ISSN 1000-1239/CN 11-1777/TP
50(11): 2315-2324, 2013

基于粒度偏移因子的支持向量机学习方法

郭虎升[1]　王文剑[2]
[1](山西大学计算机与信息技术学院　太原　030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学)　太原　030006)
(chaofei142@163.com)

A Support Vector Machine Learning Method Based on Granule Shift Parameter

Guo Husheng[1] and Wang Wenjian[2]
[1](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing* (*Shanxi University*), *Ministry of Education*, *Taiyuan* 030006)

Abstract　For practical application problems, data size and distribution density are always imbalanced. Because the probabilities of samples falling into various regions are different due to the influence of data size or density distribution, the hyperplane obtained by traditional support vector machine (SVM) based on maximum margin maybe not optimal. Combined with granular computing (GrC) theory, an improved granular support vector machine (GSVM) model based on granule shift parameter, namely S_GSVM, is presented to solve the imbalanced data classification problems. For S_GSVM model, the original data will be firstly mapped into a high-dimensional feature space by Mercer kernel, and then the mapped data will be granulated in this space. Two granule factors, support and disperse, are defined to measure the influence of sample distributions on the performance of SVM. Then, the shift parameter of each granule is computed by support and disperse. Based on these shift parameters, a new convex quadratic optimization problem is constructed and solved. Fully considering the influence of data distribution on the generalization performance, the proposed S_GSVM model can improve the obtained hyperplane which is based on maximum margin. Experiment results on benchmark datasets and database of interacting proteins demonstrate the effectiveness and efficiency of the proposed S_GSVM model.

Key words　support vector machine; granular support vector machine; imbalance distribution; granule; shift parameter; S_GSVM model

摘　要　在实际应用中,数据集样本规模、分布密度的不平衡性可能会使传统支持向量机(support vector machine, SVM)得到的分类超平面不是最优.在对传统支持向量机最优分类面分析的基础上,结合粒度计算(granular computing, GrC)理论,针对数据规模和分布密度不平衡的数据集,提出一种基于粒度偏移因子的粒度支持向量机(granular SVM, GSVM)学习方法,称为 S_GSVM 方法.该方法将原始样本用 Mercer 核映射到高维空间,然后在高维空间中对数据进行有效的粒划分,通过对不同的粒计算不同的超平面偏移因子,重新构造支持向量机的凸二次优化问题,以得到一个泛化能力更好的分类

收稿日期:2012-05-09;**修回日期**:2013-07-23
基金项目:国家自然科学基金项目(60975035,61273291,71031006);高等学校博士学科点专项科研基金项目(20091401110003);山西省自然科学基金项目(2009011017-2);山西省回国留学人员科研资助项目(2012-008);山西省优秀研究生创新项目(20103021)
通信作者:王文剑(wjwang@sxu.edu.cn)

超平面. S_GSVM 方法充分考虑了数据复杂分布对于泛化能力的影响,对基于最大间隔的分类面进行改进. 实验结果表明,S_GSVM 方法在非平衡数据集上能得到较好的泛化性能.

关键词 支持向量机;粒度支持向量机;不平衡分布;粒;偏移因子;S_GSVM 模型

中图法分类号 TP18

支持向量机(support vector machine, SVM)是 Vapnik[1] 提出的一类通用有效的机器学习方法,能够非常成功地解决分类和回归等问题,目前已成为机器学习的研究热点,并在手写数字识别、人脸图像识别、时间序列预测等诸多领域得到应用. 在基于 SVM 的应用中,其学习效率和泛化能力一直是人们关注和研究的重点[2-4],并且已取得丰富的研究成果. 众所周知,对于分类问题,SVM 训练的目的是在给定的特征空间中获得两类数据间隔最大的超平面. 而对于实际问题,数据分布往往不确定、不平衡且非常复杂,最优的核函数或参数通常不容易得到. 因此,如何在选定的核空间中改进 SVM 的最优分类面,使其获得更好的泛化性能具有重要的理论意义和实用价值.

粒度计算(granular computing, GrC)是信息处理中一种新的概念和计算范式,覆盖了所有有关粒度的理论、方法、技术和工具的研究,主要用于处理不确定的、模糊的、不完整的和海量的信息[5]. GrC 思想的实质是用简单易求、低成本的足够满意的近似解代替精确解. 将 GrC 理论与较成熟的智能计算方法(如 SVM、神经网络等)相融合,以形成对复杂问题的高效求解算法正成为机器学习及相关领域的研究热点.

1 粒度支持向量机

Tang[6] 有效结合 GrC 理论与 SVM 方法,提出粒度支持向量机模型(granular support vector machine, GSVM). 这种模型通过粒划分方法获得一系列信息粒,然后在每个信息粒上进行学习(本质上可以并行学习),最后通过聚合信息粒上的信息(或数据、规则、知识、属性等)获得最终的支持向量机决策函数. 与传统的 SVM 方法相比,GSVM 模型具有更强的线性可分能力、更好的泛化能力和更高的学习效率.

实际上,在 Tang 之前就有学者提出一些有效的 SVM 模型,它们可以看作 GSVM 的雏形. 如 Vapnik[1] 提出的选块算法,通过迭代方法排除块中非支持向量对应的训练点,并逐步把支持向量对应点选入工作集. Osuna 等人[7] 在选块算法的基础上提出分解算法,有效提高了算法效率. John[8] 提出了适合于处理稀疏样本的序贯最小优化(sequence minimum optimization, SMO)算法. Chang 等人[9] 开发了 LIBSVM 方法库并证明了其收敛性.

目前,已有很多学者在粒度支持向量机模型的设计方面做了大量工作:如 Tang 等人[10] 提出了基于关联规则的 GSVM 算法;Zhang 等人[11-12] 提出了基于聚类的 GSVM 算法;Yu 等人[13] 提出了基于树形层次结构的 GSVM 学习模型;Katagiri 等人[14] 结合几何分析方法得到逼近于最优超平面的近似最优超平面;Cheng 等人[15] 提出的 GSVM 算法的泛化能力得到了进一步的提高;Mangasarian 等人[16] 提出了具有线性收敛性的 Lagrangian SVM(LSVM)算法,大大提高了算法的训练速度;Zhang 等人[17] 通过 K 近邻方法估计数据的相对密度,设计了基于密度间隔的 SVM 方法,并证明了该方法的错误上界;文献[18]设计了处理非平衡数据分类问题的粒度支持向量机模型. 此外,还有诸如基于神经网络、粗糙集、决策树等的粒度支持向量机方法也被诸多学者研究.

传统的 GSVM 模型大都是在原空间进行粒划分,然后在核空间训练,这些模型虽然可在训练效率方面有显著改进,但泛化能力方面却有不小损失. 造成泛化能力损失的主要原因有两点:一是粒划分之后可能会导致数据在核空间分布与原空间分布不一致;其次,这些 GSVM 模型在训练之前划分粒,并用粒中部分样本点(如粒心)代替整个粒中的数据参与训练,忽略了数据分布的差异,也会降低学习器的泛化能力[12].

传统的 SVM 由于数据集样本规模、分布密度的不平衡性等,可能会使得到的最大间隔分类面并不最优. 本文结合粒度计算理论,提出一种能够针对样本规模、分布密度不平衡数据的粒度支持向量机分类方法,称为 S_GSVM 方法. 该方法将原始样本用 Mercer 核映射到高维空间,然后在高维空间中对数据进行有效的粒划分并对不同粒计算不同的超平面偏移因子,通过构造求解一个新的凸二次优化问题,从而得到一个泛化能力更好的分类超平面.

2 基于粒度偏移因子的支持向量机学习方法

由于数据集样本规模、分布密度的不平衡性,测试样本落在不同区域的概率不同,因此采用最大间隔法得到的分类超平面泛化能力不一定最优.为此,本文提出一种改进的粒度支持向量机方法.该模型首先在给定的核空间根据数据分布特性进行粒划分,并采用支持度和分散度衡量不同分布的粒对SVM泛化能力的影响,给不同粒分别赋予不同的偏移因子,从而改进超平面,以得到泛化能力更高的学习器.

2.1 基于核的粒划分

对于训练集 $X=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{l}$,其中 $\boldsymbol{x}_i\in\mathbb{R}^n$ 为样本,$y_i\in\{-1,1\}$ 为标签,经过非线性函数 Φ,映射到 N 维特征空间 $\mathbb{R}^N$ 中,样本在该空间表示为 $X=\{(\Phi(\boldsymbol{x}_i)),y_i\}_{i=1}^{l}$,在该核空间将样本划分为 k 个粒,即 $X=\{X_1,X_2,\cdots,X_k\}$,其中 $X_i=\{\Phi(\boldsymbol{x}_{i_p})\}_{p=1}^{l_i}$($l_i$ 为第 i 个粒中样本个数).每个粒都可看作一个超球,其中心和半径定义如下.

定义 1. 核空间粒超球的中心和半径.将在给定的核空间粒划分后形成的任一样本粒 X_i 称为一个粒超球(为简便起见,本文将粒超球仍记作 X_i),其中心(粒心)$\boldsymbol{\mu}_i$ 和半径 r_i 分别为

$$\boldsymbol{\mu}_i=\frac{1}{l_i}\sum_{p=1}^{l_i}\Phi(\boldsymbol{x}_{i_p})=\sqrt{\frac{1}{l_i}\Big(\sum_{p=1}^{l_i}\Phi(\boldsymbol{x}_{i_p})\Big)^2}=\frac{1}{l_i}\sqrt{\sum_{p=1}^{l_i}\sum_{q=1}^{l_i}\Phi(\boldsymbol{x}_{i_p})\cdot\Phi(\boldsymbol{x}_{i_q})}=\frac{1}{l_i}\sqrt{\sum_{p=1}^{l_i}\sum_{q=1}^{l_i}K(\boldsymbol{x}_{i_p},\boldsymbol{x}_{i_q})};$$

$$r_i=\max_{\boldsymbol{x}\in X_i}(\Phi(\boldsymbol{x}_{i_s})-\boldsymbol{\mu}_i)=\max_{\boldsymbol{x}_{i_s}\in X_i}\Big(\sqrt{(\Phi(\boldsymbol{x}_{i_s}))^2-2\Phi(\boldsymbol{x}_{i_s})\cdot\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^2}\Big)=$$

$$\max_{\boldsymbol{x}_{i_s}\in X_i}\left(\sqrt{K(\boldsymbol{x}_{i_s},\boldsymbol{x}_{i_s})-\frac{2}{l_i}\sum_{p=1}^{l_i}K(\boldsymbol{x}_{i_s},\boldsymbol{x}_{i_p})+\frac{1}{l_i^2}\sum_{p=1}^{l_i}\sum_{q=1}^{l_i}K(\boldsymbol{x}_{i_p},\boldsymbol{x}_{i_q})}\right). \tag{2}$$

根据定义 1,N 维空间中任一样本 $\Phi(\boldsymbol{x}_j)$ 到粒超球 X_i 的距离为

$$d(\Phi(\boldsymbol{x}_j),X_i)=\sqrt{K(\boldsymbol{x}_j,\boldsymbol{x}_j)-\frac{2}{l_i}\sum_{p=1}^{l_i}K(\boldsymbol{x}_j\cdot\boldsymbol{x}_{i_p})+\frac{1}{l_i^2}\sum_{p=1}^{l_i}\sum_{q=1}^{l_i}K(\boldsymbol{x}_{i_p},\boldsymbol{x}_{i_q})}. \tag{3}$$

由于本文的重点在于如何结合粒度计算理论来设计具有较好泛化能力的GSVM模型,以克服不平衡的数据分布对分类器泛化性能产生的负面影响,所以有关如何选择核函数及参数将不在本文进行具体讨论,相关内容可参考文献[3,19-23].实际上,本文提出的S_GSVM模型可以与任意的核选择方法相结合.

本文采用粒超球及其相关度量来进行迭代粒划分,粒划分的主要步骤见算法 1.

算法 1. 粒划分算法.

输入:初始聚类样本集 X 和粒划分参数 k;

输出:划分后得到的粒集 $\{X_1,X_2,\cdots,X_k\}$.

Step1. 任意选择 k 个样本点作为粒心.

Step2. 按照式(3)中样本点距离一个粒的距离公式来对所有样本点采用核空间近邻法进行粒划分.

Step3. 依据式(1)调整粒心,观察粒心是否有变化,若有变化则返回 Step2,否则转 Step4.

Step4. 算法结束,得到划分粒集 $\{X_1,X_2,\cdots,X_k\}$.

2.2 S_GSVM 模型

为有效衡量样本密度分布的不平衡性,本文引入平均密度.对于给定的数据集 X,X^+ 和 X^- 分别表示正类样本集和负类样本集.通过式(3)得到正类样本集和负类样本集的中心分别为 $\boldsymbol{\mu}^+$ 和 $\boldsymbol{\mu}^-$.正负类样本集的平均密度 ρ^+ 和 ρ^- 定义如下:

$$\rho^+=\frac{1}{\frac{1}{|X^+|}\sum_{i=1}^{|X^+|}d(\Phi(\boldsymbol{x}_i^+),\boldsymbol{\mu}^+)}=\frac{|X^+|}{\sum_{i=1}^{|X^+|}d(\Phi(\boldsymbol{x}_i^+),\boldsymbol{\mu}^+)},\boldsymbol{x}_i^+\in X^+; \tag{4}$$

$$\rho^-=\frac{1}{\frac{1}{|X^-|}\sum_{i=1}^{|X^-|}d(\Phi(\boldsymbol{x}_i^-),\boldsymbol{\mu}^-)}=\frac{|X^-|}{\sum_{i=1}^{|X^-|}d(\Phi(\boldsymbol{x}_i^-),\boldsymbol{\mu}^-)},\boldsymbol{x}_i^-\in X^-; \tag{5}$$

这里,$|\cdot|$ 表示集合的规模,$d(\Phi(\boldsymbol{x}_i^+),\boldsymbol{\mu}^+)$ 和 $d(\Phi(\boldsymbol{x}_i^-),\boldsymbol{\mu}^-)$ 分别表示核空间中样本到正负类数据集中心的距离.而样本规模分布的不平衡性直接通过正负类样本中样本的个数来衡量.

在构造模型之前,首先给出关于粒支持度(*Support*)与分散度(*Disperse*)的定义,这两个量主要用于衡量不同粒样本的分布情况,估计数据分布对分类面的影响,从而有效地校正分类超平面.

定义 2. 粒的支持度.假设样本集 X 经粒划分后得到 k 个粒 $X=\{X_1,X_2,\cdots,X_k\}$,粒 $X_i=\{\Phi(\boldsymbol{x}_{i_p})\}_{p=1}^{l_i}$ 的支持度定义为该粒所包含的样本数占总样本数的比例,即

$$Support(X_i)=|X_i|/|X|=l_i/l, \tag{6}$$

其中,l_i 为粒 X_i 的样本规模,l 为整个样本集 X 的样本规模.

定义 3. 粒的分散度.粒 X_i 的分散度定义为

$$Disperse(X_i)=\frac{1}{l_i r_i}\sum_{p=1}^{l_i}|\Phi(\boldsymbol{x}_{i_p})-\boldsymbol{\mu}_i|=\frac{1}{l_i r_i}\sum_{p=1}^{l_i}\sqrt{(\Phi(\boldsymbol{x}_{i_p}))^2-2\Phi(\boldsymbol{x}_{i_p})\boldsymbol{\mu}_i+\boldsymbol{\mu}_i^2}=$$

$$\frac{1}{l_i r_i}\sum_{p=1}^{l_i}\sqrt{K(\boldsymbol{x}_{i_p},\boldsymbol{x}_{i_p})-\frac{2}{l_i}\sum_{q=1}^{l_i}K(\boldsymbol{x}_{i_p},\boldsymbol{x}_{i_q})+\frac{1}{l_i^2}\sum_{q=1}^{l_i}\sum_{r=1}^{l_i}K(\boldsymbol{x}_{i_q},\boldsymbol{x}_{i_r})}, \tag{7}$$

其中,l_i 为粒 X_i 的样本规模,r_i 为粒超球 X_i 的半径.

一般地,若粒支持度大,在测试集中,分布于该粒区域内及附近区域的样本点就多,为提高泛化能力,分类面应适当远离该粒;若粒分散度大,分布于该粒区域范围内及附近区域中测试集样本密度较小,容易发生错误分类,分类面应适当远离该粒.在给出粒支持度和分散度基础上,定义粒 X_i 的超平面校正偏移因子 σ_i,即

$$\sigma_i=Support(X_i)Disperse(X_i), \tag{8}$$

$0\leqslant\sigma_i\leqslant 1$.

改进的硬间隔 SVM 二次优化问题可写为如下形式:

$$\begin{aligned}\min_{w,b}\quad & \frac{1}{2}\|\boldsymbol{w}\|^2\\ \text{s.t.}\quad & y_i(\boldsymbol{w}^{\mathrm{T}}\Phi(\boldsymbol{x}_i)+b)\geqslant 1-\delta_i,\\ & i=1,2,\cdots,l,\end{aligned} \tag{9}$$

其中,若 $\boldsymbol{x}_i\in X_j$,则 $\delta_i=\sigma_j$,即每个粒中所有样本的偏移因子是相同的.其对偶问题为

$$\begin{aligned}\max_{\alpha}\quad & W(\boldsymbol{\alpha})=-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)+\\ & \sum_{i=1}^{l}(\alpha_i-\alpha_i\delta_i)\\ \text{s.t.}\quad & \sum_{i=1}^{l}\alpha_i y_i=0,\\ & \alpha_i\geqslant 0,\ i=1,\cdots,l.\end{aligned} \tag{10}$$

本文提出的 S_GSVM 算法的主要步骤见算法 2.

算法 2. S_GSVM 算法.

输入:训练集 $X=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{l}$;

输出:决策函数 $f(X)$.

初始化:给出训练集 $X=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{l}$,并选择核函数及相关参数.

Step1. 基于 Mercer 核的粒划分:

给出粒划分参数 k,并按照算法 1 进行粒划,获得一系列信息粒 $\{X_1,X_2,\cdots,X_k\}$.

Step2. 相关参数计算:

按照式(6)和式(7)分别计算每个信息粒 X_i 的支持度 $Support(X_i)$ 与分散度 $Disperse(X_i)$,然后按式(8)计算超平面偏移因子.

Step3. 进行 S_GSVM 训练:

构造并求解 S_GSVM 最优化问题(10),并得到最优解 $\boldsymbol{a}^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_l^*)^{\mathrm{T}}$.计算 $w^*=\sum_{i=1}^{l}y_i\alpha_i^*\boldsymbol{x}_i$,并选择 $\boldsymbol{a}^*$ 的一个正分量 α_m^*,计算 $b^*=y_m-\frac{\delta_m}{y_m}-\sum_{i=1}^{l}y_i\alpha_i^*K(\boldsymbol{x}_i,\boldsymbol{x}_m)$.

Step4. 求得最优超平面及决策函数:

构造分类超平面 $f:(\boldsymbol{w}^{*\mathrm{T}}\cdot\boldsymbol{x})+b^*=0$,并得到决策函数 $f(x)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^*y_iK(\boldsymbol{x}_i,\boldsymbol{x})+b^*)$.

此外,可采用惩罚参数加权的方法构造软间隔 S_GSVM 模型.与硬间隔 S_GSVM 方法类似,若粒的支持度较大则增加样本的惩罚因子,使更多数据分类正确;若粒的分散度大也同样增大样本的惩罚因子,避免分散的测试样本被错误分类.利用偏移因子构造软间隔 SVM 二次优化问题如下:

$$\begin{aligned}\min_{w,b}\quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\delta_i\xi_i,\\ \text{s.t.}\quad & y_i(\boldsymbol{w}^{\mathrm{T}}\cdot\Phi(\boldsymbol{x}_i)+b)\geqslant 1-\delta_i-\xi_i,\\ & \xi_i\geqslant 0,\\ & i=1,\cdots,l.\end{aligned} \tag{11}$$

其对偶问题为

$$\begin{aligned}\max_{\boldsymbol{\alpha}}\quad & W(\boldsymbol{\alpha})=-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)+\\ & \sum_{i=1}^{l}(\alpha_i-\alpha_i\delta_i)\\ \text{s.t.}\quad & \sum_{i=1}^{l}\alpha_i y_i=0,\\ & 0\leqslant\alpha_i\leqslant\delta_i C,\\ & i=1,\cdots,l.\end{aligned} \tag{12}$$

3 数据实验及结果分析

为验证 S_GSVM 模型的有效性,本文将提出的

S_GSVM 方法与基于聚类粒划的 GSVM 模型(记作 C_GSVM)[17],Lagrangian SVM(LSVM)模型进行对比.实验在 1 台 PC 机(2.66 GHz CPU,1 GB 内存)上进行测试,实验平台是 Matlab7.0.相关参数设置如表 1 所示:

Table 1 Experiment Parameters

表 1 实验参数设置

Training Data Size l	Granulation Parameter K
$\leqslant 1\,000$	{10,20,30,40,50}
$>1\,000$	{20,40,60,80,100}

3.1 构造数据集

分别构造正态分布和棋盘状分布的数据集.对于正态分布的训练集,分别以(2,2),(2,−2)为中心,以$\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}$为协方差矩阵构造正类数据集,以(2,2)和(2,−2)为类心的数据集对应的样本规模分别为 100 与 10;同理,分别以(−2,2),(−2,−2)为中心,以$\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}$为协方差矩阵构造负类数据集,以(−2,2)和(−2,−2)为类心的数据集对应的样本规模分别为 10 与 100.测试集构造方法与训练集相同,只是规模相应地扩大 10 倍.训练时采用线性核函数,惩罚参数 C 取 200.对于棋盘状数据集,正负类数据各占棋盘的 8 格,随机生成训练集,正类数据规模为 200×8=1 600 个,负类数据规模为 20×8=160 个.测试集采用相同的方法来生成,规模为训练集的 10 倍.实验采用高斯核,参数 $\sigma=1.0$,惩罚因子 $C=1\,000$.2 种类型数据的分布分别如图 1 和图 2 所示:

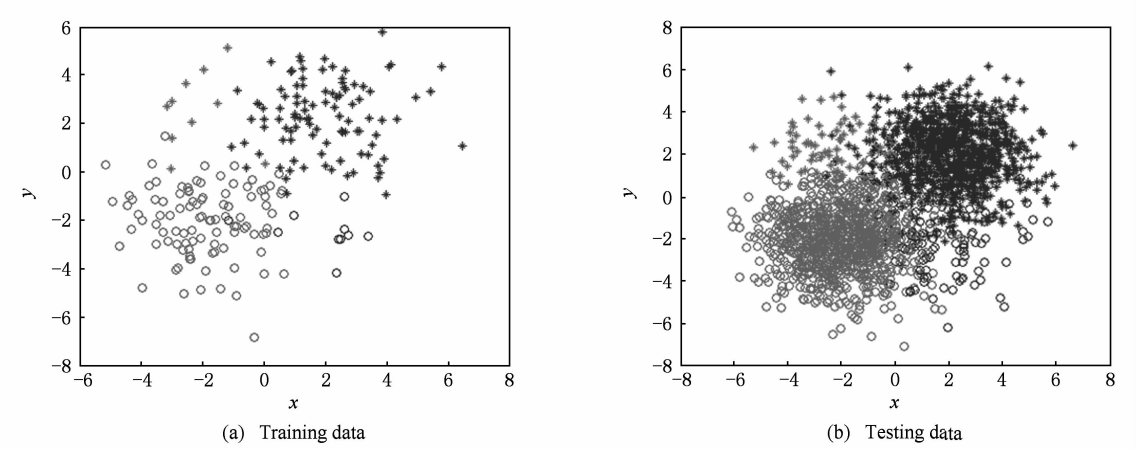

Fig. 1 Distribution of Gaussian data.

图 1 正态分布数据集

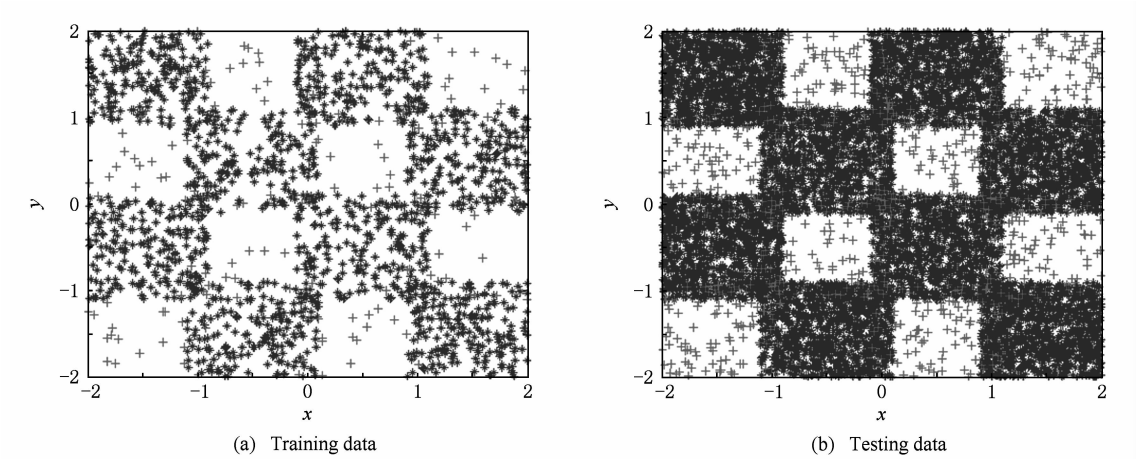

Fig. 2 Distribution of chessboard data.

图 2 棋盘状分布数据集

采用标准 SVM 方法、LSVM 方法和 S_GSVM 方法分别在这两个构造数据集上进行测试,3 种模型得到的实验结果比较分别如表 2 和表 3 所示:

Table 2 Comparison Results by Three Models for Gaussian Data

表 2 正态分布数据集上的实验结果比较

Model	#SV	Accuracy
SVM	33	0.946
LSVM	74	0.931
S_GSVM	20	0.951

Table 3 Comparison Results by Three Models for Chessboard Data

表 3 棋盘状数据集上的实验结果比较

Model	#SV	Accuracy
SVM	196	0.943
LSVM	1 160	0.785
S_GSVM	263	0.956

从表 2 和表 3 可以看出,本文提出的 S_GSVM 方法测试精度与标准 SVM 相比得到了提高. LSVM 只是将传统的 SVM 最优化问题转化为一个无上界约束的二次函数最小值的问题,并利用拉格朗日法求解,因此其在非平衡数据集上存在与 SVM 同样的问题,所以泛化能力也较低. 对于高斯分布的数据集,其样本规模是不平衡的,对于棋盘状分布的数据集其密度是不平衡的. 在同样的参数设置下,S_GSVM 方法表现更好. 因此,当数据样本规模、密度分布不平衡时,不同样本对于分类的错误代价不同,采用传统 SVM 方法得到的超平面不是最优的,而采用本文所提出的基于不同粒赋予不同分类间隔参数和惩罚因子的方法,有望在复杂数据集上得到更好的泛化能力.

3.2 标准数据集

实验采用表 4 所示的 UCI 数据库中 9 个标准数据集[24],每个数据集 5 等分,其中任意 1 份作为训练集,其他 4 份作为测试集,采用交叉验证法并取平均值以减少实验本身的误差.

Table 4 UCI Datasets Used in Experiments

表 4 实验 UCI 数据集

Dataset	Size	Features
Banana	8 800	2
Breast_cancer	2 000	9
Diabetic	4 680	8
German	3 500	20
Image	6 500	18
Spambase	2 500	57
Splice	5 000	60
Thyroid	2 800	5
Titanic	3 000	3

为验证算法性能,在标准数据集上采用 S_GSVM 方法与 C_GSVM,LSVM 方法进行对比,为比较的一致性,实验中所有模型均在进行粒划分后的训练集上训练(此时把 LSVM 记作 L_GSVM,). 以 Banana 数据集为例,数据集的规模为 8 800,维度为 2,分为 5 组,每组共 1 760 个数据,每次采用其中的 1 组作为训练集,其他 4 组作为测试集,取 5 次测试的平均结果作为最终结果. 由于训练集规模大于 1 000,所以粒度参数 K 分别取[20/40/60/80/100]. 采用高斯核函数,参数 $\sigma=1.0$,惩罚因子 $C=1\,000$.

表 5 为 3 种模型实验结果对比,其中带下划线的数据表示每种模型在不同的粒度参数设置下所得到的最优结果,而带方框的数据表示不同模型在同一粒度参数下得到的最优结果,黑体的数据表示每

Table 5 Comparisons of Training and Testing Results for Different Models on Banana Dataset

表 5 Banana 数据集上的实验结果对比

Granulation Parameter K	C_GSVM			L_GSVM			S_GSVM		
	#SV	Training Time/s	Accuracy	#SV	Training Time/s	Accuracy	#SV	Training Time/s	Accuracy
20	20	2.719	0.864	20	2.656	0.859	19	1.998	0.851
40	34	3.093	0.84	32	3.078	0.843	32	3.643	0.859
60	36	3.735	0.854	44	4.61	0.845	35	4.899	0.855
80	40	7.781	**0.878**	49	9.037	0.868	42	7.037	0.878
100	47	9.094	0.871	53	8.687	**0.871**	47	10.069	**0.883**

种模型在不同的参数组合下的最优结果.首先观测不同模型在特定的粒度参数下的对比情况,当粒度参数 $K=20$ 时,C_GSVM 得到测试精度最大,且 S_GSVM 的测试精度略低于其他 2 种模型;除此之外,在其他粒度参数下,本文提出的 S_GSVM 方法都得到了相对于其他方法更为优秀的结果.从算法运行的效率看,本文提出的 S_GSVM 方法与 C_GSVM,L_GSVM 算法的效率几乎不相上下.其次,观测特定模型在不同粒度参数设置下的对比情况,除模型 C_GSVM 外,其他模型均在粒度参数为 $K=100$ 时得到了最大的测试精度值,而 C_GSVM 模型在 $K=100$ 附近所取得的测试精度值也接近于最大测试精度值.

在其他数据集上采用相同的参数实验得到的结果如图 3 所示.图中横轴为模型的粒度参数值,纵轴为测试精度.从实验结果对比可以看出,本文提出的 S_GSVM 方法在 Banana, Breast_cancer, Diabetic, German, Spambase, Splice, Thyroid 这 7 个数据集上得到的测试精度的最优值均要大于其他模型,而在其他 2 个数据集 Image 和 Titanic 上,S_GSVM 模型所得到的测试精度最优值几乎与所有模型的测试精度最优值相差无几.

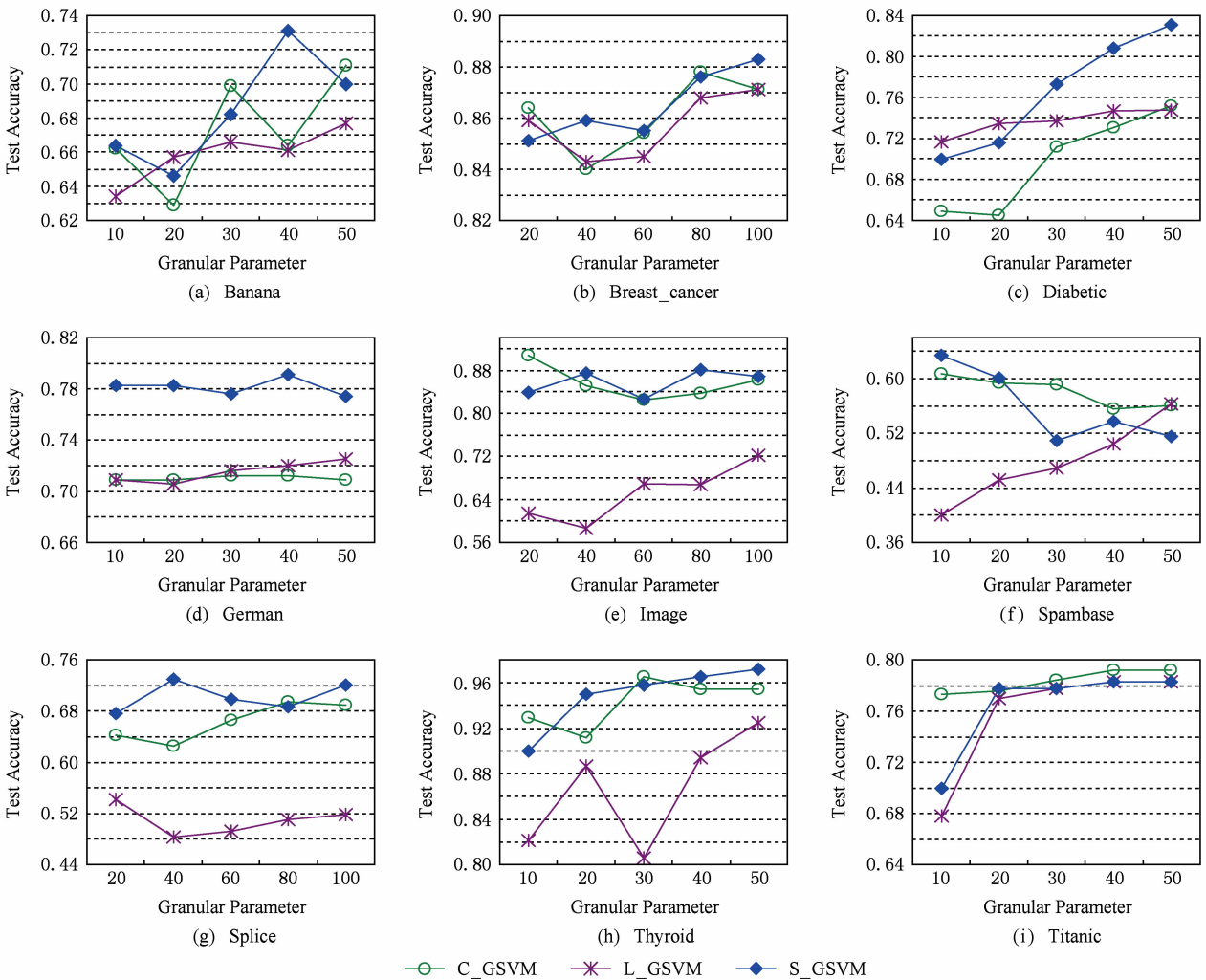

Fig. 3 Comparisons of testing results on different datasets.

图 3 不同模型在标准数据集上的测试精度比较

本文统计了实验采用的 9 个数据集的样本正负类个数及比例,如表 6 所示.

可以发现,除数据集 Splice 外,其他数据集均在样本规模上存在不同程度的不平衡性.

实验中所用数据集正负类样本的平均密度及比例如表 7 所示.从表 7 可以看出,在 5 个数据集 Banana, Breast_cancer, Spambase, Splice, Thyroid 上,正负类样本的平均密度比例超过 1.5.而只有在数据集 German 和 Image 上,正负类样本的平均密度近似相同.

Table 6 Ratios of Positive and Negative Samples

表 6 正负类样本比例

Dataset	# Positive Samples	# Negative Samples	Ratio
Banana	3 893	4 907	1:1.26
Breast_cancer	589	1411	1:2.40
Diabetic	1 639	3 041	1:1.86
German	1 061	2 439	1:2.30
Image	3 962	2 538	1:1.56
Spambase	1 000	1 500	1:1.50
Splice	2 484	2 516	1:1.01
Thyroid	849	1 951	1:2.30
Titanic	981	2 019	1:2.06

Table 7 Average Density Degree Ratios of Positive and Negative Samples

表 7 正负类样本的平均密度比例

Dataset	Average Aensity of Positive	Average Density of Negative	Ratio
Banana	0.807 2	0.524 5	1:1.539 0
Breast_cancer	0.310 5	0.472 1	1:1.520 5
Diabetic	0.351 1	0.420 0	1:1.196 2
German	0.233 1	0.236 2	1:1.013 3
Image	0.247 8	0.258 8	1:1.044 4
Spambase	0.002 2	0.004 9	1:2.227 3
Splice	0.124 6	0.216 8	1:1.740 0
Thyroid	0.287 6	1.074 4	1:3.735 7
Titanic	0.571 8	0.659 2	1:1.152 9

从前面的实验结果可知,S_GSVM 方法在 7 个数据集 Banana,Breast_cancer,Diabetic,German,Spambase,Splice,Thyroid 上得到了较好的测试精度.其中,在数据集 Breast_cancer,Spambase 和 Thyroid 上,样本规模和分布密度的不平衡比例均达到 1.5 以上.而在其他 4 个数据集上,不同类数据的数据规模或分布密度也有较大的不平衡性.从实验所采用的标准数据集来看,数据的规模及密度分布的不平衡性在大多数数据中都是客观存在的,若采用传统 SVM 模型的最大间隔法可能很难得到实际最优的超平面.因此,数据本身分布的不平衡性必须加以考虑,以进一步增强学习器的泛化性能,从实验结果可以看出本文提出的 S_GSVM 模型可以有效地解决样本规模或密度分布不均衡的问题.

3.3 蛋白质互作用数据库

本文还在蛋白质互作用数据库(database of interacting proteins, DIP)上测试 S_GSVM 模型的有效性.DIP 是有关生物体蛋白质间关系预测的数据库①.该数据库是由 D. melanogaster, S. cerevisiae, E. coli, C. elegans, H. sapiens, H. pylori, M. musculus, R. norvegicus 等物种的蛋白质间关系构成.每种物种中的蛋白质关系的数据库又由全部关系(full)及核心关系(core)构成,其中核心关系中蛋白质间的关系是经实验验证后的关系,而全部关系中的蛋白质关系包括没有经过生物实验验证的蛋白质关系.本文取该数据库中最新的(2010-10-10 更新)面包酵母 S. cerevisiae (baker's yeast)中的核心蛋白质关系来实验.

该蛋白质关系数据库包含 4 000 条正类蛋白质对(即有关系的蛋白质对)和 4 000 条负类蛋白质对(即无关系的蛋白质对),每组蛋白质均有其标号,通过蛋白质标号在另一最新的数据库(FULL,该数据库相当于一个蛋白质的氨基酸构成字典,存放着所有蛋白质的氨基酸序列)中查询出相应构成蛋白质的氨基酸序列,然后通过分段局部描述符的编码方式对其进行标准化编码.编码后每个蛋白质均由 630 个特征构成.经过编码后该数据库就由 8 000 条记录构成,每条记录包含两个蛋白质,因此每条记录的特征数为 1 260 个.在实验中,随机取 1 000 条正类样本和 1 000 条负类样本构成训练集,其他样本作为测试集.采用 PCA 方法进行特征提取,分别提取 10,20,30,40,50 个主成分进行分类,采用标准 SVM 及 S_GSVM 方法学习的结果如表 8 所示:

Table 8 Comparisons of Testing Accuracy on DIP

表 8 DIP 数据集上测试结果比较

No. of Principle Components	SVM	S_GSVM				
		$K=20$	$K=40$	$K=60$	$K=80$	$K=100$
10	0.774 2	0.743 8	0.861 0	0.920 6	0.946 2	0.955 3
20	0.807 3	0.778 0	0.890 6	0.957 8	0.978 2	0.942 9
30	0.853 7	0.753 8	0.825 3	0.935 7	0.974 3	0.954 3
40	0.876 9	0.743 5	0.843 5	0.912 4	0.964 7	0.950 8
50	0.883 5	0.805 2	0.860 2	0.925 4	0.958 3	0.930 3
1 260	0.871 2	0.755 0	0.805 5	0.847 4	0.874 3	0.832 8
Average	0.844 5	0.763 2	0.847 7	0.916 6	0.949 3	0.927 7

Note: Positive size 1 000, Negative size 1 000

① 该数据库可从网站 http://dip.doe-mbi.ucla.edu/dip/Main.cgi 下载.

在表 8 中,带下划线的值为 S_GSVM 模型在不同的粒度参数下的测试精度最大值.从表 8 可以看出,10 或 20 个主成分被提取来构造训练集时,若粒度参数 $K\geqslant 40$,本文提出的 S_GSVM 方法的测试精度要优于传统 SVM 方法.当 30,40 或 50 个主成分被提取时,若 $K\geqslant 60$,S_GSVM 比 SVM 得到更好的测试结果.若不采用 PCA 方法进行特征提取,则 S_GSVM 模型在粒度参数 $K=80$ 和 100 时优于传统的 SVM 模型.表 8 中,灰度部分均为 S_GSVM 模型优于传统 SVM 的情况.这就意味着 S_GSVM 能够获得比 SVM 更优秀的泛化能力.此外,当 $K=80$,除了主成分为 10 的情况,S_GSVM 模型均能够获得最优的结果.同时,也可以观测到当粒度参数 K 取 40,60,80,100 时,与传统 SVM 模型相比,S_GSVM 的测试精度分别高出 0.32%,7.21%,10.48% 和 8.32%.当粒划参数 $K=80$,可得到最大的平均测试精度.

在上述实验中,正负类样本的比例为 1∶1.但对于 DIP 数据集本身来说,绝大多数蛋白质组是没有联系的.因此,该数据集本质上是一个规模分布极不平衡的数据集.为进一步测试 S_GSVM 模型的有效性,随机取 100 条正类样本和 2000 条负类样本构成训练集和测试集,比较 SVM 和 S_GSVM 在不平衡正负类样本比例下的实验结果(如表 9 所示):

Table 9 Comparisons of Testing Accuracy on DIP

表 9 DIP 数据集上测试结果比较

No. of Principle Components	SVM	S_GSVM				
		$K=20$	$K=40$	$K=60$	$K=80$	$K=100$
10	0.9262	0.9274	0.9436	0.9489	0.9695	0.9658
20	0.9133	0.9298	0.9459	0.9562	0.9624	0.9695
30	0.9224	0.9206	0.9382	0.9509	0.9581	0.9597
40	0.9362	0.9153	0.9370	0.9383	0.9564	0.9569
50	0.9400	0.9176	0.9301	0.9427	0.9557	0.9497
1260	0.9495	0.9052	0.9228	0.9439	0.9543	0.9466
Average	0.9313	0.9193	0.9279	0.9468	0.9594	0.9580

Note: Positive size 100, Negative size 2000

从表 9 可以看出,当提取 10 或 20 个主成分构造训练集时,若粒度参数 $K\geqslant 20$, S_GSVM 方法测试精度要优于传统 SVM 方法;当提取 30 或 40 个主成分时,若 $K\geqslant 40$,S_GSVM 比 SVM 得到更好的测试结果;当提取 50 个主成分时,若 $K\geqslant 60$, S_GSVM 比 SVM 得到更好的测试结果;类似地,若不采用 PCA 方法进行特征提取,则 S_GSVM 模型在粒度参数 $K=80$ 时优于传统的 SVM 模型.在表 9 中,灰度部分均为 S_GSVM 模型优于传统 SVM 的情况.实验结果表明,与标准 SVM 模型相比,在非平衡数据上S_GSVM 模型具有更加明显的优势.

4 结 语

本文结合粒度计算理论,提出了一个改进的粒度支持向量机模型来解决数据的样本规模和分布密度不平衡性问题.通过为不同的粒计算不同的超平面偏移因子来解决一个新的二次优化问题,从而获得更好的泛化能力.实验结果表明 S_GSVM 模型具有较优秀的性能.在未来的工作中,我们将试图构造一种特殊的核函数,使得映射到核空间后的正类支持向量和负类支持向量能够对称或近似对称.通过这种方式将可能使得学习器得到更为优秀的泛化能力.

参 考 文 献

[1] Vapnik V N. The Nature of Statistical Learning Theory [M]. Berlin: Springer, 1998: 493-520

[2] Wang W J, Men C Q, Lu V Z. Online prediction model based on support vector machine [J]. Neurocomputing, 2008, 71(4/5/6): 550-558

[3] Liu Xiangdong, Luo Bin, Chen Zhaoqian. Optimal model selection for support vector machines [J]. Journal of Computer Research and Development, 2005, 42(4): 576-581 (in Chinese)
(刘向东,骆斌,陈兆乾. 支持向量机最优模型选择的研究[J].计算机研究与发展, 2005, 42(4): 576-581)

[4] Ding Shifei, Qi Bingjuan, Tan Hongyan. An overview on theory and algorithm of support vector machines [J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1): 2-10 (in Chinese)
(丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1): 2-10)

[5] Yao J T. A ten year review of granular computing [C] //Proc of 2007 IEEE Int Conf on Granular Computing (GrC2007). Los Alamitos, CA: IEEE Computer Society, 2007: 734-739

[6] Tang Y C. Granular support vector machines based on granular computing, soft computing and statistical learning [D]. Atlanta: College of Arts and Sciences, Georgia State University, 2006

[7] Osuna E, Freund R, Girosi F. Training support vector machines: An application to face detection [C] //Proc of IEEE Computer Society Conf on Computer Vision and Pattern Recognition (CVPR1997). Los Alamitos, CA: IEEE Computer Society, 1997: 130-136

[8] John C P. Fast Training of Support Vector Machines Using Sequential Minimal Optimization [M] //Advances in Kernel Methods—Support Vector Learning. Cambridge: MIT Press, 1999: 185-20

[9] Chang C C, Lin C J. LIBSVM—A library for support vector machines [EB/OL]. (2005) [2011-01-01]. http://www.csie.ntu.edu.tw/~cjlin/libsvm

[10] Tang Y C, Jin B, Zhang Y Q. Granular support vector machines with association rules mining for protein homology prediction [J]. Artificial Intelligence in Medicine, 2005, 35(1): 121-134

[11] Zhang X G. Using class-center vectors to build support vector machines [C] //Proc of Neural Networks for Signal Processing IX (NNSP1999). Piscataway, NJ: IEEE, 1999: 3-11

[12] Guo H S, Wang W J, Men C Q. A novel learning model—kernel granular support vector machine [C] //Proc of 2009 Int Conf on Machine Learning and Cybernetics (ICMLC2009). Berlin: Springer, 2009: 930-935

[13] Yu H, Yang J, Han J W, et al. Making SVMs scalable to large data sets using hierarchical cluster indexing [J]. Data Mining and Knowledge Discovery, 2005, 11(3): 100-128

[14] Katagiri S, Shigeo A. Incremental training of support vector machines using hyperspheres [J]. Pattern Recognition Letters, 2006, 27(13): 1495-1507

[15] Cheng S X, Shih F. An improved incremental training algorithm for support vector machines using active query [J]. Pattern Recognition, 2007, 40(3): 964-971

[16] Mangasarian O L, David R. Lagrangian support vector machines [J]. Journal of Machine Learning Research, 2001, 1: 161-177

[17] Zhang L, Zhou W D. Density-induced margin support vector machines [J]. Pattern Recognition, 2011, 44(7): 1448-1460

[18] Guo Husheng, Qi Hui, Wang Wenjian. Granular SVM learning algorithm for processing imbalanced data [J]. Computer Engineering, 2010, 36(2): 181-183 (in Chinese)
(郭虎升, 亓慧, 王文剑. 处理非平衡数据的粒度SVM学习算法[J]. 计算机工程, 2010, 36(2): 181-183)

[19] Shawkat A, Smith-Miles K A. A meta-learning approach to automatic kernel selection for support vector machines [J]. Neural Networks, 2006, 24(1/2/3): 173-186

[20] Wang W J, Xu Z B, Lu V Z, et al. Determination of the spread parameter in the Gaussian kernel for classification and regression [J]. Neurocomputing, 2003, 55(3/4): 643-663

[21] Wang W J, Guo J L, Men C Q. An approach for kernel selection based on data distribution [G] //LNAI 5009: Proc of the 3rd Int Conf on Rough Set and Knowledge Technology (RSKT2008). Berlin: Springer, 2008: 596-603

[22] Liao Shizhong, Jia Lei. Constructing a new spherical kernel function [J]. Journal of Computer Research and Development, 2007, 44(Supp Ⅱ): 398-402 (in Chinese)
(廖士中, 贾磊. 一类新的球面核函数的构造[J]. 计算机研究与发展, 2007, 44(增刊Ⅱ): 398-402)

[23] Wu Tao, He Hangen, He Mingke. Interpolation based kernel function's construction [J]. Chinese Journal of Computers, 2003, 26(8): 990-996 (in Chinese)
(吴涛, 贺汉根, 贺明科. 基于插值的核函数构造[J]. 计算机学报, 2003, 26(8): 990-996)

[24] UCI Machine Learning Repository [EB/OL]. (2009-10-16) [2010-11-05]. http://archive.ics.uci.edu/ml

Guo Husheng, born in 1986. PhD candidate. Student member of China Computer Federation. His main research interests include machine learning and data mining.

Wang Wenjian, born in 1968. PhD. Professor. Mmember of China Computer Federation. Her main research interests include machine learning, computing intelligence, etc(wjwang@sxu.edu.cn).