可补偿类别差异的加权支持向量机算法

范昕炜 杜树新 吴铁军
(浙江大学工业控制技术国家重点实验室，浙江大学智能系统与决策研究所，杭州 310027)

摘 要 支持向量机(SVM)算法在各类别样本数多少不同时，样本数量多的类别，其分类误差小，而样本数量少的类别，其分类误差大. 针对这种倾向性问题，在分析其产生原因的基础上，提出了加权 SVM 算法，从而克服了常规 SVM 算法不能灵活处理每一个样本的缺陷，同时补偿了这种倾向性造成的不利影响. 这种以牺牲大类别精度来提高小类别精度的加权支持向量机方法，可应用于诸如故障诊断等关注小类别分类精度的场合. 户外图象识别的实验结果证明，该算法是有效的.

关键词 模式识别(520·2040) 支持向量机(SVM) 分类精度 类别差异 权值 户外图象
中图法分类号：TP391.4 **文献标识码**：A **文章编号**：1006-8961(2003)09-1037-06

Weighted Support Vector Machine Based Classification Algorithm for Uneven Class Size Problems

FAN Xin-wei, DU Shu-xin, WU Tie-jun
(National Laboratory for Industrial Control Technology Institute of Intelligent Systems and Decision Making Zhejiang University, Hangzhou 310027)

Abstract When training sets with uneven class sizes are used, the classification result based on support vector machine (SVM) is undesirably biased towards the class with more samples in the training set. That is to say, the larger the sample size, the smaller the classification error, whereas the smaller the sample size, the larger the classification error. This paper proposes weighted support vector machine algorithms based on the analysis of the cause of such problem, and this algorithm overcomes the drawback which standard support vector machine algorithm can't deal with each sample flexibly and compensates for the unfavorable impact caused by this bias. Such weighted support vector machines improve classification accuracy for class with small size at the cost of accuracy reduction for large size class, and can be applied to the case of regarding small sort of classification accuracy, such as fault diagnosis. The result of outdoor image recognition shows the effectiveness of this algorithm.

Keywords Support Vector Machine(SVM), Classification accuracy, Uneven class sizes, Weight, Outdoor images

0 引 言

在解决小样本、非线性和高维模式识别问题时，支持向量机(Support Vector Machine，简称 SVM)表现出推广能力强、全局最优等许多特有的优势，并能够应用到函数拟合等其他机器学习问题中[1]. 目前，统计学习理论和 SVM 已经成为国际上机器学习领域新的研究热点，并已被应用于人脸识别、文本识别、手写体识别等领域[2~5]. 在对 SVM 的研究中，提高它的分类能力是所有研究的出发点和归宿[·]. 据文献[6]介绍，SVM 在各类别中样本数大小不同时，样本数多的类别，其得到的训练误差和预测误差较小，而样本数少的类别，其训练误差和预测误差较大，这种偏差行为在实际中应尽可能消除，例如故障诊断中，由于异常情况的样本数要比正常情况的样本数少很多，因此故障预测的误差较大，从而导致故障误诊断.

收稿日期：2002-12-05；改回日期：2003-03-24

文献[6]详细分析了 C-SVM 算法中因类别大小不平衡而造成对分类精度影响的原因,并提出了相应的解决方法. 文献[7]提出了用双 ν-SVM 算法来解决 ν-SVM 算法[8]训练类别大小不均衡带来的问题,其是用两个 ν 来修正 ν-SVM 算法,每个 ν 对应一类,以便可以对每一类的支持向量数界限分别进行调整,由于这种调节可以补偿训练类别大小不等的影响,因而可灵活地为每一类指定一个不同的误差比率. 文献[9]提出了一种 FSVM(即 Fuzzy SVM,模糊支持向量机)算法,该算法给每个样本都赋一个模糊隶属度值,这样不同的样本由于对决策函数的学习有不同的贡献,因而可以减少外部和数据中噪声的影响. 但是由于在一些实际应用中,某些重要样本要求小的训练误差,而有些样本数据对误差的要求又不是很高,因此,在优化问题描述时,对每个样本点应采用不同的惩罚系数,以得到更准确的分类,这是本文提出加权支持向量机的初衷.

本文综合了 C-SVM 和 ν-SVM 的问题描述,提出了广义支持向量机,但广义 SVM 与 C-SVM、ν-SVM 一样,在各类别样本数多少不同时,对样本数量多的类别,其训练误差和预测误差小;而对样本数量少的类别,其训练误差和预测误差大. 针对该缺陷,本文在分析了分类倾向性现象产生原因的基础上,提出了加权 SVM 算法,并对类别差异造成的影响进行了相应的补偿,从而提高了小类别的分类精度,这对于某些需要重点关注小类别精度的应用场合有重要的现实意义. 实验结果证明,加权支持向量机克服了广义 SVM 算法不能灵活处理每一个样本的缺陷. 户外图象识别的实验结果已证明了该算法的有效性.

1 广义 SVM 算法和类别大小差异对分类影响的原因分析

1.1 广义 SVM 算法

对于训练向量 $\boldsymbol{x}_i \in \boldsymbol{R}^{(n)}, i=1,\cdots,l$,属于两类,即 $y_i \in \{1,-1\}$,支持向量机的问题可描述为

$$\begin{aligned} &\min_{\omega,b,\xi_i,\rho} \frac{1}{2}\omega^{\mathrm{T}}\omega + C\left(-\nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i\right) \\ &\text{s.t.} \quad y_i(\omega^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b) \geq \rho - \xi_i \\ &\qquad \xi_i \geq 0, i=1,\cdots,l, \rho \geq 0 \end{aligned} \tag{1}$$

在上式中,s.t. 是 subject to 的缩写,表示约束条件的意思. 当 $C=1$ 时,式(1)的优化目标变为 $\min\limits_{\omega,b,\xi_i,\rho} \frac{1}{2}\omega^{\mathrm{T}}\omega - \nu\rho + \frac{1}{l}\sum\limits_{i=1}^{l}\xi_i$,也就是 Schölkoph 和 Smola 提出的 ν-SVM[8,10]. 当 $\nu=0$ 和 $\rho=1$ 时,式(1)的优化目标则变为 $\min\limits_{\omega,b,\xi_i,\rho} \frac{1}{2}\omega^{\mathrm{T}}\omega + \frac{C}{l}\sum\limits_{i=1}^{l}\xi_i$,也就是 Vapnik 提出的 SVM(为了区别,称其为 C-SVM[11]). 式(1)的 SVM 称为广义 SVM 算法. 可见,ν-SVM 算法和 C-SVM 算法都是广义 SVM 算法的一个特例.

显然,上述问题是一个二次规划问题,其最优解为下列拉格朗日函数的鞍点 L,即

$$L(\omega,\xi_i,b,\rho,\alpha_i,\beta,\delta) = \frac{1}{2}\omega^{\mathrm{T}}\omega + C\left(-\nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i\right) - \sum_{i=1}^{l}(\alpha_i(y_i(\omega^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b) - \rho + \xi_i) + \beta_i\xi_i) - \delta\rho \tag{2}$$

其中,α_i, β_i 和 δ 都是拉格朗日乘子,并且,$\alpha_i \geq 0$, $\beta_i \geq 0$ 和 $\delta \geq 0$. 计算该拉格朗日函数鞍点 L 的偏微分,并令偏微分等于零,得

$$\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{l}\alpha_i y_i \varphi(\boldsymbol{x}_i) = 0 \tag{3}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{l}\alpha_i y_i = 0 \tag{4}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{l}C - \alpha_i - \beta_i = 0 \tag{5}$$

$$\frac{\partial L}{\partial \rho} = -C\nu + \sum_{i=1}^{l}\alpha_i - \delta = 0 \tag{6}$$

将式(3)、式(4)、式(5)和式(6)代入式(2),可得到下面式(7)的广义 SVM 对偶最优化问题,这样就把最优超平面的问题转化为一个二次规划问题,然后就可以采用现有的高效算法来求解

$$\begin{aligned} &\min \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ &\text{s.t.} \quad 0 \leq \alpha_i \leq \frac{C}{l}, i=1,\cdots,l \\ &\qquad \sum_{i=1}^{l}\alpha_i y_i = 0 \quad \sum_{i=1}^{l}\alpha_i \geq C\nu \end{aligned} \tag{7}$$

其中,$\frac{C}{l}>0$ 是上界,$K(\boldsymbol{x}_i,\boldsymbol{x}_j) \equiv \varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)$ 是核函数,它有 3 种常用的表达式:多项式函数为 $(\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^{(d)}$,径向基函数为 $\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$ 和 Sigmoid 核函数为 $\tanh[g(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + p]$[1],其中,$d$ 表示多项式的次数,$\sigma>0$,p 表示常数. 支持向量(Support Vector,即 SV)是具有性质 $\alpha_i>0$ 的数据向量;标准支持向量(Normal Support Vector,即

NSV)是具有性质 $0<\alpha_i<\frac{C}{l}$ 的支持向量；边界支持向量(Boundary Support Vector，即 BSV)是具有性质 $\alpha_i=\frac{C}{l}$ 和 $\xi_i>0$ 的支持向量，也是错误分类的样本. 决策函数为

$$\mathcal{F}(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{l} y_i\alpha_i K(\boldsymbol{x}_i,\boldsymbol{x}) + b\right) \qquad (8)$$

其中，参数 b 的计算式为

$$b = \frac{1}{N_{NSV}}\sum_{\boldsymbol{x}_i\in J_{NSV}}\left(y_i - \sum_{\boldsymbol{x}_j\in J_{SV}}\alpha_j y_j K(\boldsymbol{x}_j,\boldsymbol{x}_i)\right) \qquad (9)$$

其中，N_{NSV} 为标准支持向量数，J_{NSV} 为标准支持向量的集合，J_{SV} 为支持向量的集合.

1.2　类别大小差异对分类的影响

训练好的 SVM 间隙宽度在大多数情况下是不为零的，由式(1)的约束条件得到 $\rho>0$. 根据 Kuhn-Tucher 定理[1]，对偶变量与约束的乘积在鞍点处为 0，即 $\delta\rho=0$. 如果 $\rho>0$，那么 $\delta=0$，把它代入式(6)，则可得到

$$\sum_{i=1}^{l}\alpha_i = C\nu \qquad (10)$$

由于边界支持向量具有性质 $\alpha_i=\frac{C}{l}$，对于 N_{BSV} 个边界支持向量的性质 α_i 之和总是小于 $\sum_{i=1}^{l}\alpha_i$，即

$$N_{BSV}\cdot\frac{C}{l}\le\sum_{i=1}^{l}\alpha_i = C\nu \qquad (11)$$

即

$$\frac{N_{BSV}}{l}\le\nu \qquad (12)$$

由于支持向量中的 α_i 最大值是 $\frac{C}{l}$，因此

$$C\nu = \sum_{i=1}^{l}\alpha_i \le N_{SV}\cdot\frac{C}{l} \qquad (13)$$

式中，N_{SV} 表示支持向量的数目，即

$$\nu\le\frac{N_{SV}}{l} \qquad (14)$$

根据式(12)和式(14)可以得到

$$\frac{N_{BSV}}{l}\le\nu\le\frac{N_{SV}}{l} \qquad (15)$$

由式(15)知，由于 ν 既是边界支持向量数与样本数之比的上界，同时又是支持向量数与样本数之比的下界，因此通过 ν 可以有效地控制支持向量的数目. 由于支持向量机有设定支持向量数目下界的显著特征，因此得到的 SVM 有更好的推广性能，其不仅对训练数据集性能好，而且对测试数据集同样有好的性能.

由于根据式(7)的约束条件可得到

$$\sum_i\alpha_i y_i = \sum_{i:y_i=+1}\alpha_i - \sum_{i:y_i=-1}\alpha_i = 0 \qquad (16)$$

因此，式(10)变成

$$\sum_i\alpha_i = \sum_{i:y_i=+1}\alpha_i + \sum_{i:y_i=-1}\alpha_i = 2\sum_{i:y_i=+1}\alpha_i = 2\sum_{i:y_i=-1}\alpha_i = C\nu \qquad (17)$$

把式(17)代入式(15)，由于边界支持向量有 $\alpha_i=\frac{C}{l}$，因此对于 N_{BSV} 个边界支持向量的 α_i 之和总是小于 $\sum_{i:y_i=+1}\alpha_i$，N_{BSV_+} 表示边界支持向量属于正类的数目，即

$$2N_{BSV_+}\cdot\frac{C}{l}\le 2\sum_{i:y_i=+1}^{l}\alpha_i = C\nu \qquad (18)$$

即

$$\frac{2N_{BSV_+}}{l}\le\nu \qquad (19)$$

另外，由于支持向量中的 α_i，其最大值是 $\frac{C}{l}$，N_{SV_+} 表示支持向量属于正类的数目，因此

$$C\nu = 2\sum_{i:y_i=+1}^{l}\alpha_i \le 2N_{SV_+}\cdot\frac{C}{l} \qquad (20)$$

即

$$\nu\le\frac{2N_{SV_+}}{l} \qquad (21)$$

根据式(19)和式(21)可以得到

$$\frac{2N_{BSV_+}}{l}\le\nu\le\frac{2N_{SV_+}}{l} \qquad (22)$$

同理，可以得到

$$\frac{2N_{BSV_-}}{l}\le\nu\le\frac{2N_{SV_-}}{l} \qquad (23)$$

其中，N_{BSV_-} 表示边界支持向量属于负类的数目，N_{SV_-} 表示支持向量属于负类的数目. 用 l_+ 表示正类的数目，l_- 表示负类的数目，$l=l_++l_-$，则式(22)和式(23)可变换为

$$\frac{N_{BSV_+}}{l_+}\le\frac{\nu}{2}\left(1+\frac{l_-}{l_+}\right)\le\frac{N_{SV_+}}{l_+} \qquad (24)$$

$$\frac{N_{BSV_-}}{l_-}\le\frac{\nu}{2}\left(1+\frac{l_+}{l_-}\right)\le\frac{N_{SV_-}}{l_-} \qquad (25)$$

这样，对于样本数大的类别，其错误分类率小，而样本数少的类别，其错误分类率大. 例如正样本数(l_+)大时，$\frac{\nu}{2}\left(1+\frac{l_-}{l_+}\right)$ 小，$\frac{N_{BSV_+}}{l_+}$ 小，即错误分类率

小;反之亦然.这种偏差行为在实际分类中应尽可能消除,例如故障诊断中,由于异常情况的样本数要比正常情况的样本数少很多,因此故障预测的误差较大,从而导致故障误诊断.

2 加权SVM算法

2.1 加权SVM算法的公式

针对广义SVM算法的上述缺陷,本文提出了加权SVM算法,其优化问题描述为

$$\min_{\omega,b,\xi_i,\rho} \frac{1}{2}\omega^{\mathrm{T}}\omega + C\left(-\nu\rho + \frac{1}{l}\sum_{i=1}^{l} s_i\xi_i\right)$$

$$\text{s.t.}\quad y_i(\omega^{\mathrm{T}}\varphi(x_i)+b) \geq \rho - \xi_i$$

$$\xi_i \geq 0, i=1,\cdots,l, \rho \geq 0 \tag{26}$$

式中,s_i表示各样本的权值.可以采用拉格朗日乘子法求解上述优化问题,即

$$L(\omega,\xi_i,b,\rho,\alpha_i,\beta,\delta) = \frac{1}{2}\omega^{\mathrm{T}}\omega + C\left(-\nu\rho + \frac{1}{l}\sum_{i=1}^{l}s_i\xi_i\right) - \sum_{i=1}^{l}(\alpha_i(y_i(\omega^{\mathrm{T}}\varphi(x_i)+b) - \rho + \xi_i) + \beta_i\xi_i) - \delta\rho \tag{27}$$

其中,α_i,β_i和δ都是拉格朗日乘子,并且,$\alpha_i \geq 0$,$\beta_i \geq 0$和$\delta \geq 0$.计算拉格朗日函数鞍点偏微分,并令偏微分等于零,得

$$\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{l}\alpha_i y_i \varphi(x_i) = 0 \tag{28}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{l}\alpha_i y_i = 0 \tag{29}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{C}{l}s_i - \alpha_i - \beta_i = 0 \tag{30}$$

$$\frac{\partial L}{\partial \rho} = -C\nu + \sum_{i=1}^{l}\alpha_i - \delta = 0 \tag{31}$$

把式(28)、式(29)、式(30)和式(31)代入式(27),得到的对偶表达式为

$$\min \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j \alpha_i \alpha_j K(x_i,x_j)$$

$$\text{s.t.}\quad 0 \leq \alpha_i \leq \frac{C}{l}s_i, i=1,\cdots,l$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0 \quad \sum_{i=1}^{l}\alpha_i \geq C\nu \tag{32}$$

该算法增加的系数s_i用于给每一个样本赋权值,它可以是函数,如随样本到达的时间变化的函数,也可以是$0<s_i\leq 1$的常数.如果某些样本对某一个类别来说很重要或者某些被噪声污染的样本点要舍去,则可通过给每个样本赋不同的权值来有效解决这些个性问题,例如,要舍去的样本点,可将其权值取为接近零的数,而将最重要的样本点,其权值取为1.这样就可以克服广义SVM算法在这方面的缺陷.

更重要的是,加权SVM算法可有效地补偿如第1节描述的,因各类别样本数量之间的差异大而引起的倾向性问题.

2.2 加权SVM算法的类别补偿

在加权SVM算法中,通过给不同类别的样本设置一定比例关系的权值就可有效地补偿类别大小差异对分类的影响.例如在两种类别的情况下,正类和负类可通过分别设定不同的权值来消除这种倾向性的影响.对于正类,$s_i=s_+$;对于负类,$s_i=s_-$,则式(32)的约束条件变为

$$0 \leq \alpha_i \leq \frac{C}{l}s_+, \forall\ y_i=+1, i=1,\cdots,l \tag{33}$$

$$0 \leq \alpha_i \leq \frac{C}{l}s_-, \forall\ y_i=-1, i=1,\cdots,l \tag{34}$$

然后根据式(33)和式(34)就可以得到

$$\sum_i \alpha_i = \sum_{i:y_i=+1}\alpha_i + \sum_{i:y_i=-1}\alpha_i = 2\sum_{i:y_i=+1}\alpha_i$$

$$=2\sum_{i:y_i=-1}\alpha_i = C\nu \tag{35}$$

类似式(24)和式(25),应用式(35)得到的边界为

$$2N_{\mathrm{BSV}_+}\frac{C}{l}s_+ \leq C\nu \leq 2N_{\mathrm{SV}_+}\frac{C}{l}s_+ \tag{36}$$

$$2N_{\mathrm{BSV}_-}\frac{C}{l}s_- \leq C\nu \leq 2N_{\mathrm{SV}_-}\frac{C}{l}s_- \tag{37}$$

即

$$\frac{2N_{\mathrm{BSV}_+}s_+}{l} \leq \nu \leq \frac{2N_{\mathrm{SV}_+}s_+}{l} \tag{38}$$

$$\frac{2N_{\mathrm{BSV}_-}s_-}{l} \leq \nu \leq \frac{2N_{\mathrm{SV}_-}s_-}{l} \tag{39}$$

也可以变为

$$\frac{N_{\mathrm{BSV}_+}}{l_+} \leq \frac{\nu l}{2s_+l_+} \leq \frac{N_{\mathrm{SV}_+}}{l_+} \tag{40}$$

$$\frac{N_{\mathrm{BSV}_-}}{l_-} \leq \frac{\nu l}{2s_-l_-} \leq \frac{N_{\mathrm{SV}_-}}{l_-} \tag{41}$$

为了使两类之间得到平衡的误差率,需令$\frac{\nu l}{2s_+l_+}=\frac{\nu l}{2s_-l_-}$,并可以得到如下的比例关系

$$s_+/s_- = l_-/l_+ \tag{42}$$

在加权支持向量机分类方法中,可通过对小类别样本进行加权来提高其分类精度.同时,由于消除了大类别样本的倾向性,因而使得对大类别样本的分类精度有所下降,这种以大类别精度降低作为代价来提高小类别精度的加权支持向量机应用于诸如

故障诊断等关注小类别分类精度的场合，有其重要的实际价值．第 3 节的仿真结果进一步说明了这一问题．

3 户外图象实验仿真

为了验证加权支持向量机分类方法的效果，利用一个户外图象数据集进行了仿真实验．仿真例子来自 UCI 机器学习数据库[11]中的户外图象数据集，每幅图象为一个 3×3 的区域．整个数据集包括 7 种分类，共有2 100个测试样本和 210 个训练样本，每个样本有 19 个属性(维数)．该 19 个属性依次为：区域的中心点的列号、区域的中心点的行号、区域内像素点的个数、短的线密度－5、短的线密度－2、垂直边界像素差异均值、垂直边界的标准方差值、水平边界像素差异均值、水平边界的标准方差值、密度平均值、原红色的均值、原蓝色的均值、原绿色的均值、扩展红色的均值、扩展蓝色的均值、扩展绿色的均值、总颜色的均值、饱和度的均值和色调的均值，这些属性值都是连续量．为简化起见，本文只采用了2 100个测试样本中的其中 4 类：Brickface 为第 1 类、Sky 为第 2 类、Foliage 为第 3 类和 Cement 为第 4 类，每一类为 300 个样本共1 200个样本．在1 200个样本中任意取 60%用于训练，40%用于测试．训练及测试数据的组成如表 1 所示．

表 1　4 种类别 1 200 个样本的组成

类别	Brickface	Sky	Foliage	Cement
训练数据	231	118	122	249
测试数据	69	182	178	51

由表 1 可知，训练数据的各类别样本数目不平衡．为了进行比较，下面用第 1 节中的广义 SVM 算法和第 2 节中的加权 SVM 两种算法来训练分类器，分类器的核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 采用径向基函数，其表达式为 $\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$，其中，$\sigma$ 为 3.08，ν 为 0.5，停止误差为0.000 01．训练时，采用“一对一”方法来构造 $k(k-1)/2$ 个分类器，其中，k 为类别数目，且每个分类器都从两个不同的类别取数据进行训练[12]．加权 SVM 算法根据公式(42)，取 $s_2/s_1=1.96$，$s_3/s_1=1.91$，$s_4/s_1=0.93$，s_1、s_2、s_3 和 s_4 分别代表 4 类样本的权值．分类时，首先根据求到的比例关系来选取权值，再通过训练样本来得到模型，并预测分类．表 2 列出了各类别的训练精度和预测精度，由该表可知．对于训练样本数多的类别 Brickface 和 Cement，广义 SVM 算法的分类精度要高于本文提出的加权 SVM 算法的分类精度；对于训练样本数少的类别 Foliage，广义 SVM 算法的分类精度低于加权 SVM 算法的分类精度．仿真实验结果与第 2.2 节分析相符合，即在加权支持向量机分类方法中，尽管通过对小类别样本进行加权可提高其分类精度，但这是以牺牲大类别样本分类精度为代价的．表 2 说明本文提出的加权 SVM 算法可以有效地提高小样本数目类别的分类精度．

表 2　两种算法各类别的分类精度比较

数据类别		分类精度(%)	
		广义 SVM 算法	加权 SVM 算法
Brickface	测试数据	97.10	95.65
	训练数据	99.13	98.70
Sky	测试数据	100	100
	训练数据	100	100
Foliage	测试数据	88.76	93.82
	训练数据	78.69	87.71
Cement	测试数据	50.98	43.14
	训练数据	97.19	97.19

4 结　论

本文通过广义 SVM 算法中因类别差异而造成的倾向性影响的原因分析，提出了加权 SVM 算法，从而补偿了因训练数据类别大小差异而造成的影响，并减少了训练模型的误差，尤其是提高了小样本类别的分类精度，这对于某些需要重点关注的小类别精度的应用场合有非常重要的现实意义．该算法还克服了广义 SVM 算法不能灵活处理每一个样本的缺陷．其已得到户外图象识别实验结果的印证，并证明了该算法的有效性．

参 考 文 献

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

2　Joachims T. Text categorization with support vector machines [R]. Technical Report, LS VIII Number 23, University of Dortmund, German, 1997.

3　Edgar Osuna, Robert Freund, Federico Girosi. Training support vector machines: An application to face detection[A]. In: IEEE Conference on Computer Vision and Pattern Recognition[C], Puerto Rico, 1997: 130～136.

4　Schmidt M. Identifying speaker with support vector networks [A]. In: Interface'96 Proceedings [C], Sydney, Australia, 1996.

5 Cai Yu-Dong, Liu Xiao-Jun, Xu Xue biao *et al.* Prediction of protein structural classes by support vector machines [J]. Computers and Chemistry, 2002,26(3):293~296.

6 Chew Hong-Gunn, Crisp D J, Bogner R E *et al.* Target detection in radar imagery using support vector machines with training size biasing [A]. In: Proceedings of the Sixth International Conference on Control, Automation, Robotics and Vision[C], Singapore, 2000.

7 Chew Hong-Gunn, Bogner Robert E, Lim Cheng-Chew. Dual nu-support vector machine with error rate and training size biasing[A]. In: Proceedings of 26th IEEE ICASSP (International Conference on Acoustics, Speech, and Signal Processing) 2001 [C], Salt Lake City, UT, USA, 2001:1269~1272.

8 Schölkopf B, Smola A, Williamson R C *et al.* New support vector algorithms[J]. Neural Computation, 2000, 12(5): 1207~1245.

9 Lin Chun Fu, Wang Sheng-De. Fuzzy support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464~471.

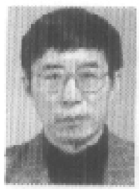
10 Chang Chih-Chung, Lin Chih-Jen. Training nu-support vector classifiers: theory and algorithms [J]. Neural Computation, 2001, 13(9): 2119~2147.

11 UCI 机器学习数据库[EB/OL]. http://www.ics.uci.edu/~mlearn/MLRrepository.htm

12 Hsu Chih Wei, Lin Chih-Jen. A comparison of methods for multi-class support vector machines[J]. IEEE Transactions on Neural Networks 2002, 13(2):415~425.

范昕炜 1973 年生,浙江大学智能系统与决策研究所博士生.现主要研究领域为数据挖掘、模式识别、复杂对象建模和智能控制.

杜树新 1967 年生,1995 年获西北工业大学飞行器控制专业博士学位,现为浙江大学智能系统与决策研究所副研究员.主要研究领域为智能控制和环境工程自动化等领域.



吴铁军 1950 年生,1988 年获浙江大学工业自动化专业博士学位,现为该校智能系统与决策研究所所长,教授,博士生导师.现主要研究领域为智能系统控制与决策及其在动态生产调度、微机器人控制、网络流量控制、环境工程自动化和智能交通系统中的应用.