向量机中最小化留一法误差率上界的方法.这些误差界包括半径-间隔界(radius-margin 界)[7]、Joachims 上界[8]等.以半径-间隔界为例[9],该方法借助基于梯度的二步迭代算法来优化上界以得到最优核参数.相比交叉验证方法,这类方法能明显地降低计算量,但优化算法中支持向量机训练过程的多次调用也带来较大的计算消耗.

另一种数据依赖的核参数优化方法是设计核参数选择准则,直接从训练数据出发来寻找最优的核参数及对应的最优学习模型.不经过学习器训练使得该方法更加高效,从而弥补了前面 2 种方法计算消耗较大的缺点.具体说来,这些类核选择准则又可划分为基于距离度量的优化准则、基于 Fisher 准则的优化准则和基于核校准的优化准则等.

常见的基于距离度量的优化准则有最大化 2 类样本的质心间距[10]及最大化有代表性的较大样本间距与较小样本间距之差[11]等.这些准则的优点是对应算法易于实现,但是由于该类准则建立在样本服从多维正态分布的前提下,当违背此前提条件时这些准则通常表现较差.基于 Fisher 准则的核优化准则由于兼顾类间分散度和类内集中度的度量[3,12],在很多应用中取得了不错的结果.但和基于距离度量的准则一样,这些准则也仅适用于满足多维正态分布的样本数据.文献[12]提出一种等价于 Fisher 准则的高斯核参数优化准则 $\mathrm{tr}(\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}})$(其中 $\boldsymbol{S}_{\mathrm{w}}$,$\boldsymbol{S}_{\mathrm{b}}$ 分别为类内离散度矩阵和类间离散度矩阵)

能成功地应用于二分类、多分类及小样本集情形中. 但该方法的实现除了需要借助分解算法和高通 Butterworth 滤波器保证准则对应目标函数的连续可微性外,还需选择不同的初始点实现梯度下降法来确定最优值,这些操作无疑增加了运算时间及误差. 基于核校准的优化准则是通过度量核矩阵与理想核矩阵之间的相似性来优化核参数[13-14]. 理论分析与实验结果表明,该核参数优化准则是合理、高效的. 在最初的核校准准则被提出后,非均衡数据的核校准准则[15]、核极化准则[16]、中心核校准准则[17]等相继被提出. 近来,文献[18]在证明核极化准则对应目标函数近似最优解的存在唯一性后,利用拟牛顿算法高效搜索最优高斯核参数取得了很好的分类效果. 综合观察上述 3 类核参数选择准则,可以看出优化准则对应目标函数的连续可微性及最优解的存在唯一性是获得优化算法高效性的保证,而目标函数对样本数据结构信息的准确刻画则是保证该准则优化性能的关键.

为了更好地刻画样本数据的结构,各类核优化准则现正经历着"细化"进程. 文献[19]针对一类数据依赖的核函数,构建了利用局部 Fisher 准则、子类 Fisher 准则、边界 Fisher 准则等描述局部信息的手段修正优化目标函数的统一框架,以解决二分类及多分类问题中的核参数优化问题. 文献[20]通过将类内局部结构信息引入核极化准则,提出了局部核极化准则,并通过实验进行了验证,但文中未对该准则对应目标函数最优解的存在性或唯一性给出说明.

本文提出了一类更一般的核极化准则——广义核极化准则. 同核极化准则相比,该准则不仅能部分保持同类样本间的局部结构,还通过中心化核矩阵来增强目标函数值与分类性能间的匹配. 文中给出了广义核极化准则对应目标函数的近似最优解的存在唯一性证明,并将文献[20]提出的局部核极化准则作为广义核极化准则的一个特例,明确给出其对应目标函数的近似最优解满足存在唯一性的结论. 此外,本文还分别将广义核极化准则及局部核极化准则推广至多分类情形. 最后以支持向量机为例,通过标准数据集上的实验结果验证新提出的高斯核参数优化准则在二分类及多分类问题中的有效性,并进一步通过实验探讨广义核极化准则和局部核极化准则及其对应的多分类扩展形式中调节参数 t 的取值对分类性能的影响,给出了参数 t 的取值建议.

1 预备知识

高斯核函数是核方法中一类被广泛应用的核函数,其表达形式为

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\mathrm{e}^{-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sigma^2}},\quad \sigma>0.$$

对于高斯核函数而言,当核参数 $\sigma\to 0$ 时,$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)\to 0$,这意味着向量 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 正交也即任意 2 样本都可以被分开从而使得算法陷入"过拟合". 另一种情形,当核参数 $\sigma\to+\infty$ 时,$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)\to 1$,这时学习算法将所有样本点判为一类,促使算法面临"欠拟合". 正因如此,高斯核参数 σ 不易取得过大也不易取得过小[21].

设给定一组样本点 $\{(\boldsymbol{x}_i,y_i)\mid i=1,2,\cdots,n\}$,其中 $\boldsymbol{x}_i\in X\subset\mathbb{R}^m$,$y_i\in\{-1,1\}$. 将特征空间中异类样本点尽可能地分开同时将同类样本点尽可能地聚集在一起是正确分类最朴素的想法. 借用物理学概念,文献[16]提出核极化准则(kernel polarization, KP),即

$$KP(\boldsymbol{K},\boldsymbol{Y})=\langle\boldsymbol{K},\boldsymbol{Y}\rangle_{\mathrm{F}}=\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j),$$

其中,$\boldsymbol{K}$ 为核矩阵,$\boldsymbol{Y}=\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ 为理想核矩阵且 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$,$\langle\cdot,\cdot\rangle_{\mathrm{F}}$ 表示矩阵之间的 Frobenius 内积. 核极化准则可以改写成:

$$KP(\boldsymbol{K},\boldsymbol{Y})=\sum_{y_i=y_j}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{y_i\neq y_j}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j).\quad(1)$$

由于核函数可以视为特征空间中相似性的一种度量,因此式(1)可以解释为:当同类样本点很相似(对应的 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 取较大值)且异类样本点很不相似(对应的 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 取较小值)时可以得到较大的核极化值,这时最优核参数 σ 可以通过最大化 $KP(\boldsymbol{K},\boldsymbol{Y})$ 得到. 等式 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=y_iy_j$ 对应理想的核极化值,这意味着达到理想核极化值需要同类样本点映射到特征空间中的同一个向量(即 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=1$),且特征空间中异类样本点对应的映射向量之间成反比(即 $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=-1$)[22]. 对上述任意一条的违背都会促使 $KP(\boldsymbol{K},\boldsymbol{Y})$ 取不到理想值,自然这样的要求过于苛刻.

考虑到分类的最终目的,文献[20]保持异类样本间强可分性,而引入关联系数来弱化同类样本点尽可能映射到特征空间中同一个点的要求,构造出局部核极化准则(local kernel polarization, LKP),其中关联系数 G_{ij} 定义如下:

$$G_{ij}=\begin{cases}e^{-t\|x_i-x_j\|^2}, & y_i=y_j,\\ 1, & y_i\neq y_j,\end{cases}$$

其中,调节参数 $t\geqslant 0$ 用来控制关联系数的作用范围.局部核极化准则可表述为[20]

$$LKP(\boldsymbol{K},\boldsymbol{Y})=\langle \boldsymbol{K}_L,\boldsymbol{Y}\rangle_F, \tag{2}$$

这里称 $\boldsymbol{K}_L$ 为局部修正核矩阵,定义为

$$(\boldsymbol{K}_L)_{ij}=G_{ij}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=$$

$$\begin{cases}e^{-t\|x_i-x_j\|^2}e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}}, & y_i=y_j,\\ e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}}, & y_i\neq y_j.\end{cases}$$

显然,当 $t=0$ 时,局部核极化准则退化为标准核极化准则.局部核极化准则的直观意义是不再强制要求同类样本点必须投影到相同点,而是依据输入空间中 2 个样本点间的距离对特征空间中这 2 个样本点对应的映射向量间的相似性进行加权修正;对异类样本点而言,广义核极化准则同核极化准则一样仍旧要求它们尽可能地分开.由 e^{-x} 的单调性知同类样本点间距越大则允许它们对应映射向量间的相似性越小,这样就部分保留了同类样本间的局部结构信息.文献[20]通过实验验证了准则的有效性,但并未给出该准则下最优解的存在性证明.

文献[17]指出一个较小的核校准值也可能对应一个较高的测试精度.为增强核校准值与分类性能间的相关性,文献[17]引入中心化核矩阵思想,提出了中心核校准准则(centered kernel alignment, *CKA*),并给出了该准则的一系列理论结果及实验验证.中心核校准准则描述为

$$CKA(\boldsymbol{K},\boldsymbol{Y})=\frac{\langle \boldsymbol{K}_C,\boldsymbol{Y}_C\rangle_F}{\|\boldsymbol{K}_C\|_F\|\boldsymbol{Y}_C\|_F}, \tag{3}$$

其中,$\boldsymbol{K}_C=\boldsymbol{HKH}$ 且 $\boldsymbol{H}=\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^T$,$\boldsymbol{I}_n$ 为 n 阶单位矩阵,$\boldsymbol{1}_n$ 为分量均为 1 的 n 维列向量.由于 $E_{x,x'}[\boldsymbol{K}_C(\boldsymbol{x},\boldsymbol{x}')]=0$,称 $\boldsymbol{K}_C$ 为中心化后的核矩阵,文献[23-24]均指出相比核校准准则,中心核校准值 $CKA(\boldsymbol{K},\boldsymbol{Y})$ 与分类性能之间的相关性更强,并分别将中心核校准准则成功地应用于多核学习及聚类算法.

2　广义核极化准则

本文给出一种新的核极化准则——广义核极化准则(generalized kernel polarization, *GKP*),即

$$GKP(\boldsymbol{K},\boldsymbol{Y})=\langle(\boldsymbol{K}_L)_C,\boldsymbol{Y}_C\rangle_F,$$

其中,$(\boldsymbol{K}_L)_C$ 与 $\boldsymbol{Y}_C$ 分别为中心化后的局部修正核矩阵与理想核矩阵,即 $(\boldsymbol{K}_L)_C=\boldsymbol{HK}_L\boldsymbol{H}$,$\boldsymbol{Y}_C=\boldsymbol{HYH}$.可以说,广义核极化准则 $GKP(\boldsymbol{K},\boldsymbol{Y})$ 是依据矩阵 $(\boldsymbol{K}_L)_C$ 与 $\boldsymbol{Y}_C$ 之间相似性的最大化来优化核参数.由于 $\boldsymbol{H}$ 为幂等矩阵,即 $\boldsymbol{H}^2=\boldsymbol{H}$,结合矩阵迹的性质,得:

$$\langle(\boldsymbol{K}_L)_C,\boldsymbol{Y}_C\rangle_F=\mathrm{tr}((\boldsymbol{K}_L)_C\boldsymbol{Y}_C)=$$
$$\mathrm{tr}(\boldsymbol{HK}_L\boldsymbol{HHYH})=\mathrm{tr}(\boldsymbol{HK}_L\boldsymbol{HYH})=$$
$$\mathrm{tr}(\boldsymbol{K}_L\boldsymbol{HYH})=\langle\boldsymbol{K}_L,\boldsymbol{Y}_C\rangle_F,$$

所以 $GKP(\boldsymbol{K},\boldsymbol{Y})$ 可以等价定义为

$$GKP(\boldsymbol{K},\boldsymbol{Y})=\langle\boldsymbol{K}_L,\boldsymbol{Y}_C\rangle_F. \tag{4}$$

为讨论方便,不妨设样本集合中前 n_1 个样本为其正类样本,后 n_2 个样本为其负类样本,有 $n_1+n_2=n$ 成立.整理化简 $\boldsymbol{Y}_C$ 可以得到[3]:

$$\boldsymbol{Y}_C=4\begin{bmatrix}\left(\frac{n_2^2}{n^2}\right)_{n_1\times n_1} & \left(-\frac{n_1n_2}{n^2}\right)_{n_1\times n_2}\\ \left(-\frac{n_2n_1}{n^2}\right)_{n_2\times n_1} & \left(\frac{n_1^2}{n^2}\right)_{n_2\times n_2}\end{bmatrix}_{n\times n}, \tag{5}$$

则广义核极化准则可以写作:

$$GKP(\boldsymbol{K},\boldsymbol{Y})=$$
$$4\sum_{y_i=y_j=1}\frac{n_2^2}{n^2}e^{-t\|x_i-x_j\|^2}e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}}+$$
$$4\sum_{y_i=y_j=-1}\frac{n_1^2}{n^2}e^{-t\|x_i-x_j\|^2}e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}}-$$
$$4\sum_{y_i\neq y_j}\frac{2n_1n_2}{n^2}e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}}.$$

注意到当 $n_1=n_2$ 且 $t=0$ 时,该准则等价于标准核极化准则;当 $n_1=n_2$ 时,该准则退化成局部核极化准则;当 $t=0$ 时,该准则退化成未归一化的中心核校准准则.因此本文称该准则为广义核极化准则,最大化 $GKP(\boldsymbol{K},\boldsymbol{Y})$ 将得到广义核极化准则下的最优核参数,即 $\sigma_{opt}=\max_\sigma GKP(\boldsymbol{K},\boldsymbol{Y})$.

2.1　目标函数最优解的存在唯一性

本节将讨论广义核极化准则下最优高斯核参数的存在唯一性.

为讨论方便,记:

$$S(\sigma)=\sum_{y_i\neq y_j}\frac{2n_1n_2}{n^2}e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}}-$$
$$\sum_{y_i=y_j=1}\frac{n_2^2}{n^2}e^{-(t+\frac{1}{\sigma^2})\|x_i-x_j\|^2}-$$
$$\sum_{y_i=y_j=-1}\frac{n_1^2}{n^2}e^{-(t+\frac{1}{\sigma^2})\|x_i-x_j\|^2},$$

这样就把最大化 $GKP(\boldsymbol{K},\boldsymbol{Y})$ 问题转化为最小化 $S(\sigma)$ 的问题,即:

$$\sigma_{opt}=\min_\sigma S(\sigma). \tag{6}$$

令 $T_{ij}=\|x_i-x_j\|^2$,则:

$$S(\sigma)=\sum_{y_i\neq y_j}\frac{2n_1n_2}{n^2}e^{-\frac{1}{\sigma^2}T_{ij}}-\sum_{y_i=y_j=1,i\neq j}\frac{n_2^2}{n^2}e^{-(t+\frac{1}{\sigma^2})T_{ij}}-$$
$$\sum_{y_i=y_j=-1,i\neq j}\frac{n_1^2}{n^2}e^{-(t+\frac{1}{\sigma^2})T_{ij}}-\frac{n_1n_2}{n}. \tag{7}$$

欧拉-麦克劳林求和公式是一个十分有力的解析工具,许多比较困难的级数求和问题借助此公式往往能得到更为一般的结论[25]. 欧拉-麦克劳林求和公式等价表述为

$$\sum_{z=u}^{v} f(z) = \int_u^v f(z)\mathrm{d}z + \frac{1}{2}[f(v)+f(u)] + \frac{B_2}{2!}[f'(v)-f'(u)] + \frac{B_4}{4!}[f^{(3)}(v)-f^{(3)}(u)] + \cdots + \frac{B_{2k}}{2k!}[f^{(2k-1)}(v)-f^{(2k-1)}(u)] + \cdots,$$

其中,B_{2k}是 Bernoulli 数,令 $f(z)=\mathrm{e}^z$ 并将等式右端展开至 $k=1$,得:

$$\sum_{z=u}^{v} \mathrm{e}^z \approx \frac{19}{12}\mathrm{e}^v - \frac{7}{12}\mathrm{e}^u.$$

不妨设:

$$\max_{y_i=y_j=1,i\neq j} T_{ij} = A,\ \min_{y_i=y_j=1,i\neq j} T_{ij} = B,$$
$$\max_{y_i=y_j=-1,i\neq j} T_{ij} = C,\ \min_{y_i=y_j=-1,i\neq j} T_{ij} = D,$$
$$\max_{y_i\neq y_j} T_{ij} = E,\ \min_{y_i\neq y_j} T_{ij} = F.$$

容易推想多数情形下在固定数据集上,变量 A,B,C,D,E,F 一般满足:

$$\begin{cases} A \gg B, C \gg D, E \gg F, \\ \max\{A,C,E\} - \min\{A,C,E\} \text{ 较小.} \end{cases} \tag{8}$$

这样,对于广义核极化准则有以下结论:

定理 1. 在式(8)成立的前提下,广义核极化准则对应的优化目标函数 $S(\sigma)$存在唯一的近似最小值点.

证明. 考虑到虽然变量 B,D,F 之间有 6 种大小关系,但由 B,D,F 的定义可推断多数数据集上将存在关系 $F\geqslant\min\{B,D\}$成立,故而仅需讨论以下 4 种情形:$F\geqslant D\geqslant B$, $B\geqslant F\geqslant D$, $F\geqslant B\geqslant D$, $D\geqslant F\geqslant B$.

下列讨论中将假定广义核极化准则中的参数 t 是一个定值. 对于第 1 种情形 $F\geqslant D\geqslant B$,不防先讨论 $F>D>B$.

依据前面的假设、定义以及欧拉-麦克劳林求和公式,$S(\sigma)$经化简得:

$$S(\sigma) \approx \sum_{F}^{E} \frac{2n_1n_2}{n^2}\mathrm{e}^{-\frac{1}{\sigma^2}T_{ij}} - \sum_{B}^{A} \frac{n_2^2}{n^2}\mathrm{e}^{-(t+\frac{1}{\sigma^2})T_{ij}} - \sum_{D}^{C} \frac{n_1^2}{n^2}\mathrm{e}^{-(t+\frac{1}{\sigma^2})T_{ij}} - \frac{n_1n_2}{n} = \sum_{-\frac{E}{\sigma^2}}^{-\frac{F}{\sigma^2}} \frac{2n_1n_2}{n^2}\mathrm{e}^z - \sum_{-(\frac{1}{\sigma^2}+t)A}^{-(\frac{1}{\sigma^2}+t)B} \frac{n_2^2}{n^2}\mathrm{e}^z - \sum_{-(\frac{1}{\sigma^2}+t)C}^{-(\frac{1}{\sigma^2}+t)D} \frac{n_1^2}{n^2}\mathrm{e}^z - \frac{n_1n_2}{n} \approx \frac{2n_1n_2}{n^2}[\frac{19}{12}\mathrm{e}^{-\frac{F}{\sigma^2}} - \frac{7}{12}\mathrm{e}^{-\frac{E}{\sigma^2}}] - \frac{n_2^2}{n^2}[\frac{19}{12}\mathrm{e}^{-(\frac{1}{\sigma^2}+t)B} - \frac{7}{12}\mathrm{e}^{-(\frac{1}{\sigma^2}+t)A}] - \frac{n_1^2}{n^2}[\frac{19}{12}\mathrm{e}^{-(\frac{1}{\sigma^2}+t)D} - \frac{7}{12}\mathrm{e}^{-(\frac{1}{\sigma^2}+t)C}] - \frac{n_1n_2}{n}.$$

显然,$S(\sigma)$关于 σ 的一、二阶导数 $S'(\sigma)$与 $S''(\sigma)$存在且连续. 为表示方便,令:

$$\bar{S}(\sigma) = \frac{2n_1n_2}{n^2}[19\mathrm{e}^{-\frac{F}{\sigma^2}} - 7\mathrm{e}^{-\frac{E}{\sigma^2}}] - \frac{n_2^2}{n^2}[19\mathrm{e}^{-(\frac{1}{\sigma^2}+t)B} - 7\mathrm{e}^{-(\frac{1}{\sigma^2}+t)A}] - \frac{n_1^2}{n^2}[19\mathrm{e}^{-(\frac{1}{\sigma^2}+t)D} - 7\mathrm{e}^{-(\frac{1}{\sigma^2}+t)C}]. \tag{9}$$

$\bar{S}(\sigma)$称为 $S(\sigma)$的近似目标函数,这样,原优化问题(式(6))可以转化为 $\bar{S}(\sigma)$的优化问题,即 $\sigma_{\mathrm{opt}}\approx\min_{\sigma}\bar{S}(\sigma)$. 令$\frac{\mathrm{d}\bar{S}(\sigma)}{\mathrm{d}\sigma}=0$,可以得到:

$$\frac{2n_1n_2}{n^2\sigma^3}[19F\mathrm{e}^{-\frac{F}{\sigma^2}} - 7E\mathrm{e}^{-\frac{E}{\sigma^2}}] - \frac{n_2^2}{n^2\sigma^3}[19B\mathrm{e}^{-(\frac{1}{\sigma^2}+t)B} - 7A\mathrm{e}^{-(\frac{1}{\sigma^2}+t)A}] - \frac{n_1^2}{n^2\sigma^3}[19D\mathrm{e}^{-(\frac{1}{\sigma^2}+t)D} - 7C\mathrm{e}^{-(\frac{1}{\sigma^2}+t)C}] = 0.$$

由高斯核参数的性质分析知,高斯核参数 σ 不易取得过大,故而舍去 $\sigma\to+\infty$的情形而仅讨论下面的等式:

$$\frac{2n_1n_2}{n^2}[19F\mathrm{e}^{-\frac{F}{\sigma^2}} - 7E\mathrm{e}^{-\frac{E}{\sigma^2}}] - \frac{n_2^2}{n^2}[19B\mathrm{e}^{-(\frac{1}{\sigma^2}+t)B} - 7A\mathrm{e}^{-(\frac{1}{\sigma^2}+t)A}] - \frac{n_1^2}{n^2}[19D\mathrm{e}^{-(\frac{1}{\sigma^2}+t)D} - 7C\mathrm{e}^{-(\frac{1}{\sigma^2}+t)C}] = 0.$$

由于已经假设 $F>D>B$,在等式两端同除 $\mathrm{e}^{-\frac{B}{\sigma^2}}$可以得到:

$$\frac{38n_1n_2F}{n^2}\mathrm{e}^{\frac{B-F}{\sigma^2}} - \frac{14n_1n_2E}{n^2}\mathrm{e}^{\frac{B-E}{\sigma^2}} - \frac{19n_2^2B}{n^2}\mathrm{e}^{-tB} + \frac{7n_2^2A}{n^2}\mathrm{e}^{\frac{B-A}{\sigma^2}}\mathrm{e}^{-tA} - \frac{19n_1^2D}{n^2}\mathrm{e}^{\frac{B-D}{\sigma^2}}\mathrm{e}^{-tD} + \frac{7n_1^2C}{n^2}\mathrm{e}^{\frac{B-C}{\sigma^2}}\mathrm{e}^{-tC} = 0.$$

整理得:

$$\left(\frac{38n_1n_2F}{n^2} - \frac{19n_1^2D}{n^2}\right)\mathrm{e}^{\frac{B-F}{\sigma^2}} - \frac{19n_1^2D}{n^2}(\mathrm{e}^{\frac{B-D}{\sigma^2}}\mathrm{e}^{-tD} - \mathrm{e}^{\frac{B-F}{\sigma^2}}) -$$

$$\frac{19n_2^2B}{n^2}\mathrm{e}^{-tB}-\Big(\frac{14n_1n_2E}{n^2}-\frac{7n_2^2A}{n^2}\mathrm{e}^{-tA}-\frac{7n_1^2C}{n^2}\mathrm{e}^{-tC}\Big)\mathrm{e}^{\frac{B-A}{\sigma^2}}-\frac{14n_1n_2E}{n^2}(\mathrm{e}^{\frac{B-E}{\sigma^2}}-\mathrm{e}^{\frac{B-A}{\sigma^2}})+\frac{7n_1^2C}{n^2}\mathrm{e}^{-tC}(\mathrm{e}^{\frac{B-C}{\sigma^2}}-\mathrm{e}^{\frac{B-A}{\sigma^2}})=0. \tag{10}$$

引入辅助变量:

$$\frac{n_1^2C}{n^2}\mathrm{e}^{-tC}(\mathrm{e}^{\frac{B-C}{\sigma^2}}-\mathrm{e}^{\frac{B-A}{\sigma^2}})=\varepsilon_1,$$

$$\frac{n_1n_2E}{n^2}(\mathrm{e}^{\frac{B-E}{\sigma^2}}-\mathrm{e}^{\frac{B-A}{\sigma^2}})=\varepsilon_2,$$

$$\Big(\frac{2n_1n_2E}{n^2}-\frac{n_2^2A}{n^2}\mathrm{e}^{-tA}-\frac{n_1^2C}{n^2}\mathrm{e}^{-tC}\Big)\mathrm{e}^{\frac{B-A}{\sigma^2}}=\varepsilon_3,$$

$$\frac{n_1^2D}{n^2}(\mathrm{e}^{\frac{B-D}{\sigma^2}}\mathrm{e}^{-tD}-\mathrm{e}^{\frac{B-F}{\sigma^2}})=\varepsilon_4.$$

基于变量 $A\sim F$ 满足式(8)的假设以及 $F>D>B$,可以推导出 $\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4$ 均为接近于 0 的小数.为简化表示,进一步记 $19\varepsilon_4+7\varepsilon_3+14\varepsilon_2-7\varepsilon_1=19\varepsilon$,化简求解式(10),得:

$$\sigma_0=\sqrt{\frac{B-F}{\ln\Big(\dfrac{n_2^2B\mathrm{e}^{-tB}+n^2\varepsilon}{2n_1n_2F-n_1^2D}\Big)}}.$$

所得 σ_0 为 $\bar{S}(\sigma)$的唯一驻点.

观察式(9),由于:

$$\lim_{\sigma\to0^+}\bar{S}(\sigma)=0,$$

$$\lim_{\sigma\to+\infty}\bar{S}(\sigma)=\frac{24n_1n_2}{n^2}-\frac{n_2^2}{n^2}(19\mathrm{e}^{-tB}-7\mathrm{e}^{-tA})-\frac{n_1^2}{n^2}(19\mathrm{e}^{-tD}-7\mathrm{e}^{-tC}):=S^*.$$

考虑到 A,C,E 之间相差不大且它们相应地远远大于 B,D,F 以及 $F>D>B$,可以推导出 $\bar{S}(\sigma_0)<0$ 且 $\bar{S}(\sigma_0)<S^*$.由此可知,σ_0 即为 $\bar{S}(\sigma_0)$的最小值点.

用同样的方法分析式(10),可以得到 3 种特殊情形:1)$D=B$ 且 $F>D$;2)$F=D$ 且 $F>B$;3)$F=D=B$,有如下结论:

当情形 1 时,$\bar{S}(\sigma_0)$的最小值点为

$$\sigma_{\mathrm{opt}}=\sqrt{\frac{B-F}{\ln\Big(\dfrac{n^2\eta+B(n_1^2+n_2^2)\mathrm{e}^{-tB}}{2n_1n_2F}\Big)}};$$

当情形 2 时,$\bar{S}(\sigma_0)$的最小值点为

$$\sigma_{\mathrm{opt}}=\sqrt{\frac{B-F}{\ln\left(\dfrac{n^2\eta+n_2^2B\mathrm{e}^{-tB}}{2n_1n_2F-n_1^2F\mathrm{e}^{-tF}}\right)}};$$

当情形 3 时,$\bar{S}(\sigma_0)$的最小值点为

$$\sigma_{\mathrm{opt}}=\sqrt{\frac{B-A}{\ln\Big(\dfrac{38n_1n_2B-19B(n_1^2+n_2^2)\mathrm{e}^{-tB}}{14n_1n_2E-7n_2^2A\mathrm{e}^{-tA}-7n_1^2C\mathrm{e}^{-tC}}\Big)}},$$

且 $\eta=\dfrac{7}{19}\Big(\dfrac{2n_1n_2E}{n^2}\mathrm{e}^{\frac{B-E}{\sigma^2}}-\dfrac{n_2^2A}{n^2}\mathrm{e}^{-tA}\mathrm{e}^{\frac{B-A}{\sigma^2}}-\dfrac{n_1^2C}{n^2}\mathrm{e}^{-tC}\mathrm{e}^{\frac{B-C}{\sigma^2}}\Big)$,详细的推导过程不再赘述.

同理,可以得到在剩余 3 种情形($F\geqslant B\geqslant D$,$D\geqslant F\geqslant B$,$B\geqslant F\geqslant D$)下,最优参数 σ_{opt}存在唯一性的证明.综上所述,定理 1 得证. 证毕.

这样,广义核极化准则目标函数的一、二阶导函数的连续性以及上述定理 1 的结论使得借用成熟优化算法来寻找高斯核参数成为可能.

2.2　局部核极化准则对应最优解分析

通过分析式(7),我们注意到,当正负样本个数相等,即 $n_1=n_2=n/2$ 时,广义核极化准则可退化成局部核极化准则,因此局部核极化准则可视为广义核极化准则的 1 个特例.下面仍旧仅以 $F>D>B$ 为例说明.在式(8)成立的前提下,沿用同样的证明方法可以得出:局部核极化准则对应目标函数的近似最小值点是:

$$\sigma_0=\sqrt{\frac{B-F}{\ln\Big(\dfrac{B\mathrm{e}^{-tB}+\varepsilon}{F}\Big)}},$$

其中,$\varepsilon=\dfrac{7}{19}(E\mathrm{e}^{\frac{B-E}{\sigma^2}}-A\mathrm{e}^{-tA}\mathrm{e}^{\frac{B-A}{\sigma^2}})$.

特殊地,当 $F=D=B$ 时,局部核极化准则对应目标函数的近似最小值点为

$$\sigma_0=\sqrt{\frac{B-A}{\ln\Big(\dfrac{19B-19B\mathrm{e}^{-tB}}{7E-7A\mathrm{e}^{-tA}}\Big)}}.$$

这样,上述结论可以总结如下:

定理 2. 在式(8)成立的前提下,局部核极化准则对应的优化目标函数存在唯一的近似最小值点.

上面结论补充说明了文献[20]提出的局部核极化准则对应目标函数最优解的存在唯一性.

2.3　多类广义核极化准则

广义核极化准则是针对二分类问题提出的,本节探讨如何将其推广至多分类情形.文献[26]曾提出多类核极化准则(multiclass kernel polarization,MKP),其表达式为

$$MKP(\boldsymbol{K},\boldsymbol{Y})=\langle\boldsymbol{K},\boldsymbol{Y}_{\mathrm{M}}\rangle_{\mathrm{F}}, \tag{11}$$

其中,

$$(\boldsymbol{Y}_{\mathrm{M}})_{ij}=\begin{cases}1, & y_i=y_j,\\ -1, & y_i\neq y_j.\end{cases}$$

这里 $y_i,y_j=1,2,\cdots,l$,其中 l 为样本类别个数.观察可知,$\boldsymbol{Y}_{\mathrm{M}}$ 的定义去除了 2 类的限制,直接将

多类信息综合在一个理想矩阵中. 文献[10]在用“一对一”法解决多类问题时,通过最小化所有任意 2 类样本间的 δ_{4F} 指标之和来优化核参数. 相比这种办法,通过 $\boldsymbol{Y}_M$ 将多类样本间的信息融入一个理想矩阵能明显节省计算时间.

沿袭文献[19,26]中将核优化准则由二类问题拓广至多类问题的策略,多类广义核极化准则(multiclass generalized kernel polarization, *MGKP*)定义为

$$MGKP(\boldsymbol{K},\boldsymbol{Y}) = \langle(\boldsymbol{K}_L)_C,(\boldsymbol{Y}_M)_C\rangle_F, \quad (12)$$

其中,局部修正核矩阵 $\boldsymbol{K}_L$ 的定义见式(2),$(\boldsymbol{K}_L)_C$ 与 $(\boldsymbol{Y}_M)_C$ 分别指对矩阵 $\boldsymbol{K}_L$ 与 $\boldsymbol{Y}_M$ 进行中心化处理后所得矩阵. 显然,当 $l=2$ 时,该准则退化为广义核极化准则.

由 Frobenius 内积定义及矩阵迹的性质,得:

$$\langle(\boldsymbol{K}_L)_C,(\boldsymbol{Y}_M)_C\rangle_F =$$
$$\mathrm{tr}((\boldsymbol{K}_L)_C(\boldsymbol{Y}_M)_C) = \mathrm{tr}(\boldsymbol{H}\boldsymbol{K}_L\boldsymbol{H}\boldsymbol{H}\boldsymbol{Y}_M\boldsymbol{H}) =$$
$$\mathrm{tr}(\boldsymbol{H}\boldsymbol{K}_L\boldsymbol{H}\boldsymbol{Y}_M\boldsymbol{H}) = \langle\boldsymbol{K}_L,(\boldsymbol{Y}_M)_C\rangle_F,$$

则多类广义核极化准则可等价表示为

$$MGKP(\boldsymbol{K},\boldsymbol{Y}) = \langle\boldsymbol{K}_L,(\boldsymbol{Y}_M)_C\rangle_F.$$

类似地,多类局部核极化准则(multiclass local kernel polarization, *MLKP*)可以定义为

$$MLKP(\boldsymbol{K},\boldsymbol{Y}) = \langle\boldsymbol{K}_L,\boldsymbol{Y}_M\rangle_F. \quad (13)$$

注意到,多类广义核极化准则与多类局部核极化准则均引入调节参数 t,第 3 节实验中将具体讨论参数 t 的取值对分类性能的影响.

3 实 验

本节实验以支持向量机为载体来验证广义核极化准则及多类广义核极化准则的有效性. 具体来说,就是比较核极化准则、局部核极化准则、广义核极化准则及其对应多类扩展形式和交叉验证方法在高斯核参数 σ 选择中的性能表现. 实验分为 4 部分:1)验证优化目标函数 $S(\sigma)$(式(7))的全局极值点的存在唯一性;2)在二分类数据集上应用不同核极化准则进行核参数优化,并与交叉验证方法的实验结果进行比较;3)在多分类数据集上,给出基于不同多类核极化准则的核选择方法与交叉验证方法的实验比较结果;4)参数 t 对广义核极化准则与局部核极化准则及其多类形式所对应的分类性能的影响.

3.1 数据集及实验设置

本文从 UCI 机器学习数据库[27]及 Delve 数据库[28]中选取 19 个数据集,其中 12 个为二分类数据集,其余为多分类数据集. 所使用数据集的基本属性见表 1 所示:

Table 1 Data Sets for Experiments

表 1 实验中使用的数据集

Datasets	Number of Features	Number of Samples	Number of Classes
Sonar	60	208	2
Liverdisorder	6	345	2
Heart	13	270	2
Ionosphere	34	351	2
Monks1	6	432	2
Monks2	6	432	2
Monks3	6	432	2
Wdbc	30	569	2
Australian	14	690	2
Ringnorm	20	1 000	2
Twonorm	20	1 000	2
German	24	1 000	2
Iris	4	150	3
Wine	13	178	3
Svmguide2	20	391	3
Balance	4	625	3
Vehicle	18	846	4
Satimage	36	2 000	6
Segment	19	2 310	7

对于每一个数据集,选择 2/3 的样本组成训练集,且保持训练集中每类样本所占比例与原样本集中的对应比例保持一致;剩余 1/3 的样本组成测试集. 实验中采用 3 种核选择准则来优化高斯核参数,分别是核极化准则、局部核极化准则以及本文所提出的广义核极化准则,并依次用 *KP*,*LKP*,*GKP* 表示,它们所对应的多类形式则相应地用 *MKP*,*MLKP*,*MGKP* 表示,10-折交叉验证方法简记为 CV.

利用核选择准则优化核参数的算法是:令搜寻算法初始点为样本集中两两样本间距离的平均值,步长为 1,最大迭代次数为 20. 首先利用进退法求解极值区间,然后在该区间内利用 fminbnd 函数来搜寻极值点. 实验中,若未加特殊说明,局部核极化准则及多类局部核极化准则中参数 t 的取值和文献[20]中参数 t 的设置一样,即令 $t=1$;广义核极化准则及多类广义核极化准则中的参数 $t=Q$,其中 $Q=1/\min\{B,D\}$,其中 B 与 D 的定义见 2.1 节所述.

实验中支持向量机的惩罚系数遍历集合$\{2^{-2},$

$2^0,2^2,2^4,2^6,2^8\}$. 10-折交叉验证中,高斯核参数的搜索集合为$\{2^{-5},2^{-4},2^{-3},2^{-2},2^{-1},2^0,2^1,2^2,2^3,2^4,2^5\}$.

本文的实验环境是 Window 7.0,E7500 CPU,2GB RAM 以及 Matlab 2008a,并调用 Libsvm 工具包[29]实现算法.

3.2　GKD 目标函数解的存在唯一性验证

为验证式(8)的合理性,首先给出本文 12 个二分类数据集上变量 $A\sim F$(定义见 2.1 节)的取值,见表 2 所示. 从表 2 可以看出,分别描述正类样本集、负类样本集及全体样本集上任意 2 个样本间距平方最大值的变量 A,C,E 要远大于描述对应样本集上的样本间距平方最小值的 B,D,F. 在 Ionosphere,Wdbc,Ringnorm 数据集上,变量 A,C,E 之间差距较大,但考虑到函数 e^{-x} 在 x 较大时的快速衰减性,A,C,E 之间的差距对 2.1 节结论的影响是有限的. 除上述 3 个数据集外,其余数据集上 A,C,E 之间相差不大,满足式(8)的前提假设. 在固定数据集上,描述数据集上 3 种最小间距的 B,D,F 之间均满足 $\min\{B,D,F\}=\min\{B,D\}$,也即 $F\geqslant\min\{B,D\}$始终成立. 综上所述,2.1 节证明中的假设是合理的.

Table 2　Values of $A\sim F$ of Datasets

表 2　数据集上参数 $A\sim F$ 的取值

Datasets	A	B	C	D	E	F
Sonar	10.8569	0.0392	12.4571	0.0532	11.5366	0.2573
Liverdisorder	7.0028	0.0089	9.1179	0.0016	8.7002	0.0047
Heart	31.4439	0.0877	29.1991	0.0069	32.6465	0.1403
Ionosphere	59.8720	0.0100	98.0000	0.0382	76.9753	0.2178
Monks1	24.0000	0.4444	21.7778	0.4444	24.0000	0.4444
Monks2	18.7778	0.4444	24.0000	0.4444	24.0000	0.4444
Monks3	21.7778	0.4444	24.0000	0.4444	24.0000	0.4444
Wdbc	37.1670	0.1007	30.6440	0.0739	48.0031	0.1674
Australian	31.1831	0.0064	26.2098	0.0020	29.6062	0.0023
Ringnorm	3.2293	0.1751	12.3441	0.5665	7.9836	0.4377
Twonorm	9.6409	0.4781	11.1077	0.5185	13.6038	0.6048
German	55.8311	0.0011	55.4530	0.0121	57.8617	0.2697

为直观展示 12 个二分类数据集上优化目标函数最小值点的存在唯一性,图 1 给出了这些数据集上 $S(\sigma)$随 σ 变化的曲线图. 从图 1 可以看出,所有数据集上 $S(\sigma)$的曲线都有 1 个明显的波谷,进而直观说明了最小值点的存在唯一性.

3.3　二分类数据集上的实验结果比较

为了验证广义核极化准则的核参数优化效果,本节实验将对比在二分类数据集上分别利用核极化准则、局部核极化准则和广义核极化准则获取核参数后得到的测试精度以及利用交叉验证技术得到的测试精度. 在每个数据集上都进行 10 次独立的实验.

图 2 展示了 10 次实验中测试精度的平均值及标准差. 从图 2 可以看出,在 12 个数据集中,利用 3 类核极化准则所能达到的最高测试精度在 10 个数据集上都优于交叉验证法所得结果. 在 3 种核极化准则中,广义核极化准则与局部核极化准则在 7 个数据集上相比标准核极化准则有比较明显的精度提升(精度增幅超过 0.05),且广义核极化准则在 10 个数据集上获得最高的测试精度. 这说明通过挖掘数据集自身更多的信息来修正核极化准则对于提高核优化准则的分类性能是有效的.

为了保证实验结果的客观性,对测试精度采用显著性水平为 0.05 的成对 t 假设检验作统计检验,得到的结果列在表 3. 从表 3 的统计结果来看,在 12 个数据集上广义核极化准则分别在 10,7,4 个数据集上有显著性差异地优于核极化准则、局部核极化准则以及交叉验证法,且分别在 1,5,8 个数据集上相应 2 个比较对象间没有显著性差异;仅在 Ringnorm 数据集上广义核极化准则劣于核极化准则,但观察图 2 可知在此数据集上 4 种方法所得测试精度的绝对差异小于 0.01.

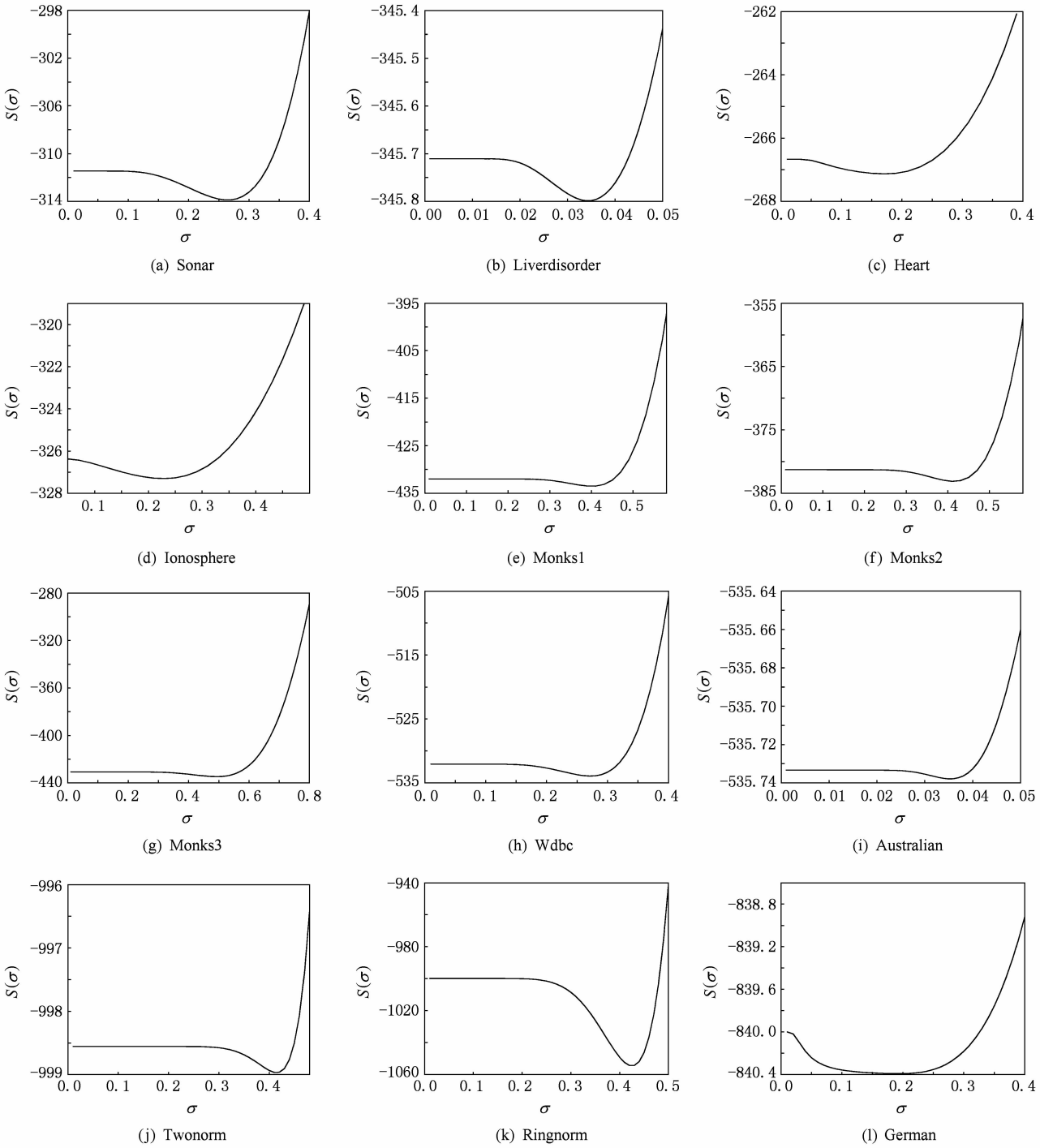


Fig. 1 $S(\sigma)$ on twelve standard data sets.

图 1 12 个标准数据集上 $S(\sigma)$的曲线图形

Table 3 Results of *t*-test on Binary Classification Data Sets

表 3 二分类数据集上测试精度的 *t* 假设检验结果

Comparison Object	Win:Tie:Loss
GKP/*KP*	10:1:1
GKP/*LKP*	7:5:0
GKP/*CV*	4:8:0

图 3 给出的是相应方法在每个数据集上平均运行时间的对数值(以 e 为底). 对于 3 种准则而言,运行时间记录的是寻找最优核参数的时间与支持向量机的训练时间之和. 图 3 说明利用 3 类核极化准则来优化核参数训练支持向量机相比交叉验证法能明显地提高时间效率. 对比 3 类核极化准则的时间复杂度,在 6 个数据集上广义核极化准则消耗时间稍

长,这可能是由于它需要同时计算 $\boldsymbol{K}_{\mathrm{L}}$ 与 $\boldsymbol{Y}_{\mathrm{C}}$ 花费时间较多的缘故,但是总体上说来这 3 种准则之间训练时间损耗的差异并不大.

从本节的实验结果可以看出对二分类问题而言,广义核极化准则是一种有效的高斯核参数选择准则.

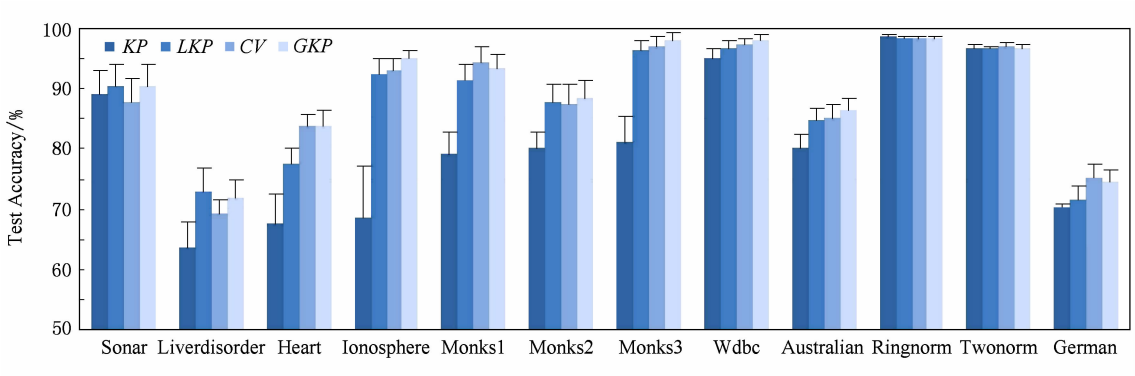


Fig. 2　Comparison of testing results on binary classification datasets.

图 2　二分类数据集上测试结果比较

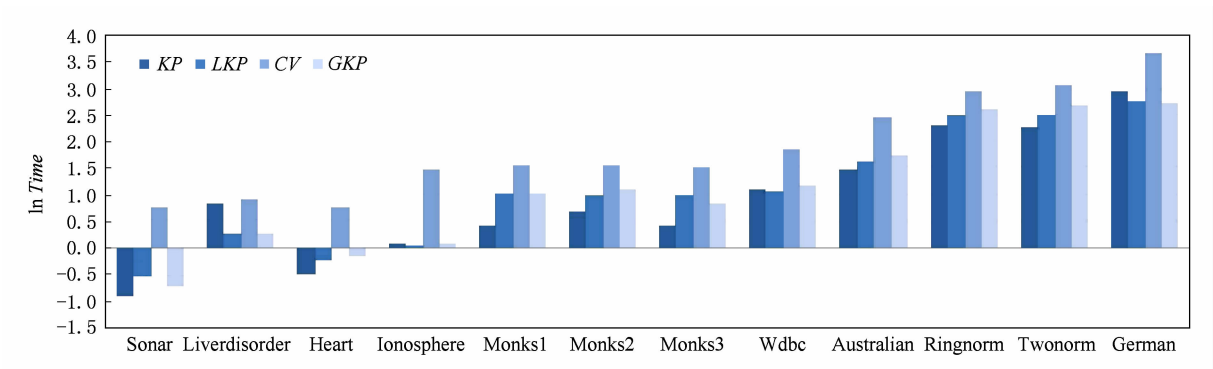


Fig. 3　Comparison of logarithm value of average training time on binary classification datasets.

图 3　二分类数据集上平均运行时间的对数值比较

3.4　多分类数据集上的实验比较

在多分类数据集上,本节对比利用多类核极化准则、多类局部核极化准则和多类广义核极化准则优化核参数所得测试精度与利用交叉验证方法得到的测试精度.与二分类数据集上的实验设计相同,每个数据集上都进行了 10 次独立的实验.

图 4 给出 7 个多分类数据集上测试精度的平均值及标准差.从图 4 可以看出,应用各种核极化准则得到的测试精度多数情形下不逊色于交叉验证技术的遍历寻参,且多类广义核极化准则在 5 个数据集上得到最高的测试精度.

表 4 列出的是对测试精度采用显著性水平为 0.05 的成对 t 假设检验所得到的统计结果.表 4 显示,多类广义核极化准则分别在 3 个及 2 个数据集上显著优于多类核极化准则及交叉验证方法,在 4 个及 5 个数据集上多类广义核极化准则与后 2 种方法的分类性能相当;多类广义核极化准则与多类局部核极化准则在 7 个数据集上均没有明显的差异.

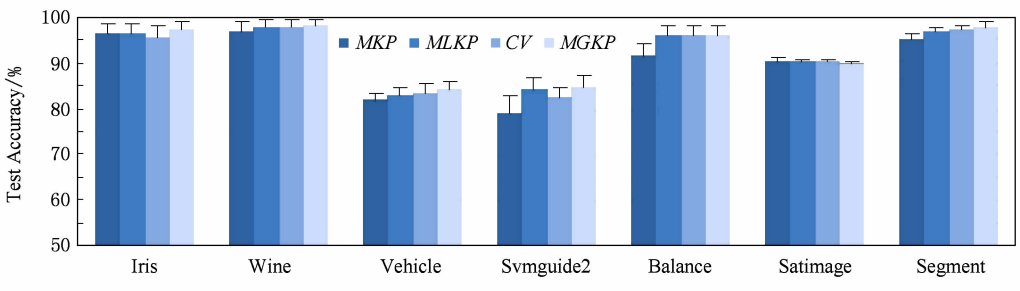


Fig. 4　Comparison of testing results on multiclass classification datasets.

图 4　多分类数据集上测试结果比较

Table 4 Results of *t*-test on Multiclass Classification Data Sets

表 4 多分类数据集上测试精度的 *t* 假设检验结果

Comparison Object	Win:Tie:Loss
MGKP/*MKP*	3:4:0
MGKP/*MLKP*	0:7:0
MGKP/*CV*	2:5:0

这说明对广义核极化准则及局部核极化准则进行多分类情形下的推广(2.3 节)是有效的.

图 5 展示的是平均运行时间以 e 为底的对数值,其中运行时间的定义与 3.3 节相同. 图 5 展示出与二分类问题相同的结论,即采用独立于学习器的核选择准则学习核参数相比交叉验证方法能较明显地提高时间效率,且 3 种准则间训练的时间复杂度差异不大.

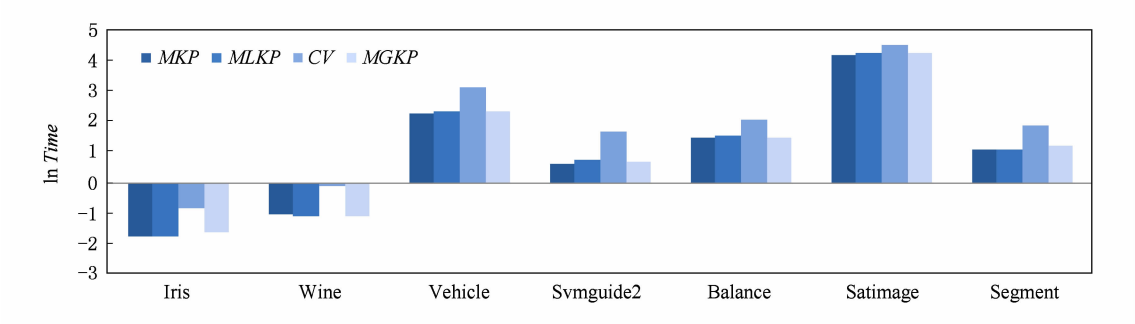


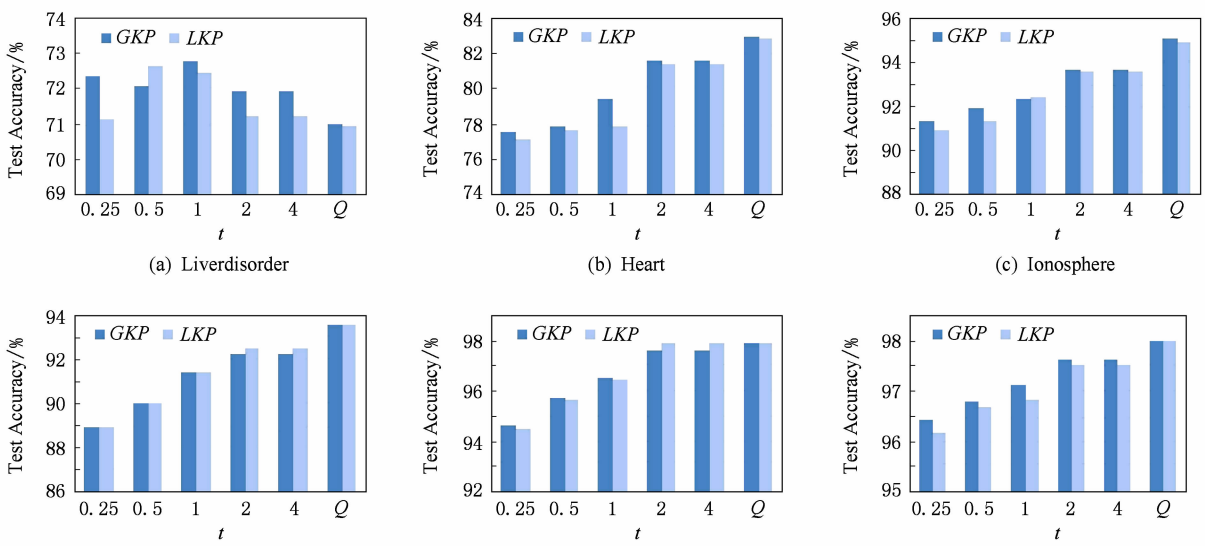
Fig. 5 Comparison of logarithm value of average training time on multiclass classification datasets.

图 5 多分类数据集上平均运行时间的对数值比较

总之,数值实验说明多类广义核极化准则相比交叉验证方法能大幅地节省优化时间,且多类广义核极化准则的分类性能稍优胜于多类核极化准则,与多类局部核极化准则的分类精度相差不大. 简言之,多类广义核极化准则与多类局部核极化准则是有效的核选择准则.

3.5 参数 *t* 对分类精度的影响

为了刻画局部结构,局部核极化准则与广义核极化准则及其相应的多类扩展形式都引入了一个调节参数 t,文献[20]曾指出该参数的取值不宜过大也不宜过小,并简单赋值 $t=1$. 本节试图通过实验直观说明 t 的取值对测试精度的影响. 在表 1 所提供的 19 个数据集上,分别令 t 遍历集合 $\{2^{-2},2^{-1},2^{0},2^{1},2^{2},Q\}$,其中 $Q=1/\min\{B,D\}$,B 与 D 的定义见 2.1 节. 对涉及到参数 t 的 4 种核极化准则(局部核极化准则、广义核极化准则、多类局部核极化准则与多类广义核极化准则)进行分组实验. 实验设计与 3.3 节及 3.4 节对应相同. 由于在二分类数据集 Sonar, Monks3, Ringnorm 和 Twonorm 以及多分类数据集 Iris, Satimage 和 Segment 上,测试精度的变化区间长度小于 0.01,故而仅列出其余 12 个数据集上的测试结果,如图 6 所示:



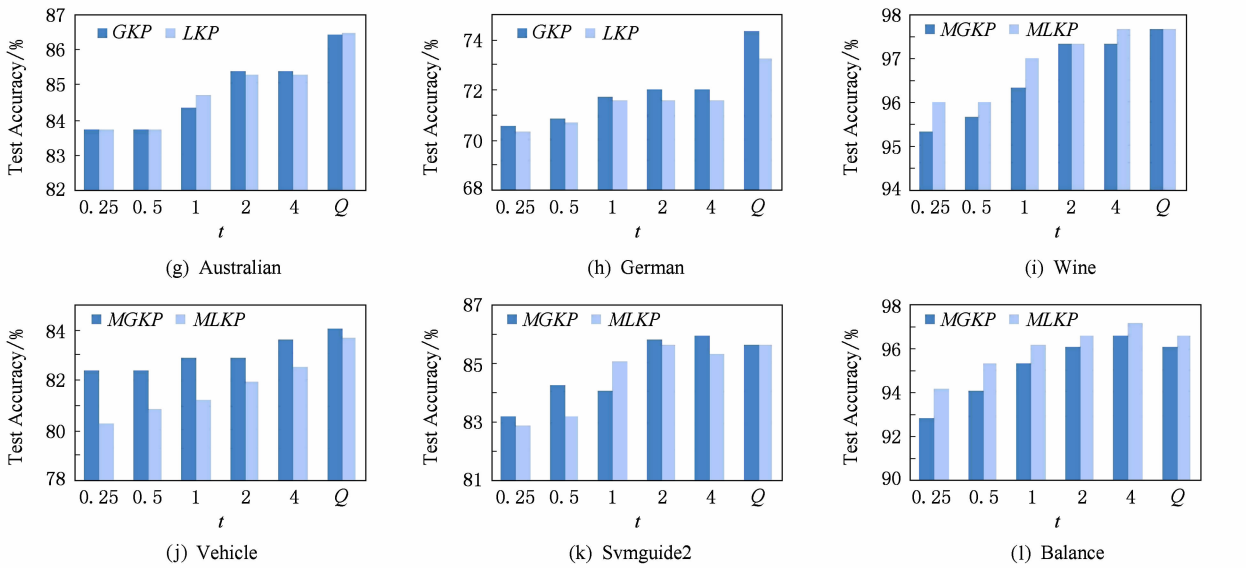


Fig. 6　Influence of t on the test accuracy.

图 6　参数 t 对测试精度的影响

图 6 中前 8 个数据集(图 6(a)～(h))是二分类数据集,在除 Liverdisorder 之外的 7 个数据集上,无论是局部核极化准则还是广义核极化准则都在 $t=Q$ 处取到最高的测试精度. 对于多分类数据集,多类局部核极化准则与多类广义核极化准则在 Wine 和 Vehicle 数据集上均于 $t=Q$ 处取到最高精度;在另 2 个多分类数据集上,虽然这 2 种准则在 $t=Q$ 处没有得到最好的测试精度,但所得测试精度与 t 取遍 $\{2^{-2},2^{-1},2^{0},2^{1},2^{2}\}$ 得到的最高精度相差并不大. 由此,可以说无论对二分类问题还是多分类问题而言,$t=Q$ 是一个可以接受的较好赋值.

此外,由图 6 也注意到,在 8 个二分类数据集中,当对 t 赋相同值时,广义核极化准则多数情况下都优于局部核极化准则;而在多分类数据集上,多类广义核极化准则与多类局部核极化准则的分类性能相当,这和 3.3 节与 3.4 节得出的结论是吻合的.

4　结　　语

本文提出利用广义核极化准则及多类广义核极化准则来分别解决核方法中的二分类以及多分类问题中的高斯核参数优化问题,并利用支持向量机验证所提优化准则的有效性. 由于广义核极化准则独立于学习算法,且其目标函数的最优解存在且唯一,因此利用许多成熟的优化算法可以快速地搜索到最优参数值,进而相比交叉验证方法能明显地节省寻参时间. 与标准核极化准则相比,广义核极化准则能更好地刻画特征空间中类别间的分离度,因此能搜索到对应更高分类精度的高斯核参数. 本文补充证明了局部核极化准则对应目标函数最优解的存在唯一性,并将其拓展至多分类情形. 标准数据集上的数值实验验证了所提准则的有效性,且本文所提的准则及所采用的解析方法可拓广应用于其他类径向基核函数.

为保持局部结构,广义核极化准则和局部核极化准则及其对应的多类扩展形式都引入了调节系数 t,本文通过实验对比展示 t 对测试精度的影响,并给出了参数 t 的一个建议取值,但如何从理论上说明这个取值的合理性还应进一步的研究. 此外,在当前大数据背景下,广义核极化准则与多类广义核极化准则的有效性还值得进一步探讨.

参　考　文　献

[1] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis [M]. Cambridge, USA: Cambridge University Press, 2004: 25-46

[2] Vapnik V N. The Nature of Statistical Learning Theory [M]. Berlin: Springer, 1998: 493-520

[3] Xiong Huilin, Swamy M N S, Ahmad M O. Optimizing the kernel in the empirical feature space [J]. IEEE Trans on Neural Networks, 2005, 16(2): 460-474

[4] Daoud E A, Turabieh H. New empirical nonparametric kernels for support vector machine classification [J]. Applied Soft Computing, 2013, 13(4): 1759-1765

[5] Tsuda K, Kawanabe M, Rätsch G, et al. A new discriminative kernel from probabilistic models [J]. Neural Computation, 2002, 14(10): 2397-2414

[6] Liao Shizhong, Jia Lei. Constructing a new spherical kernel function [J]. Journal of Computer Research and Development, 2007, 44(Suppl Ⅱ): 398-402 (in Chinese)
(廖士中, 贾磊. 一类新的球面核函数的构造[J]. 计算机研究与发展, 2007, 44(增刊Ⅱ): 398-402)

[7] Chapelle O, Vapnik V N, Bousquet O, et al. Choosing multiple parameters for support vector machines [J]. Machine Learning, 2002, 46(1): 131-159

[8] Joachims T. Estimating the generalization performance of a SVM efficiently [C] //Proc of the 17th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2000: 431-438

[9] Wang Lei, Xue Ping, Chan K L. Two criteria for model selection in multiclass support vector machines [J]. IEEE Trans on Systems, Man, and Cybernetics—Part B: Cybernetics, 2008, 38(6): 1432-1448

[10] Wu Kuo-Ping, Wang Sheng-de. Choosing the kernel parameters for support vector machine by the inter-cluster distance in the feature space [J]. Pattern Recognition, 2009, 42(5): 710-717

[11] Xu Zongben, Dai Mingwei, Meng Deyu. Fast and efficient strategies for model selection of Gaussian support vector machine [J]. IEEE Trans on Systems, Man, and Cybernetics—Part B: Cybernetics, 2009, 39(5): 1292-1307

[12] Wang Jie, Lu Haiping, Plataniotis K N, et al. Gaussian kernel optimization for pattern classification [J]. Pattern Recognition, 2009, 42(7): 1237-1247

[13] Cristianini N, Shawe-Taylor J, Elisseeff A, et al. On kernel-target alignment [C] //Proc of Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002: 367-373

[14] Liu Xiangdong, Luo Bin, Chen Zhaoqian. Optimal model selection for support vector machines [J]. Journal of Computer Research and Development, 2005, 42(4): 576-581 (in Chinese)
(刘向东, 骆斌, 陈兆乾. 支持向量机最优模型选择的研究[J]. 计算机研究与发展, 2005, 42(4): 576-581)

[15] Kandola J, Shawe-Taylor J, Cristianini N. On the extensions of kernel alignment, 120 [R]. London: University of London, 2002

[16] Baram Y. Learning by kernel polarization [J]. Neural Computation, 2005, 17(6): 1264-1275

[17] Cortes C, Mohri M, Rostamizadeh A. Algorithms for learning kernels based on centered alignment [J]. Journal of Machine Learning Research, 2012, 13(1): 795-828

[18] Zhong Shangping, Chen Daya, Xu Qiaofen, et al. Optimizing the Gaussian kernel function with the formulated kernel target alignment criterion for two-class pattern classification [J]. Pattern Recognition, 2013, 46(7): 2045-2054

[19] Chen Bo, Liu Hongwei, Bao Zheng. Optimizing the data-dependent kernel under a unified kernel optimization framework [J]. Pattern Recognition, 2008, 41(6): 2107-2119

[20] Wang Tinghua, Tian Shengfeng, Huang Houkuan, et al. Learning by local kernel polarization [J]. Neurocomputing, 2009, 72(13/14/15): 3077-3084

[21] Wang Wenjian, Xu Zongben, Lu Weizhen, et al. Determination of the spread parameter in the Gaussian kernel for classification and regression [J]. Neurocomputing, 2003, 55(10): 643-663

[22] Nguyen C H, Ho T B. An efficient kernel matrix evaluation measure [J]. Pattern Recognition, 2008, 41(11): 3366-3372

[23] Afkanpour A, Szepesvári C, Bowling M. Alignment based kernel learning with a continuous set of base kernels [J]. Machine Learning, 2013, 91(3): 305-324

[24] Lu Yanting, Wang Liantao, Lu Jianfeng, et al. Multiple kernel clustering based on centered kernel alignment [J]. Pattern Recognition, 2014, 47(11): 3656-3664

[25] Kuang Jichang. Applied Inequalities [M]. Jinan: Shandong Science and Technology Press, 2010: 590-592 (in Chinese)
(匡继昌. 常用不等式[M]. 济南: 山东科技出版社, 2010: 590-592)

[26] Wang Tinghua, Zhao Dongyan, Feng Yansong. Two-stage multiple kernel learning with multiclass kernel polarization [J]. Knowledge-Based Systems, 2013, 48: 10-16

[27] Lichman M. UCI machine learning repository [EB/OL]. 2013 [2015-02-04]. http://archive.ics.uci.edu/ml

[28] Delve. Delve datasets [EB/OL]. 2002 [2015-02-04]. http://www.cs.toronto.edu/~delve/data/datasets.html

[29] Chang Chih-Chung, Lin Chih-Jen. LIBSVM-A library for support vector machines [EB/OL]. 2014 [2015-02-04]. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Tian Meng, born in 1979. PhD candidate, associate professor. Her research interests include pattern recognition, machine learning and kernel methods.

Wang Wenjian, born in 1968. PhD, professor. Senior member of China Computer Federation. Her research interests include neural networks, support vector machines, machine learning theory.