软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2522−2534 [doi: 10.3724/SP.J.1001.2013.04483]

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

L_p 范数约束的多核半监督支持向量机学习方法*

胡庆辉[1,2], 丁立新[1], 何进荣[1]

[1](软件工程国家重点实验室(武汉大学 计算机学院),湖北 武汉 430072)

[2](桂林航天工业学院 信息工程系,广西 桂林 541004)

通讯作者: 胡庆辉, E-mail: huqinghui2004@126.com

摘 要: 在机器学习领域,核方法是解决非线性模式识别问题的一种有效手段.目前,用多核学习方法代替传统的单核学习已经成为一个新的研究热点,它在处理异构、不规则和分布不平坦的样本数据情况下,表现出了更好的灵活性、可解释性以及更优异的泛化性能.结合有监督学习中的多核学习方法,提出了基于 L_p 范数约束的多核半监督支持向量机(semi-supervised support vector machine,简称 S^3VM)的优化模型.该模型的待优化参数包括高维空间的决策函数 f_m 和核组合权系数 θ_m.同时,该模型继承了单核半监督支持向量机的非凸非平滑特性.采用双层优化过程来优化这两组参数,并采用改进的拟牛顿法和基于成对标签交换的局部搜索算法分别解决模型关于 f_m 的非平滑及非凸问题,以得到模型近似最优解.在多核框架中同时加入基本核和流形核,以充分利用数据的几何性质.实验结果验证了算法的有效性及较好的泛化性能.

关键词: 半监督;支持向量机;拟牛顿法;多核学习;半监督支持向量机

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 胡庆辉,丁立新,何进荣.L_p 范数约束的多核半监督支持向量机学习方法.软件学报,2013,24(11):2522−2534. http://www.jos.org.cn/1000-9825/4483.htm

英文引用格式: Hu QH, Ding LX, He JR. L_p norm constraint multi-kernel learning method for semi-supervised support vector machine. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2522−2534 (in Chinese). http://www.jos.org.cn/1000-9825/4483.htm

L_p Norm Constraint Multi-Kernel Learning Method for Semi-Supervised Support Vector Machine

HU Qing-Hui[1,2], DING Li-Xin[1], HE Jin-Rong[1]

[1](State Key Laboratory of Software Engineering (School of Computer, Wuhan University), Wuhan 430072, China)

[2](School of Information Engineering, Guilin University of Aerospace Technology, Guilin 541004, China)

Corresponding author: HU Qing-Hui, E-mail: huqinghui2004@126.com

Abstract: Kernel method is an effective approach to solve the nonlinear pattern recognition problems in the field of machine learning. At present, multiple kernel method has become a new research focus. Compared with the traditional single kernel method, multiple kernel method is more flexible, more interpretable and has better generalization performance when dealing with heterogeneous, irregular and non-flat distribution samples. A multi-kernel S^3VM optimization model based on L_p norm constraint is presented in this paper in accordance with kernel method of supervised learning. Such model has two sets of parameters including decision functions f_m in reproducing kernel Hilbert space and weighted kernel combination coefficients, and inherits the non-smooth and non-convex properties from single-kernel based S^3VM. A two-layer optimization procedure is adopted to optimize these two groups of parameters, and an improved Quasi-Newton method named subBFGS as well as a local search algorithm based on label switching in pair are used to solve

* 基金项目: 国家自然科学基金(60975050); 广东省省部产学研结合专项(2011B090400477); 珠海市产学研合作专项资金(2011A050101005, 2012D0501990016); 珠海市重点实验室科技攻关项目(2012D0501990026); 中央高校基本科研业务费专项资金(2012211020209); 桂林航天工业学院科研基金(Y12Z028)

收稿时间: 2013-05-21; 修改时间: 2013-07-17; 定稿时间: 2013-08-27

non-smooth and non-convex problems respectively with respect to f_m. Base kernels and manifold kernels are added into the multi-kernel framework to exploit the geometric properties of the data. Experimental results show that the proposed algorithm is effective and has excellent generation performance.

Key words: semi-supervised; support vector machine (SVM); quasi-Newton method; multiple kernel learning; semi-supervised support vector machine (S^3VM)

在传统的监督学习中,分类器通过对大量有标记的训练样本进行学习,从而建立模型用于预测未知数据的标记.在实际应用中,如文本文类、图像检索等等,收集大量未标记的样本很容易,但对这些样本进行正确的标记,则需要花费很大的代价,这使得监督学习受到一定的限制.近年来,结合了监督学习和无监督学习特点的半监督学习方法得到了广泛关注.半监督学习主要研究样本在只有少量被标记的情况下,通过使用大量未标记样本来获得具有理想性能和推广能力的学习器.它通常基于两种假设[1]:

- 一是聚类假设,即同一聚类中的样本点很可能具有同样的类别标记.这个假设要求决策边界所穿过的应当是数据点较为稀疏的区域.
- 二是流形假设,即高维中的数据存在着低维的特性.它要求处于一个很小的局部邻域内的样本具有相似的性质,反映了决策函数的局部平滑性.

半监督学习对于减少标注代价、提高分类器泛化性能具有非常重大的实际意义.

半监督支持向量机(semi-supervised support vector machine,简称 S^3VM)是一种基于聚类假设的半监督学习方法,其优化目标是利用已标记和未标记的样本构建分类器,使其包含的分类间隔能够最大地分隔原始已标记和未标记样本,新找到的最优分类边界满足对原始未标记样本的分类具有最小泛化误差.S^3VM 最初是应用于文本分类的直推式支持向量机 TSVM[2].它实际上是一个非凸非平滑函数的优化问题.为了解决这一问题,研究者提出了各种各样的方法[3],常见的有:局部组合搜索法[2]、梯度下降法[4]、凹凸法[5]、确定性退火方法[6]、连续优化方法[7]、半定规划方法[8]、分支定界法[9]等等.这些算法根据参数特点的不同可以分为两大类:一种是基于组合的 S^3VM,另一种是连续方法的 S^3VM.

当样本数据为非线性可分时,需要使用核技巧(kernel trick),通过一个非线性映射$\phi: X\to H$ 把输入数据映射到高维再生核希尔伯特特征空间 H,然后在新的特征空间里寻找线性决策边界.核函数把非线性映射和特征空间中两个向量的内积两者结合起来,使得非线性映射ϕ不需要明确指定,而高维空间中的线性超平面可以由所有训练样本与测试样本在特征空间中的内积的线性组合表示,从而有效避免了维数灾难[10].

这种采用核技巧的方法是解决非线性模式识别问题的一种有效方法.自从支持向量机(support vector machine,简称 SVM)学习方法被提出来以后,核方法已经成功地应用到模式识别的诸多方面,如回归估计[11]、概率密度估计[12]、模式分类[13]以及子空间分析[14]等等.但是这些方法绝大多数都是基于单个特征空间映射的单核方法.在不同的应用场合,核函数的选择和参数的设置不同,往往使得模型的性能表现有很大的差异;并且对于高维、包含有异构信息、分布不平坦以及不规则的样本数据,采用单核进行映射对样本的处理不一定合适.于是,一些学者提出了基于核组合的多核学习算法[15−18].构造多核学习模型首先要考虑多核组合方法,常见的有线性组合方法[13]、扩展合成方法[19]、非平稳组合方法[16]及局部多核学习方法[20]等等,其中,线性凸组合多核学习是最常见的一种.大量研究结果表明,多核学习方法是一种灵活性更强的基于核的学习方法,用它代替传统的单核学习,可以大幅度提高模型的可解释性和泛化性能[21].

多核线性组合学习过程就是求取核组合权系数的过程.针对如何获取核组合的权系数,目前已经有了多种有效的学习手段,如 Boosting 多核组合模型学习方法[22]、超核学习方法[17]、半定规划多核学习方法[23]、二次约束型二次规划多核学习方法[21]以及简单多核学习方法[15]等等.以上的所有多核学习方法对权系数的约束条件都是$\|\theta\|\leqslant 1$(即 L_1 约束),最终得到的是权系数的稀疏解,即核组合中多数核的权系数为 0.稀疏核组合的好处是模型易于解释,易于处理.它表明:不相关的或代价较大的核应该从模型中去掉,以减少冗余,提高运算效率.但稀疏解可能会导致模型有用信息的丢失和泛化性能变差.有研究表明,稀疏核组合在性能上很可能还不如直接使用非加权和核组合 $K=\sum_m K_m$ 的标准 SVM[24].因此,Kloft 等人提出了一种非稀疏的多核学习方法[25],通过对

权系数施加L_2范数约束,得到权系数的非稀疏解.相对于L_1约束,L_2约束的多核学习有更强的抗噪声能力和更好的鲁棒性.随后,Kloft 等人又将 L_2 范数约束推广到任意 L_p 范数约束[26,27],其中,p>1,以进一步增强模型的泛化性能,并对 L_p 范数约束的收敛特性进行了讨论[28].

S^3VM 学习算法是基于聚类假设的,通常在具有这种结构的数据集上表现有较好的分类精度,但是对于具有流形结构特征的数据集以及数据包含有多个子类的问题,考虑了数据几何性质的 LapSVM[29]往往表现更为出色.事实上,任何一种半监督学习算法都是在特定假设条件下有较好的性能,没有证据表明,基于一种假设的算法就一定优于基于另外一种假设的算法.而组合两种假设,在 S^3VM 中考虑数据本身的几何特征,可以使算法得到一定的改善[3,30].

本文结合传统的 S^3VM 学习以及多核学习的理论和方法,提出了基于 L_p 范数约束的多核 S^3VM 学习模型(L_p-MKL-S^3VM),在多核框架中同时学习基本核和流形核,以实现两种假设的结合.通过对核组合系数θ施加 p 范数约束,得到θ的非稀疏解,以提高 S^3VM 的可解释性和泛化性能.

本文的模型继承了 S^3VM 的非凸非平滑特性.我们采用一种改进的拟牛顿法[31]和成对标签交换的局部搜索法来解决其非凸非平滑问题,并将提出的算法应用在二分类问题上,在人工数据集和真实数据集上进行仿真,通过实验证明了算法的有效性和较好的泛化性能.

本文第 1 节概要介绍 S^3VM 的基本形式.第 2 节介绍一种改进的拟牛顿法(用于解决模型非平滑问题).第 3 节提出基于 L_p 范数约束的多核 S^3VM(L_p-MKL-S^3VM)的优化模型.第 4 节推导出 L_p-MKL-S^3VM 的求解过程.第 5 节是本文算法与其他算法的实验分析.最后是结论和展望.

1 半监督支持向量机(S^3VM)

为了引导出 L_p-MKL-S^3VM 模型,首先介绍 S^3VM 的优化问题.

考虑二分类问题,输入数据集为 $D=\{x_1,x_2,\ldots,x_{l+u}\}$,不失一般性,设数据集的前 l 个为已标记样本 $\{x_i,y_i\}_1^l$,y_i=±1,紧接着是 u 个未标记样本,其中,$l<<u$,$l+u=n$ 且 $x_i\in R^d$.通常,S^3VM 解决如下目标函数的最小化问题:

$$J(f,b)=\frac{\lambda}{2}\|f\|_H^2+C\sum_{i=1}^{l}V(y_i,g(x_i))+C^*\sum_{i=l+1}^{l+u}U(|g(x_i)|) \tag{1}$$

公式(1)右边第 1 项定义了一个标准 SVM,超参数 C 和 C^*用于平衡已标记样本和未标记样本的拟合误差.

为避免出现将所有未标记样本分配到同一类的极端情况,最小化公式(1)需要满足平衡约束条件:

$$\frac{1}{u}\sum_{i=l+1}^{n}\max(0,y_i)=r,$$

其中,r 为未标记数据中正样本所占比例.对于未标记数据而言,r 是未知的,通常通过已标记数据中正样本所占的比例进行计算,也可以利用具体分类问题的先验知识来设置.

式(1)中,$y_i=sign(g(x_i))$表示第 i 个样本的类别标签,取值为+1 或−1.U 和 V 通常采用如下的损失函数:

(1) $V(y_i,g(x_i))=\max(0,1-y_ig(x_i))$,用于度量已标记样本的真实类别标签 y_i 与计算值 $g(x_i)$之间的偏差;

(2) $U(|g(x_i)|)=\begin{cases}1-|g(x_i)|, & |g(x_i)|<1\\ 0, & \text{其他}\end{cases}$,用于对未标记样本被错误分类后实施惩罚.

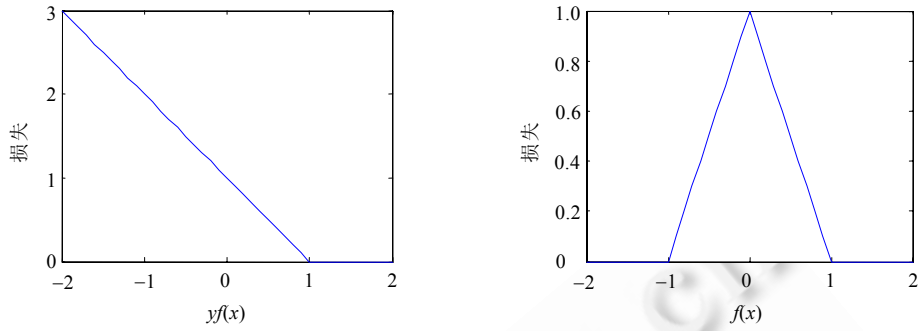
决策函数定义为 $g(x)=f(x)+b$,其中,标量 b 为偏差项,为了简化计算,本文中省略了该项;f 是再生核希尔伯特空间 H 中的函数.对于非线性可分的样本数据,首先通过“核技巧”将数据映射到高维再生核希尔伯特空间 H 中,然后在 H 中构建线性决策边界.设ϕ:$X\to H$ 为一个映射,$k(x,y)=\phi(x)^T\phi(y)$为 H 对应的核,一旦采用了适当的核 k,就隐含定义了原始样本到高维空间 H 的映射函数ϕ,根据再生核希尔伯特空间的再生属性,有:

$$f(x_i)=\langle f,k(x_i,\cdot)\rangle,$$

于是,公式(1)可写成

$$J(f)=\frac{\lambda}{2}\|f\|_H^2+C\sum_{i=1}^{l}V(y_i,fk(x_i,\cdot))+C^*\sum_{i=l+1}^{l+u}U(|fk(x_i,\cdot)|) \tag{2}$$

函数 V 和 U 的函数曲线图分别如图 1(a)和图 1(b)所示.由图可知,目标函数(2)是非凸(图 1(a))、非平滑(图 1(b))的.



(a) 损失函数 V (b) 损失函数 U

Fig.1 Characteristics of loss functions V and U

图 1 损失函数 V 和 U 的特征

2 一种改进的拟牛顿法 subBFGS

第 3 节推导出的 L_p-MKL-S^{3}VM 的优化函数继承了公式(2)的非凸非平滑特性.本文采用文献[31]中提出的一种改进拟牛顿法(subBFGS)来解决函数的非平滑问题,这里首先对该方法做简要介绍.

拟牛顿法是一种求解非线性优化问题的有效方法,具有超线性收敛性,用不包含二阶导数的矩阵来近似牛顿法中 Hessian 矩阵的逆.根据构造近似矩阵的方法不同,有不同的拟牛顿法,常见的有 BFGS 公式和 LBFGS 公式.

拟牛顿法只能优化平滑的目标函数,但是在机器学习领域,很多目标函数在某些点是非平滑的,如公式(2)中的损失函数 V 在 $yf(x)=1$ 处是非平滑的(如图 1(a)所示),但是目标函数在这些点的次梯度 g 总是存在.对于平滑函数,负梯度方向始终是函数的下降方向;但是对于非平滑函数,在非平滑点的任意负次梯度方向不一定是函数的下降方向,从而导致拟牛顿法失效.文献[31]提出的 subBFGS 方法通过利用束方法(bundle method)的束搜索(bundle search)产生下降方向 p 的迭代过程,以寻找非平滑点确切的下降方向.其基本思想如下:

给定一个平滑目标函数 $J:R^d\to R$ 及当前迭代 $\omega_t\in R^d$,构造一个局部二次型:

$$Q_t(p)=J(\omega_t)+\frac{1}{2}p^TB_t^{-1}p+\nabla J(\omega_t)^Tp \tag{3}$$

其中,正定矩阵 $B_t\succ 0$ 是函数 J 的 Hessian 矩阵的近似,∇J 是 J 的梯度.将二次型公式(3)改为如下形式:

$$Q_t(p)=J(\omega_t)+M_t(p) \tag{4}$$

$$M_t(p)=\frac{1}{2}p^TB_t^{-1}p+\sup_{g\in\partial J(\omega_t)}g^Tp \tag{5}$$

通过最小化 $Q_t(p)$可以导出ω_t的迭代公式,这等价于最小化 $M_t(p)$,即

$$\min M_t(p)=\min_{p\in R^d}\left(\frac{1}{2}p^TB_t^{-1}p+\sup_{g\in\partial J(\omega_t)}g^Tp\right) \tag{6}$$

通过一个迭代过程,最小化 $M_t(p)$的凸下界来渐进逼近 $M_t(p)$.第 i 次迭代时,$M_t(p)$的凸下界为

$$M_t^{(i)}(p)=\frac{1}{2}p^TB_t^{-1}p+\sup_{j\leqslant i}g^{(j)T}p \tag{7}$$

其中,$i,j\in N,g^{(j)}\in\partial J(\omega_t)$.给定一个迭代 $p^{(i)}$,公式(7)的下界通过计算 $g^{(i+1)}=\arg\sup_{g\in\partial J(\omega_t)}g^Tp^{(i)}$ 而逐渐收紧,使得 $M_t^{(i)}(p)\leqslant M_t^{(i+1)}(p)\leqslant M_t(p),\forall p\in R^d$ 成立.

$$\begin{cases}\min\limits_{p\in R^d,\xi} M_t^{(i)}(p)=\min\limits_{p,\xi}\left(\dfrac{1}{2}p^TB_t^{-1}p+\xi\right)\\ \text{s.t. } g^{(j)T}p\leqslant\xi,\ \forall j\leqslant i\end{cases} \tag{8}$$

则

$$\begin{cases}p_t^{(i)}=\arg\min\limits_{p\in R^d} M_t^{(i)}(p)=\arg\min\limits_{p\in R^d}\left(\dfrac{1}{2}p^TB_t^{-1}p+\xi\right)\\ \text{s.t. } \forall g_1,...,g_k\in\partial J(\omega_t),\ g_i^Tp\leqslant\xi\end{cases} \tag{9}$$

我们的目标是寻找一个次梯度方向 p,使得$\forall g_1,\ldots,g_k\in\partial J(\omega_t)$,有 $g_i^Tp\leqslant 0$ 成立,subBFGS 从当前迭代的任意次梯度 g 和给定的下降方向 p(最开始不一定是真实的下降方向)开始,产生 p 的迭代过程,最终找到确切的下降方向,具体算法见文献[31].

3 基于 L_p 范数约束的多核 S^{3}VM(L_p-MKL-S^{3}VM)

本文考虑多核线性组合形式,给定 M 个不同特征映射 ϕ_m:$X\to H_m$,m=1,…,M,H_m 对应的再生核为 k_m,我们的目标是学习这 M 个核的线性组合,即 $k(x_i,x_j)=\sum_{m=1}^M\theta_mk_m(x_i,x_j)$, M 是基本核总个数,θ_m 为第 m 个核的权系数,是待优化参数.在多核学习情况下,决策函数定义为

$$f(x)=\sum_{m=1}^M f_m(x) \tag{10}$$

则目标函数公式(2)在多核学习的情况下定义为如下形式:

$$J(f_m,\theta_m)=\frac{\lambda}{2}\left(\sum_{m=1}^M\|f_m\|_{H_m}\right)^2+C\sum_{i=1}^l V\left(y_i,\sum_{m=1}^M f_mk_m(x_i,\cdot)\right)+C^*\sum_{i=l+1}^{l+u}U\left(\left|\sum_{m=1}^M f_mk_m(x_i,\cdot)\right|\right) \tag{11}$$

当 f_m=0 时,公式(11)右边的第 1 项是不可微的,非平滑部分包含了第 1 项和第 3 项,根据文献[26]提出的基于 L_p 范数约束的监督学习优化代价函数,将公式(11)简化为如下形式:

$$\begin{cases}J(f_m,\theta_m)=\dfrac{\lambda}{2}\sum\limits_{m=1}^M\dfrac{1}{\theta_m}\|f_m\|_{H_m}^2+C\sum\limits_{i=1}^l V\left(y_i,\sum\limits_{m=1}^M f_mk_m(x_i,.)\right)+C^*\sum\limits_{i=l+1}^{l+u}U\left(\left|\sum\limits_{m=1}^M f_mk_m(x_i,.)\right|\right)\\ \text{s.t. } \dfrac{1}{u}\sum_{i=l+1}^{l+u}\sum_{m=1}^M K_{m(i)}^Tf_m=r,\ \text{约定}\dfrac{t}{0}=\begin{cases}0, & t=0\\ \infty, & \text{其他}\end{cases}\end{cases} \tag{12}$$

本文对 θ_m 施加了 L_p 范数约束,即 $\left(\sum_{m=1}^M\theta_m^p\right)^{1/p}\leqslant 1, p\geqslant 1,\theta_m\geqslant 0$.当 p=1 时,得到的是 θ_m 稀疏解;当 p>1 时,得到的是 θ_m 非稀疏解.我们称该优化模型为 L_p-MKL-S^{3}VM.

4 优化 L_p-MKL-S^{3}VM 模型

直接优化目标函数 $J(f_m,\theta_m)$很难收敛,本文采用了双层优化过程对参数进行交替优化[26,30,32−34].把参数分成两组:第 1 组为 f_m,第 2 组为核混合系数 θ_m.首先固定 θ_m,则目标函数是关于 f_m 的非凸非平滑函数,本文采用 subBFGS 算法、退火方法和成对标签交换的局部搜索算法分别解决非平滑和非凸问题;然后固定 f_m,目标函数是关于 θ_m 的非线性单调函数,在 L_p 范数约束条件下,我们采用一阶导数求极值方法进行求解.这两个过程交替进行,直到满足事先设定的收敛准则.下面介绍具体的求解过程.

4.1 求解 f_m

固定 θ_m,目标函数是关于 f_m 的最小化问题 min$J(f_m)$.为了使用 subBFGS 算法,首先设 E,R,W 分别表示样本集中被错误分类、被分在边界区域和被正确分类的样本的下标集合,即

$$E=\left\{i\in\{1,2,...,l+u\}:1-y_i\sum_{m=1}^M f_mk_m(x_i,\cdot)>0\right\} \tag{13}$$

$$R=\left\{i\in\{1,2,...,l+u\}:1-y_i\sum_{m=1}^{M}f_m k_m(x_i,\cdot)=0\right\} \tag{14}$$

$$W=\left\{i\in\{1,2,...,l+u\}:1-y_i\sum_{m=1}^{M}f_m k_m(x_i,\cdot)<0\right\} \tag{15}$$

则关于f_m的导数(次微分)为

$$\frac{\partial J}{\partial f_m}=\frac{\lambda}{\theta_m}f_m-C\sum_{i=1}^{l}\beta_i y_i K_{m(i)}^T-C^*\sum_{i=l+1}^{l+u}\beta_i y_i K_{m(i)}^T=\overline{w}-\left(C\sum_{i=1,i\in R}^{l}\beta_i y_i K_{m(i)}^T+C^*\sum_{i=l+1,i\in R}^{l+u}\beta_i y_i K_{m(i)}^T\right) \tag{16}$$

其中,$\overline{w}=\frac{\lambda}{\theta_m}f_m-\left(C\sum_{i=1,i\in E}^{l}\beta_i y_i K_{m(i)}^T+C^*\sum_{i=l+1,i\in E}^{l+u}\beta_i y_i K_{m(i)}^T\right)$,$\beta_i=\begin{cases}1, & i\in E\\ [0,1], & i\in R\\ 0, & i\in W\end{cases}$,$K_m$表示第 m 个核对应的核矩阵,$K_{m(i)}$表示K_m的第 i 列.

在 subBFGS 算法中,对于给定的下降方向 p,需要计算该方向与次微分内积的上确界,即

$$\sup_{g\in\partial J(f_m)}g^T p=\sup\left(\frac{\partial J}{\partial(f_m)}\right)^T p=\overline{w}^T p-\inf\left(C\sum_{i=1,i\in R}^{l}\beta_i y_i K_{m(i)}^T+C^*\sum_{i=l+1,i\in R}^{l+u}\beta_i y_i K_{m(i)}^T\right)p \tag{17}$$

对于一个给定的方向 p,当 $i\in R$ 时,有$\beta_i\in[0,1]$,因此可以通过下面的约定得到 $\sup_{g\in\partial J(f_m)}g^T p$:

$$\beta_i=\begin{cases}0, & y_i K_{m(i)}^T p\geqslant 0\\ 1, & y_i K_{m(i)}^T p<0\end{cases} \tag{18}$$

得到了确切的下降方向 p 以后,需进行一维线性搜索确定最佳的步长η,通过求解下面的最小化问题得到:

$$\eta=\arg\min_{\eta>0}\phi(\eta)=\arg\min_{\eta>0}J(f_m+\eta p) \tag{19}$$

下面给出求解f_m的算法描述.

算法 1. 求解f_m.

输入:样本集$(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1}),\ldots,(x_n)$;参数$\lambda$,$C$, $\hat{C}^*$, θ_m,Hessian 矩阵 $B_m=I$,当前的f_m.

输出:f_m.

算法步骤:

1. $C^*=10^{-5}\hat{C}^*$;
2. WHILE $C^*\leqslant\hat{C}^*$
   2.1. 针对已标记和未标记的样本,利用 subBFGS 求解函数 $J(f_m)$;
   2.2. t=1;
   2.3. WHILE ($t\leqslant u/2$)
      IF 在未标记样本中,$\exists x_i,x_j$ 满足临时标签互换的条件
         则交换两个样本 i,j 的临时标签,并且 $t=t+1$;
      ELSE BREAK;
      END IF
   END WHILE
   2.4. $C^*=10\times C^*$;
   END WHILE
3. RETURN f_m;

公式(12)中的超参数 C^*需要事先设定,用于反映未标记样本在优化模型中的重要性.目标函数的非凸性质使得拟牛顿法得到的f_m将陷入局部最优解.S^{3}VM 的求解本身是一个 NP 问题,因此我们的目标是要获取全局次优解.在算法 1 中,首先让 C^*从一个比较小的值开始逐步增长,目的是避免目标函数在优化过程中过快陷入局部最优解,同时,第 2.3 步循环实现的是针对当前的局部最优解 f_m,通过成对标签交换方法进行局部搜索,在未标记

样本集合中,如果存在一对临时标签分别为+1 和−1 的样本 x_i,x_j,则互换这对临时标签后可以降低目标函数的值,即满足如下条件:

$$\begin{aligned} &Loss\left(y_i=+1,\sum_{m=1}^{M}K_{m(i)}^{T}f_m\right)+Loss\left(y_j=-1,\sum_{m=1}^{M}K_{m(j)}^{T}f_m\right)> \\ &\qquad Loss\left(y_i=-1,\sum_{m=1}^{M}K_{m(i)}^{T}f_m\right)+Loss\left(y_j=+1,\sum_{m=1}^{M}K_{m(j)}^{T}f_m\right) \end{aligned} \tag{20}$$

则将这两个样本的临时标签互换,通过这种成对临时标签互换的局部搜索算法可得到目标函数的次优解.

4.2 求解 θ_m

对于 θ_m,本文采用与文献[26]相同的求解方法.

固定 f_m,此时要求解的是如下关于 θ_m 函数的最小化问题:

$$\begin{cases} J(\theta_m)=\dfrac{\lambda}{2}\sum_{m=1}^{M}\dfrac{1}{\theta_m}\|f_m\|_{H_m}^2+C\sum_{i=1}^{l}V\left(y_i,\sum_{m=1}^{M}f_m k_m(x_i,\cdot)\right)+C^*\sum_{i=l+1}^{l+u}U\left(\left|\sum_{m=1}^{M}f_m k_m(x_i,\cdot)\right|\right) \\ \text{s.t. } \left(\sum_{m=1}^{M}\theta_m^p\right)^{1/p}\leqslant 1,\ \theta_m\geqslant 0 \end{cases} \tag{21}$$

最小化公式(21)等价于:

$$\min_{\theta_m:\theta_m\geqslant 0}\left\{\frac{\lambda}{2}\sum_{m=1}^{M}\frac{1}{\theta_m}\|f_m\|_{H_m}^2+C\sum_{i=1}^{l}V\left(y_i,\sum_{m=1}^{M}f_m(x_i)\right)+C^*\sum_{i=l+1}^{l+u}U\left(\left|\sum_{m=1}^{M}f_m(x_i)\right|\right)+\frac{\mu}{2}\left(\left(\sum_{i=1}^{M}\theta_i^p\right)^{1/p}\right)^2\right\} \tag{22}$$

其中,$\mu>0$.对 θ_m 求导并令其等于 0,得到:

$$-\frac{\lambda}{2\theta_m^2}\|f_m\|_{H_m}^2+\mu\theta_m^{p-1}\left(\sum_{i=1}^{M}\theta_i^p\right)^{2-p}=0,\ \forall m=1,\ldots,M \tag{23}$$

进一步将上式转化为如下的优化条件:

$$\exists\xi,\forall m=1,\ldots,M:\theta_m=\xi\left\|\sqrt{\lambda}f_m\right\|_{H_m}^{2/(p+1)} \tag{24}$$

由于 $f_m\neq 0,\left(\sum_{i=1}^{M}\theta_i^p\right)^{1/p}\leqslant 1$. 当 $\left(\sum_{i=1}^{M}\theta_i^p\right)^{1/p}=1$ 时,能得到 $\min J(\theta_m)$的最优解.将公式(24)带入 $\left(\sum_{i=1}^{M}\theta_i^p\right)^{1/p}=1$,得到:

$$\xi=\left(\sum_{i=1}^{M}\left\|\sqrt{\lambda}f_i\right\|_{H_i}^{2p/(p+1)}\right)^p \tag{25}$$

将公式(25)代入公式(24),得到:

$$\theta_m=\left\{\|f_m\|_{H_m}^{2/(p+1)}\right\}\Big/\left\{\left(\sum_{i=1}^{M}\|f_i\|_{H_i}^{2p/(p+1)}\right)^{1/p}\right\} \tag{26}$$

4.3 算法描述

最后,我们得到 L_p-MKL-S^{3}VM 的算法描述如下:

算法 2. L_p-MKL-S^{3}VM 优化算法.

输入:样本集$(x_1,y_1),\ldots,(x_l,y_l),(x_{l+1}),\ldots,(x_n)$.

输出:f_m,θ_m.

算法步骤:

1. 初始化:M=核的个数,$f_m=O,\theta_m=\sqrt[p]{1/M}$;
2. 利用已标记样本学习多核线性组合的 SVM,得到初始的 $f_m,\theta_m,m=1,\ldots,M$;
3. 对于未标记的样本集 $X_u(X_u=x_{l+1},\ldots,x_n)$,计算 $\sum_{m=1}^{M}K_{m(i)}^{T}f_m$, 并按降序进行排序;
4. 按正负样本的比例 r 给未标记样本赋值临时标签+1 或−1;
5. REPEAT

   固定 θ_m,利用算法 1 计算 f_m,直到 f_m 收敛;

固定f_m,利用公式(26)计算θ_m;

UNTIL θ_m收敛或者满足其他收敛条件.

4.4 时间复杂度分析

算法1、算法2之间是属于嵌套关系,算法1可以在很少迭代次数内得到确切的下降方向,算法2的循环也可以在较少的次数内使θ_m收敛.因此,算法2的时间复杂度体现在算法1的运算上,而算法1的时间复杂度体现在两个嵌套的循环,即时间复杂度为$O(n_u(n_{switches}+n_{S^3VM\text{-}subLBGFS}))$,其中,$n_u$为算法1中第1个循环的次数,$n_{switches}$为成对标签互换次数,最多为$u/2$次,$n_{S^3VM\text{-}subLBGFS}$为用subLBGFS算法求解$f_m$的迭代次数.另外,在进行标签互换和计算$f_m$时,都需要做$\sum_{m=1}^{M}K_{m(i)}^{T}f_m$运算,因此,总的时间复杂度还与核的个数$M$、输入样本集大小$l+u$有关,总的时间复杂度约为$O(M(l+u)^3)$.

文献[30]提出了一种基于L_1约束的多核半监督学习算法,并将该算法成功运用到BCI的数据分析.作者采用了DC Programming过程解决目标函数的非凸非平滑问题,通过对偶问题和梯度下降法求解核的组合系数.函数U为$U(|g(x_i)|)=R_s(g(x_i))+R_s(-g(x_i))-(1-s)$,其中,$R_s(g(x_i))=\max(0,1-y_ig(x_i))+\max(0,s-y_ig(x_i))$.这直接导致目标函数关于未标记样本被计算了2次,相当于未标记样本增加了1倍,从而算法的计算复杂度最坏情况下为$O(M(l+2u)^3)$.另外,该算法还增加了一个超参数s.

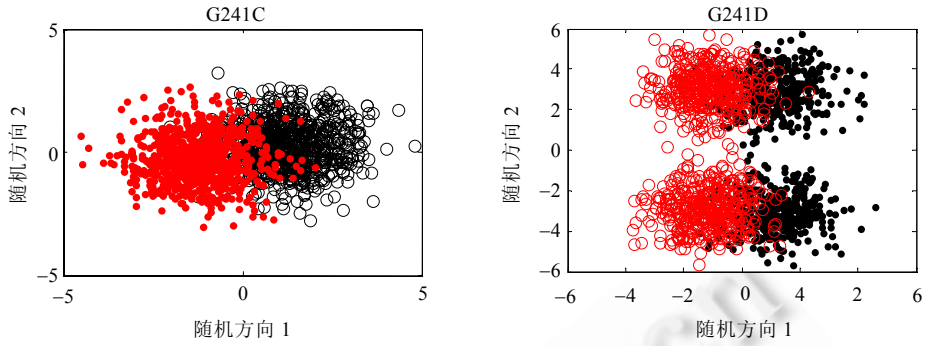
5 实验分析

为了验证算法的性能,实验分别与一些有代表性的半监督学习算法进行比较,包括S^3VM^{light}[2],LapSVM[29],S^3VM^{CCCP}[5]及TSVM-MKL[30],除TSVM-MKL外,其余的都是单核半监督学习算法.S^3VM^{light}通过解决一系列的SVM问题来优化目标函数,利用成对标签交换来改善优化结果,每次迭代交换两个样本的标签,以满足所施加的正负样本平衡约束.LapSVM是一种基于流形学习的SVM扩展方法.它将基于图的拉普拉斯作为正则化项应用到半监督学习中,以此提高分类的准确性.S^3VM^{CCCP}使用CCCP(concave-convex procedure)方法来优化目标函数,将目标函数分解为一个凸函数和一个非凸函数,再通过迭代优化这两个函数,最终达到优化目标函数的目的.TSVM-MKL是一个半监督多核学习框架,采用与S^3VM^{CCCP}相同的方法来解决目标函数的非凸非平滑问题,该算法对权系数施加的是L_1约束,因此得到的是核组合的稀疏解.

我们分别在3组人工数据集2MOONS,G241C,G241D及3组真实数据集COIL100,USPST,TEXT上进行实验.其中,后5种数据集来自文献[1]:

- G241C:241维,该数据集具有聚类假设的结构,每类样本为750个,分别来自具有各向同性、单位方差的两个高斯分布.在一个随机方向上,设置这两个高斯分布的中心点之间的距离为2.5,即$\|\mu_1-\mu_2\|=2.5$,在这个随机方向上的二维投影图如图2(a)所示.
- G241D:241维,具有易产生误导的聚类假设结构,从两个各向同性、单位方差的高斯分布中分别产生375个点(共750个点),并且在一个随机方向上,两个高斯分布的中心点的距离为6,这些点构成+1类;–1类的产生方式和+1类相同,两类在另一个随机方向上的高斯分布中心点之间的距离为2.5.在两个随机方向上的二维投影图如图2(b)所示.
- COIL100:来自COIL-100图形数据库,该数据库共有100个对象,每个对象从0~360°的水平方向旋转,每隔5度采集一副图片,则每个对象包含72副图像,图片大小为128×128像素.先从100个对象中随机选择24个对象(共24×72=1728副图片),划分为+1和–1两类,每类12个对象.然后将图片按比例压缩为16×16像素,并随机删除每类中的114副图片.最后,随机选择241列并做一定的图像模糊处理.
- USPST:来自于美国邮政的手写数字(0~9)数据集,随机选择每个数字的150副图片,“2”和“5”被分为+1类,其余的为–1类,与COIL00数据集类似,对该数据集进行了模糊处理.该数据集是非平衡的,同时具有聚类假设和流形假设的结构.
- TEXT数据集:用于文本分类的新闻组数据集,数据项是0或1,具有高维稀疏的特性.实验中随机选择了

其中的 1 000 个数据,每类样本各占 500.



(a) G241C (b) G241D

Fig.2 Two-Dimensional projections of G241C and G241D

图 2 G241C 和 G241D 数据集的二维投影图

5.1 参数设置

超参数λ=1,C=C^*,并在[0 10 100 500]上采用格点搜索选取最优参数,r 取值为表 1 中实际正负样本的比例,LapSVM 算法的超参数γ_A和γ_I采用了文献[1]中的设置.实验使用的是高斯核 $K(x_1,x_2)=\exp(-\|x_1-x_2\|^2/2\sigma^2)$,$\sigma$的取值见表 2.对于单核学习算法,采用 10 折交叉验证选择最优的σ作为参数.为了利用样本数据分布的几何特征,两个多核学习算法中同时加入了带高斯权重的流形核,采用如下公式生成:

$$\tilde{k}(x_i,x_j)=k(x_i,x_j)+K_{x_i}^T(I+MK)^{-1}MK_{x_j},$$

其中,$M=(\gamma_I/\gamma_A)L^p$,$L=D-W$ 为一个拉普拉斯矩阵.如果 x_i 和 x_j 为 K 近邻(TEXT 的 K 取值为 50,其余的为 5),则 $W_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)$;否则,$W_{ij}=0$.$D_{ii}=W_{ij}$ 为一个对角矩阵.核的总数为 6,基本核和流形核各 3 个.

Table 1 Characteristics of datasets

表 1 数据集的特征

数据集	维数	样本 n	r	数据集	维数	样本 n	r
2MOONS	2	200	0.5	COIL100	241	1 500	0.5
G241C	241	1 500	0.5	USPST	241	1 500	0.2
G241D	241	1 500	0.5	TEXT	11 960	1 000	0.5

Table 2 Settings of parameter σ

表 2 σ参数设置

数据集	σ取值	数据集	σ取值
2moons	(0.1,0.5,1)	COIL100	(0.5,2,4)
G241C	(0.5,1,2)	USPST	(0.1,0.5,2)
G241D	(0.5,1,2)	TEXT	(0.1,0.5,2)

当约束范数 p 的取值为 1 时,部分权系数趋近于 0,得到的是θ的最稀疏解;当 p 趋于∞时,目标函数是关于 $K=\sum_m K_m$ 的非加权和多核学习.文献[26,28]通过实验表明,最优的性能通常介于两者之间.本文的实验中 p 最终取值为 5,这样,θ既有一定的稀疏度,又可以使较好的核能够保留较高的权系数.

5.2 实验结果

(1) Transductive 实验

训练集包含 $l+u=n$ 个样本,通过预测未标记样本的标签来评估算法的性能.实验时,l 分别取值为训练集的 5%和 10%.将数据集分为 l/n 折,每次选取一折作为已标记样本,采用交叉验证统计各算法的平均误分类情况,最后一折已标记样本不足的,采用从训练集中随机选择补足.表 3 是平均误分类率及标准差.

Table 3 Misclassification rates for transductive experiment

表 3 Transductive 实验的误分类情况

数据集	2MOONS	G241C	G241D	COIL100	USPST	TEXT	l
S^3VM^{light}	5.27±1.05	18.94±3.01	25.36±2.25	24.27±3.13	13.42±1.02	27.72±1.34	5%
LapSVM	4.65±0.32	23.43±3.14	26.76±2.67	18.53±2.52	8.81±1.73	29.45±2.54	
$TSVM^{CCCP}$	4.72±0.67	17.65±1.21	22.54±2.14	21.54±2.33	11.23±1.37	26.53±3.52	
TSVM-MKL	4.56±0.53	17.14±1.34	22.25±1.67	21.21±2.06	10.57±0.56	25.47±2.31	
L_p-MKL-S^3VM	4.23±0.41	15.32±1.15	21.23±1.35	18.44±1.07	9.52±0.51	25.36±2.35	
S^3VM^{light}	3.52±0.81	17.03±2.17	18.19±2.43	21.73±1.54	8.05±1.24	24.67±3.21	10%
LapSVM	2.45±0.33	20.15±1.67	22.41±2.18	15.74±2.54	4.25±0.73	26.06±2.54	
$TSVM^{CCCP}$	2.57±0.56	14.65±1.28	15.84±1.28	17.26±1.08	5.41±0.52	24.43±2.92	
TSVM-MKL	2.63±0.25	14.09±1.42	15.76±1.36	16.45±1.35	5.14±0.46	23.78±2.05	
L_p-MKL-S^3VM	2.04±0.42	13.22±1.07	14.58±1.05	15.42±1.24	4.43±0.43	22.43±2.65	

(2) Inductive 实验

将数据集分成训练集和测试集两部分,利用训练集中已标记和未标记样本训练分类器,以预测测试集样本标签的精度来评估算法的性能.其中,训练集和测试集参数设定见表 4,训练集和测试集随机生成.

Table 4 Parameter settings for inductive experiment

表 4 Inductive 实验的参数设置

数据集	训练集 n	测试集	数据集	训练集 n	测试集
2MOONS	100	100	USPST	800	600
G241C	800	600	COIL100	800	600
G241D	800	600	TEXT	600	400

在这个实验中,我们选取了 3 种算法与本文的算法进行比较,已标记样本 l 分别取值为训练集的 5%,10%,15%及 20%,采用 l/n 折交叉验证训练分类器,最后一折已标记样本不足的,采用从训练集中随机选择补足.图 3(a)~图 3(f)是各算法在测试集上的分类情况.

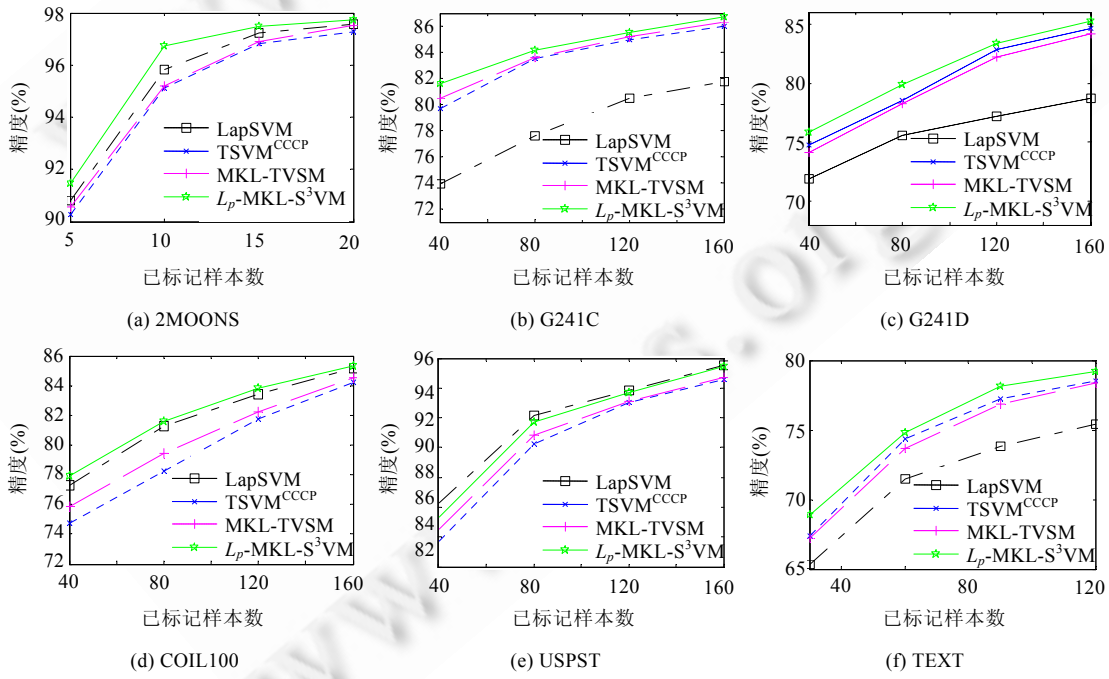


Fig.3 Classification accuracies on different datasets

图 3 不同数据集上的分类精度

从两种实验的结果可以看出:大多数情况下,本文提出的算法比其他算法具有有更好的分类精度.本算法通过在多核框架中加入流形核共同学习,避免了单一假设算法的局限,在一定程度上提高了算法在具有流形结构数据集上的学习性能.由于采用了相同的优化方法,S^3VM^{CCCP} 和 TSVM-MKL 的结果比较接近,TSVM-MKL 要略微优于 S^3VM^{CCCP}.LapSVM 算法在 G241C,G241D 以及 Text 上表现较差,但在两个图像数据集上表现却比较突出,这表明 LapSVM 适合处理具有流形结构特征的数据集.

在非平衡数据集 USPST 上,本文的算法要稍差于 LapSVM,特别是在 l 较小的情况下,这种差距比较明显.但是当 l 增加到 120 时,基本上达到了 LapSVM 的分类精度.这可能是流形核在 l 较小的情况下没有起到改善的作用.但本文的算法明显比 S^3VM^{CCCP} 和 TSVM-MKL 要好.

稀疏解会导致模型有用信息的丢失,这是因为 θ_m 多数为 0 或者系数很低,从而使得可能有用的核,如混合的流形核被从模型中去掉,失去了其应有的改善作用,从而导致泛化性能变差.因此,实验数据表明,L_1 范数约束的 TSVM-MKL 算法在个别情况下出现了不如单核学习算法 S^3VM^{CCCP} 和 LapSVM 的情况.

相反的是,非稀疏解能够尽可能地保留模型中的有用核,模型倾向于选择最好的核,但对模型有改善作用的核也能够以一定的系数得到保留,从而提高了算法的性能.因此,本文提出的算法能够表现出比其他算法更好的分类性能.

6 结论和展望

过去的相关研究主要集中在学习多核的凸组合及如何提高算法的执行效率,稀疏的核组合在实际应用中并不一定表现得比单核的要好.因此,本文结合传统的半监督支持向量机学习以及多核学习的理论和方法,提出了基于 L_p 范数约束的多核半监督支持向量机学习模型 L_p-MKL-S^3VM,并采用了一种改进的拟牛顿法和成对标签交换的局部搜索法来优化模型.在多核学习框架中,我们同时加入了基本核和流形核,以在学习过程中利用数据的几何属性,这样可以改善单一假设算法的局限,使算法在具有流行假设结构或聚类假设结构的数据集上都表现出较好的性能.人工数据集和真实数据集上的仿真实验结果验证了算法的有效性和较好的分类精度.将来可以把该算法扩展到多类问题上,研究流形核嵌入对算法性能的影响.另外,p 的取值与数据集大小、数据的稀疏性、数据的几何性质以及核个数之间的关系,也是值得研究的内容.

References:

[1] Chapelle O, Scholkopf B, Zein A. Semi-Supervised Learning. London: MIT Press, 2006.

[2] Joachims T. Transductive inference for text classification using support vector machines. In: Bratko I, Dzeroski S, eds. Proc. of the 16th Int'l Conf. on Machine Learning (ICML'99). Morgan Kaufmann Publishers, 1999. 200−209. http://www.informatik.uni-trier.de/~ley/db/conf/icml/icml1999.html

[3] Chapelle O, Sindhwani V, Keerthi SS. Optimization techniques for semi-supervised support vector machines. Journal of Machine Learning Research, 2008,9:203−233.

[4] Chapelle O, Zent A. Semi-Supervised classification by low density separation. In: Cowell R, Ghahramani Z, eds. Proc. of the 10th Int'l Workshop on Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics, 2005. 57−64. http://eprints.pascal-network.org/archive/00001144/01/aistats2005.pdf

[5] Collobert R, Sinz F, Weston J, Bottou L. Large scale transductive SVMs. Journal of Machine Learning Research, 2006,7(8): 1687−1712.

[6] V Sindhwani, S Keerthi, O Chapelle. Deterministic annealing for semi-supervised kernel machines. In: Cohen WW, Moore A, eds. Proc. of the 23rd Int'l Conf. on Machine Learning. ACM, 2006. 841−848. http://dblp.uni-trier.de/rec/bibtex/conf/icml/SindhwaniKC06 [doi: 10.1145/1143844.1143950]

[7] Chapelle O, Chi M, Zien A. A continuation method for semi-supervised SVMs. In: Cohen WW, Moore A, eds. Proc. of the 23rd Int'l Conf. on Machine Learning. ACM, 2006. 185−192. http://dblp.uni-trier.de/rec/bibtex/conf/icml/ChapelleCZ06 [doi: 10.1145/1143844.1143868]

[8] De Bie T, Cristianini N. Semi-Supervised learning using semi-definite programming. In: Chapelle O, Schöelkopf B, Zien A, eds. Semi-supervised Learning. MIT Press, 2006.

[9] Chapelle O, Sindhwani V, Keerthi S. Branch and bound for semi-supervised support vector machine. In: Schölkopf B, Platt J, Hoffman T, eds. Proc. of the 20th Annual Conf. on Neural Information Processing Systems. Cambridge: MIT Press, 2006. 217−224. http://nips.cc/Conferences/2006/Committees/

[10] Wang HQ, Sun FC, Cai YN, Chen N, Ding LG. On multiple kernel learning methods. Acta Automatica Sinica, 2010,36(8): 1037−1050 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2010.01037]

[11] Smola AJ, Scholkopf B. A tutorial on support vector regression. Statistics and Computing, 2004,14(3):199−222. [doi: 10.1023/B: STCO.0000035301.49549.88]

[12] Kerm PV. Adaptive kernel density estimation. Stata Journal, 2003,3(2):148−156.

[13] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998,2(2): 121−167. [doi: 10.1023/A:1009715923555]

[14] SchÄolkopf B, Mika S, Smola A, Ratsch G, Muller KR. Kernel PCA pattern reconstruction via approximation pre-images. In: Proc. of the Int'l Conf. on Artificial Neural Networks. IEEE, 1998. 147−152. http://openlibrary.org/books/OL367885M/ICANN_98

[15] Rakotomamonjy A, Bach F, Canu S, Grandvalet Y. SimpleMKL. Journal of Machine Learning Research, 2008,9(11):2491−2521.

[16] Lewis DP, Jebara T, Noble WS. Nonstationary kernel combination. In: Cohen WW, Moore A, eds. Proc. of the 23rd Int'l Conf. on Machine Learning. ACM, 2006. 553−560. http://dblp.uni-trier.de/rec/bibtex/conf/icml/LewisJN06 [doi: 10.1145/1143844.1143914]

[17] Ong CS, Smola AJ, Williamson RC. Learning the kernel with hyperkernels. Journal of Machine Learning Research, 2005,6(7): 1043−1071.

[18] Yuan Y, Shao J, Wu F, Zhuang YT. Image annotation by the multiple kernel learning with group sparsity effect. Ruan Jian Xue Bao/Journal of Software, 2012,23(9):2500−2509 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4154.htm [doi: 10.3724/SP.J.1001.2012.04154]

[19] Lee WJ, Verzakov S, Duin RP. Kernel combination versus classifier combination. In: Haindl M, Kittler J, Roli F, eds. Proc. of the 7th Int'l Workshop on Multiple Classifier Systems. LNCS 4472, Springer-Verlag, 2007. 22−31. http://www.informatik.uni-trier.de/~ley/db/conf/mcs/mcs2007.html [doi: 10.1007/978-3-540-72523-7_3]

[20] GÄonen M, Alpaydm E. Multiple kernel machines using localized kernels. In: Proc. of the 4th Int'l Conf. on Pattern Recognition in Bioinformatics. Sheffield: University of Sheffield, 2009. 1−10. http://www.springer.com/computer/bioinformatics/book/978-3-642-04030-6

[21] Lanckriet BG, Jordan M. Multiple kernel learning, conic duality, and the SMO algorithm. In: Brodley CE, ed. Proc. of the 21st Int'l Conf. on Machine Learning. ACM, 2004. 41−48. http://dblp.uni-trier.de/rec/bibtex/conf/icml/BachLJ04

[22] Bennett KP, Momma M, Embrechts MJ. MARK: A boosting algorithm for heterogeneous kernel models. In: Proc. of the 8th ACM-SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Edmonton: New York: ACM, 2002. 24−31. http://dl.acm.org/citation.cfm?id=775047&picked=prox&CFID=254452625&CFTOKEN=75751662 [doi: 10.1145/775047. 775051]

[23] Lanckriet GRG, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI. Learning the kernel matrix with semi-definite programming. The Journal of Machine Learning Research, 2004,5(1):27−72.

[24] Cortes C, Gretton A, Lanckriet G, Mohri M, Rostamizadeh A. Automatic selection of optimal kernels. In: Proc. of the NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels. 2008. http://www.cs.nyu.edu/learning_kernels

[25] Kloft M, Brefeld U, Laskov P, Sonnenburg S. Non-Sparse multiple kernel learning. In: Proc. of the NIPS Workshop on Kernel Learning: Automatic Selection of Kernels. MIT Press, 2008. 1−4. http://eprints.pascal-network.org/archive/00004977/

[26] Kloft M, Brefeld U, Sonnenburg S, Laskov P, Muller KR, Zien A. l_p-Norm multiple kernel learning. Journal of Machine Learning Research, 2011,12(5):953−997.

[27] Kloft M, Brefeld U, Sonnenburg S, Laskov P, Muller KR, Zien A. Efficient and accurate l_p-norm multiple kernel learning. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, eds. Proc. of the Conf. on Neural Information Processing Systems. MIT Press, 2009. 997−1005. http://books.nips.cc/nips22.html

[28] Kloft M, Blanchard G. On the convergence rate of l_p-norm multiple kernel learning. Journal of Machine Learning Research, 2012, 13(8):2465−2501.

[29] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006,6(11):2399−2434.

[30] Tian XL, Gasso G, Canu S. A multiple kernel framework for inductive semi-supervised SVM learning. Neurocomputing, 2012(90): 46−58. [doi: 10.1016/j.neucom.2011.12.036]

[31] Yu J, Vishwanathan SVN, Gunter S, Schraudolph NN. A quasi-Newton approach to nonsmooth convex optimization problems in machine learning. Journal of Machine Learning Research, 2010,11(5):1145−1200.

[32] Chapelle O, Rakotomamonjy A. Second order optimization of kernel parameters. In: Proc. of the NIPS Workshop on Kernel Learning: Automatic Selection of Kernels. 2008. http://www.citeulike.org/user/m8j/article/9501849

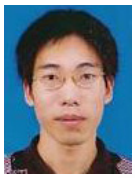
[33] Xu Z, Jin R, King I, Lyu M. An extended level method for efficient multiple kernel learning. Advances in Neural Information Processing Systems, 2009,21:1825−1832.

[34] Nath JS, Dinesh G, Ramanand S. On the algorithmics and applications of a mixed-norm based kernel learning formulation. Advances in Neural Information Processing Systems, 2009,22:844−852.

附中文参考文献:

[10] 汪洪桥,孙富春,等.多核学习方法.自动化学报,2010,36(8):1037−1050. [doi: 10.3724/SP.J.1004.2010.01037]

[18] 袁莹,邵健,吴飞,庄越挺.结合组稀疏效应和多核学习的图像标注.软件学报,2012,23(9):2500−2509. http://www.jos.org.cn/1000-9825/4154.htm [doi: 10.3724/SP.J.1001.2012.04154]



胡庆辉(1976－),男,重庆人,博士生,副教授,主要研究领域为半监督学习,智能信息处理.
E-mail: huqinghui2004@126.com

何进荣(1984－),男,博士生,主要研究领域为半监督学习与降维.
E-mail: 282857208@qq.com

丁立新(1967－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为智能计算,智能信息处理,统计学习,云计算.
E-mail: lxding@whu.edu.cn