软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(6):1458−1471 [doi: 10.3724/SP.J.1001.2012.04071]

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

大间隔最小压缩包含球学习机*

陶剑文[1,2], 王士同[1+]

[1](江南大学 信息工程学院,江苏 无锡 214122)

[2](浙江工商职业技术学院 信息工程学院,浙江 宁波 315012)

Large Margin and Minimal Reduced Enclosing Ball Learning Machine

TAO Jian-Wen[1,2], WANG Shi-Tong[1+]

[1](School of Information Engineering, Jiangnan University, Wuxi 214122, China)

[2](School of Information Engineering, Zhejiang Business Technology Institute, Ningbo 315012, China)

+ Corresponding author: E-mail: wxwangst@yahoo.com.cn

Tao JW, Wang ST. Large margin and minimal reduced enclosing ball learning machine. *Journal of Software*, 2012,23(6):1458−1471. http://www.jos.org.cn/1000-9825/4071.htm

Abstract: In this paper, inspired by the support vector machines for classification and the small sphere and large margin method, the study presents a novel large margin minimal reduced enclosing ball learning machine (LMMREB) for pattern classification to improve the classification performance of gap-tolerant classifiers by constructing a minimal enclosing hypersphere separating data with the maximum margin and minimum enclosing volume in the Mercer induced feature space. The basic idea is to find two optimal minimal reduced enclosing balls by adjusting a reduced factor parameter q such that each of binary classes is enclosed by them respectively and the margin between one class pattern and the reduced enclosing ball is maximized. Thus the idea implements implementing both maximum between-class margin and minimum within-class volume. Experimental results obtained with synthetic and real data show that the proposed algorithms are effective and competitive to other related diagrams.

Key words: generalization; support vector data description; support vector machine; minimum enclosing hypersphere

摘 要: 为了提高球形分类器的分类性能,受支持向量机和小球体大间隔等方法的启发,提出一种大间隔最小压缩包含球(large margin and minimal reduced enclosing ball,简称 LMMREB)学习机,其在 Mercer 核诱导的特征空间,通过优化一个最小包含球,以寻求两个同心的分别包含二类模式的压缩包含球,且使二类模式分别与压缩包含球间最小间隔最大化,从而可以同时实现类间间隔和类内内聚性的最大化.分别采用人工数据和实际数据进行实验,结果显示, LMMREB 的分类性能优于或等同于相关方法.

关键词: 泛化;支持向量数据描述;支持向量机;最小包含超球体

中图法分类号: TP181 **文献标识码**: A

* 基金项目: 国家自然科学基金(60975027, 60903100); 宁波市自然科学基金(2009A610080)
收稿时间: 2011-01-14; 修改时间: 2011-04-11; 定稿时间: 2011-06-20

模式分类旨在通过有限的训练样本学习一个分类器,且该分类器需对未来数据具有良好的泛化能力[1,2].类间间隔和类内内聚性是影响分类器分类性能的两种重要因素,增大类间间隔和提高类内聚类性有利于分类能力的提高[2,3].已有多种用于模式分类的方法提出,其中,建立在结构风险最小化(SRM)[4]理论基础上的支持向量机(support vector machine,简称 SVM)[5]及其相关变体是目前实现模式分类的主流方法之一[6].为了实现有效模式分类,SVM 构建一个线性或非线性的模式分割超平面,需满足如下 VC 维复杂度[7,1,3]:

$$VC_{SVM} \leqslant \min\left\{\left[\frac{D_{\max}^2}{\Delta_{\min}^2}\right], m\right\}+1,$$

其中,D 指覆盖所有样本的最小超球的直径,Δ指两类间的间隔,m 指样本维数.因此,为了使得 VC 维尽量最小化,需同时使得Δ最大化和 D 最小化.前者导出经典的 SVM 方法;而后者导出 SVM 的对偶版本,即超球 SVM,或称 gap-tolerant 分类机[8].一般来说,在类球形数据分布(或某个核函数映射后呈球形数据分布)情况下,gap-tolerant 分类机较无约束线性分类机具有更严格的界[8].SVM 具有完善的理论支撑和稳定优良的性能表现,其通过最大化类间间隔来达到强泛化能力,在机器学习和模式识别领域得到了广泛的应用[2].

为了解决一类分类(one class classification,简称 OCC)问题,Tax 等人[9]受到支持向量机的启发,提出一种支持向量数据描述(support vector data description,简称 SVDD)的球形学习机,其寻求一个包含所有目标样本的最小超球体,本质上等价于类内聚类性的最大化.文献[10]提出分隔超球模型(seperating hypersphere,简称 SH),其建立在 SVM 的类间间隔最大化和 SVDD 的类内聚类性最大化的思想基础上,采用类似支持向量机的建模结构,力图通过一个超球将正负两类样本分隔.文献[11]对 SH 进行改进,得到最大间隔球形支持向量机(maximal-margin support vector machine,简称 MSSVM),能够实现正负类类间间隔的增大和正类类内体积的减小,但没有考虑负类体积的减小,即不能提高负类的聚类性,使之在对测试样本分类时还存在较大的混淆可能性[2].对此,文献[2]运用 SVM 基本理论和 SVDD 超球结构,提出一种最大间隔最小体积球形支持向量机(maximal-margin minimal-volume hypersphere support vector machine,简称 MMHSVM),其构造两个大小不一的同心超球,小超球将正类样本包裹其中,大超球将负类样本排斥在外.MMHSVM 模型目标函数优化两个超球间隔,同时实现正负两类类内体积的缩小.文献[12]在 SVM 和 SVDD 的基础上提出一种用于离群检测的最小包含与最大排斥(minimum enclosing and maximum excluding machine,简称 MEMEM)的球形分类机,本质上来说,MEMEM 与 MMHSVM 方法相同,只是 MMHSVM 隐含了大间隔的思想.Ye 等人[13]综合 SVDD 中的小球体和 OCSVM(one class SVM)[14]中的大间隔思想,提出一种新颖的解决 OCC 问题的小球体和大间隔(samll sphere and large margin,简称 SSLM)方法,该方法在训练样本中利用少量异常样本信息.SSLM 在 SVDD 方法基础上引入参数 v 以细化分类边界,当 v=0 时,SSLM 变成 SVDD,故 SSLM 是 SVDD 方法的一个泛化版本.

上述 gap-tolerant 球形分类机实质上是一类基于 SVM 的大间隔小球体的方法,即采用一个(或两个)最小包含超球来最大可能地包含二类模式,而使二类模式尽可能地相互远离,从而实现二类模式的最大分割.同样基于此思想,本文通过引入一个包含球压缩因子,提出一种新颖的大间隔最小压缩包含球(large margin and minmal reduced enclosing ball,简称 LMMREB)学习机,其本质上是通过调节包含球压缩因子寻求二个同心的最小压缩包含球,以分别包含二类模式,同时使二类模式分别与所属压缩包含球间最小间隔最大化,使得类间间隔和类内内聚性同时最大化,从而增强 gap-tolerant 分类器泛化性能.与现有方法思想相比,所提方法的创新之处在于:

(1) 引入压缩包含球及其压缩因子等概念,LMMREB 将 MEMEM,MMHSVM 和 SSLM 等方法中二类模式间分割间隔变为一个与压缩因子相关的可适应调节的优化变量,增强了 LMMREB 对不同分类问题的自适应性;

(2) 引入一个可调参数 v 来细化分类边界,使二类模式分别与所属最小压缩包含球间间隔最大化,从而能够动态控制分类超平面的间隔误差,进一步提升了 LMMREB 方法的泛化能力;

(3) 在满足一定的参数变换条件(如压缩因子 q=1 或(和)v=0)下,LMMREB 方法可等价为 SVDD,MEMEM,MMHSVM 和 SSLM 等 gap-tolerant 方法.

1 LMMREB

1.1 相关概念

为了更好地描述 LMMREB,首先引入如下概念:

定义 1(二类模式分类问题). 设给定训练集 $T=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_N,\boldsymbol{y}_N)\}$,其中:$\boldsymbol{x}_i\in\boldsymbol{\chi}\subset R^d(1\leqslant i\leqslant m_1+m_2=N)$为输入数据;$\boldsymbol{y}_i\in\{+1,-1\}$为类标签,且当 $1\leqslant i\leqslant m_1$ 时,$\boldsymbol{y}_i=1$;当 $m_1+1\leqslant i\leqslant N$ 时,$y_i=-1$.又设其中一类(或正常类)含有 m_1 个模式,另一类(或异常类)含有 m_2 个模式,则对上述 N 个模式进行分类的问题称为二类分类问题 $Q(m_1,m_2,d)$.

定义 2(包含球). 对于一个二类模式分类问题 $Q(m_1,m_2,d)$,一个内包含球定义[8]为

$$B_{\leqslant}(a_{\leqslant},R_{\leqslant})=\{x\in R^d:\|x-a_{\leqslant}\|^2\leqslant R_{\leqslant}^2\},$$

其中,$a_{\leqslant}\in\chi$为内包含球球心,$R_{\leqslant}>0$ 为内包含球半径,$\|\cdot\|$指向量的欧式范数.同理,定义外包含球为

$$B_{\geqslant}(a_{\geqslant},R_{\geqslant})=\{x\in R^d:\|x-a_{\geqslant}\|^2\geqslant R_{\geqslant}^2\},$$

其中,$a_{\geqslant}\in\chi$为外包含球球心,$R_{\geqslant}>0$ 为外包含球半径.内包含球与外包含球统称为包含球.

定义 3(压缩包含球). 对于一个内包含球 $B_{\leqslant}(a_{\leqslant},R_{\leqslant})$,其同心压缩包含球(或称内压缩包含球) $B_{\leqslant}^r(a_{\leqslant},R_{\leqslant}^r,q)$ 定义为

$$B_{\leqslant}^r(a_{\leqslant},R_{\leqslant}^r,q)=\{x\in R^d:\|x-a_{\leqslant}\|^2\leqslant qR_{\leqslant}^2\},$$

其中,$q\in(0,1]$, $R_{\leqslant}^r$ 为内压缩包含球半径, $q=\dfrac{R_{\leqslant}^r}{R_{\leqslant}}$ 为内包含球压缩因子.当 $q=1$ 时, $B_{\leqslant}^r(a_{\leqslant},R_{\leqslant}^r,q)=B_{\leqslant}(a_{\leqslant},R_{\leqslant})$.同理,定义外包含球 $B_{\geqslant}(a_{\geqslant},R_{\geqslant})$的同心压缩包含球(或称外压缩包含球) $B_{\geqslant}^r(a_{\geqslant},R_{\geqslant}^r,p)$ 为

$$B_{\geqslant}^r(a_{\geqslant},R_{\geqslant}^r,p)=\{x\in R^d:\|x-a_{\geqslant}\|^2\geqslant pR_{\geqslant}^2\},$$

其中,$p\geqslant 1$, $R_{\geqslant}^r$ 为外压缩包含球半径,压缩因子 $p=\dfrac{R_{\geqslant}^r}{R_{\geqslant}}$.$p=1$ 时, $B_{\geqslant}^r(a_{\geqslant},R_{\geqslant}^r,p)=B_{\geqslant}(a_{\geqslant},R_{\geqslant})$.内压缩包含球与外压缩包含球统称为压缩包含球.

根据以上定义,可引出如下定理:

定理 1. 对于一个外包含球 $B_{\geqslant}(a_{\geqslant},R_{\geqslant})$及其压缩包含球 $B_{\geqslant}^r(a_{\geqslant},R_{\geqslant}^r,p)$ $(p\geqslant 1)$,$B_{\geqslant}(a_{\geqslant},R_{\geqslant})$可以同构为一个同心的内包含球,且 $B_{\geqslant}^r(a_{\geqslant},R_{\geqslant}^r,p)$ 为该内包含球的同心压缩包含球.

证明:设 $x\in R^d$ 为 $B_{\geqslant}(a_{\geqslant},R_{\geqslant})$所包含区域内任意一点,即 x 满足 $\|x-a_{\geqslant}\|^2\geqslant R_{\geqslant}^2$,连接球心 $a_{\geqslant}$与点 x,并向外延伸为射线 l,在 l 上选取一点 x',使得 $\|x'-a_{\geqslant}\|=\dfrac{1}{\|x-a_{\geqslant}\|}\leqslant\dfrac{1}{R_{\geqslant}}$,则点 x 与 x'是一一对应的.由于 x 为 $B_{\geqslant}(a_{\geqslant},R_{\geqslant})$所表示的区域内任意点,故对于 $B_{\geqslant}(a_{\geqslant},R_{\geqslant})$所包含区域内所有点均存在对应的像 x',使得其满足公式:

$$\left\{x'\in R^d:\|x'-a_{\geqslant}\|^2\leqslant\frac{1}{R_{\geqslant}^2}\right\}=B_{\leqslant}\left(a_{\geqslant},\frac{1}{R_{\geqslant}}\right),$$

式中等号由定义 2 可得,从而可知 $B_{\leqslant}\left(a_{\geqslant},\dfrac{1}{R_{\geqslant}}\right)$为一个与 $B_{\geqslant}(a_{\geqslant},R_{\geqslant})$同心的内包含球.

故存在双射 $f:x\to x'$,使得 $B_{\geqslant}(a_{\geqslant},R_{\geqslant})$与同心内包含球 $B_{\leqslant}\left(a_{\geqslant},\dfrac{1}{R_{\geqslant}}\right)$同构.设 $B_{\leqslant}\left(a_{\geqslant},\dfrac{1}{R_{\geqslant}}\right)$的同心压缩包含球为 $B_{\leqslant}^r\left(a_{\geqslant},\dfrac{1}{R_{\geqslant}},q'\right)$,令 $q'=\dfrac{1}{p}(p\geqslant 1)$,另根据定义 3 可得:

$$B_{\leqslant}^r\left(a_{\geqslant},\frac{1}{R_{\geqslant}},q'\right)=\left\{x'\in R^d:\|x'-a_{\geqslant}\|^2\leqslant\frac{1}{pR_{\geqslant}^2}\right\}=\{x\in R^d:\|x-a_{\geqslant}\|^2\geqslant pR_{\geqslant}^2\}=B_{\geqslant}^r(a_{\geqslant},R_{\geqslant}^r,p).$$

即 $B_{\leqslant}\left(a_{\geqslant},\dfrac{1}{R_{\geqslant}}\right)$的同心压缩包含球为 $B_{\geqslant}^r(a_{\geqslant},R_{\geqslant}^r,p)$,从而定理得证. □

近来,核学习方法已成功应用于机器学习的许多不同方面[14].对于一个给定的输入空间$\chi\subset\Re^d$和一个正定核函数$K:\chi\times\chi\to\Re$,输入空间中的数据被隐式地映射到一个高维特征空间F.设$\phi(\cdot)$为从χ到F的映射,则对任意的$u,v\in\chi$,$K(u,v)=\langle\phi(u),\phi(v)\rangle$.本文关注核学习方法,在高维特征空间$F$,上述包含球(或压缩包含球)变成包含超球(或压缩包含超球).

1.2 问题描述

根据定义2,考虑以下3种情况:

(1) 当$a_\leqslant=a_\geqslant$且$R_\leqslant\neq R_\geqslant$时,内外包含球同心且半径不同.MMHSVM即为该种情况;

(2) 当$a_\leqslant=a_\geqslant$且$R_\leqslant=R_\geqslant$时,内外包含球同心且球面重合.SVDD,SSLM,MEMEM等方法属于该种情况,这也是目前大多球性学习机采纳的情形;

(3) 当$a_\leqslant\neq a_\geqslant$时,内外包含球不同心,从而形成两个完全相异的包含超球.双元超球学习机(twin support vector hypersphere,简称TSVH)即属于这种情况[15].

本文方法基于上述第(2)种情况,即$a_\leqslant=a_\geqslant=a$且$R_\leqslant=R_\geqslant=R$.为了简单起见,下文将内包含(超)球和外包含(超)球统称为包含(超)球.

对于一个二类模式分类问题,LMMREB旨在高维特征空间F中寻求一个最小包含超球$B_\leqslant(a,R)(a_\leqslant\in F)$,使得二类模式分别被包含于两个同心压缩包含超球$B_\leqslant^r(a,R_\leqslant^r,q)$ $(q\in(0,1))$和$B_\geqslant^r(a,R_\geqslant^r,p)$ $(p\geqslant1)$内,且使两类模式分别与所在压缩包含超球间最小间隔最大化.

图1所示为LMMREB方法的二维几何解释,其中,实线表示$B_\leqslant(a,R)$(或$B_\geqslant(a,R)$),虚线分别为两个同心压缩包含超球$B_\leqslant^r(a,R_\leqslant^r,q)$ $(q\in(0,1))$和$B_\geqslant^r(a,R_\geqslant^r,p)$ $(p\geqslant1)$,ρ^2为二类模式分别与压缩包含超球间最小间隔,包含超球$B_\leqslant(a,R)$(或$B_\geqslant(a,R)$)与其压缩包含超球间间隔分别为$(1-q)R$.由$q=\frac{R_\leqslant^r}{R}$得:$p=\frac{R_\geqslant^r}{R}\propto\frac{R_\leqslant^r+2(1-q)R}{R}=2-q$.

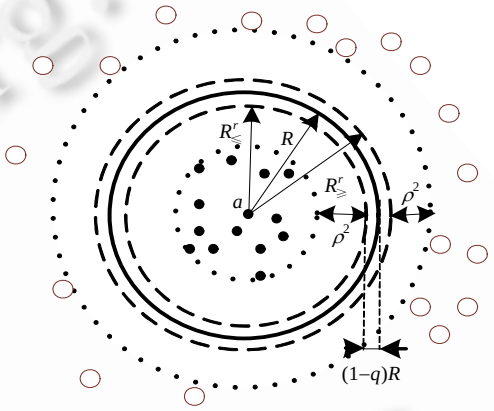


Fig.1 Illustration of LMMREB in 2D

图1 LMMREB的二维几何示意

基于以上思想,在某个Mercer核映射的高维特征空间F,LMMREB原始问题可描述为

$$\min f(R,\rho,\xi,a)=R^2-\nu\rho^2+C_1\sum_{i=1}^{m_1}\xi_i+C_2\sum_{j=m_1+1}^{N}\xi_j \tag{1}$$

s.t.

$$\|\phi(x_i)-a\|^2\leqslant qR^2-\rho^2+\xi_i,1\leqslant i\leqslant m_1 \tag{2}$$

$$\|\phi(x_j)-a\|^2\geqslant(2-q)R^2+\rho^2-\xi_j,m_1+1\leqslant j\leqslant N \tag{3}$$

$$\xi_k\geqslant0,1\leqslant k\leqslant N \tag{4}$$

其中,$\phi(\cdot)$为从χ到F的核映射;$a\in F$,$R>0$分别为超球体的球心和半径;$\xi=[\xi_1,\xi_2,\ldots,\xi_N]^T$为松弛向量;$C_1,C_2$为两个正常数;$q\in(0,1)$为包含超球压缩因子,是可调参数.公式(1)中成本函数最小化使得包含超球的体积尽可能地小,同时二类模式间最小间隔尽可能地大.同时我们注意到,当利用正定核函数时,上面的原始问题(1)~问题(4)与一类和二类SVM优化形式以及文献[13]中的优化形式在数学表达上非常相似,故我们可以借鉴,即用求其对偶目标解来求解上述问题.

定理2. LMMREB的压缩超球存在且不唯一.

证明:由于LMMREB训练集中一类(或正常类)只有有限(m_1)个训练点,总能寻求到一个包含球$B_\leqslant(a,R)$以包含所有这一类(或正常类)样本.根据定义3,对于某个包含球压缩因子$q\in(0,1)$,存在该包含球的压缩包含球$B_\leqslant^r(a,R_\leqslant^r,q)$,且随着$q$值调节变化,该压缩包含球自适应变化.由此可知,LMMREB的压缩超球存在且不唯一.□

1.3 LMMREB算法

为导出原始问题(1)~问题(4)的对偶问题,引入 Lagrange 函数:

$$L(R,\rho,a,\xi,\alpha,\beta)=f+\sum_{i=1}^{m_1}\alpha_i(\|\phi(x_i)-a\|^2-qR^2+\rho^2-\xi_i)-\sum_{j=m_1+1}^{N}\alpha_j(\|\phi(x_j)-a\|^2-(2-q)R^2-\rho^2+\xi_j)-\sum_{k=1}^{N}\beta_k\xi_k \quad (5)$$

其中,$\alpha_i\geqslant 0,\beta_i\geqslant 0$ 分别为 Lagrange 乘子向量.从而可得如下定理:

定理 3. 最优化问题:

$$\min_{\alpha\in R^n}\frac{1}{\Delta}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i y_i K(x_i,x_i) \quad (6)$$

s.t.

$$0\leqslant\alpha_i\leqslant C_1,1\leqslant i\leqslant m_1 \quad (7)$$

$$0\leqslant\alpha_j\leqslant C_2,m_1+1\leqslant j\leqslant N \quad (8)$$

$$\sum_{i=1}^{N}\alpha_i=v \quad (9)$$

是原始问题(1)~问题(4)的对偶问题,其中,$\Delta=\sum_{i=1}^{N}\alpha_i y_i=1+(1-q)v$,$\alpha_i\geqslant 0(i=1,\ldots,m_1),\alpha_j\geqslant 0(j=m_1+1,\ldots,N)$,$K(u,v)$为某个符合 Mercer 条件的核函数.

证明:在 $L(R,\rho,a,\xi,\alpha,\beta,\lambda)$方程中分别对 R,ρ,a,ξ等原始变量求偏导数,可得:

$$\frac{\partial L}{\partial R}=2R\left(1-q\sum_{i=1}^{m_1}\alpha_i+(2-q)\sum_{j=m_1+1}^{N}\alpha_j\right)=0 \quad (10)$$

$$\frac{\partial L}{\partial\rho}=2\rho\left(-v+\sum_{i=1}^{m_1}\alpha_i+\sum_{j=m_1+1}^{N}\alpha_j\right)=0\rightarrow\sum_{i=1}^{N}\alpha_i=v \quad (11)$$

$$\frac{\partial L}{\partial\xi_i}=C_1-\alpha_i-\beta_i=0 \quad (12)$$

$$\frac{\partial L}{\partial\xi_j}=C_2-\alpha_j-\beta_j=0 \quad (13)$$

$$\frac{\partial L}{\partial a}=2a\sum_{i=1}^{N}\alpha_i y_i-2\sum_{i=1}^{N}\alpha_i y_i\phi(x_i)=0 \quad (14)$$

将公式(10)~公式(14)代入方程(5)中,得到原始优化问题(1)~问题(4)的对偶形式:

$$\min_{\alpha\in R^n}\frac{1}{\Delta}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i y_i K(x_i,x_i) \quad (15)$$

s.t.

$$0\leqslant\alpha_i\leqslant C_1,1\leqslant i\leqslant m_1 \quad (16)$$

$$0\leqslant\alpha_j\leqslant C_2,m_1+1\leqslant j\leqslant N \quad (17)$$

$$\Delta=\sum_{i=1}^{N}\alpha_i y_i=1+(1-q)v \quad (18)$$

$$\sum_{i=1}^{N}\alpha_i=v \quad (19)$$

其中,公式(18)由公式(10)、公式(11)综合推导可得.综上,定理得证. □

对偶式(6)~对偶式(9)为一个二次规划问题,其与标准的 v-SVM[6]具有相同的优化形式,因此,LMMREB 能够直接利用解 *v-SVM* 的软件包来实现,使得 LMMREB 算法实现简单方便.从对偶形式(6)~对偶式(9)可以看出,在一定的参数变换条件下,LMMREB 可分别转化为对应的 SVDD,MEMEM,SSLM 和 MMHSVM 等 gap-tolerant 方法.

由公式(18)与公式(19)可得 $\sum_{i=1}^{m_1}\alpha_i=\frac{1}{2}(v(2-q)+1)$,考虑两个支持向量集合:

$$SV_1=\left\{x_i\left|\sum_{i=1}^{m_1}\alpha_i=\frac{1}{2}((2-q)v+1),0<\alpha_i<C_1,1\leqslant i\leqslant m_1\right.\right\},$$

$$SV_2=\left\{x_j\left|\sum_{j=m_1+1}^{N}\alpha_j=\frac{1}{2}(vq-1),0<\alpha_j<C_2,1+m_1\leqslant j\leqslant N\right.\right\}.$$

根据 K.K.T.条件[1],对于输入 $x_i\in SV_1(1\leqslant i\leqslant m_1)$,公式(2)变成一个所有松弛变量等于 0 的等式:

$$\|\phi(x_i)-a\|^2=qR^2-\rho^2,1\leqslant i\leqslant m_1.$$

同理,对于 $x_j\in SV_2(m_1+1\leqslant j\leqslant N)$,公式(3)也变成一个等式:

$$\|\phi(x_j)-a\|^2=(2-q)R^2+\rho^2,m_1+1\leqslant j\leqslant N.$$

设 $n_1=|SV_1|,n_2=|SV_2|$,$|\cdot|$表示集合基数,则可得优化包含球半径为

$$R^{*2}=\frac{1}{2}\left[\min_{x_i\in SV_1}\left(\frac{1}{q}\|\phi(x_i)-a\|^2\right)+\max_{x_j\in SV_2}\left(\frac{1}{2-q}\|\phi(x_j)-a\|^2\right)\right].$$

根据公式(14)、公式(18)可得优化球心 a^*为

$$a^*=\frac{1}{1+(1-q)v}\sum_{i=1}^{N}\alpha_i y_i\phi(x_i).$$

可得优化的类间间隔为

$$\rho^{*2}=\frac{1}{2}[\min_{x_i\in SV_1}(qR^{*2}-\|\phi(x_i)-a^*\|^2)+\min_{x_j\in SV_2}(\|\phi(x_j)-a\|^2-(2-q)R^{*2})].$$

在某个核函数诱导的特征空间,为了测试某个未知数据 x 所属类别,LMMREB 只需判别其是否被最小内包含球所包含,即 LMMREB 的决策函数定义为

$$f(x)=\mathrm{sgn}(R^{*2}-\|\phi(x)-a^*\|^2)=\mathrm{sgn}\left(R^{*2}-\left(K(x,x)-\frac{2}{\Delta}\sum_{i=1}^{N}\alpha_i y_i K(x,x_i)+\frac{1}{\Delta^2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\right)\right).$$

2 算法分析

2.1 泛化能力分析

所提方法 LMMREB 通过引入一个可调的包含球压缩因子 q,使传统 gap-tolerant 分类机中二类模式间间隔变为一个关于包含球压缩因子 q 的可调变量 $2(\rho^2+(1-q)R)(0<q\leqslant 1)$.其中,$R$ 指 LMMREB 中分割超球半径,ρ^2 为二类模式与所属压缩包含球间间隔.根据 SVM 的间隔误差界理论[7],有如下泛化误差界定理:

定理 4. LMMREB 的 VC 维满足 $VC_{\mathrm{LMMREB}}\leqslant\min\left\{\left\lceil\frac{D^2_{\max}}{(2\rho^2+2(1-q)R^2)_{\min}}\right\rceil,d\right\}+1$,其中,$D$ 为覆盖所有样本的最小超球的直径,R 指 LMMREB 中分割超球半径,$0<q\leqslant 1$,d 指样本维数.

推论 1. 在给定的训练样本集下,LMMREB 的 VC 维上界由参数 q 和 v 决定.

从原始问题(1)~问题(4)可知,LMMREB 通过寻求一个最小包含球 $B_{\leqslant}(a,R)$,使得二类模式分别被包含于两个同心压缩包含球 $B^r_{\leqslant}(a,R^r_{\leqslant},q)$ $(q\in(0,1))$和 $B^r_{\geqslant}(a,R^r_{\geqslant},p)$ $(p\geqslant 1)$内,且使两类模式分别与压缩包含球间最小间隔最大化;另根据定理 1 可知,LMMREB 在一定程度上充分考虑了二类模式类内分布结构的最小化和类间间隔的最大化.由此可引出定理 5.

定理 5. LMMREB 算法能同时实现类间间隔最大化和二类模式类内分布结构最小化.

定理 5 充分说明 LMMREB 方法具有增强的模式分类泛化能力,这可从第 3 节的实验结论进一步得以验证.

2.2 复杂度分析

SVM,SVDD,MMHSVM,MEMEM,LMMREB 的求解都归结于具有线性约束条件的二次凸规划问题,线性约束二次规划问题的算法复杂度主要取决于规划中变量的个数和约束方程的个数[2].下面引出一个关于LMMREB 算法复杂度的定理:

定理 6. LMMREB 与 SVM,SVDD,MMHSVM 及 MEMEM 算法复杂度同级.

证明:采取文献[2]中符号标记,$O(d,s)$表示一个线性约束二次凸优化问题,$C_Q^T(d,s),C_Q^S(d,s)$ 分别表示对应的时间复杂度和空间复杂度,其中,d 为变量个数,s 为线性约束方程个数.设训练样本数为 n,则 SVM,SVDD,MEMEM,MMHSVM,LMMREB 算法的复杂度分别表示为

$$C_Q^{(*)}(n,2n+1),C_Q^{(*)}(n,2n+1),C_Q^{(*)}(n,2n+2),C_Q^{(*)}(n,2n+2),C_Q^{(*)}(n,2n+2),$$

其中(*)代表 T 或 S.文献[16]分析指出,SVM 的求解在时间和空间上的复杂度分别为 $O(n^3)$和 $O(n^2)$,即

$$C_Q^T(n,2n+1)=O(n^3),C_Q^S(n,2n+1)=O(n^2).$$

令 $n=m+1$,则有 $C_Q^T(m+1,2m+3)=O((m+1)^3)=O(m^3)$.显然有 $C_Q^T(n,2n+1)\leqslant C_Q^T(n,2n+2)\leqslant C_Q^T(n+1,2n+3)$,从而 $C_Q^T(n,2n+2)=O(n^3)$.同样道理可推得 $C_Q^S(n,2n+2)=O(n^2)$.综上,定理得证. □

根据定理 6,LMMREB 的空间复杂度和时间复杂度与 SVM 相同,分别为 $O(n^2)$和 $O(n^3)$,随着样本数的增加,所提方法的二次规划问题求解的计算复杂度明显增加.对此,可通过最小包含球(MEB)算法[16,17]来近似求解,具体细节这里不再赘述.

2.3 参数q,v属性分析

按照文献[13]中的术语,本文称对应于 Lagrange 乘子$\alpha_i>0$ 的训练样本 $x_i(1\leqslant i\leqslant N)$为支持向量(SV),对应于松弛变量$\xi_j>0$ 的训练样本 $x_j(1\leqslant j\leqslant N)$称为间隔误差(ME).

定理 7. 设 m^+,m^-分别指正类和负类的间隔误差数,s^+,s^-分别指正类和负类的支持向量数,则参数 v,q,C_1,C_2 间存在关系:

$$m^+\leqslant\frac{1}{2C_1}[(2-q)v+1]\leqslant s^+ \tag{20}$$

$$m^-\leqslant\frac{1}{2C_2}(qv-1)\leqslant s^- \tag{21}$$

证明:根据 K.K.T.条件,当$\xi_i>0$ 时,$\beta_i=0$.由公式(8)知$\alpha_i=C_1$ 对所有正间隔误差成立,则下式成立:

$$\sum_{i=1}^{m_1}\alpha_i=\frac{1}{2}((2-q)v+1)\geqslant m^+C_1 \tag{22}$$

另外,由公式(14)知,每个正支持向量对 $\sum_{i=1}^{m_1}\alpha_i$ 至多贡献 C_1,故

$$\sum_{i=1}^{m_1}\alpha_i\leqslant C_1s^+ \tag{23}$$

综合公式(22)、公式(23)可得证公式(20),同理可证公式(21). □

定理 7 说明,LMMREB 中参数 v 和 q 具有和 v-SVM 中参数 v 同样的性质,即能同时控制支持向量的下界和间隔误差的上界.定理 7 对本文方法实验参数的优化选取具有指导意义.

3 实验结果及分析

本文重点关注模式识别中二类和一类分类问题,对于多类问题,可通过“一对多(OAA)”和“一对一(OAO)”等方法化为多个二类分类问题来解决.为了验证 LMMREB 的有效性,通过与其他几个核学习方法进行比较来评价所提方法的优化性能.分别在人造数据集和 UCI 机器学习数据库以及人脸数据库上进行了一系列实验,分别将 LMMREB 与 v-SVM,SVDD,MEMEM,MMHSVM 及 SSLM 等方法在二类分类和一类分类上进行测试比较,

实验结果显示,LMMREB 具有优化的模式分类性能.

3.1 实验参数设置

所有实验样本首先归一化为[−1,1],实验中所有参数的协调通过最优搜索策略来选取[18].实验算法均采用高斯核函数(RBF): $K(x,y)=\exp\left(-\frac{1}{g}\|x-y\|^2\right)$,其中,$g$ 值在网格{$\sigma^2/8,\sigma^2/4,\sigma^2/2,\sigma^2,\sigma^2*2,\sigma^2*4$}中搜索选取.其中,$\sigma$为训练样本平均范数的平方根.

对于 v-SVM,参数 v 在{0.01k,0.1k}中搜索选取,其中,k 为 1~9 之间的整数.对于 SVDD,参数 C_1 在{0.01,0.05,0.1,0.5,1,5,10,20,30,40,50,100}中搜索选取;参数 C_2 从 $\left\{\frac{1}{4}\times\frac{m_1}{m_2},\frac{1}{2}\times\frac{m_1}{m_2},2\times\frac{m_1}{m_2},4\times\frac{m_1}{m_2}\right\}$ 中搜索.对于 MEMEM,参数 C 在{0.1,0.2,...,1}中搜索选取,γ在{0.05,0.1,0.2,...,1}中搜索选取.对于 MMHSVM,其涉及 3 个可调参数:(M,C_1,C_2)[2],按照文献[2]中做法,取折中参数 C=C_1=C_2,分别在{2,2^2,…,2^{10}}中搜索选取,m_1 在{1.1,1.5,2,2.5,3}中搜索选取.对于 SSLM,参数 v 在网格{10,30,50,70,90}中搜索选取,v_1,v_2 在集合{0.001,0.01}中选取[13].对于 LMMREB,根据定理 7,参数 v 在网格{1,5,10,15,20,25,30,35,40,45,50,60,80,90}中搜索选取; $C_1=\frac{1}{v_1m_1}$,$C_2=\frac{1}{v_2m_2}$,其中,v_1,v 从{0.001,0.01}中选取[13];根据参数 q 的定义域(0,1),确定 q 在[ε,1]区间搜索选取,其中,ε>0 为一极小正数.关于参数 q 值的调节策略将在第 3.2 节详述.所有方法均采取交叉验证法选择参数值,实验结果均在 Matlab2009B 运行环境下取得.

3.2 参数q,v对LMMREB分类精度影响

为了研究参数 q,v 对 LMMREB 分类精度影响,随机生成一个包含二类的二维人造数据集作为实验数据,测试在不同 q,v 参数值下 LMMREB 的分类精度变化.数据集样本数 N=200,二类样本数分别为 m_1=98,m_2=102,实验中,70%样本作为训练集,剩余样本作为测试集,q 在{0.05,0.1,0.2,0.4,0.6,0.8,.09,1}中搜索选取.每对 q,v 参数值实验 5 次,实验结果取其平均值,表 1 记录了根据 q 取值的实验结果.从表 1 可以看出,当 q=0.8,v=2 时,LMMREB 取得最好分类性能.由此可知,参数 q,v 不同值的优化选取对 LMMREB 的分类性能起到关键作用,由此进一步验证了推论 1 结论.根据文献[13]和定理 7,确定参数 v 的搜索范围如第 3.1 节所述,下面将重点探讨参数 q 的选择策略.

Table 1 Performance comparisons of LMMREB on different q and v

表 1 参数 q,v 对 LMMREB 分类精度影响

参数值(g=16.7)	q=0.05 v=12	q=0.1 v=25	q=0.2 v=11	q=0.4 v=9	q=0.6 v=38	q=**0.8** $\boldsymbol{v}$=**2**	q=0.9 v=3	q=1 v=2
分类精度(%)	73.5	73.5	75	75	75.5	**79.4**	75	76.5

为了进一步研究参数 q 的优化选取对 LMMREB 分类精度的影响,通过数据生成器[9]随机生成 20 个包含二类的二维人造数据集作为实验数据,测试在不同 q 参数选取策略下 LMMREB 的分类精度变化.所有数据集样本数 N=200,其中,正类样本数 m_1 分别为{90,95,100,…,185}.实验中,70%样本作为训练集,剩余样本作为测试集.q 值在定义区间[ε,1]内进行搜索选取,分别划分 4 种由小到大的搜索集:G_0=[ε_0:0.0001:0.001],其中 ε_0=0.0001;G_1=[ε_1:0.01:0.1 0.1:0.1:1],其中 ε_1=0.01;G_2=[ε_2:0.01:1],其中 ε_2=0.01;G_3=[ε_3:0.001:1],其中 ε_3=0.001.很显然,$G_0\subset G_1\subset G_2\subset G_3$.对于每种搜索集中的每一个 q 值,每个数据集实验一次,然后计算相同 q 值所对应的所有数据集的分类精度的平均值,得到 4 个分别对应于 4 种搜索集的所有数据集的平均分类精度集,再对每个集合按照值大小从大到小排序,分别取前 5 个(top-n,n=1,2,…,5)最大分类精度值,并记录于表 2,其中,参数 v 值根据搜索集中每个 q 值采取 5 重交叉验证法选择.

Table 2 Performance comparisons of LMMREB on different q searching grid

表 2 在不同搜索集下参数 q 对 LMMREB 分类精度影响

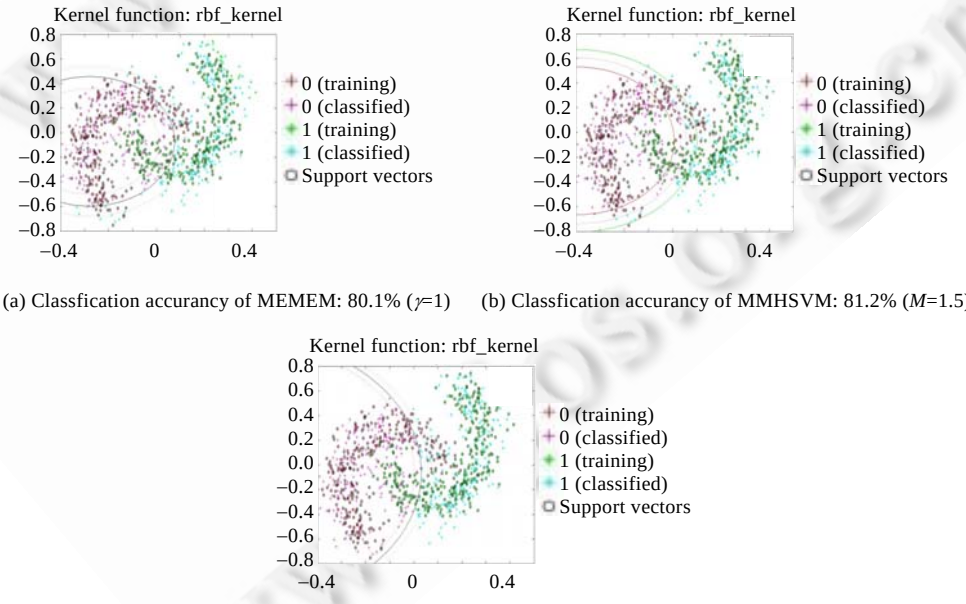
搜索集	Top-n 精度				
	Top-1 精度(q)	Top-2 精度(q)	Top-3 精度(q)	Top-4 精度(q)	Top-5 精度(q)
G_0 (m=10^{-4})	3.11% (1m)	3.11% (2m)	4.41% (3m)	4.1% (4m)	3.6% (5m)
G_1 (n=0.1)	87.1% (8n)	82.36% (7n)	82.31% (6n)	78.3% (9n)	75.5% (10n)
G_2	88.54% (0.85)	88.42% (0.86)	88.19% (0.84)	87.73% (0.83)	87.1% (0.82)
G_3	88.57% (0.858)	88.57% (0.859)	88.57% (0.857)	88.56% (0.856)	88.56% (0.855)

从表 2 可看出,在搜索集 G_0 策略下,LMMREB 对于所有数据集的最高平均分类精度不超过 10%.由此可见,当 $q\in(0,0.001)$时,所提方法 LMMREB 的分类结果失去意义.原因在于,当压缩因子 q 值趋于极小时,压缩包含球的半径也趋于极小,其所能包含的一类(或正类)样本数也将趋于极小,根据定理 7,学习机的分类间隔误差数趋于增大,导致分类精度下降.另外,从表 2 还看出,G_1 策略的分类结果均相应低于 G_2 和 G_3 策略的分类结果.由此可见,q 值的搜索集越细粒化,q 值的搜索取值越接近优化 q 值,导致分类结果也愈趋优化.G_2 策略的分类结果与 G_3 策略的相应分类结果基本相当,但是 G_3 策略的训练时间是 G_2 策略训练时间的 10 倍,G_3 策略明显不利于大样本分类情况,故本文以下所有实验均采用 G_2 策略的 q 值搜索方式.

3.3 人造数据集

首先通过一个二维香蕉型人造数据集来比较 LMMREB 方法与 MEMEM 和 MMHSVM 方法的二类模式分类性能.人造数据集样本数 N=1000,正类数#pos=477,负类数#neg=523,核函数参数 g=28.实验中,70%样本为训练集,余下样本作为测试集.两种分类方法的性能分别如图 2(a)~图 2(c)所示,其中,实曲线为各方法寻求的优化包含球.

从图 2 中显示的情况可知:(1) 对于非球形状分布的数据集,3 种球性学习机均能取得较好的分类性能,其中,本文所提方法由于采用了最小压缩包含的思想,使得泛化性能明显优于其他两种方法;(2) 参数 q 和 v 的优化取值能增强所提方法的分类性能,验证了推论 1 所提结论.



(a) Classfication accurancy of MEMEM: 80.1% (γ=1) (b) Classfication accurancy of MMHSVM: 81.2% (M=1.5)

(c) Classfication accurancy of LMMREB: 83.2% (q=0.9, v=1.5)

Fig.2 Classification results on the 2D artificial dataset

图 2 人造香蕉型数据集分类结果比较

3.4 实际数据集

本节通过 UCI 数据集和人脸数据集来评价所提方法分别在低维和高维数据集情况下,与 *v*-SVM,MEMEM 及 MMHSVM 等方法的模式分类性能比较.对于多类数据集采取一对一(OAO)方式将其化为多个二类分类问题实现.

3.4.1 UCI 数据实验

A. 二类模式分类

为了评价所提方法的二类分类性能,选取 14 个 UCI 实际数据集作为实验数据[13],实验中所采用数据集见表 3.实验中各类数据集随机分割成训练集和测试集,每个数据集分别实验 10 次,取其平均值作为各自实验结论.表 4 记录了 3 种方法各自实验参数值和相应的分类精度.

Table 3 Binary classification datasets

表 3 二元模式分类数据集

Datasets	#pos	#neg	m_1	m_2	d
Iris-Setosa	50	100	50	100	4
Sonar-Mines	111	97	111	97	60
Imports	71	88	71	88	25
Ecoli-Periplasm	284	52	284	52	7
Cancer-Wpbc	151	47	151	47	33
Breast	458	241	458	241	9
Heart-Healthy	164	139	164	139	13
Abalone class	1 407	2 770	1 407	2 000	10
Glass building	76	138	76	138	9
Glass vehicle	17	197	17	197	9
Arrhythmia	237	183	237	183	278
Ionosphere good	225	126	225	126	34
Waveform 0	300	600	300	600	21
Vowel 0	48	480	48	480	10

Table 4 Average classification accuracy rates for binary pattern classification (%)

表 4 二元模式分类平均分类精度 (%)

Dataset	*v*-SVM	MEMEM	MMHSVM	LMMREB
	Acc. rate (v,g)	Acc. rate (g,γ)	Acc. rate (M,C)	Acc. rate (v,q)
Iris-Setosa	98.3% (0.4,0.001)	98.5% (1.7,0.6)	98.6% (1.5,2^9)	**100%** (2,0.4)*
Sonar-Mines	52.57% (0.3,0.002)	52.5% (8.8,0.8)	54.2% (2.5,2^9)	**57.6%** (2,0.1)
Imports	63.79% (0.5,0.0005)	73.1% (0.5,0. 4)	66.7% (1.5,2^7)	**74.4%** (2,0.5)
Ecoli-Periplasm	83.2% (0.2,0.002)	98.2% (8.5,0.4)	90.9% (2.5,2^2)	**100%** (30,0.6)*
Cancer-Wpbc	86.4% (0.2,0.0001)	66.7% (8.5,0.8)	86.8% (1.5,2^9)	**87.7%** (3,0.2)
Breast	96.5% (0.6,0.003)	98.6% (0.5,0.5)	98.1% (1.5,2^4)	**99.6%** (10,0.8)
Heart	73.8% (18,0.7)	79.4% (0.7,0.5)	79.3% (1.1,2^{10})	**82.4%** (8,0.95)*
Abalone class	76.6% (0.8,0.0005)	78.4% (0.5,0.6)	79.1% (2.5,2^{10})	**80.2%** (12,0.7)
Glass building	64.5% (0.4,0.4)	45.6% (0.7,0.5)	**65.7%** (2,2^9)	65% (2,0.3)
Glass vehicle	**93.2%** (0.06,0.5)	92.2% (0.7,0.01)	90.3% (2.5,2^4)	**93.2%** (16,0.6)
Arrhythmia	68.9% (0.4,0.04)	73.4% (0.5,0.01)	73.7% (1.5,2^8)	**74.4%** (2,1)
Ionosphere good	16.67% (0.8,0.004)	73.6% (0.5,0.4)	74.8% (1.5,2^7)	**75.7%** (6,0.8)
Waveform 0	75.2% (0.6,0.9)	67.2% (0.4,1)	70.8% (3,2^9)	**79.6%** (1,0.8)*
Vowel 0	88.18% (0.08,0.05)	83.8% (0.6,1)	54.8% (2,2^5)	**91.6%** (18,0.8)*

B. 一类模式分类

为了评价 LMMREB 在 OCC 问题中分类的性能,实验采用 8 个 UCI 数据集(Ball-Bearing,Delft-Pump,Spectf0, Liver,Biomed,Housing MEDV,Hepatitis,Diabetes)作为实验数据.按照文献[13]中实验设置,对这些数据集分别随机抽取 70%正类样本和一小部分负类样本用于训练,以使得 95%训练样本属于正类,仅有 5%的训练样本属于负类,余下样本用作测试.8 个数据集的详细信息见表 5.

Table 5 One-Class classification datasets

表 5 一类模式分类数据集

Datasets	Normal examples	Abnormal examples	Target (m_1)	m_2	d
Ball-Bearing	913	3 237	640	40	32
Delft-Pump	376	1 124	263	14	64
Spectf0	95	254	67	5	44
Liver	200	145	140	7	6
Biomed	127	67	90	5	5
Housing	458	48	300	15	13
Hepatitis	123	32	90	5	19
Diabetes	500	268	200	10	8

采用 g-means 度量来评价算法性能,所有实验独立执行 10 次,实验结果取平均值.采用几何平均度量方法评价算法性能: $g=\sqrt{a^+\cdot a^-}$,其中,a^+和 a^-分别指正类和负类分类精度.该方法被广泛应用于处理不平衡数据集问题[16],并同时考虑了正类与负类的分类效果.

本节实验重点比较所提方法 LMMREB 与 SVDD,MEMEM,MMHSVM 及 SSLM 等方法的 OCC 分类性能,实验中各类数据集随机分割成训练集和测试集,每个数据集分别实验 10 次,取其平均值作为各自实验结论.表 6 记录了 3 种方法对 8 个实际数据集的 OCC 分类精度及其相应实验参数值.

Table 6 Average classification accuracy rates for one-class classification (%)

表 6 一类模式分类平均分类精度 (%)

Dataset	SVDD	MEMEM	MMHSVM	SSLM	LMMREB
	Acc. rate (g)	Acc. rate (γ)	Acc. rate (M,C)	Acc. rate (v)	Acc. rate (v,q)
Ball-Bearing	96.2% (0.8)	91.5% (0.1)	93.7% (2.5,2^5)	96.69% (2.5)	**96.7%** (2,0.95)
Delft-Pump	79.7% (11.5)	80.5**%** (0.3)	80.2% (2.5,2^{10})	**80.7% (35)**	80.5% (58,1)
Spectf0	75.9% (3.5)	77.8% (0.3)	66.7% (3,2^{10})	72.73% (2.5)	**79.7%** (1,0.7)*
Liver	53.5% (11.7)	59.7% (0.4)	61.3% (3,2^6)	68.5% (15)	**68.8%** (18,0.1)
Biomed	78.7% (2.7)	80.8% (0.8)	80.2% (3,2^8)	**81.1% (5)**	80.6% (1,1)
Housing EDV	79.9% (9.7)	81.1% (0.9)	80.8% (2.5,2^8)	81.2% (1.2)	**81.3%** (1,1)
Hepatitis	56.4% (4.9)	55.9% (0.95)	56.5% (2.5,2^{10})	56.1% (1.5)	**56.4%** (0.1,1)
Diabetes	69.8% (2.2)	71.6% (0.9)	69.2% (2,2^6)	69.9% (1)	**71.6%** (0.1,1)

3.4.2 人脸识别实验

为了进一步评价所提方法在高维数据集下的二类分类性能,选取 2 个标准人脸数据库:ORL 数据库(http://www.research.att.com/facedatabase.html)[19]和 Yale 数据库(http://cvc.yale.edu/projects/yalefaces/yalefaces.html)[20]作为测试数据,以评价所提方法在高维数据集下的分类性能.其中,ORL 人脸数据集包含 40 个对象的人脸数据,每类对象由不同表情的 10 张图片构成;Yale 人脸数据集包括 15 个类别的 165 幅灰度级图像,同一类由不同光照条件和脸部表情的 11 张人脸数据组成.实验前,对上述图像集进行预处理,使其缩放到 32×32 像素大小,且每个像素为 256 灰度级,则在图像空间,每张图像由一个 1 024 维向量表示.图 3(a)、图 3(b)分别显示了经预处理后的 ORL 数据集和 Yale 数据集中某一类对象的人脸图像.

(a) Yale datasets for one object

(b) ORL datasets for one object

Fig.3 Samples cropped face image in the face image datasets after being preprocessed

图 3 经过预处理后的人脸数据库部分样本数据

在 ORL 人脸数据集中随机抽取 5 个二分类的数据子集,即 ORL.1-2(指 ORL 库中第 1 类和第 2 类,以下类同),ORL.3-5,ORL.6-10,ORL.20-32,ORL.18-40.对于每个数据子集,随机选取 10 个数据作为训练集,其余作为测

试数据集.实验结果取自 10 重交叉验证实验的平均值.在 Yale 人脸数据集中随机抽取 3 个二分类的数据子集,即 Yale.3-5,Yale.6-10,Yale.8-12,在每个数据子集中分别对每类对象随机抽取 p={2,3,5}个图像数据作为训练子集,其余数据作为测试子集.实验中,以符号 P_m/T_n 表示每个数据子集中每类对象抽取 m 个图像用于训练,n 个图像用于测试.所有数据子集的实验结果分别取自 10 重交叉验证的平均值.表 7 记录了 ORL 数据集在不同二分类数据集下的实验结果,表 8~表 10 分别记录了 Yale 数据库中 3 个随机二分类数据集在抽取类对象不同训练图像数据下的实验结果.

Table 7 Recognition rate comparison on ORL datasets (%)

表 7 ORL 数据库识别性能比较 (%)

数据集		v-SVM	MEMEM	MMHSVM	LMMREB (v,q)
ORL	ORL.1-2	85.29	88.41	89.72	**91.60 (12,0.7)**
	ORL.3-5	88.82	87.26	89.96	**94.11 (7,0.95)**
	ORL.6-10	90.0	90.37	89.62	**93.53 (18,0.9)**
	ORL.20-32	**100**	97.90	98.70	99.32 (22,0.7)
	ORL.18-40	96.5	97.64	97.72	**98.80 (15,0.96)**

Table 8 Recognition rate comparison on Yale datasets P2/T9 (%)

表 8 Yale 数据库 P_2/T_9 识别性能比较 (%)

P_2/T_9		v-SVM	MEMEM	MMHSVM	LMMREB (v,q)
Yale	Yale.3-5	60.0	74.28	76.0	**82.3 (4,0.68)**
	Yale.6-10	90.0	89.37	91.12	**94.46 (7,0.85)**
	Yale.8-12	80.0	82.51	82.14	**84.74 (11,0.7)**

Table 9 Recognition rate comparison on Yale datasets P3/T8 (%)

表 9 Yale 数据库 P_3/T_8 识别性能比较 (%)

P_3/T_8		v-SVM	MEMEM	MMHSVM	LMMREB (v,q)
Yale	Yale.3-5	62.74	78.27	80.33	**84.28 (4,0.7)**
	Yale.6-10	88.33	89.43	91.36	**94.74 (5,0.9)**
	Yale.8-12	78.50	82.62	82.50	**84.92 (13,0.75)**

Table 10 Recognition rate comparison on Yale datasets P5/T6 (%)

表 10 Yale 数据库 P_5/T_6 识别性能比较 (%)

P_5/T_6		v-SVM	MEMEM	MMHSVM	LMMREB (v,q)
Yale	Yale.3-5	58.89	78.44	81.50	**85.15 (4,0.74)**
	Yale.6-10	76.58	89.48	92.28	**95.26 (9,0.98)**
	Yale.8-12	79.00	82.0	84.42	**88.27 (13,0.8)**

3.4.3 实验结论

通过对上述实际数据集的实验结果进行分析,可得出如下几点结论:

(1) 从 UCI 数据集实验结果可看出,对于模式分类(包括二类分类和一类分类)问题,在优化选取参数 q,v 值对的情况下,LMMREB 对所有数据集均具有优于或可比较的模式分类性能,甚至对某些数据集(表 4 和表 6 中* 号标注),LMMREB 的分类性能明显优于相关方法.

(2) 当参数 q 值趋近于 1 时,LMMREB 模式分类性能与 MEMEM,SVDD,MMHSVM 及 SSLM 等方法相当.这进一步证实,当 q=1 时,在一定的参数对应条件下,LMMREB 方法等价为 MEMEM,SVDD,MMHSVM 及 SSLM 等 gap-tolerant 分类方法.为此引出如下定理:

定理 8. 在满足一定的参数条件(如 q=1 或(和)v=0)下,LMMREB 方法等价于 SVDD,MEMEM,MMHSVM 和 SSLM 等 gap-tolerant 方法.

证明:当 q=1 时,LMMREB 变成了 MEMEM,从而可知 MEMEM 是 LMMREB 方法的一个特例,即 LMMREB 是 MEMEM 方法的一个泛化.另外,MMHSVM 直接优化两个分别包含二类模式的最小包含球,实现二类模式最大化分割,其隐含了模式间大间隔思想,令 R=(R_1+R_2)/2,其中,R 为 MEMEM 和 MMHSVM 的分割包含球半径,R_1,

R_2 分别为 MMHSVM 的两个最小包含球半径.可见,MEMEM 和 MMHSVM 的分割球具有一一对应关系,MMHSVM 本质上等价于 MEMEM 方法,从而可知 LMMREB 也等价于 MMHSVM 方法.SVDD 是 SSLM 在 v=0 下的一种特例,下面只需证明在一定的参数条件下,LMMREB 等价于 SSLM 方法.

假设(R_s^*, ρ_s^{*2})为 SSLM 的优化解,令 $R_L = R_s^* + \rho_s^{*2}/2, \rho_L^2 = \rho_s^{*2}$,其中,R_L 为 LMMREB 的分割球半径,$\rho_L^2 = 2\rho^2 + 2(1-q)R_L^2$ 为 LMMREB 的二类模式间间隔.当 q=1 时, $\rho_L^2 = 2\rho^2$,即 $\rho_s^{*2} = 2\rho^2$.由此可知,当 q=1 时,SSLM 的分割球半径和二类模式间间隔分别与 LMMREB 的分割球半径和二类模式与其所属压缩包含球间间隔一一对应,即 LMMREB 与 SSLM 的优化解等价.

可以推论,在满足一定的参数条件(如 q=1 或(和)v=0)下,LMMREB 方法等价于 SVDD,MEMEM, MMHSVM 和 SSLM 等 gap-tolerant 方法.由此定理得证. □

定理 8 进一步说明,LMMREB 通过引入可调的压缩因子参数 q,具有相比其他 gap-tolerant 方法更强的学习泛化能力.

(3) 从人脸识别实验结果(表 7~表 10)可以看出,对于高维数据分类(人脸识别)问题,LMMREB 在一定程度上具有明显的统计学习优势.v-SVM 方法在大多数二分类图像数据集上的识别性能均低于其他几种方法,这说明对于图形识别这类具有高维非线性结构的数据集,考虑数据类间的绝对最大间隔在一定程度上已不能很好地满足模式识别性能的要求.而考虑了数据类内分布结构最小化的球形学习机方法均取得了相对较好的识别性能,尤其是在考虑了数据类内结构压缩最小化的 LMMREB 方法的识别性能明显优于其他几种相关方法.尤其值得说明的是,本文方法在充分考虑数据类内分布结构压缩最小化和类间间隔最大化的情况下,在所有数据集上均取得了最优或相当的人脸识别性能.

(4) 从表 8~表 10 可看出,在各类对象训练样本较少时,由于不能较好地呈现数据类内的分布结构性状,从而导致 4 种方法均不能取得较好的模式识别率,但是相比较之下,本文方法在一定程度上依然取得了较好的识别性能.随着各类对象训练数据集大小的增加,数据的类内分布结构呈现复杂,传统的 v-SVM 方法的识别性能有所下降;其他几种方法的识别性能虽然在一定程度上有所增强,但本文方法性能上升更明显,从而说明不管是在简单的或是复杂的高维非线性分布结构的数据下,由于充分考虑了数据的类内分布结构压缩最小化和类间间隔最大化,导致本文方法均具备较强的模式识别能力.

(5) 另外,从表 8~表 10 还可以看出,随着训练样本数的增大,类内数据分布体积趋于增大,使得压缩包含球的体积增大,从而使参数 q 值呈现上升趋势.

4 结束语

受 SVM 和大间隔小球体等方法的启发,提出了一种大间隔最小压缩包含球学习机 LMMREB.通过引入一个可调节的包含球压缩因子 q,显著增强 gap-tolerant 球形分类机的模式分类性能.另外,通过引入一个控制间隔误差的参数 v,进一步细化了 LMMREB 模式分类边界,使得 LMMREB 在最小化二类模式内聚性的同时最大化类间间隔,提升了 LMMREB 的泛化能力.人工数据和实际数据实验结果均显示,LMMREB 具有相较于其他相关方法明显或相比较的模式分类性能优势.如何更有效地协调选取参数 q,v 的优化取值,有待进一步研究.

致谢 在此,我们向对本文工作给予支持和建议的同行,尤其是各位审稿专家表示衷心的感谢.

References:

[1] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998,2(2): 955−974. [doi: 10.1023/A:1009715923555]

[2] Wen CJ, Zhan YZ, Chen CJ. Maximal-Margin minimal-volume hypersphere support vector machine. Control and Decision, 2010, 25(1):79−83 (in Chinese with English abstract).

[3] Tao JW, Wang ST, Hu WJ, Ying WH. ρ-Margin kernel learning machine with magnetic field effect for both binary classification and novelty detection. Int'l Journal of Software and Informatics, 2010,4(3):305−324.

[4] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[5] Cortes C, Vapnik V. Support vector networks. Machine Learning, 1995,20(3):273−297. [doi: 10.1007/BF00994018]
[6] Schölkopf B, Smola AJ, Williamson R, Bartlett PL. New support vector algorithms. Neural Computation, 2000,12(5):1207−1245. [doi:10.1162/089976600300015565]
[7] Vapnik V, Chapelle O. Bounds on error expectation for support vector machines. Neural Computation, 2000,12(9):2013−2036. [doi:10.1162/089976600300015042]
[8] Shivaswamy P, Jebara T. Ellipsoidal kernel machines. In: Meila M, Shen XT, eds. Proc. of the 11th Int'l Conf. on Artificial Intelligence and Statistics. San Juan: Omni Press, 2007.
[9] Tax DMJ, Duin RPW. Support vector data description. Machine Learning, 2004,54(1):45−66.I. [doi: 10.1023/B:MACH.0000008084.60811.49]
[10] Wang JG, Neskovic P, Cooper LN. Pattern classification via single sphere. Lecture Notes in Computer Science, Discovery Science, 2005,37(35):241−252. [doi: 10.1007/11563983_21]
[11] Hao PY, Chiang JH, Lin YH. A new maximal margin spherical structured multi-class support vector machine. Applied Intelligence, 2009,30(2):98−111. [doi: 10.1007/s10489-007-0101-z]
[12] Liu Y, Zheng YF. Minimum enclosing and maximum excluding machine for pattern description and discrimination. In: Tang YY, *et al*., eds. Proc. of the ICPR 2006 Conf. Hongkong: IEEE Computer Society Press, 2006. 129−132. [doi: 10.1109/ICPR.2006.799]
[13] Wu MR, Ye JP. A small sphere and large margin approach for novelty detection using training data with outliers. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(11):2088−2092. [doi: 10.1109/TPAMI.2009.24]
[14] Müller KR, Mika S, Ratsch G, Tsuda K, Schölkopf B. An introduction to kernel-based learning algorithms. IEEE Trans. on Neural Networks, 2001,12(2):181−201. [doi: 10.1109/72.914517]
[15] Peng XJ. Least squares twin support vector hypersphere (LS-TSVH) for pattern recognition. Expert Systems with Applications, 2010,37(12):8371−8378. [doi: 10.1016/j.eswa.2010.05.045]
[16] Tsang W, Kwok JT, Cheung PM. Core vector machines: Fast SVM training on very large datasets. Journal of Machine Learning Research, 2005,6(1):363−392.
[17] Deng ZH, Chung FL, Wang ST. FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation. Pattern Recognition, 2008,41(4):1363−1372. [doi: 10.1016/j.patcog.2007.09.013]
[18] Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines. Machine Learning, 2002,46(1):131−159. [doi: 10.1023/A:1012450327387]
[19] Wang HX, Chen SB, Hu ZL, Zheng WM. Locality-Preserved maximum information projection. IEEE Trans. on Neural Networks, 2008,19(4): 571−585. [doi: 10.1109/TNN.2007.910733]
[20] Gao QX, Xie DY, Xu H, Li YZ, Gao XQ. Supervised feature extraction based on information fusion of local structure and diversity information. Acta Automatica Sinica, 2010,36(8):1107−1114 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2010.01107]

附中文参考文献:

[2] 文传军,詹永照,陈长军.最大间隔最小体积球形支持向量机.控制与决策,2010,25(1):79−83.
[20] 高全学,谢德燕,徐辉,李远征,高西全.融合局部结构和差异信息的监督特征提取算法.自动化学报,2010,36(8):1107−1114. [doi: 10.3724/SP.J.1004.2010.01107]

陶剑文(1973−),男,湖北武汉人,博士生,副教授,主要研究领域为模式识别,Web挖掘.

王士同(1964−),男,教授,博士生导师,主要研究领域为人工智能,机器学习.