隐私保护的 SVM 快速分类方法

胡文军[1,2],王士同[1]
(1.江南大学数字媒体学院,江苏无锡 214122; 2.湖州师范学院信息与工程学院,浙江湖州 313000)

摘 要: 许多核分类方法的决策函数可以表示为支持向量的组合,如 SVM,而支持向量含有非常重要的隐私信息,因此,在分类决策时可能会暴露此类信息,同时分类速度受限于支持向量的个数,如 SVM 的分类复杂度为 $O(|SVs|)$.为解决上述两个问题,本文基于最小包含球球心在原始空间中的代理原像,提出了一种隐藏支持向量信息并能快速实现分类的 SVM 方法,称为隐私保护的快速 SVM 分类方法(Fast Classification Approach of SVM with Privacy Preservation,FCA-SVM$_{WPP}$).同时提供了两种求解代理球心原像的方法,分别称为 QP 解法和直接解法.UCI 和 PIE 人脸数据集的实验结果表明,本文方法可解决上述两个问题并具有较好的效果.

关键词: 分类; 支持向量机; 快速分类; 最小包含球; 代理球心; 原像
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2012) 02-0280-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.02.012

Fast Classification Approach of Support Vector Machine with Privacy Preservation

HU Wen-jun[1,2], WANG Shi-tong[1]
(1. *School of Digital Media, Jiangnan University, Wuxi, Jiangsu* 214122, *China*;
2. *School of Information and Engineering, Huzhou Teachers College, Huzhou, Zhejiang* 313000, *China*)

Abstract: The decision functions of various kernelized classification methods can be expressed as a combination of Support Vectors (SVs), i.e. SVM, which contain the individual privacy information, so this information will be released during detecting unknown samples. Meanwhile, the amount of SVs limits classification speed, i.e. the computational time complexity of SVM is $O(|SVs|)$. For overcoming the above drawbacks, a fast classification approach of SVM with privacy preservation is proposed, which is based on the agent preimage of the center of minimum enclosing ball (MEB), and two preimage-finding methods are presented in this paper, called QP-based solution and direct solution respectively. Experimental results on UCI and PIE face image demonstrate that two drawbacks as above can not only be solved, but also the obtained effectiveness of the proposed method is competitive.

Key words: classification; support vector machine; fast classification; minimum enclosing ball (MEB); the agent of sphere center; preimage

1 引言

分类是模式识别和机器学习中一个重要研究内容,广泛应用于现实生活中,如药物检测,门禁系统,医疗诊断等[1~3].近几十年,得到了广泛研究,多种分类方法如支持向量机(Support Vector Machine, SVM)[4,5],最小包含球(Minimum Enclosing Ball, MEB),支持向量数据描述(Support Vector Data Description, SVDD)[2],最近邻方法[6,7],核密度估计器[8,9]以及基于模糊理论的分类器[10]等被提出.像 SVM,MEB 和 SVDD 实质是利用支持向量获得最优分类界面,即其决策函数是通过支持向量扩展生成,如线性组合,因此在决策过程中支持向量起着决定性的作用.而支持向量含有此类样本的本质信息,决策过程可能会暴露此信息,这是因为窃取隐私信息的攻击者既可通过训练样本进行攻击,也可通过统计未知样本的决策输出攻击.在某些应用领域这是绝对不允许的,如药物检测,医疗数据处理,患者疾病信息,弹药成份信息等,因此从决策函数输出保护数据隐私也有重要意义.

本文主要针对二类 SVM,重点关注决策过程中的隐私保护和分类速度.因 $\boldsymbol{w}_\Phi^{\mathrm{T}}\phi(\boldsymbol{x})+b=0$ 是其分离超平面,故决策 1 个未知样本的复杂度是 $O(n)$(n 是训练样本数),显然 SVM 很难适用于高实时性的数据监测,特别是在线数据监测或离线处理大数据等.实际上,快速决策是分类中的一个重要研究课题,如约简集 SVM[11],支持向

收稿日期:2010-12-07;修回日期:2011-07-21
基金项目:国家自然科学基金(No.60903100,No.60975027,No.90820002);江苏省自然科学基金(No.BK2009067);江苏省普通高校研究生科研创新计划项目(No.CXZZ11_0483)

量回归法[12], Separable Case Approximation(SCA)[13], Dynamic Decay Adjustment SVM[14]和聚类 SVM[15]等方法,此类方法是利用逼近准则(如 ISE)减少支持向量个数使其复杂度降至 $O(|SVs|)$,若 $|SVs|$ 过大也不能适用于在线数据监测.最近,原像问题得到了广泛地关注[16~20],其实质是核空间到原始空间的逆映射问题.因此,本文利用 $\boldsymbol{w}_\Phi$ 的原像提出一种隐藏支持向量信息并能快速决策的 SVM 方法,称为隐私保护的快速 SVM 分类方法(Fast Classification Approach of SVM with Privacy Preservation, FCA-SVM$_{WPP}$).

2 SVM

定义 1　设原始样本空间为 $O\subseteq\boldsymbol{R}^d$,输入样本 $X=X^+\cup X^-=\{\boldsymbol{x}_i\mid\boldsymbol{x}_i\in O,1\leqslant i\leqslant n\}$,其中 X^+ 和 X^- 是正负类样本集,$\boldsymbol{x}_i$ 是列向量.$\boldsymbol{y}=(y_1,\cdots,y_n)^{\mathrm{T}}$是类标签,$y_i=+1(\boldsymbol{x}_i\in X^+)$;$y_i=-1(\boldsymbol{x}_i\in X^-)$.设 $\phi:\boldsymbol{x}\in O\to\phi(\boldsymbol{x})\in\Phi$,其中 Φ 是一个高维特征空间,且其内积由正定核 $k:\boldsymbol{R}^d\times\boldsymbol{R}^d\in\boldsymbol{R}$ 诱导,即 $\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

2.1 C-SVM 算法

设 $\boldsymbol{w}_\Phi^{\mathrm{T}}\phi(\boldsymbol{x})+b=0$ 是 SVM 在 Φ 中最优超平面,其可由现成软件包,如 SVM tool box, LibSVM[21]获得.本文沿用文献[5]将软间隔 SVM 算法称之为 C-SVC,其原始问题为:

$$\min_{\boldsymbol{w}_\Phi,b,\xi_i}\frac{1}{2}\|\boldsymbol{w}_\Phi\|^2+\frac{C}{n}\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{s.t.}\quad y_i(\boldsymbol{w}_\Phi^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,\xi_i\geqslant 0,1\leqslant i\leqslant n$$

其中,C 是惩罚因子,ξ_i 是松弛变量.其对偶形式为:

$$\max_{\alpha}2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{K}}\boldsymbol{\alpha} \tag{2}$$

$$\text{s.t.}\quad\sum_{i=1}^{n}\alpha_i y_i=0,0\leqslant\alpha_i\leqslant C/n,1\leqslant i\leqslant n$$

其中,$\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_n)^{\mathrm{T}}\geqslant 0$(指各元素$\geqslant 0$,下同)是拉格朗日乘子向量,$\mathbf{1}$是 n 维单位列向量,$\widetilde{\boldsymbol{K}}=[\tilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_j)]=[y_iy_j\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)]$是对应的核矩阵.因 $\alpha_i>0$ 所对应的样本称为支持向量(Support Vectors)[4,5],故有如下集合:未错分的正类支持向量 $SV_{uerr}^+=\{\boldsymbol{x}_i\mid 0<\alpha_i<C/n,\boldsymbol{x}_i\in X^+,1\leqslant i\leqslant n\}$和错分的正类支持向量 $SV_{err}^+=\{\boldsymbol{x}_i\mid C/n\leqslant\alpha_i,\boldsymbol{x}_i\in X^+,1\leqslant i\leqslant n\}$;同理有 $SV_{uerr}^-=\{\boldsymbol{x}_i\mid 0<\alpha_i<C/n,\boldsymbol{x}_i\in X^-,1\leqslant i\leqslant n\}$和 $SV_{err}^-=\{\boldsymbol{x}_i\mid C/n\leqslant\alpha_i,\boldsymbol{x}_i\in X^-,1\leqslant i\leqslant n\}$;以及正类,负类和所有支持向量 $SVs^+=SV_{uerr}^+\cup SV_{err}^+$, $SVs^-=SV_{uerr}^-\cup SV_{err}^-$和 $SVs=SVs^+\cup SVs^-$.根据对偶问题的推导过程[4]和 KKT 条件,有

$$\boldsymbol{w}_\Phi=\sum_{i=1}^{n}\alpha_i y_i\phi(\boldsymbol{x}_i)=\sum_{\boldsymbol{x}_i\in SV}\alpha_i y_i\phi(\boldsymbol{x}_i) \tag{3}$$

$$b=\frac{1}{|SV_{uerr}^+\cup SV_{uerr}^-|}\sum_{\boldsymbol{x}_i\in SV_{uerr}^+\cup SV_{uerr}^-}\left(\frac{1}{y_i}-\boldsymbol{w}_\Phi^{\mathrm{T}}\phi(\boldsymbol{x}_i)\right) \tag{4}$$

2.2 决策函数及其复杂度

给定未知样本 $\boldsymbol{x}\in\boldsymbol{R}^d$,可以通过下式决策:

$$f(\boldsymbol{x})=\boldsymbol{w}_\Phi^{\mathrm{T}}\phi(\boldsymbol{x})+b=\sum_{\boldsymbol{x}_i\in SVs}\alpha_i y_i k(\boldsymbol{x}_i,\boldsymbol{x})+b \tag{5}$$

若 $f(\boldsymbol{x})\geqslant 0$,则 $\boldsymbol{x}$ 属于正类;否则为负类.显然,SVM 决策 1 个未知样本的复杂度为 $O(|SVs|)$,当 N 个未知样本时,复杂度为 $O(N\times|SVs|)$.若 $|SVs|$ 或 N 很大则计算量很大,不能适合实时数据监测的应用场合.

2.3 支持向量的隐私

一般地,支持向量含有区别于其他类样本数据的重要特征,如药品和药物中的重要成份以及成份的比重数据等.以药物成份比重数据为例,由于 SVM 决策过程需要所有支持向量,而支持向量和其相应的权值 α_i 综合体现了药物的成份比重,显然药物成份以及药物成份比重是需要保护的信息.而对于目前的 SVM,当用户使用 SVM 决策函数时就会知道训练样本中的此类重要信息,这与样本隐私保护相违背.因此,本文所关注的隐私信息是从决策函数输出方面考虑,这与以往的隐私保护模型不同,如个体隐私保护模型[22],因为此类模型往往考虑训练过程,即通过训练样本中的若干个不可区分个体来保护隐私信息.然而攻击者不但可以通过训练过程也可通过统计未知样本的决策过程进行攻击,因此从决策函数输出方面考虑隐私保护也具有一定的实际意义.为此,本文提出一种隐藏支持向量信息并实现快速决策的 FCA-SVM$_{WPP}$方法.

3 MEB 及球心原像

3.1 MEB 算法

MEB 常用于一类问题,如一类 SVDD[2],其目标是在 Φ 中找到一个最小超球体 $B(\boldsymbol{c}_\Phi,R)$包含目标样本,其中 $\boldsymbol{c}_\Phi$ 和 R 分别是 Φ 空间下的球心和半径.设输入样本 $\chi=\{\boldsymbol{x}_i\mid\boldsymbol{x}_i\in\boldsymbol{R}^d,1\leqslant i\leqslant n\}$,其原始问题为:

$$\min_{R,\boldsymbol{c}_\Phi,\xi_i}R^2+C\sum_{i=1}^{n}\xi_i \tag{6}$$

$$\text{s.t.}\quad\|\phi(\boldsymbol{x}_i)-\boldsymbol{c}_\Phi\|^2\leqslant R^2+\xi_i,\xi_i\geqslant 0,1\leqslant i\leqslant n$$

其中,$C>0$ 是控制参数,ξ_i 是松弛变量.利用拉格朗日技巧得其对偶形式

$$\max_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha} \tag{7}$$

$$\text{s.t.}\quad\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,0\leqslant\alpha_i\leqslant C,1\leqslant i\leqslant n$$

其中,$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_n)^{\mathrm{T}}\geqslant 0$ 是拉格朗日乘子向量,$\boldsymbol{K}=[k(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{n\times n}=[\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)]_{n\times n}$是对应核矩阵.将 $\alpha_i>0$ 对应样本称为支持向量[2],则类似 2.1 节有以下集合:$SV_{uerr}=\{\boldsymbol{x}_i\mid 0<\alpha_i<C,1\leqslant i\leqslant n\}$未错分的支持向量;$SV_{err}=\{\boldsymbol{x}_i\mid C\leqslant\alpha_i,1\leqslant i\leqslant n\}$错分的支持向量;所有支持向量 $SVs=SV_{uerr}\cup SV_{err}$.则球心和半径为:

$$\boldsymbol{c}_{\Phi}=\sum_{i=1}^{n}\alpha_i\phi(\boldsymbol{x}_i)=\sum_{\boldsymbol{x}_i\in SVs}\alpha_i\phi(\boldsymbol{x}_i) \tag{8}$$

$$R=\left(\begin{array}{l}k(\boldsymbol{x}_{su},\boldsymbol{x}_{su})-2\sum\limits_{\boldsymbol{x}_i\in SVs}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}_{su})\\+\sum\limits_{\boldsymbol{x}_i\in SVs}\sum\limits_{\boldsymbol{x}_j\in SVs}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\end{array}\right)^{1/2}\Bigg|_{\boldsymbol{x}_{su}\in SV_{uerr}} \tag{9}$$

给定如 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/h)$的高斯核函数，$h$ 是其带宽参数，其决策函数为：

$$f(\boldsymbol{x})=R^2-\|\phi(\boldsymbol{x})-\boldsymbol{c}_{\Phi}\|^2=2\sum_{\boldsymbol{x}_i\in SVs}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x})+q \tag{10}$$

其中 $q=R^2-1-\sum\limits_{\boldsymbol{x}_i\in SVs}\sum\limits_{\boldsymbol{x}_j\in SVs}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)$是一常量，若 $f(\boldsymbol{x})\geqslant 0$，则 $\boldsymbol{x}$ 是目标类，否则是异常点．显然 MEB 决策一个未知样本的复杂度也为 $O(|SVs|)$．

3.2 几何属性

当给定高斯核时，$\forall\,\boldsymbol{x}\in\boldsymbol{R}^d$，那么在 Φ 空间下有 $\|\phi(\boldsymbol{x})\|=\sqrt{k(\boldsymbol{x},\boldsymbol{x})}=1$，故 $\boldsymbol{R}^d$ 被 ϕ 映射到 Φ 中的单位球上[20]，如图 1 所示．图中 $B(\boldsymbol{c}_{\Phi},R)$是 Φ 空间中获得的最小超球，则有以下结论：

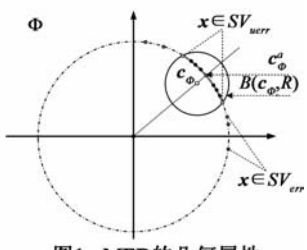

图1 MEB的几何属性

结论 1 由式(8)知，$\|\boldsymbol{c}_{\Phi}\|\leqslant\sum\limits_{\boldsymbol{x}_i\in SVs}\alpha_i\|\phi(\boldsymbol{x}_i)\|\leqslant\sum\limits_{\boldsymbol{x}_i\in SVs}\alpha_i=1$，故 $\boldsymbol{c}_{\Phi}$ 在单位球内或球面上；

结论 2 单位球上离 $\boldsymbol{c}_{\Phi}$ 最近点是 $\boldsymbol{c}^a_{\Phi}$，此点是 $\boldsymbol{c}_{\Phi}$ 沿其方向在单位球上的投影，称为 $\boldsymbol{c}_{\Phi}$ 的代理球心(Agent Center)，故有 $\boldsymbol{c}^a_{\Phi}=\boldsymbol{c}_{\Phi}/\|\boldsymbol{c}_{\Phi}\|$；

结论 3 根据结论 1 可知，$\boldsymbol{c}_{\Phi}$ 在单位球内的可能性较大，因此 $\boldsymbol{R}^d$ 中的任意点 $\boldsymbol{x}$ 均不能满足 $\phi(\boldsymbol{x})=\boldsymbol{c}_{\Phi}$，即 $\boldsymbol{c}_{\Phi}$ 原像 $\phi^{-1}(\boldsymbol{c}_{\Phi})$不可能在 $\boldsymbol{R}^d$ 中，同时根据结论 2 可知 $\boldsymbol{c}^a_{\Phi}$ 原像 $\phi^{-1}(\boldsymbol{c}^a_{\Phi})\in\boldsymbol{R}^d$(记为 $\boldsymbol{c}^a$)的可能性较大．

综上，若能在 $\boldsymbol{R}^d$ 中获得 $\boldsymbol{c}^a_{\Phi}$ 的准确原像 $\boldsymbol{c}^a\in\boldsymbol{R}^d$，即 $\phi(\boldsymbol{c}^a)=\boldsymbol{c}^a_{\Phi}$，则根据结论 2 有 $\phi(\boldsymbol{c}^a)=\boldsymbol{c}_{\Phi}/\|\boldsymbol{c}_{\Phi}\|$，此时式(10)为：

$$f(\boldsymbol{x})=R^2-1-\|\boldsymbol{c}_{\Phi}\|^2+2\|\boldsymbol{c}_{\Phi}\|k(\boldsymbol{c}^a,\boldsymbol{x}) \tag{11}$$

可见 MEB 的决策复杂度降为 $O(1)$．下文将重点讨论代理球心 $\boldsymbol{c}^a_{\Phi}$ 在 $\boldsymbol{R}^d$ 中的原像．

3.3 代理球心原像

3.3.1 若干定义和定理

定义 2 Φ 中的单位球面被 $B(\boldsymbol{c}_{\Phi},R)$分割成两个子空间，如图 1 所示的 $B(c_{\Phi},R)$球内和球外单位球面，分别定义为 $B_{\Phi 1}$和 $B_{\Phi 2}$，可知 $\boldsymbol{c}_{\Phi}$ 的代理点 $\boldsymbol{c}^a_{\Phi}\in B_{\Phi 1}$．

定义 3 根据 $\phi(\boldsymbol{R}^d)$属于 $B_{\Phi 1}$还是 $B_{\Phi 2}$，将 $\boldsymbol{R}^d$ 分为两个子空间，分别为 $O_{in}=\{\boldsymbol{x}\,|\,\boldsymbol{x}\in\boldsymbol{R}^d,\phi(\boldsymbol{x})\in B_{\Phi 1}\}$和 $O_{out}=\{\boldsymbol{x}\,|\,\boldsymbol{x}\in\boldsymbol{R}^d,\phi(\boldsymbol{x})\in B_{\Phi 2}\}$且 $\boldsymbol{R}^d=O_{in}\cup O_{out}$．

定义 4 给定训练样本 $\chi=\{\boldsymbol{x}_i\,|\,\boldsymbol{x}_i\in\boldsymbol{R}^d,1\leqslant i\leqslant n\}$，同理定义 $\chi_{in}=\{\boldsymbol{x}\,|\,\boldsymbol{x}\in O_{in}\}$和 $\chi_{out}=\{\boldsymbol{x}\,|\,\boldsymbol{x}\in O_{out}\}$，可知 $SV_{uerr}\in\chi_{in}$和 $SV_{err}\in\chi_{out}$．

定理 1 若 $\boldsymbol{R}^d$ 存在 $\boldsymbol{c}^a_{\Phi}$ 的准确原像 $\boldsymbol{c}^a$，那么 $\boldsymbol{c}^a\in O_{in}$．

证明 根据结论 3 可知，$\boldsymbol{c}^a$ 属于 $\boldsymbol{R}^d$ 的概率可能性较大，因此可以假定 $\boldsymbol{c}^a\in\boldsymbol{R}^d$，因为 $\phi(\boldsymbol{c}^a)=\boldsymbol{c}^a_{\Phi}\in B_{\Phi 1}$，根据定义 3 可知 $\boldsymbol{c}^a\in O_{in}$．证毕．

3.3.2 逼近准则

为获得代理球心的原像，可构造其损失函数 $\|\hat{\boldsymbol{c}}^a-\boldsymbol{c}^a\|$并 $\min\|\hat{\boldsymbol{c}}^a-\boldsymbol{c}^a\|$．根据中值定理，可知

$$\begin{aligned}&\phi(\hat{\boldsymbol{c}}^a)-\phi(\boldsymbol{c}^a)\approx\phi'(\xi)(\hat{\boldsymbol{c}}^a-\boldsymbol{c}^a)\\&\Leftrightarrow\|\phi(\hat{\boldsymbol{c}}^a)-\phi(\boldsymbol{c}^a)\|\approx\|\phi'(\xi)\|\times\|(\hat{\boldsymbol{c}}^a-\boldsymbol{c}^a)\|\\&\Rightarrow\|\phi(\hat{\boldsymbol{c}}^a)-\phi(\boldsymbol{c}^a)\|\geqslant\|(\hat{\boldsymbol{c}}^a-\boldsymbol{c}^a)\|\times\min(\|\phi'(\xi)\|)\end{aligned} \tag{12}$$

从式(12)可知，$\min\|\hat{\boldsymbol{c}}^a-\boldsymbol{c}^a\|$可通过$\|\phi(\hat{\boldsymbol{c}}^a)-\phi(\boldsymbol{c}^a)\|$的下界来近似求解，根据三角不等式有

$$\begin{aligned}\|\phi(\hat{\boldsymbol{c}}^a)-\phi(\boldsymbol{c}^a)\|&=\|\phi(\hat{\boldsymbol{c}}^a)-\boldsymbol{c}_{\Phi}+\boldsymbol{c}_{\Phi}-\phi(\boldsymbol{c}^a)\|\\&\leqslant\|\phi(\hat{\boldsymbol{c}}^a)-\boldsymbol{c}_{\Phi}\|+1-\|\boldsymbol{c}_{\Phi}\|\end{aligned} \tag{13}$$

故 $\min\|\hat{\boldsymbol{c}}^a-\boldsymbol{c}^a\|$可使 $\min\|\phi(\hat{\boldsymbol{c}}^a)-\boldsymbol{c}_{\Phi}\|$来求解代理球心原像．因此，构造 $\hat{\boldsymbol{c}}^a$ 的累积平方误差 ISE[23]，因此 $ISE(\hat{\boldsymbol{c}}^a)=\|\phi(\hat{\boldsymbol{c}}^a)-\boldsymbol{c}_{\Phi}\|^2$，则

$$\hat{\boldsymbol{c}}^a=\arg\min\|\phi(\hat{\boldsymbol{c}}^a)-\boldsymbol{c}_{\Phi}\|^2 \tag{14}$$

3.3.3 QP 解法

一个众所周知的假设：空间中任意点 $\boldsymbol{x}$ 可以通过其某个邻域内的点经过线性组合来进行逼近，如局部线性嵌入(Locally Linear Embedding, LLE)[24,25]等．因 $\boldsymbol{c}^a\in O_{in}$，$\chi_{in}\subset O_{in}$，故将 χ_{in}选为 $\boldsymbol{c}^a$ 的邻域，则

$$\hat{\boldsymbol{c}}^a=\sum_{\boldsymbol{x}_i\in\chi_{in}}^{|\chi_{in}|}\mu_i\boldsymbol{x}_i \tag{15}$$

其中 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_{|\chi_{in}|})^{\mathrm{T}}\geqslant 0$ 是权向量，且 $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{1}=1$，

$$\begin{aligned}\boldsymbol{\mu}&=\underset{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{1}=1}{\operatorname{argmin}}ISE(\boldsymbol{\mu})=\underset{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{1}=1}{\operatorname{argmin}}\|\phi(\sum_{\boldsymbol{x}_i\in\chi_{in}}\mu_i\boldsymbol{x}_i)-\boldsymbol{c}_{\Phi}\|^2\\&\approx\underset{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{1}=1}{\operatorname{argmin}}\|\sum_{\boldsymbol{x}_i\in\chi_{in}}\mu_i\phi(\boldsymbol{x}_i)-\boldsymbol{c}_{\Phi}\|^2\\&=\underset{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{1}=1}{\operatorname{argmin}}\left\{\begin{array}{l}\sum\limits_{\boldsymbol{x}_i\in\chi_{in}}\sum\limits_{\boldsymbol{x}_j\in\chi_{in}}\mu_i\mu_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\\-2\sum\limits_{\boldsymbol{x}_i\in\chi_{in}}\mu_i\sum\limits_{\boldsymbol{x}_j\in SVs}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}_j)\\+\sum\limits_{\boldsymbol{x}_i\in SVs}\sum\limits_{\boldsymbol{x}_j\in SVs}\alpha_i\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\end{array}\right\}\end{aligned} \tag{16}$$

式(16)右侧第 3 项与 $\boldsymbol{\mu}$ 无关，故 $\boldsymbol{\mu}$ 可通过下式求解：

$$\begin{aligned}&\max_{\boldsymbol{\mu}}2\sum_{\boldsymbol{x}_i\in\chi_{in}}\mu_i\sum_{\boldsymbol{x}_j\in SVs}\alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{\boldsymbol{x}_i\in\chi_{in}}\sum_{\boldsymbol{x}_j\in\chi_{in}}\mu_i\mu_j k(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&\text{s.t.}\quad\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{1}=1,\mu_i\geqslant 0,1\leqslant i\leqslant|\chi_{in}|\end{aligned} \tag{17}$$

显然，上式是 QP 问题，其时间复杂度不小于 $O(|\chi_{in}|^2)$，

空间复杂度为 $O(|\chi_{in}|^2)$[26~29],若 $|\chi_{in}|$ 较大,则耗时很长.为此,下面给出一种直接解法.

3.3.4 直接解法

根据式(14)且为高斯核函数时,则

$$\frac{\partial ISE(\hat{\boldsymbol{c}}^a)}{\partial \hat{\boldsymbol{c}}^a} = 0 \Rightarrow \hat{\boldsymbol{c}}^a = \frac{\sum\limits_{\boldsymbol{x}_i \in SVs} \alpha_i k(\hat{\boldsymbol{c}}^a, \boldsymbol{x}_i)\boldsymbol{x}_i}{\sum\limits_{\boldsymbol{x}_i \in SVs} \alpha_i k(\hat{\boldsymbol{c}}^a, \boldsymbol{x}_i)} \tag{18}$$

式(18)等号两边均出现 $\hat{\boldsymbol{c}}^a$,故可通过迭代算法实现求解,本文重点不讨论迭代方法. $\forall \boldsymbol{x}_i \in SVs$,则

$$\|\phi(\hat{\boldsymbol{c}}^a) - \phi(\boldsymbol{x}_i)\|^2 = 2 - 2k(\hat{\boldsymbol{c}}^a, \boldsymbol{x}_i) \tag{19}$$

若 $\hat{\boldsymbol{c}}^a$ 是 $\boldsymbol{c}_\Phi^a$ 的准确原像,即 $\phi(\hat{\boldsymbol{c}}^a) = \boldsymbol{c}_\Phi^a = \boldsymbol{c}_\Phi / \|\boldsymbol{c}_\Phi\|$,则

$$\left\| \frac{\boldsymbol{c}_\Phi}{\|\boldsymbol{c}_\Phi\|} - \phi(\boldsymbol{x}_i) \right\|^2 = 2 - \frac{2}{\|\boldsymbol{c}_\Phi\|} \sum_{\boldsymbol{x}_j \in SVs} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_i) \tag{20}$$

根据式(19)和(20),可知

$$k(\hat{\boldsymbol{c}}^a, \boldsymbol{x}_i) = \frac{1}{\|\boldsymbol{c}_\Phi\|} \sum_{\boldsymbol{x}_j \in SVs} \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_i) \tag{21}$$

将式(21)代入式(18),得

$$\hat{\boldsymbol{c}}^a = \sum_{\boldsymbol{x}_i \in SVs} \sum_{\boldsymbol{x}_j \in SVs} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)\boldsymbol{x}_i \Big/ \sum_{\boldsymbol{x}_i \in SVs} \sum_{\boldsymbol{x}_j \in SVs} \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{22}$$

3.3.5 解法讨论

QP 解法是从代理球心原像子空间出发,利用局部线性逼近的一种解法,该法优点是对核函数没有特殊要求,缺点是涉及 QP 问题.而直接解法须满足 2 个条件,一是假设 $\hat{\boldsymbol{c}}^a$ 是 $\boldsymbol{c}_\Phi^a$ 的准确原像;二是必须是高斯核,这限制了它的适用性,但其不涉及 QP 故求解速度会快于 QP 解法,第 5 节实验也验证了此结论.

4 FCA-SVM$_{WPP}$

4.1 基本思想

根据 SVM 的决策函数式(5),有

$$\begin{aligned} f(\boldsymbol{x}) &= \phi(\boldsymbol{x})^{\mathrm{T}} \sum_{\boldsymbol{x}_i \in SV^+} \alpha_i \phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x})^{\mathrm{T}} \sum_{\boldsymbol{x}_i \in SV^-} \alpha_i \phi(\boldsymbol{x}_i) + b \\ &= \sum_{\boldsymbol{x}_i \in SV^+} \alpha_i \phi(\boldsymbol{x})^{\mathrm{T}} \sum_{\boldsymbol{x}_i \in SV^+} \beta_i \phi(\boldsymbol{x}_i) \\ &\quad - \sum_{\boldsymbol{x}_i \in SV^-} \alpha_i \phi(\boldsymbol{x})^{\mathrm{T}} \sum_{\boldsymbol{x}_i \in SV^-} \beta_i \phi(\boldsymbol{x}_i) + b \\ &= a_1 \phi(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{c}_{1\Phi} - a_2 \phi(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{c}_{2\Phi} + b \end{aligned} \tag{23}$$

其中 $a_1 = \sum\limits_{\boldsymbol{x}_i \in SV^+} \alpha_i$, $a_2 = \sum\limits_{\boldsymbol{x}_i \in SV^-} \alpha_i$, $\boldsymbol{c}_{1\Phi} = \sum\limits_{\boldsymbol{x}_i \in SV^+} \beta_i \phi(\boldsymbol{x}_i)$, $\boldsymbol{c}_{2\Phi} = \sum\limits_{\boldsymbol{x}_i \in SV^-} \beta_i \phi(\boldsymbol{x}_i)$,且 $\sum\limits_{\boldsymbol{x}_i \in SV^+} \beta_i = 1$, $\sum\limits_{\boldsymbol{x}_i \in SV^-} \beta_i = 1$ 和 $\beta_i > 0$.故 $\boldsymbol{c}_{1\Phi}$ 和 $\boldsymbol{c}_{2\Phi}$ 可认为是 SV^+ 和 SV^- 对应超球的球心,即 SVM 决策函数是由两个 MEB 球心的线性组合构成,但注意此时并非是包络 SV^+ 和 SV^- 的最小超球.根据第 3 节可知,若能分别获得 $\boldsymbol{c}_{1\Phi}$ 和 $\boldsymbol{c}_{2\Phi}$ 代理球心的原像 $\boldsymbol{c}_1^a$ 和 $\boldsymbol{c}_2^a$,则式(5)为

$$\begin{aligned} f(\boldsymbol{x}) &= a_1 \phi(\boldsymbol{x})^{\mathrm{T}} \|\boldsymbol{c}_{1\Phi}\| \phi(\boldsymbol{c}_1^a) - a_2 \phi(\boldsymbol{x})^{\mathrm{T}} \|\boldsymbol{c}_{2\Phi}\| \phi(\boldsymbol{c}_2^a) + b \\ &= a_1 \|\boldsymbol{c}_{1\Phi}\| k(\boldsymbol{c}_1^a, \boldsymbol{x}) - a_2 \|\boldsymbol{c}_{2\Phi}\| k(\boldsymbol{c}_2^a, \boldsymbol{x}) + b \end{aligned} \tag{24}$$

显然,SVM 决策复杂度降至 $O(2)$,这将大大提高分类速度;同时,SV^+ 和 SV^- 中的重要信息隐藏于 $\boldsymbol{c}_1^a$ 和 $\boldsymbol{c}_2^a$ 中,因此从决策方面讲,保护了用户关键数据.

4.2 算法实现

综上,FCA-SVM$_{WPP}$算法总结为如下两个过程:

FCA-SVM$_{WPP}$算法:

(训练过程)

step1　初始化核宽度参数 h 和调节参数 C;

step2　求解式(2)获得 SVM 的 SV_{err}^+, SV_{err}^-, SV^+, SV^-;

step3　取 $\chi_{in}^+ = X^+ - SV_{err}^+$, $\chi_{in}^- = X^- - SV_{err}^-$,求解式(17)得到正负类权向量 $\boldsymbol{\mu}^+$ 和 $\boldsymbol{\mu}^-$,利用式(15)计算 $\boldsymbol{c}_1^a$ 和 $\boldsymbol{c}_2^a$;或利用式(22)基于 SV^+ 和 SV^- 计算 $\boldsymbol{c}_1^a$ 和 $\boldsymbol{c}_2^a$;

(决策过程)

step4　对于未知样本 $\boldsymbol{x}$,根据式(24)进行决策,若 $f(\boldsymbol{x}) \geqslant 0$,则 $\boldsymbol{x}$ 属于 +1 类,否则属于 -1 类.

可见,上述算法首先完成 SVM,然后利用 QP 解法或直接解法求得原像 $\boldsymbol{c}_1^a$ 和 $\boldsymbol{c}_2^a$,因此在训练速度上会慢于 SVM,但对于直接解法,因其不涉及 QP 故训练速度略慢于 SVM,第 5 节亦验证了此结论.

5 实验结果与分析

实验环境:CPU 2.6GHz,2G RAM,Intel Core(TM),XP OS,Matlab 2009a.选高斯核函数,并从测试精度,训练和分类时间等 3 方面比较算法性能.考虑到数据的不平衡,精度采用几何精度 g,此法常用于评价不平衡数据集[3,30],即分别统计正负类的精度 a^+ 和 a^-,则 $g = \sqrt{a^+ \times a^-}$,其中

$$a^+ = \frac{\#\ \text{positive samples correctly classified}}{\#\ \text{total positive samples classified}} \times 100\%$$

$$a^- = \frac{\#\ \text{negative samples correctly classified}}{\#\ \text{total negative samples classified}} \times 100\%$$

为简洁,称 QP 解法对应为 FCA-SVM$_{WPP}$-I,而称直接解法对应为 FCA-SVM$_{WPP}$-II.同时,为了公平比较,每次实验均从数据集的正负类中各随机抽取 50% 构成训练样本,剩余 50% 构成测试样本.选好参数后,依次运行 SVM,FCA-SVM$_{WPP}$-I/II 算法,随机运行 10 次后统计性能并以均值和标准差给出.

5.1 参数实验

本节实验分析核参数 h 和惩罚因子 C 对本文算法精度的影响,数据集为 Iris(见表 1).当进行核参数实验时,C 取训练样本个数,h 从网格 $\{s^2/256, s^2/128, s^2/64, s^2/32, s^2/16, s^2/8, s^2/4, s^2/2, s^2, 2s^2, 4s^2, 8s^2, 16s^2, 32s^2, 64s^2, 128s^2, 256s^2\}$ 中依次选择,其中 s 是训练样本 2 范数的平均值,随机运行 10 次,其结果如图 2 所示,注意图中横坐标 1,…,17 依次对应网格中的核参数;当

进行惩罚因子 C 实验时，$h = s^2$，C 从 $\{0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\} \times n$ 依次选择（n 是训练样本数），结果如图 3 所示，横坐标 1，…，14 依次对应上述网格. 从图 2 和 3 可看出，选择合适参数时，算法获得了较好的逼近效果；同时，从图 2 可知 h 越大算法越逼近 SVM；而图 3 表明算法对参数 C 比较不敏感，当 C 越大算法也越逼近 SVM.

表 1 UCI 数据集

数据集	维数	样本数	+1 类	−1 类
Wine	13	178	59	119
Iris	4	150	50	100
Biomed	5	194	67	127
Ionosphere	34	351	225	126
Hepatitis	19	155	123	32
C. Bench	60	208	111	97
S. Heart	44	267	212	55
B. Cancer	9	699	241	458
PBRH-0-1	16	1559	779	780

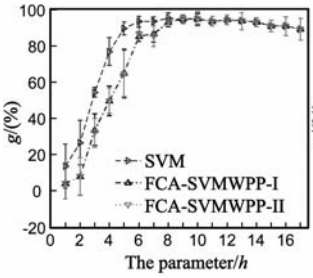

图2 核函数带宽参数实验

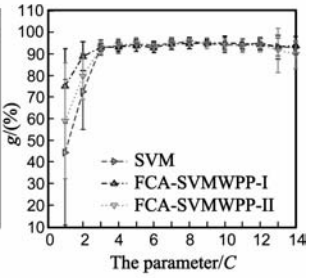

图3 惩罚因子实验

5.2 测试 UCI 数据集

利用表 1 中 9 种 UCI 数据集比较 3 种算法性能，其中 C. Bench，S. Heart，B. Cancer 和 PBRH(0,1)分别指 Connectionist Bench，SPECTF Heart，Breast Cancer 和 Pen Based Recognition of Handwritten Digits(数字 0，1 构成). 核参数 h 从 $\{s^2/128, s^2/64, s^2/32, s^2/16, s^2/8, s^2/4, s^2/2, s^2, 2s^2, 4s^2, 8s^2\}$ 中选择，C 从 $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50\} \times n$ 中选择. 表 2，表 3 和表 4 给出了实验结果.

表 2 SVM 和 FCA-SVM$_{WPP}$精度比较

数据集	几何精度 g(%)		
	SVM	FCA-SVM$_{WPP}$-I	FCA-SVM$_{WPP}$-II
Wine	90.09 ± 3.34	90.09 ± 3.34	90.09 ± 3.34
Iris	95.34 ± 2.57	95.27 ± 1.42	95.45 ± 2.75
Biomed	86.06 ± 3.46	84.07 ± 4.54	86.79 ± 2.37
Ionosphere	91.41 ± 2.56	86.24 ± 7.02	89.26 ± 2.56
Hepatitis	38.11 ± 8.83	42.79 ± 12.50	35.66 ± 17.57
C. Bench	80.32 ± 3.21	73.32 ± 4.57	73.56 ± 4.07
S. Heart	56.16 ± 9.12	73.23 ± 5.07	67.20 ± 7.84
B. Cancer	96.02 ± 1.42	96.22 ± 1.23	95.96 ± 1.44
PBRH(0,1)	99.94 ± 0.07	99.94 ± 0.07	99.94 ± 0.07
平均 AV	**81.49 ± 3.84**	**82.35 ± 4.42**	**81.55 ± 4.67**

表 2 给出了几何精度的比较结果，最后 1 行给出了整个数据集上的平均结果，从表中可看出，除了数据集 C. Bench 外，本文算法均获得了较佳效果，在有些数据集上精度甚至高于 SVM，如 Biomed，Hepatitis，S. Heart，最后一行的平均精度表明了 SVM 和本文算法的几何精度基本一致.

表 3 SVM 和 FCA-SVM$_{WPP}$训练速度比较

数据集	训练时间(s)		
	SVM	FCA-SVM$_{WPP}$-I	FCA-SVM$_{WPP}$-II
Wine	0.2854 ± 0.1488	0.3833 ± 0.1694	0.2869 ± 0.1521
Iris	0.2402 ± 0.1755	0.4790 ± 0.2809	0.2414 ± 0.1778
Biomed	0.5008 ± 0.1701	1.3174 ± 0.7634	0.5021 ± 0.1722
Ionosphere	1.7834 ± 0.2167	5.2136 ± 0.5306	1.7860 ± 0.2188
Hepatitis	0.1322 ± 0.1549	4.0032 ± 2.9885	0.1337 ± 0.1572
C. Bench	0.4267 ± 0.1596	1.8852 ± 0.5552	0.4284 ± 0.1619
S. Heart	0.2194 ± 0.1582	5.3927 ± 3.1494	0.2220 ± 0.1603
B. Cancer	23.3257 ± 2.0570	35.8527 ± 5.2960	23.3277 ± 2.0568
PBRH(0,1)	98.6157 ± 4.3693	172.2522 ± 6.7336	98.6191 ± 4.3717
平均 AV	**13.9477 ± 0.8456**	**25.1977 ± 2.2741**	**13.9497 ± 0.8476**

表 4 SVM 和 FCA-SVM$_{WPP}$分类速度比较

数据集	训练时间(s)		
	SVM	FCA-SVM$_{WPP}$-I	FCA-SVM$_{WPP}$-II
Wine	0.0134 ± 0.0031	0.0008 ± 0.0002	0.0008 ± 0.0002
Iris	0.0110 ± 0.0019	0.0006 ± 0.0000	0.0006 ± 0.0000
Biomed	0.0252 ± 0.0045	0.0009 ± 0.0002	0.0009 ± 0.0002
Ionosphere	0.1339 ± 0.0079	0.0015 ± 0.0000	0.0015 ± 0.0001
Hepatitis	0.0298 ± 0.0041	0.0007 ± 0.0001	0.0007 ± 0.0001
C. Bench	0.0505 ± 0.0056	0.0012 ± 0.0003	0.0011 ± 0.0003
S. Heart	0.1122 ± 0.0075	0.0013 ± 0.0002	0.0013 ± 0.0002
B. Cancer	0.1149 ± 0.0140	0.0029 ± 0.0002	0.0029 ± 0.0006
PBRH(0,1)	0.1644 ± 0.0114	0.0066 ± 0.0001	0.0065 ± 0.0001
平均 AV	**0.0728 ± 0.0067**	**0.0018 ± 0.0001**	**0.0018 ± 0.0002**

从表 3 的训练速度可看出，SVM，FCA-SVM$_{WPP}$-II 和 I 依次变慢. 这是由于本文算法需要运行 SVM 获得支持向量 SV^+_{err}，SV^-_{err}，SV^+ 和 SV^-，然后再获得两球心原像. FCA-SVM$_{WPP}$-I/II 分别采用 QP 解法和直接解法求解代理球心原像，故 FCA-SVM$_{WPP}$-I 最慢，而 FCA-SVM$_{WPP}$-II 接近于 SVM 算法.

从表 4 的分类时间可看出，SVM 的分类速度明显慢于 FCA-SVM$_{WPP}$-I/II，而且本文两种算法的分类速度基本一致，其原因是：对于 1 个未知样本，本文算法的决策复杂度均为 $O(2)$. 同时，SVM 分类时间标准差较大，本文方法很小，其原因是 SVM 决策复杂度与支持向量数有关，而每次运行时支持向量数不完全相同，但本文方法与其无关. 总体而言，本文方法在 UCI 数据集上获得了较好效果.

5.3 测试 PIE 数据集

本节利用 PIE 人脸图像[31]比较 3 种算法性能. 从 PIE 中随机选择 2 男性和 2 女性，每人为 170 张图像，男

性和女性各构成一个二类数据集,如图 4 和 5 所示.实验中男性标签对应 1 和 2 号,女性对应 35 和 38 号.参数选择同 UCI 实验,表 5 给出了实验结果.

图4　PIE-男性人脸

图5　PIE-女性人脸

表 5　PIE 数据集的比较结果

算法	数据集	几何精度 g (%)	训练时间 (s)	分类时间 (s)
SVM	MAN	99.76 ± 0.30	2.1920 ± 0.1867	0.4821 ± 0.0199
	WOMAN	99.94 ± 0.19	2.0444 ± 0.2812	0.4106 ± 0.0109
	平均 AV	**99.85 ± 0.25**	**2.1182 ± 0.2340**	**0.4464 ± 0.0154**
FCA-SVM_{WPP}-I	MAN	98.46 ± 1.20	5.8168 ± 1.0253	0.0074 ± 0.0002
	WOMAN	96.68 ± 2.70	6.0367 ± 0.5492	0.0076 ± 0.0004
	平均 AV	**97.57 ± 1.95**	**5.9268 ± 0.7873**	**0.0075 ± 0.0003**
FCA-SVM_{WPP}-II	MAN	99.76 ± 0.30	2.1956 ± 0.1875	0.0056 ± 0.0002
	WOMAN	99.88 ± 0.26	2.0476 ± 0.2833	0.0056 ± 0.0002
	平均 AV	**99.82 ± 0.28**	**2.1216 ± 0.2354**	**0.0056 ± 0.0002**

从表 5 可以看出,3 种算法在解决次 2 类问题时均获得了较好的精度;同时,表 5 也表明了 SVM, FCA-SVM_{WPP}-II 和 I 的训练速度依次变慢,其原因是本文方法需要先运行 SVM 才能获得两个逼近球心的原像;在分类速度上,FCA-SVM_{WPP}-I/II 明显快于 SVM,这正是因为本文方法的测试复杂度降到了 $O(2)$.因此,本文算法可用于此数据集,特别是对实时性要求很高的场合,如上班高峰期的门禁系统等.

6　结论

SVM 的决策函数分解为正负类支持向量,从而转换为两个 MEB 球心的线性组合,并利用 MEB 代理球心的原像,提出了一种隐藏支持向量信息并能快速实现决策的 FCA-SVM_{WPP}方法.同时,基于 ISE 准则导出两种代理球心原像的求解方法:QP 解法和直接解法.虽然实验结果表明这两种解法均获得了较好逼近效果及较快的分类速度,但这两种解法的区别是:(1)QP 解法训练速度慢,而直接解法速度快;(2)QP 解法适用于任何核函数,而直接解法只能适用于 Gaussian 核函数.UCI 和 PIE 两个数据集的实验结果表明,本文 FCA-SVM_{WPP}-I/II 与原始 SVM 相比,不但获得较好精度,而且分类速度明显提高.但如何提高 FCA-SVM_{WPP}-I 的训练速度? 以及代理球心原像在 $\boldsymbol{R}^d$ 中并不是全概率事件,因此如何更加准确地获取其原像? 这些将作为我们近期的研究工作.

参考文献

[1] 孙即祥.现代模式识别(第 2 版)[M].北京:高等教育出版社,2008.1 – 3.
Sun Jixiang. Modern Pattern Recognition[M]. Beijing: Higher Education Press, 2008.1 – 3.(in Chinese)

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

[3] Wu M R, Ye J P. A small sphere and large margin approach for novelty detection using training data with outliers[J]. IEEE Trans on PAMI, 2009, 31(11): 2088 – 2092.

[4] Cortes C, Vapnik V N. Support vector networks[J]. Machine Learning, 1995, 20(3): 273 – 297.

[5] Schölkopf B, Smola A, Williamson RC, Bartlett PL. New support vector algorithms[J]. Neural Computation, 2000, 12: 1207 – 1245.

[6] Cover TM, Hart PE. Nearest neighbor pattern classification[J]. IEEE Trans on Information Theory, 1967, 13(1): 21 – 27.

[7] Hastie T, Tibshirani R. Discriminant adaptive nearest neighbor classification[J]. IEEE Trans on PAMI, 1996, 18(6): 607 – 616.

[8] Babich G A, Camps O I. Weighted Parzen windows for pattern classification[J]. IEEE Trans on PAMI, 1996, 18(5): 567 – 570.

[9] Marzio M DI, Taylor C C. Kernel density classification and boosting: an L2 analysis[J]. Statistics and Computing, 2005, 15: 113 – 123.

[10] Angelov P P, Xiao W Z. Evolving fuzzy-rule-based classifiers from data streams[J]. IEEE Trans on Fuzzy Systems, 2008, 16(6): 1462 – 1475.

[11] Tang B, Mazzoni D. Multiclass reduced-set support vector machines[A]. Proc 23rd ICML[C]. New York: ACM Press, 2006.921 – 928.

[12] Osuna E, Girosi F. Reducing the run-time complexity of support vector machines[A]. In Advances in Kernel Methods: Support Vector Learning[C]. MA: MIT Press, 1999. 271 – 283.

[13] Dries G, Johan A K Suykens, Joos V. Reduce the amount of support vectors of SVM classifiers using Separable Case Approximation[DB/OL] ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/dgeebele/pub/SCA.pdf, 2010-11-08.

[14] Renzo P, Elisa R. Reduced complexity RBF classifiers with support vector centres and dynamic decay adjustment[J]. Neurocomputing, 2006, 69(16 – 18): 2446 – 2450.

[15] Quang-Anh T, Qian-Li Z, Xing L. Reduce the number of support vectors by using clustering techniques[A]. ICMLC[C]. CA: IEEE Computer Society Press, 2003.2: 1245 – 1248.

[16] Bakir G H, Weston J, Sch? lkopf B. Learning to find pre-images[A]. Advances in Neural Information Processing Systems

[C]. MA: MIT Press, 2004. 16: 449 – 456.
[17] Kwok J T, Tsang I W. The pre-image problem in kernel methods[J]. IEEE Trans on Neural Networks, 2004, 15(6): 1517 – 1525.
[18] Schölkopf B, Mika S, Burges C. J. C, et al. Input space vs. feature space in kernel-based methods[J]. IEEE Trans on Neural Networks, 1999, 10(5): 1000 – 1017.
[19] Bakir G, Zien A, Tsuda K. Learning to find graph pre-images [A]. Proc 26th Pattern Recognition[C]. Berlin: Springer, 2004. 253 – 261.
[20] Yi H L, Yan C L, Yen J C. Fast support vector data descriptions for novelty detection[J]. IEEE Trans on Neural Networks, 2010, 21(8): 1296 – 1313.
[21] Chang C. C, Lin C. J. LIBSVM: a library for support vector machines[CP/OL]. http://www.csie.ntu.edu.tw/~cjlin/, 2010-10-06.
[22] 韩建民,于娟,虞慧群,贾 ▮. 面向敏感值的个性化隐私保护[J]. 电子学报, 2010, 38(7): 1723 – 1728.
Han Jianming, Yu Juan, Yu Huiqun, Jia Jiong. Individuation privacy preservation oriented to sensitive values[J]. Acta Electronica Sinica, 2010, 38(7): 1723 – 1728. (in Chinese)
[23] Alan J I. Recent developments in nonparametric density estimation[J]. Journal of the American Statistical Association, 1991, 86(413): 205 – 224.
[24] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323 – 2326.
[25] Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319 – 2323.
[26] Tsang I W, Kwok J T, Cheung P M. Core vector machines: Fast SVM training on very large data sets[J]. JMLR, 2005, 6: 363 – 392.
[27] Deng Z H, Chung F L, Wang S T. FRSDE: fast reduced set density estimator using minimal enclosing ball approximation [J]. Pattern Recognition, 2008, 41: 1363 – 1372.
[28] Chung F L, Deng Z H, Wang S T. From minimum enclosing ball to fast fuzzy inference system training on large datasets [J]. IEEE Trans on Fuzzy Systems, 2009, 17(1): 173 – 184.
[29] Tsang I W, Kwok, J T, Zurada J M. Generalized core vector machines[J]. IEEE Trans on Neural Networks, 2006, 17(5): 1126 – 1140.
[30] Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection[A]. Proc 14th ICML[C]. Nashville: Morgan Kaufmann Publishers, 1997. 179 – 186.
[31] Xiaofei H, Deng C, Partha N. Laplacian Score for Feature Selection[A]. Advances in Neural Information Processing Systems 18[C]. MA: MIT Press, 2006. 507 – 514.

作者简介

胡文军 男, 1977 年生于安徽绩溪. 2000 年、2003 年分别在安徽工程大学、山东理工大学获得工学学士、硕士学位, 2009 年进入江南大学攻读博士学位, 主要从事模式识别、人工智能等方面的研究.
E-mail: hoowenjun@yahoo.com.cn

王士同 男, 1964 年生于江苏扬州. 教授、博士生导师、中国计算机学会高级会员. 1984 年、1987 年在南京航空航天大学获得工学学士、硕士学位. 主要从事人工智能、模式识别、模糊系统、医学图像处理和生物信息学等方面的研究.