二阶环形间隔分类器

方景龙[1,2],王万良[1],梁东升[2],周其力[2]
(1.浙江工业大学计算机科学与技术学院,浙江杭州 310023;2.杭州电子科技大学图形图像研究所,浙江杭州 310018)

摘 要: 采用二阶损失函数,提出了一种二阶环形间隔分类器,它运用超球面将两类训练样本隔开,同时最大化两类间隔.实验结果表明,不管是对平衡问题还是不平衡问题,所提出的二阶环形间隔分类方法都获得了比较好的分类效果.

关键词: 模式分类;支持向量数据描述;二阶损失函数;环形间隔
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2011) 07-1634-05

A Two-Norm Annular Margin Classifier

FANG Jing-long[1,2], WANG Wan-liang[1], LIANG Dong-sheng[2], ZHOU Qi-li[2]
(1. *School of Computer Science and Technology*, *Zhejiang University of Technology*, *Hangzhou*, *Zhejiang* 310023, *China*;
2. *Institute of Graphic and Image*, *Hangzhou Dianzi University*, *Hangzhou*, *Zhejiang* 310018, *China*)

Abstract: A two-norm annular margin classifier is put forward based on two-norm cost function. The classifier defines a margin between two categories of training data with super sphere-structure and maximizes the margin. Experiment shows that the two-norm annular margin classifier improves the effects of classification for both balance and non-balance problems.

Key words: pattern classification; support vector data description; two-norm cost function; annular margin

1 引言

1999 年,Tax 等人[1]在最小包围球(Minimum Enclosing Ball, MEB)理论和支持向量机(Support Vector Machine,SVM)理论基础上提出了一种支持向量区域描述方法 SVDD (Support Vector Domain Description),其主要思想是通过计算包含一组数据的最小超球来对该组数据进行描述,主要用来进行单类分类及剔除噪点和奇异点,这一方法现在更多地被称为支持向量数据描述方法(Support Vector Data Description,仍然简称为 SVDD).为了利用负类样本的信息,2004 年 Tax 等人[2]又提出了考虑负类样本的 SVDD 方法 NSVDD(SVDD with Negative examples),该方法针对两类样本问题进行数据描述,将正类样本约束在超球内,负类样本约束在超球外,实际上已经是一个二分类器.

2003 年,Zhu 等人[3]利用 SVDD 对多类问题进行分类,提出了一种球结构支持向量机方法 SSVM(Sphere-structured Support Vector Machine),该方法对每一类样本构造一个最小超球并将该类样本约束在超球里,然后根据测试样本离各个球心的距离来判断一个样本的类别属性.随后,2005 年,Wu Q 等人[4]提出了一个与 SSVM 方法类似的方法并将其用于脑电图源定位.2007 年,Wang 等人[5]提出了基于最小包围球的贝叶斯分类方法 BCMBS(Bayes Classication based on Minimum Bounding Spheres),该方法即是在训练样本服从高斯分布的假设下利用贝叶斯最优判别准测对 Zhu 等人提出的 SSVM 方法的决策函数进行修改;同年,Lee 等人[6]提出了一个求解多类问题的区域描述支持向量分类方法 DDSVC (Domain Described Support Vector Classier for multi-classication problems),该方法即是在 Zhu 等人提出的方法基础上根据贝叶斯公式计算出后验概率对测试样本进行判别.2008 年,梁锦锦等人[7]提出了一种求解两类问题的空间支持向量域分类器 SSVDC(Space Support Vector Domain Classifier),在求出包含每类样本的最小超后,利用测试样本离各个球心距离进行类别判断,当两个球同时包含或者同时不包含测试样本时改用 KNN 方法进行判别.同文献[7]类似,2009 年,Lei 等人[8]提出了一种 SSVM-NDR 方法(The SSVM with a Novel Decision Rule),对 Zhu 等人提出的 SSVM 方法的判别函数进行修改,当用测试样本离各个球心距离进行类别判断不明确时就在不明确区域使用最近邻方法进行判别.

2005 年,Wang 等人[9]结合支持向量机最大间隔和

收稿日期:2010-06-19;修回日期:2010-09-25
基金项目:国家自然科学基金(No.60874074);浙江省科技计划(No.2009C14032)

SVDD球面分类的思想,提出一种最大化两类超球面分离比的方法 SSPC(Pattern Classification via Single Spheres),该方法将一类样本约束在超球内,另一类样本约束在超球外,并最小化超球半径和最大化两类间隔.随后,2009 年,Mingrui Wu 等人[10]针对不平衡问题分类提出了一个与 SSPC 类似的最小球最大间隔方法 SSLM(Small Sphere and Large Margin);Hao 等人[11]则借助 SSPC 思想对多类问题提出了一个最大间隔球结构支持向量机 MSM-SVM(Maximal-margin Spherical-structured Multi-class Support Vector Machine).

另外,2007 年,Hao 等人[12]提出了一个求解多分类问题的球型支持向量机 SMSVM(Spherical-shaped Multiple-class Support Vector Machine),该方法以所有超球半径的平方和最小为目标函数对每一类构造一个超球以使该类样本约束在超球里而将其他样本约束在超球外,然后根据测试样本离各个球心的距离的一个函数来判断该样本属于哪一类.2009 年,朱孝开、杨德贵[13]提出了一种基于推广能力测度的多类 SVDD 模式识别方法,该方法训练多层的 SVDD 来对测试样本进行判别,即先用推广能力强的外层 SVDD 进行判别,当不行时再用推广能力相对较弱的内层 SVDD 进行判别;Mu 等人[14]则在 Tax 等人提出的 NSVDD 方法基础上提出了二阶 NSVDD 方法和 ν-NSVDD 方法用来对每类样本构造一个超球,然后利用线性判别分析(LDA)和最邻近准则(NN Rule)判断测试样本应该属于哪一类,文中的 ν-NSVDD 方法与 SSPC 也类似.

2　早先的工作

设给定一组 l 个训练样本 $x_i \in \mathbf{R}^n$, $i=1,2,\cdots,l$, SVDD 方法[1]就是在特征空间中寻找最小超球 $S(c,R)$,c,R 分别为球心和半径,使该球包围所有训练样本,即求解问题 1.

问题 1　(SVDD 方法)

$$\min_{R,c,\xi} \quad R^2 + C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t.} \quad \|\varphi(x_i)-c\|^2 \leqslant R^2+\xi_i, \xi_i \geqslant 0, i=1,2,\cdots,l$$

如果 l 个训练样本分成两类,$y_i \in \{1,-1\}$ 为相应的类标,l_1 和 l_2 分别是正类样本和负类样本的个数,$l = l_1 + l_2$.则 NSVDD 方法[2]就是在特征空间中寻找最小超球 $S(c,R)$,使该球包围所有正类训练样本,但不包含所有负类训练样本,即求解问题 2.

问题 2　(NSVDD 方法)

$$\min_{R,c,\xi} \quad R^2 + C_1\sum_{y_i=+1}\xi_i + C_2\sum_{y_i=-1}\xi_i$$
$$\text{s.t.} \quad y_i(R^2 - \|\varphi(x_i)-c\|^2) \geqslant -\xi_i,$$
$$\xi_i \geqslant 0, \quad i=1,2,\cdots,l$$

如果要考虑多类问题的分类,则我们假设 l 个训练样本被分成 k 类,$y_i \in \{1,2,\cdots,k\}$ 为相应的类标,每类样本的个数分别是 l_j, $l = l_1 + l_2 + \cdots + l_k$.

2003 年,Zhu 等人[3]利用 SVDD 对多类问题进行分类,提出了一种球结构支持向量机方法 SSVM,该方法通过对每一类训练样本求解问题 1 来构造包含该类样本的最小超球 $S_j(c_j,R_j)$,然后根据测试样本离各个球心的距离来判断样本该属于哪一类.2005 年,Wu Q 等人[4]提出了一个与 SSVM 完全类似的方法并将其用于脑电图源定位,该方法与 SSVM 的唯一区别是利用相对距离而不是绝对距离来进行类别判断.2007 年,Wang 等人[5]提出了基于最小包围球的贝叶斯分类方法 BCMBS,该方法实际上就是在训练样本服从高斯分布的假设下利用贝叶斯最优判别准则对 Zhu 等人提出的方法的决策函数进行修改;同年,Lee 等人[6]基于贝叶斯决策准则提出了一个求解多类问题的区域描述支持向量分类方法 DDSVC,该方法实际上就是在 Zhu 等人提出的方法基础上根据贝叶斯公式计算后验概率来对测试样本进行判别.2009 年,Lei 等人[8]提出了一种 SSVM-NDR 方法,对 Zhu 等人提出的方法的判别函数进行修改,当用测试样本对各个球心距离进行类别判断不明确时就在不明确区域使用最近邻方法.

2005 年,Wang 等人[9]结合支持向量机最大间隔和 SVDD 球面分类的思想,提出一种最大化两类超球面分离比方法 SSPC,该方法将一类样本约束在超球内,另一类样本约束在超球外,并最小化超球半径和最大化两类间隔,即求解问题 3.2009 年, Mingrui Wu 等人[10]针对不平衡二类问题提出一个与 SSPC 类似的最小球最大间隔方法 SSLM,即求解问题 4;同年,Hao 等人[11]又借助 SSPC 思想对多类问题提出了一个最大间隔球结构支持向量机 MSM-SVM,它对每一类样本求解问题 5,并采用多种决策函数来判别测试样本应该属于哪一类.

问题 3　(SSPC)

$$\min_{R,c,d,\xi} \quad R^2 - Kd^2 + C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t.} \quad y_i(R^2 - \|\varphi(x_i)-c\|^2) \geqslant d^2 - \xi_i,$$
$$\xi_i \geqslant 0, \quad i=1,2,\cdots,l$$

问题 4　(SSLM 方法)

$$\min_{R,c,d,\xi} R^2 - \upsilon d^2 + \frac{1}{l_1\upsilon_1}\sum_{y_i=1}\xi_i + \frac{1}{l_2\upsilon_2}\sum_{y_i=-1}\xi_i$$
$$\text{s.t.} \quad \|\phi(x_i)-c\|^2 \leqslant R^2+\xi_i, \quad \text{当 } y_i=+1;$$
$$\|\phi(x_i)-c\|^2 \geqslant R^2+d^2-\xi_i, \quad \text{当 } y_i=-1$$

问题 5　(MSM-SVM 方法)

$$\min_{R,c,d,\xi} R_j^2 - Md^2 + \frac{C}{l_j}\sum_{y_i=j}\xi_i + \frac{C}{l-l_j}\sum_{y_i\neq j}\xi_i$$

$$\text{s.t.}\quad \|\phi(x_i)-c_j\|^2 \leqslant R^2+\xi_i,\qquad \text{当 } y_i=j;$$
$$\|\phi(x_i)-c_j\|^2 \geqslant R^2+d_j^2-\xi_i,\quad \text{当 } y_i\neq j;\ \xi_i\geqslant 0$$

2007 年,Hao 等人[12]提出了一个求解多分类问题的球型支持向量机 SMSVM,该方法以所有超球半径的平方和最小为目标函数对每一类构造一个超球以使该类样本约束在超球里而将其他样本约束在超球外,然后根据测试样本离各个球心的距离的一个函数来判断该样本属于哪一类.2009 年,Mu 等人[14]在 Tax 等人提出的 NSVDD 方法基础上提出了二阶 NSVDD 方法和 ν-NSVDD 方法用来对每类样本构造一个超球,然后根据训练样本距球心的距离,利用线性判别分析(LDA)和最邻近准则(NN Rule)判断测试样本应该属于哪一类,文中的 ν-NSVDD 方法与 SSPC 类似.

问题 6 (二阶 NSVDD 方法)

Mu 等提出的二阶 NSVDD 方法求解如下问题[14]:

$$\min_{R,c,\xi} R^2+\frac{C_1}{2}\sum_{y_i=+1}\xi_i^2+\frac{C_2}{2}\sum_{y_i=-1}\xi_i^2$$
$$\text{s.t.}\quad y_i(R^2-\|\phi(x_i)-c\|^2)\geqslant -\xi_i,\ i=1,2,\cdots,l$$

问题 7 (ν-NSVDD 方法)

Mu 等人提出的 ν-NSVDD 方法求解如下问题[14]:

$$\min_{R,c,d,\xi} R^2-\upsilon d+\frac{1}{l_1}\sum_{y_i=1}\xi_i+\frac{1}{l_1}\sum_{y_i=-1}\xi_i$$
$$\text{s.t.}\quad y_i(R^2-\|\phi(x_i)-c\|^2)\geqslant d-\xi_i,$$
$$\xi_i\geqslant 0,\quad i=1,2,\cdots,l,\quad d\geqslant 0$$

3 二阶环形间隔分类方法

结合 Wang 等提出的 SSPC 方法[5],Mingrui Wu 等提出的 SSML 方法[11],以及 Mu 等提出的二阶 NSVDD 方法[14]的思想,提出二阶环形间隔分类方法 2-norm AMPC (2-norm Pattern Classification via Annular Margin).

问题 8 (2-norm AMPC)

二阶环形间隔分类器优化问题为:

$$\min_{R,c,d,\xi} R^2-Md^2+\frac{1}{2l\upsilon_0}\sum_{i=1}^{l}\xi_i^2$$
$$\text{s.t.}\quad y_i(R^2-\|\phi(x_i)-c\|^2)\geqslant d^2-\xi_i$$

其中 M 是调节超球半径和环形间隔的常数,υ_0 是惩罚参数.引入 Lagrange 因子 $\alpha_i\geqslant 0$,相应的 Lagrange 函数为:

$$L=R^2-Md^2+\frac{1}{2l\upsilon_0}\sum_{i=1}^{l}\xi_i^2 -\sum_{i=1}^{l}\alpha_i(y_i(R^2-\|\phi(x_i)-c\|^2)-d^2+\xi_i)$$

对 L 关于 R,d,c,ξ_i 求极小,得到:

$$\sum_{i=1}^{l}y_i\alpha_i=1,\ \sum_{i=1}^{l}\alpha_i=M,$$
$$c=\sum_{i=1}^{l}\alpha_i y_i\phi(x_i),\ \xi_i=l\upsilon_0\alpha_i$$

根据 Wolfe 对偶原理,将上述极值条件代入 Lagrange 函数,得到相应的对偶问题:

$$\max_{\alpha} L=\sum_{i=1}^{l}\alpha_i y_i K(x_i,x_i)-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j[y_i y_j K(x_i,x_j)+\delta_{ij}]$$
$$\text{s.t.}\quad \sum_{y_i=+1}\alpha_i=\frac{M+1}{2},\ \sum_{y_i=-1}\alpha_i=\frac{M-1}{2};\alpha_i\geqslant 0$$

其中 $\delta_{ij}=\begin{cases} l\upsilon_0/2 & ,i=j\\ 0 & ,i\neq j\end{cases}$,$K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$为核函数.如采用高斯核函数 $k(x,y)=\exp(-\gamma\|x-y\|^2)$,则 $K(x,x)=\text{const}$,这样对偶问题可等价为:

$$\max_{\alpha} L=-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j[y_i y_j K(x_i,x_j)+\delta_{ij}]$$
$$\text{s.t.}\quad \sum_{y_i=+1}\alpha_i=\frac{M+1}{2},\ \sum_{y_i=-1}\alpha_i=\frac{M-1}{2};\alpha_i\geqslant 0$$

定理 1 对于问题 8,有$\frac{1}{l}\sum_{i=1}^{l}\xi_i=M\upsilon_0$.

证明 根据优化问题的 KKT 条件,有 $\xi_i=l\upsilon_0\alpha_i$,所以 $\sum_{i=1}^{l}\xi_i=l\upsilon_0\sum_{i=1}^{l}\alpha_i=l\upsilon_0 M$,即$\frac{1}{l}\sum_{i=1}^{l}\xi_i=M\upsilon_0$.

如果是对不平衡问题进行分类,则需对问题 8 进行变形,下面给出两种这样的变形,即问题 9 和问题 10.

问题 9 (2-norm AMPCID1)

求解不平衡问题的第一类二阶环形间隔分类器优化问题为:

$$\min_{R,c,d,\xi} R^2-Md^2+\frac{1}{2l_1\upsilon_1}\sum_{y_i=1}\xi_i^2+\frac{1}{2l_2\upsilon_2}\sum_{y_i=-1}\xi_i^2$$
$$\text{s.t.}\quad \|\phi(x_i)-c\|^2\leqslant R^2+\xi_i,\qquad \text{当 } y_i=+1;$$
$$\|\phi(x_i)-c\|^2\geqslant R^2+d^2-\xi_i,\quad \text{当 } y_i=-1$$

其对偶问题为:

$$\max_{\alpha} L=-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j[y_i y_j K(x_i,x_j)+\delta_{ij}]$$
$$\text{s.t.}\quad \sum_{y_i=+1}\alpha_i=M+1,\ \sum_{y_i=-1}\alpha_i=M;\alpha_i\geqslant 0$$

其中 $\delta_{ij}=\begin{cases} l_1\upsilon_1/2, & i=j,\quad y_i=1\\ l_2\upsilon_2/2, & i=j,\quad y_i=-1\\ 0, & i\neq j\end{cases}$

定理 2 对于问题 9 有$\frac{1}{l_1}\sum_{y_i=+1}\xi_i=(M+1)\upsilon_1$,$\frac{1}{l_2}\sum_{y_i=-1}\xi_i=M\upsilon_2$.

证明 根据优化问题的 KKT 条件,对于正类样本有 $\xi_i=l_1\upsilon_1\alpha_i$,对于负类样本有 $\xi_i=l_2\upsilon_2\alpha_i$,所以

$$\sum_{y_i=+1}\xi_i=l_1\upsilon_1\sum_{y_i=+1}\alpha_i=l_1\upsilon_1(M+1),$$
$$\sum_{y_i=-1}\xi_i=l_2\upsilon_2\sum_{y_i=-1}\alpha_i=l_2\upsilon_2 M,$$

即$\frac{1}{l_1}\sum_{y_i=+1}\xi_i=(M+1)\upsilon_1,\frac{1}{l_2}\sum_{y_i=-1}\xi_i=M\upsilon_2$

问题 10　(2-norm AMPCID2)

求解不平衡问题的第二类二阶环形间隔分类器优化问题为:

$$\min_{R,c,d,\xi} R^2-Md^2+\frac{1}{2l_1\upsilon_1}\sum_{y_i=1}\xi_i^2$$

$$\text{s.t.}\quad \|\phi(x_i)-c\|^2\leqslant R^2-d^2+\xi_i,\ \text{当 } y_i=+1;$$

$$\|\phi(x_i)-c\|^2\geqslant R^2+d^2,\quad \text{当 } y_i=-1$$

其对偶问题为:

$$\max_\alpha L=-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j[y_iy_jK(x_i,x_j)+\delta_{ij}]$$

$$\text{s.t.}\quad \sum_{y_i=+1}\alpha_i=\frac{M+1}{2},\ \sum_{y_i=-1}\alpha_i=\frac{M-1}{2},\alpha_i\geqslant 0$$

其中 $\delta_{ij}=\begin{cases} l_1\upsilon_1/2, & i=j \quad \text{and} \quad y_i=1\\ 0, & \text{others}\end{cases}$

定理 3　对于问题 10 有$\frac{1}{l_1}\sum_{y_i=+1}\xi_i=\frac{M+1}{2}\upsilon_1$.

证明　根据优化问题的 KKT 条件,对于正类样本有 $\xi_i=l_1\upsilon_1\alpha_i$,所以 $\sum_{y_i=+1}\xi_i=l_1\upsilon_1\sum_{y_i=+1}\alpha_i=l_1\upsilon_1\frac{M+1}{2}$,即 $\frac{1}{l_1}\sum_{y_i=+1}\xi_i=\frac{M+1}{2}\upsilon_1$.

问题 8、9、10 的决策函数为 $f(x)=\text{sgn}(R^2-\|\varphi(x)-c\|^2)$,即对一个给定的测试样本,如果 $f(x)=R^2-\|\varphi(x)-c\|^2\geqslant 0$ 则判断为正类样本点,否则就判断为负类样本点.

4　实验与分析

本文实验采用的七组数据为 Breast-Cancer, Ionosphere, Sonar, Diabetes, Hepatitis, Liver, Spectf,前三个数据集取自于 http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/binary.html,后四个数据集取自于 http://archive.ics.uci.edu/ml/machine-learning-databases/,这七个数据集的特征如表 1 所示,在使用之前我们对这些数据进行了[-1,+1]的规范化.

对前三个数据集,我们采用五倍交叉验证(5-fold cross validation)来获得模型中的最优参数,即将训练样本分成五份,轮流将其中四份作为训练样本一份作为测试样本,五次结果的均值作为对算法精度的估计.为了更科学和精确一点,我们共进行了 20 次五倍交叉验证,然后求均值,最后根据这个均值选取最优参数,表 2 是这三种数据集在不同算法下交叉验证的最优分类精度.在计算中,所有方法均采用高斯核函数 $k(x,y)=\exp(-\gamma\|x-y\|^2)$,核参数 γ 值搜寻范围为$\{2^{-12},2^{-11},\cdots,2^4\}$.同时,SSML 方法中 υ_1,υ_2 参数取了同一个值,它们与 2-norm AMPC 方法中的 υ_0 一样,其搜寻范围为$\{2^{-12},2^{-11},\cdots,2^0\}$;NSVDD 方法和 2-norm NSVDD 方法中参数 C_1,C_2 也取了同一个值,其搜寻范围为$\{2^{-2},2^{-1},\cdots,2^{10}\}$.同时,SSPC 方法、SSLM 方法、MSM-SVM 方法和 2-norm AMPC 方法中的 K、υ 和 M 值的搜寻范围为$\{2^4,2^5,\cdots,2^{10}\}$,BCMBS 方法中的 ΔT 值的搜寻范围为$\{-0.2,-0.15,\cdots,0.2\}$,SSVM-NDR 方法中的 θ 值的搜寻范围为$\{0.8,0.85,\cdots,1.0\}$.

表 1　实验中使用的数据集

Dataset	维数	正类样本总数	负类样本总数	训练中正类样本数	训练中负类样本数
Breast-Cancer	10	444	239	355	191
Ionosphere	34	111	97	88	77
Sonar	60	225	126	180	100
Diabetes	10	444	239	350	5
Hepatitis	60	111	97	86	5
Live	34	225	126	140	5
Spectf	34	225	126	38	5

表 2　十一种方法在 3 个数据集上获得的最好分类精度和标准差

方法 \ 数据	Breast-Cancer	Ionosphere	Sonar
SVM	**97.3060(0.0735)**	**94.8148(0.6640)**	**89.6635(1.2032)**
SSVM[3]	**97.0864(0.2064)**	88.7179(0.9445)	83.8702(0.2454)
BCMBS[5]	**97.3206(0.2332)**	**95.0855(0.3442)**	84.2067(2.5880)
DDSVC[6]	**96.2811(0.2145)**	90.9829(0.7694)	85.8413(1.6246)
SSVM-NDR[8]	**96.5593(0.3106)**	**94.0171(0.4029)**	85.0962(1.3598)
SSPC[9]	**97.1742(0.1432)**	**95.4558(0.3517)**	**89.7356(1.2110)**
SSLM[10]	**97.1230(0.5)**	92.2507(0.6444)	78.2933(0.2140)
MSM-SVM[11]	**97.4378(0.1609)**	**95.1140(0.4264)**	**89.3990(1.172)**
NSVDD[14]	**96.8228(0.3455)**	87.0513(3.1311)	70.8894(3.2438)
2-norm NSVDD[14]	**97.1523(0.0883)**	85.8974(3.5841)	71.1058(3.2454)
2-norm AMPC	**97.0351(0.2724)**	**95.6695(0.2943)**	**90.8654(0.0000)**

对后四个数据集,我们主要测试不平衡数据对结果的影响,所以我们随机抽取 70% 的正类样本和 5 个负类样本进行训练,而把剩余的所有样本用于测试,共进行 20 次然后求均值,最后根据这个均值选取最优参数,表 3 是这四种数据集在不同算法下的最优几何分类精度 $g=\sqrt{a^+a^-}$(这里 a^+,a^- 分别是正类样本和负类样本的分类精度).在计算中,各参数的搜寻范围同前面一样,不过,这次 NSVDD 方法和 2-norm NSVDD 方法采用了两个参数 C_1,C_2,对于 SVM 方法,SSPC 方法,我们也针对不平衡问题进行了适当的变形,即采用了两个惩罚参数 C_1,C_2.

在表 2 和表 3 中，对于每个数据集，我们将分类精度较高(离最优的不超过 2%)的结果用黑体加粗的方式进行了显示，从中可以看出，对于平衡问题，SVM 方法、SSPC 方法、MSM-SVM 方法、以及我们提出的 2-norm AMPC 方法较好，对所有 3 个数据集都有较高的分类精度；对不平衡问题，SSLM 方法，以及我们提出的 2-norm AMPCID1 方法和 2-norm AMPCID2 方法较好，对所有的 4 个数据集都有较好的分类精度.

表 3 八种方法在 4 个不平衡数据集上获得的最好几何分类精度和标准差

数据 / 方法	Diabetes	Hepatitis	Liver	Spectf
SVM	**70.876(4.491)**	**74.386(4.7526)**	60.265(4.8197)	71.017(4.5305)
SSVM[3]	13.891(10.329)	7.1827(9.4599)	26.76(11.256)	20.439(9.3515)
DDSVC[6]	25.811(7.2445)	33.88(3.2197)	22.53(6.7585)	26.212(3.2896)
SSPC[9]	67.807(2.8506)	**73.485(5.5153)**	62.89(5.5271)	**73.718(5.3117)**
SSLM[10]	**71.9045(2.9861)**	**74.8758(5.5492)**	**65.0971(7.4163)**	**75.7178(5.8639)**
MSM-SVM[11]	46.7(11.502)	52.751(7.463)	50.165(6.3326)	**74.79(2.7735)**
2-norm AMPCID1	**71.92(2.6108)**	**75.093(6.5603)**	**65.318(3.4669)**	**75.339(5.5095)**
2-norm AMPCID2	**71.385(2.3203)**	**75.275(5.3156)**	**63.338(6.1555)**	**74.727(5.854)**

5 结束语

采用二阶损失函数，本文提出了一种二阶环形间隔分类器，它运用两个同心超球面将两类训练样本隔开，同时最大化两类间隔. 实验验结果表明，不管是对平衡问题还是不平衡问题，我们提出的二阶环形间隔分类方法都获得了比较好的分类效果.

参考文献

[1] D M J Tax, R P W Duin. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20(11 – 13): 1191 – 1199.

[2] D M J Tax, R P W Duin. Support vector data description[J]. Machine Learning, 2004, 54(1): 45 – 66.

[3] Meilin Zhu, Yue Wang, et al. Sphere-structured support vector machines sphere-structured support vector machines [A]. G Wang, et al. (eds.) RSFDGrC 2003[M]. LNAI 2639, Berlin/Heidelberg: Springer-Verlag, 2003. 589 – 593.

[4] Wu Q, et al. Classifying the multiplicity of the EEG source models using sphere-shaped support vector machines[J]. IEEE Transactions on Magnetics, 2005, 41(5): 1912 – 1915.

[5] Jigang Wang, Predrag Neskovic, Leon N Cooper. Bayes classification based on minimum bounding spheres[J]. Neurocomputing, 2007, 70(4 – 6): 801 – 808.

[6] D Lee, J Lee. Domain described support vector classifier for multi-classification problems[J]. Pattern Recognition, 2007, 40 (1): 41 – 51.

[7] 梁锦锦，刘三阳，吴德. 空间支持向量域分类器[J]. 西安电子科技大学学报(自然科学版)，2008, 35(6): 1080 – 1084.
LIANG Jin-jin, LIU San-yang, WU De. Space support vector domain classifier [J] Journal of Xidian University, 2008, 35 (6): 1080 – 1084. (in Chinese)

[8] Lei Gu, Hui-zhong Wu. Applying a novel decision rule to the sphere-structured support vector machines algorithm[J]. Neural Computing & Applications, 2009, 18(3): 275 – 282.

[9] Wang J, Neskovic P, Cooper LN. Pattern classification via single spheres[A]. A Hoffmann, H Motoda, T Soheffer (eds). Proceeding of Ds 2005 Conference. LNAI 3735 [C]. Berlin/Heidelberg: Springer-Verlag, 2005. 241 – 252.

[10] Mingrui Wu, Jieping Ye. A small sphere and large margin approach for novelty detection using training data with outliers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088 – 2092.

[11] Pei-Yi Hao, Jung-Hsien Chiang, Yen-Hsiu Lin. A new maximal-margin spherical-structured multi-class support vector machine[J]. Applied Intelligence, 2009, 30(2): 98 – 111.

[12] Pei-Yi Hao, Yen-Hsiu Lin. A new multi-class support vector machine with multi-sphere in the feature space [A]. H G Okuno, M Ali (eds). Proceeding of IEA/AIE 2007 Conference. LNAI 4570 [C]. Berlin/Heidelberg: Springer-Verlag, 2007. 756 – 765.

[13] 朱孝开，杨德贵. 基于推广能力测度的多类 SVDD 模式识别方法[J]. 电子学报，2009, 37(3): 464 – 469.
ZHU Xiao-kai, YANG De-gu. Multi-class support vector domain description for pattern recognition based on a measure of expansibility[J]. Acta Electronica Sinica, 2009, 37(3): 464 – 469. (in Chinese)

[14] Tingting Mu, Asoke K Nandi. Multiclass classification based on extended support vector data description[J]. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 2009, 39(5): 1206 – 1216.

作者简介

方景龙 男，研究员，1964 年 1 月出生于江西景德镇. 1985 年和 1988 年在华中科技大学获理学学士和硕士学位. 现在杭州电子科技大学图形图像研究所工作，主要从事机器学习、信息处理和目标探测等方面的研究工作.
E-mail: fjl@hdu.edu.cn

王万良 男，教授，博士生导师，1957 年 6 月出生于江苏高邮. 2001 年在同济大学获控制理论与控制工程专业博士学位. 现为浙江工业大学计算机学院院长，主要从事计算机控制与智能自动化，计算机网络控制系统，智能控制与智能管理等方面的研究工作.
E-mail: wwl@zjut.edu.cn