自适应上界的相对最大分离比单球面分类器

张 伟 柳先辉
(同济大学电子与信息工程学院 上海 201804)

摘 要 单球面分类器(RSS)以最大分离比为目标,对负类样本的分布情况缺乏考虑。根据Fisher判别准则,将相对间隔的思想引入到单球面分类器中,对特征空间中负类样本的分布上界进行约束来增强其内聚度,以提高分类器判别的准确性。由于分布上界的不可预测,为避免问题不可解,建立了自适应上界的最大相对分离比单球面分类器模型(ARRSS),并对模型参数进行了分析。实验证明,与单球面分类器相比,该方法表现出更好的泛化能力。

关键词 单球面分类器,Fisher判别,相对间隔,上界约束,自适应上界
中图法分类号 TP181 **文献标识码** A

Maximum Relative Separation Ratio Single Spherical Classifier with an Adaptive Upper Bound

ZHANG Wei LIU Xian-hui
(College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract Without taking the spread of negative class samples into account, the objective of single spherical classifier (RSS) is only to maximize the separation ratio. According to the Fisher discriminant analysis, this paper introduced relative margin into RSS to enhance the cohesion of negative class samples and improve the discriminant accuracy by the upper bound constraint in the feature space. Because the upper bound is unpredictable, a maximum relative separation ratio single spherical classifier with an adaptive upper bound (ARRSS) was built to avoid no solution and its parameters were researched afterwards. Experiments show the proposed method achieves better generalization performance compared with RSS.

Keywords RSS, Fisher discriminant analysis, Relative margin, Upper bound constraint, Adaptive upper bound

1 引言

支持向量机(Support Vector Machine, SVM)[1]采用最大间隔超平面对样本进行分类。SVM中超球面思想首次出现在1995年,Vapnik等通过计算最小闭包球的半径和最大间隔对SVM的风险上界进行估计[2]。受Vapnik启发,Tax等将SVM应用于一类问题,在特征空间中构造了一个包含所有样本的最小闭包超球,建立了支持向量域描述方法(Support Vector Domain Description, SVDD)[3]。与超平面SVM一样,SVDD方法建立的超球面边界仅取决于部分支持向量,保持了解的稀疏性。Schölkopf等证明了SVDD方法与采用高斯核函数的超平面一类SVM等价,为超球面SVM提供了理论依据[4]。随后,一类超球面SVM得到了广泛的研究和应用[5-7]。

与超平面SVM相比,超球面SVM的优势主要体现在两方面:1)采用了空间划分不对等的超球面边界:超球面划分的内部空间为一个有限的闭包球,外部空间保持了开放性,这种不平衡的空间划分方法尤其适用于不平衡样本集;2)便于建立多分类模型:针对多分类问题,只需通过SVDD方法对各类样本建立最小闭包球,根据测试样本到各类球心的距离来判断其类别,极大地降低了多类分类问题的计算复杂度。

但是,超球面SVM起源于一类问题,缺乏对类间间隔大小的考虑。类间间隔在SVM中发挥着至关重要的作用,它的大小反应了VC维度和模型的泛化能力。为了加入类间间隔的思想,Hao等提出了最大分离比超球面模型(RSS)[8],其使用最大分离比超球面取代最小闭包球,将类间间隔引入到球面SVM中。试验证明,在选取合理参数的情况下,RSS可以达到与SVM相当的性能。Wang等从最大球面间隔的几何模型出发得到了与RSS一致的范式,并将其应用于多类分类问题[9],取得了良好的试验结果。为了保证超球面间间隔最大,文传军等采用双球面实现了最大间隔、最小体积单球面分类器[10]。

单球面分类器与SVM一样,以类间最大间隔为目标,忽略了正负类内样本分布情况。根据Fisher判别准则,样本的类内方差大小对分类器的性能产生重大影响,最大间隔并不能保证SVM达到最佳的分类效果[11]。为了解决最大间隔超平面SVM中的样本分布问题,Pannagadata提出了相对最大间隔的方法(RMM)[12],其通过预先设定的阈值对特征空间

到稿日期:2011-11-10 返修日期:2012-02-29 本文受国家高新技术研究发展计划(863)项目(2009AA043503),国家科技支撑计划项目(2012BAF10B05)资助。

张 伟(1985-),男,博士生,主要研究方向为机器学习,E-mail:zhangwei036@126.com;柳先辉(1979-),男,博士,讲师,主要研究方向为数据挖掘、软件工程。

中样本分布的上界进行限制。将相对间隔方法应用于结构化数据预测[13]，取得了较好的结果。RSS 方法虽然将正类样本分布限制在闭包球内，但对负类样本的分布情况同样没有考虑。本文将相对间隔引入到最大分离比超球面模型 RSS 中，以达到增强样本内聚性、提高分类性能的目的。与超平面 SVM 不同，RSS 中球面半径及间隔大小未知，阈值大小的选取缺乏参考依据。为了避免错误的阈值导致问题无解，将阈值加入到目标函数中，建立自适应上界的相对最大分离比单球面分类器模型(ARRSS)，并对模型进行了分析和验证。

本文首先介绍最大间隔比单球面分类器 RSS；然后在 RSS 中引入相对间隔的思想，建立 ARRSS 模型并对模型参数进行讨论；最后通过试验对 ARRSS 和 RSS 进行比较，验证了自适应上界方法在单球面 SVM 分类器中的有效性。

2 最大间隔比单球面分类器 RSS

假定训练样本集 $S=\{(x_i,y_i)_{i=1}^n\}$，其中样本特征 $x_i\in R^m$，样本类别 $y_i\in\{-1,+1\}$。如图 1 所示，“+”表示正类样本，“−”表示负类样本，“○”表示支持向量。

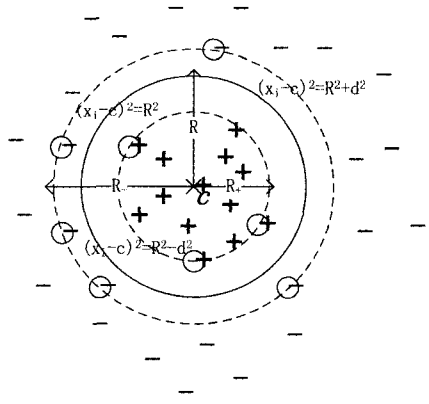

图 1 最大分离比超球面

超球面 $S(c,R)$将正、负样本分开，正类样本位于超球面 $S(c,R_+)$内部，负类样本位于超球面 $S(c,R_-)$外部。

1)当 $y_i=-1$ 时，$(x_i-c)'(x_i-c)\geqslant R^2+d^2$；

2)当 $y_i=+1$ 时，$(x_i-c)'(x_i-c)\leqslant R^2-d^2$。

即满足：$y_i(R^2-(x_i-O)'(x_i-O))\geqslant d^2$。

超球面间距 $D=R_--R_+=\sqrt{R^2+d^2}-\sqrt{R^2-d^2}$。

与 SVM 的最大间隔超平面不同，最大分离比单球面分类器 RSS 的目标函数不是最大化超球面间距 D，而是最大化超球面间距与超球半径的比值 $r=(R_--R_+)/R_+$。

由于 $R_+>R_->0$，$R^2-d^2>0$，且

$\max(R_--R_+)/R_+\Leftrightarrow\max R_-/R_+\Leftrightarrow\max R_-^2/R_+^2\Leftrightarrow\max(R^2+d^2)/(R^2-d^2)\Leftrightarrow\max d^2/(R^2-d^2)\Leftrightarrow\min(R^2-d^2)/d^2\Leftrightarrow\min R^2/d^2$

即最大化间隔比要求同时最小化 R、最大化 d，因此，原问题可以表示为：

$$\begin{aligned}&\min R^2-Md^2\\&\text{s. t. } y_i(R^2-(x_i-c)'(x_i-c))\geqslant d^2\end{aligned}\tag{1}$$

式中，惩罚因子 $M>0$。上式对应的拉格朗日函数为：

$$L_p(R,d,c)=R^2-Md^2-\sum_{i=1}^n\alpha_i[y_i(R^2-(x_i-c)'(x_i-c))-d^2]$$

式中，拉格朗日乘子 $\alpha_i\geqslant0$。对 L_p 中自变量 R,d,c 分别求偏导并置 0，得：

$$\frac{\partial L_p}{\partial R}=0\Rightarrow\sum_{i=1}^n\alpha_iy_i=1$$

$$\frac{\partial L_p}{\partial d}=0\Rightarrow\sum_{i=1}^n\alpha_i=M$$

$$\frac{\partial L_p}{\partial c}=0\Rightarrow\sum_{i=1}^n\alpha_iy_ix_i=c$$

将其代入 L_p 得到对偶式：

$$L_d=\frac{1}{2}\sum_{i,j=1}^n\alpha_i\alpha_jy_iy_jx_i'x_j-\frac{1}{2}\sum_{i,j=1}^n\alpha_iy_ix_i'x_i$$

因此，原问题的对偶问题为：

$$\min\frac{1}{2}\sum_{i,j=1}^n\alpha_i\alpha_jy_iy_jx_i'x_j-\frac{1}{2}\sum_{i=1}^n\alpha_iy_ix_i'x_i$$

$$\text{s. t. } \alpha_i\geqslant0,\sum_{i=1}^n\alpha_i=M,\sum_{i=1}^n\alpha_iy_i=1$$

这是一个凸二次规划问题，具有全局最优解 α^*。可以求出样本(x_k,y_k)的球心距平方为：

$$R_k^2=x_k'x_k-2\sum_{j=1}^n\alpha_j^*y_jx_j'x_k+\sum_{i,j=1}^n\alpha_i^*\alpha_j^*y_iy_jx_i'x_i$$

根据 KKT 互补条件，最优解满足：

$\alpha_i^*[y_i(R^2-(x_i-c^*)'(x_i-c^*))-d^2]=0$。

1)当 $y_i=+1,\alpha_i>0$ 时，$R_+^2=R_i^2$，即样本(x_k,y_k)处在超球面 $S(c,R_+)$上；

2)当 $y_i=-1,\alpha_i>0$ 时，$R_-^2=R_i^2$，即样本(x_k,y_k)处在超球面 $S(c,R_-)$上。

得到决策函数：

$$f(x_k)=\operatorname{sgn}(R^2-R_k^2)=\operatorname{sgn}\left(\frac{R_+^2+R_-^2}{2}-R_k^2\right)$$

3 自适应上界相对最大间隔比单球面分类器

根据 Fisher 判别准则，分类性能取决于类间距离及类内方差。SVM 以最大类间间隔为目标，缺乏对正负样本分布情况的考虑。文献[8]将球样本到超平面的距离上界 B 添加到约束条件中，用以限制样本在特征空间中的分布，然后在这种约束下求出最大间隔超平面，即相对最大间隔 SVM 分类器(RMM)。

3.1 相对最大分离比单球面分类器(RRSS)

球面分类器的本质是通过特征空间中样本与球心的距离来判断样本的类别，所以负类样本球心距离的分布情况直接影响了分类器的性能。因此，本文将特征空间中负类样本的球心距上界 B 添加到约束中，用于限制负类样本在特征空间中的分布。从图 1 中可以看出，正类样本球心距小于负类，所以带分布上界约束的 RSS 可以表示为：

$$\begin{aligned}&\min R^2-Md^2+C_1\sum_{i=1}^n\xi_i\\&\text{s. t. } y_i(R^2-(\phi(x_i)-c)'(\phi(x_i)-c))\geqslant d^2-\xi_i\\&\quad(\phi(x_i)-c)'(\phi(x_i)-c)\leqslant B^2\\&\quad\xi_i\geqslant0\end{aligned}\tag{2}$$

式中，C_1 是训练误差的惩罚因子，ξ_i 为松弛因子，$\phi(x_i)$表示从样本空间到特征空间的映射函数。若球心距上界 $B^2\geqslant\{(\phi(x_i)-c)'(\phi(x_i)-c)\}$，则负类样本边界约束无效，问题(2)的解与问题(1)等价；随着 B 的减小，问题(2)与问题(1)的解的差异逐渐增大。当 $B^2<R^2+d^2$ 时，问题(2)无解。与 RMM 不同，由于无法提前获知 R 与 d 的大小，B 的最小值无法确定，可能引发错误 B 值带来的无解问题。为了避免无解，令 B

$=R^2+td^2$，代入问题(2)得：

$$\begin{aligned}&\min R^2-Md^2+C_1\sum_{i=1}^{n}\xi_i\\&\text{s.t. } y_i(R^2-(\phi(x_i)-c)'(\phi(x_i)-c))\geqslant d^2-\xi_i\\&\quad(\phi(x_i)-c)'(\phi(x_i)-c)\leqslant R^2+td^2\\&\quad\xi_i\geqslant 0\end{aligned}\tag{3}$$

式中，$t>1$，保证了必然可解。t 的大小反映了对样本分布约束的紧密程度，t 越大对负类样本分布的约束越弱，问题(3)的解越接近于式(1)。问题(3)对应的拉格朗日函数为：

$$\begin{aligned}L_p=&R^2-Md^2+C_1\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(R^2-(\phi(x_i)-c)'(\phi(x_i)-c))-d^2+\xi_i]-\\&\sum_{i=1}^{n}\beta_i[(R^2+td^2)-(\phi(x_i)-c)'(\phi(x_i)-c)]-\sum_{i=1}^{n}\lambda_i\xi_i\end{aligned}$$

式中，拉格朗日乘子 $\alpha_i,\beta_i\geqslant 0$。对 L_p 求偏导并置 0，得到：

$$\frac{\partial L_p}{\partial R}=0\Rightarrow\sum_{i=1}^{n}(\alpha_i y_i+\beta_i)=1$$

$$\frac{\partial L_p}{\partial d}=0\Rightarrow\sum_{i=1}^{n}\alpha_i-t\sum_{i=1}^{n}\beta_i=M$$

$$\frac{\partial L_p}{\partial c}=0\Rightarrow c=\sum_{i=1}^{n}(\alpha_i y_i+\beta)\phi(x_i)$$

$$\frac{\partial L_p}{\partial d}=0\Rightarrow\sum_{i=1}^{n}\alpha_i-t\sum_{i=1}^{n}\beta_i=M$$

将其代入 L_p 得到对偶式：

$$L_d=\sum_{i=1}^{n}(\alpha_i y_i+\beta_i)\phi(x_i)'\phi(x_i)-\sum_{i,j=1}^{n}(\alpha_i y_i+\beta_i)(\alpha_j y_j+\beta_j)\phi(x_i)'\phi(x_j)$$

所以式(3)的对偶问题可以表示为：

$$\begin{aligned}&\max\sum_{i=1}^{n}(\alpha_i y_i+\beta_i)\phi(x_i)'\phi(x_i)-\sum_{i,j=1}^{n}(\alpha_i y_i+\beta_i)(\alpha_j y_j+\beta_j)\phi(x_i)'\phi(x_j)\\&\text{s.t. } C_1\geqslant\alpha_i\geqslant 0\\&\quad\sum_{i=1}^{n}\alpha_i-t\sum_{i=1}^{n}\beta_i=M,\sum_{i=1}^{n}(\alpha_i y_i+\beta_i)=1\end{aligned}$$

3.2 自适应上界 RRSS(ARRSS)

根据 Fisher 判别准则，样本内部方差的大小是决定分类器性能的关键因素之一。因此将边界约束 B 加入到目标函数之中，通过最小化 B 来限制负类样本的分布，以提高分类器的性能。用权值 N 自适应调节分布上界 B 的大小，来避免无解情况。问题可以表示为：

$$\begin{aligned}&\min R^2-Md^2+NB^2+C_1\sum_{i=1}^{n}\xi_i+C_2\sum_{i=1}^{n}\delta_i\\&\text{s.t. } y_i(R^2-(\phi(x_i)-c)'(\phi(x_i)-c))\geqslant d^2-\xi_i\\&\quad(\phi(x_i)-c)'(\phi(x_i)-c)\leqslant B^2+\delta_i\\&\quad\xi_i\geqslant 0,\delta_i\geqslant 0\end{aligned}\tag{4}$$

其拉格朗日函数为：

$$\begin{aligned}L_p=&R^2-Md^2+NB^2+C_1\sum_{i=1}^{n}\xi_i+C_2\sum_{i=1}^{n}\delta_i-\sum_{i=1}^{n}\alpha_i[y_i(R^2-(\phi(x_i)-c)'(\phi(x_i)-c))-d^2+\xi_i]-\sum_{i=1}^{n}\beta_i[B^2+\delta_i-\\&(\phi(x_i)-c)'(\phi(x_i)-c)]-\sum_{i=1}^{n}\lambda_i\xi_i-\sum_{i=1}^{n}\gamma_i\delta_i\end{aligned}$$

对 L_p 的自变量分别求偏导并置 0，得：

$$\frac{\partial L_p}{\partial d}=0\Rightarrow\sum_{i=1}^{n}\alpha_i=M\tag{5}$$

$$\frac{\partial L_p}{\partial B}=0\Rightarrow\sum_{i=1}^{n}\beta_i=N\tag{6}$$

$$\frac{\partial L_p}{\partial c}=0\Rightarrow c=\frac{1}{1+N}\sum_{i=1}^{n}(\alpha_i y_i+\beta_i)\phi(x_i)\tag{7}$$

$$\frac{\partial L_p}{\partial\xi_i}=C_1-\alpha_i-\lambda_i=0\Rightarrow 0\leqslant\alpha_i\leqslant C_1\tag{8}$$

$$\frac{\partial L_p}{\partial\delta_i}=C_2-\beta_i-\lambda_i=0\Rightarrow 0\leqslant\beta_i\leqslant C_2\tag{9}$$

将式(20)－式(24)代入 L_p 得到其对偶式：

$$L_d=\sum_{i=1}^{n}(\alpha_i y_i+\beta_i)K(x_i,x_i)-\frac{1}{1+N}\sum_{i,j=1}^{n}(\alpha_i y_i+\beta_i)(\alpha_j y_j+\beta_j)K(x_i,x_j)$$

式中，核函数 $K(x_i,x_j)=\phi(x_i)'\phi(x_j)$。则原问题的对偶问题可以表示为：

$$\begin{aligned}&\min\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i y_i+\beta_i)(\alpha_j y_j+\beta_j)K(x_i,x_j)-\frac{1+N}{2}\sum_{i=1}^{n}(\alpha_i y_i+\beta_i)K(x_i,x_i)\\&\text{s.t. } \sum_{i=1}^{n}\alpha_i=M,\sum_{i=1}^{n}\alpha_i y_i=1,C_1\geqslant\alpha_i\geqslant 0\\&\quad\sum_{i=1}^{n}\beta_i=N,C_2\geqslant\beta_i\geqslant 0\end{aligned}\tag{10}$$

这是一个具有 $2n$ 个变量的凸二次规划问题，可以通过多种方法计算全局最优解 α 和 β，从而得到每个样本对应的参数 $\alpha_j y_j+\beta_j$。

由式(7)得样本 (x_k,y_k) 的球心距平方：

$$R_i^2=K(x_i,x_i)-\frac{2}{1+N}\sum_{j=1}^{n}(\alpha_j y_j+\beta_j)K(x_i,x_j)+\frac{1}{(1+N)^2}\sum_{i,j=1}^{n}(\alpha_i y_i+\beta_i)(\alpha_j y_j+\beta_j)K(x_i,x_j)$$

根据 KKT 互补条件，最优解满足：

$\alpha_i[y_i(R^2-(x_i-c)'(x_i-c))-d^2+\xi_i]=0$

即 $\alpha_i[y_i(R^2-R_i^2)-d^2+\xi_i]=0$。

1)若 $y_i=+1,\alpha_i>0$，则 $R_+^2=R_i^2-\xi_i$。由于 ξ_i 未知，因此不可以由此式直接求解。但此时样本 (x_k,y_k) 处在超球面 $S(c,R_+)$ 上或者外部，$R_+^2=\min\{R_i^2\mid y_i=+1,\alpha_i>0\}$。

2)若 $y_i=-1,\alpha_i>0$，则 $R_-^2=R_i^2+\xi_i$。同理，不可以由此式直接求解。此时样本 (x_k,y_k) 处在超球面 $S(c,R\ )$ 上或者内部，$R_-^2=\max\{R_i^2\mid y_i=-1,\alpha_i>0\}$。

得到决策函数 $f(x_k)=\operatorname{sgn}\left(\frac{R_+^2+R_-^2}{2}-R_k^2\right)$。

3.3 参数的选择范围

定理 1 若训练样本集 $S=\{(x_i,y_i)_{i=1}^{n}\}$ 中正样本数为 $n^+>0$，则参数 M、N、C_1 的取值范围必须满足：

1)$M\in[1,2n^+C_1-1]$

2)$N\in[0,nC_2]$

3)$C_1\geqslant 1/n^+$

证明：假设正类样本集 S^+ 对应的拉格朗日乘子为 $\{\alpha_1^+,\alpha_2^+,\cdots,\alpha_{n^+}^+\}$，负类样本集 S^- 对应的拉格朗日乘子为 $\{\alpha_1^-,\alpha_2^-,\cdots,\alpha_{n^-}^-\}$，负类样本容量 $n^-=n-n^+$。根据式(5)、式(6)有：

$$\sum_{i=1}^{n^+}\alpha_i^+-\sum_{i=1}^{n^-}\alpha_i^-=\sum_{i=1}^{n}\alpha_i=M\tag{11}$$

$$\sum_{i=1}^{n^+}\alpha_i^+-\sum_{i=1}^{n^-}\alpha_i^-=\sum_{i=1}^{n}\alpha_i y_i=1\tag{12}$$

1)由式(6)、式(9)有 $0\leqslant N\leqslant nC_2$，即 $N\in[0,nC_2]$。

2)由式(11)、式(12)得 $M=2n^+\sum_{i=1}^{n^+}\alpha_i^+-1$；由于 $\alpha_i^+\leqslant C_1$，因此 $M\leqslant 2n^+C_1-1$。

因为 $\alpha_i\geqslant 0$，所以 $\sum_{i=1}^{n^+}\alpha_i^+-\sum_{i=1}^{n^-}\alpha_i^-\geqslant\sum_{i=1}^{n^+}\alpha_i^+-\sum_{i=1}^{n^-}\alpha_i^-$，即 $M\geqslant 1$。

综上所得，$M\in[1,2n^+C_1-1]$。

3)由于 $1\leqslant M\leqslant 2C_1-1$，因此 $2n^+C_1-1\geqslant 1$，即 $C_1\geqslant 1/n^+$。证毕。

从定理 1 可以看出，M 的取值范围依赖于训练样本集中正样本的数量及 C_1 的大小，而 C_1 的取值必须大于正类样本数目的倒数；N 的取值范围依赖于样本的容量及 C_2 的大小。选取不同的参数组合对模型的性能会产生重大影响，现实中往往通过交叉验证和网格搜索的方法选取最佳参数组合。定理 1 为限定不同训练样本集的参数搜索范围提供了依据。

4 实验对比与分析

虽然求解自适应最大相对间隔球面是一个二次约束的二次规划问题，但其对偶问题(10)是一个凸二次规划问题，存在多种方法快速求解。本文试验采用 C# 建立 CPlex 凸二次规划模型对试验样本进行训练。首先采用不同的参数进行重复试验，对比试验的结果用来分析参数对 ARRSS 性能的影响。随后，针对多组 UCI 数据集分别采用最大分离比单球面分类器 RSS 和 ARRSS 分类器进行训练，以验证方法的有效性。

4.1 参数的影响

为了直观展示不同参数对训练结果的影响，采用简单的二维数据集进行了训练，本文已经证明，M、N 的取值范围与 C_1、C_2 大小有关。为了便于讨论，试验中将 C_1、C_2 设置为正无穷大以取消对 M、N 最大值的限制。此时，分类面为硬间隔球面，所有正类训练样本都在分类面的内侧，负类训练样本都在分类面的外侧。

图 2 展示了在不同参数情况下的训练结果。正样本用“+”表示，负样本用“−”表示，实线圆(红色)表示 $S(c,R)$ 分类面，内外的虚线圆(蓝色)分别表示 $S(c,R_+)$ 和 $S(c,R_-)$。第一行表示在 $N=0.2$ 的情况下分别取选取不同的 M 值得到的不同球面。可以看出，随着 M 不断增大，d 也不断增大，分类面越来越接近于 SVM。M 值从 1 到 5，球面间隔增大相对较多；但 M 从 5 到 125，球间距 D 的改变并不明显，甚至出现了间隔比下降的情况。第二行表示在 $M=25$ 时 N 值的改变对分类面造成的影响。当 $N=0$ 时，对负类样本的分布上界没有约束，等价于 RSS；随着 N 的增大，分界面半径不断减小，但减小幅度逐渐减慢。这是因为 N 限制了负类样本的分布，随着 N 的增大，样本分布的空间逐步减小，样本分布愈加紧密。

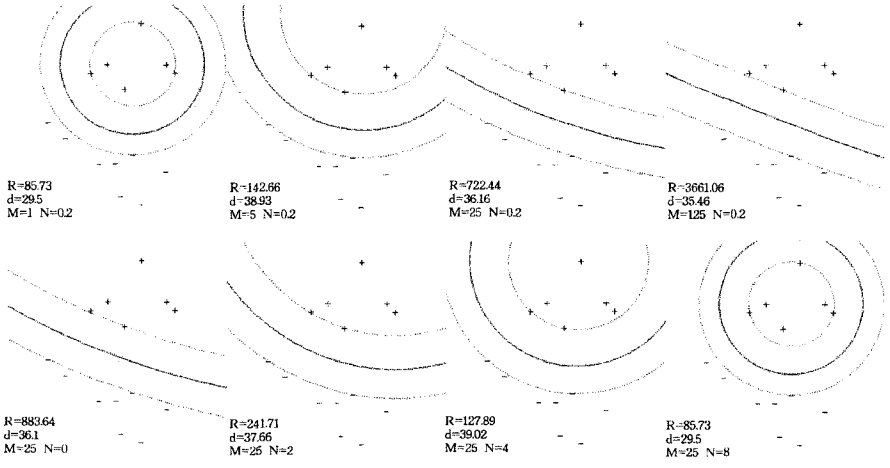

图 2 二维样本在不同参数下的硬间隔超球面

图 3 反应了参数对分类性能的影响。试验采用 UCI 的 IRIS 样本集，将 setosa 当作正类，versicolor 和 virginica 为负类。IRIS 数据集共 150 个样本，随机选取 50 个样本作为训练样本，包括 15 个正样本和 35 个负样本，另外 100 个样本为测试样本。选取径向基核函数，设置 $\delta=0.5$，$C_1=10$，$C_2=10$。

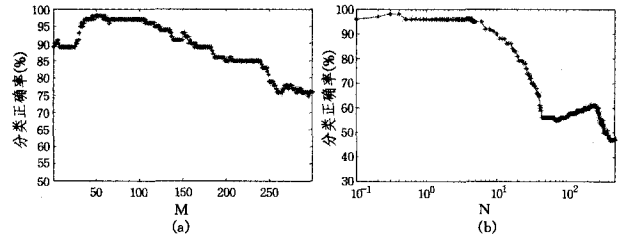

图 3 不同参数下的分类正确率的变化

根据定理 1 可知，$1\leqslant M\leqslant 299$，$0\leqslant N\leqslant 500$。图 3(a)反映了 $N=0.5$ 时不同 M 值的分类准确率。随着 M 的增大，分类准确率不断提高，当 M 在 40 到 130 之间时，准确率维持在 95%以上，随后不断下降。图 3(b)为 $M=50$ 时不同 N 值下的分类准确性。当 $N<10$ 时，分类准确性稳定，最高达到了 98%；之后随着 N 的增加，分类准确性急剧下降。可见，M、N 的取值对分类器的性能产生了重大的影响，过大或者过小的值都会降低分类器的性能。尤其当 N 值过大时，由于极度压缩了样本空间，导致分类性能急剧降低。

4.2 性能比较

为了验证自适应上界约束方法的有效性，针对 6 组不同的 UCI 数据集分别采用 RSS 和 ARRSS 分类器进行训练，比较两类分类器的分类正确率。试验采用径向基核函数 $\delta=0.5$，$C_1=10$，$C_2=10$，通过 5 重交叉验证的方式验证模型的泛化能力，搜索最佳参数组合。试验结果如表 1 所列。

表 1 RSS 与 ARRSS 分类正确率的比较

样本集	RSS	ARRSS
BREAST CANCER	94.6%	**97.3%**
IRIS	97.0%	**98.0%**
PIMA	96.8%	**97.3%**
VERTEBRAL	82.6%	82.6%
YEAST	84.9%	**87.2%**

从表 1 中可以看出，RSS 和 ARRSS 都表现出了良好的学习性能。与 RSS 相比，ARRSS 在某些数据集上表现出了更好的分类正确率，这与 RSS 和 ARRSS 的数学模型相符。

(下转第 214 页)

类发音过程的差异，较好地实现了塞擦音和摩擦音的分类，提高了低信噪比下的分类性能。文中较多错误是由塞擦音/zh/、/ch/与摩擦音/sh/引起的，因此后续的研究可以根据其他特征参数来提高塞擦音/zh/、/ch/与摩擦音/sh/间的分类准确率，进一步提高塞擦音与摩擦音的分类准确率。

参 考 文 献

[1] Lee Chin-Hui. From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition[C]// Proceedings Of ICSLP Keynote Speech. 2004

[2] Geiger J T, Lakhal M A, Schuller B, et al. Learning new acoustic events in an HMM-based system using MAP adaptation[C]// Proceedings of INTERSPEECH. 2011: 293-296

[3] David M-N, Ascensión G-A, Carmen P-M. Feature Extraction Assessment for an Acoustic-Event ClassificationTask Using the Entropy Triangle[C]// Proceedings of INTERSPEECH. 2011: 309-312

[4] 张宝奇，张连海，屈丹. 基于听觉事件检测的汉语语音声韵切分[J]. 声学学报，2010，35(6)：701-707

[5] Almpanidis G, Kotropoulos K M. Robust Detection of Phone Boundaries Using Model Selection Criteria With Few Observations[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 17(2): 287-298

[6] 陈斌，张连海，王波. 基于 Seneff 听觉谱特征的汉语连续语音声韵母边界检测[J]. 声学学报，2012，37(1)：104-112

[7] Forrest K, Weisme G. Statistical analysis of word-initial voiceless obstruents: Preliminary data[J]. Acoustical Society of America, 1996, 84(1): 115-123

[8] Jongman A, Wayland R, Wong S. Acoustic characteristics of English fricatives[J]. Journal of the Acoustical Society of America, 2000, 108: 1252-1263

[9] Sussman H M, Bessell N, Dalston E, et al. An investigation of stop place of articulation as a function of syllable position: A locus equation perspective[J]. Journal of the Acoustical Society of America, 1998, 101: 2826-2838

[10] Kluender K R, Walsh M A. Amplitude rise time and the perception of the voiceless affricate/fricative distinction[J]. Perception and Psychophysics, 2002, 51: 328-333

[11] Hu Guo-ning, Wang De-liang. Separation of fricatives and affricates[C]// Proceedings of ICASSP. 2005: 1001-1004

[12] Seneff S. A joint synchrony/mean-rate model of auditory speech processing[J]. Journal of Phonetics, 1988, 16: 55-76

[13] Seneff S. Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model[M]. Cambridge, Massachusetts Institute of Technology, 1985

[14] Ahmed M, Abdelatty A, Jan Van der S, et al. Robust Auditory-Based Speech Processing Using the Average Localized Synchrony Detection[J]. IEEE Transaction on Signal and Audio Processing, 2001, 10: 279-292

[15] AAhmed M. Abdelatty A, Jan Van der S. Paul MuellerAcoustic-Phonetic Features for the Automatic Classification of Stop Consonants[J]. IEEE Transaction on Signal and Audio Processing, 2001, 9(8): 833-841

[16] Young S. The HTK Book(for HTK Version 3.4) [M]. Cambridge University Engineering Department, 2006: 289

(上接第 191 页)

使用一个合理的 N 值限制负类样本分布，可以增强负类样本的内聚度，提高 RSS 的分类准确性。当 $N=0$ 时，负类样本分布上界约束无效，ARRSS 与 RSS 等价。

结束语 相对间隔被用于限制 SVM 中样本的分布上界，能有效克服特征空间内样本分布不合理引起的性能下降问题。本文将相对间隔加入到单球面分类器 RSS 中，建立了自适应上界约束的最大相对间隔比单球面分类器 ARRSS，通过限制负类样本分布上界的方式提高了分类性能。ARRSS 本质上是一个凸二次规划问题，是对 RSS 的一个拓展，当 $N=0$ 时，ARRSS 等价于 RSS。因此，ARRSS 表现出了比 RSS 更好的性能。ARRSS 使用单球面进行分类，只适用于单模态分布样本集。因此，对于机器学习中普遍存在的多模态分布问题，需要通过构造一系列单球面来进行逼近求解。

参 考 文 献

[1] Cortes C, Vapnik V. Support-vector network[J]. Machine Learning, 1995, 20(3): 273-297

[2] Vapnik V. The nature of statistical learning theory [M]. New York: Springer, 1995

[3] Tax M J, Duin P W. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20: 1191-1199

[4] Schölkopf B, Platt J C, Taylor J S, et al. Estimating the support of a high-dimensional distribution [J]. Neural Computation, 2001, 13: 1443-1471

[5] Larry M M, Malik Y. One-class SVMs for document classification [J]. Journal of Machine Learning Research, 2001, 2: 139-154

[6] Chen Y Q, Zhou X S, Huang T S. One-class SVM for learning in image retrieval[C]// Proceedings of the 2001 IEEE International Conference on Image Processing. Greece: Thessaloniki, 2001, 1: 34-37

[7] Zhu M L, Chen S F, Liu X D. Sphere-structured support vector machines for multi-class pattern recognition [J]. Lecture Notes in Computer Science, 2003, 2639: 589-593

[8] Wang J G, Neskovic P, Cooper L N. Pattern classification via single spheres [J]. Lecture Notes in Computer Science, 2005, 3735: 241-252

[9] Hao P Y, Chiang J H, Lin Y H. A new maximal margin spherical structured multiclass support vetor machine [J]. Applied Intelligence, 2009, 30(2): 98-111

[10] 文传军，詹永照，陈长军. 最大间隔最小体积球形支持向量机[J]. 控制与决策，2010，25(1)：79-83

[11] Shivaswamy P K, Jebara T. Relative margin machines [C]// Advances in Neural Information Processing System. 2008, 21: 1481-1488

[12] Shivaswamy P K, Jebara T. Maximum relative margin and data-dependent regularization [J]. Journal of Machine Learning Research, 2010, 11: 747-788

[13] Shivaswamy P K, Jebara T. Structured prediction with relative margin [C]// International Conference on Machine Learning and Applications. 2009: 281-287