第35卷 第19期 计 算 机 工 程 2009年10月
Vol.35 No.19 Computer Engineering October 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)19—0181—03 文献标识码：A 中图分类号：TP391.4

模糊支持向量机的偏移量计算方法

陈家德，吴小俊
(江南大学信息工程学院，无锡 214122)

摘 要：偏移量确定了支持向量机和模糊支持向量机(FSVM)的最优分类面位置，对分类性能具有较大影响。为提高模糊支持向量机的识别率，基于Fisher判别分析方法提出一种新的偏移量计算方法，将其用于FSVM多类分类器设计。对3种数据集的测试结果表明，使用新偏移量的FSVM识别率高于使用标准偏移量的FSVM识别率。

关键词：偏移量；支持向量机；模糊支持向量机；机器学习

Offset Calculation Method for Fuzzy Support Vector Machine

CHEN Jia-de, WU Xiao-jun
(School of Information Technology, Jiangnan University, Wuxi 214122)

【Abstract】 Offset determines the position of optimal separating plane of Support Vector Machine(SVM) and Fuzzy Support Vector Machine(FSVM) and affects the performance of classification greatly. In order to improve the recognition rate of FSVM, this paper proposes a new calculation approach for offset based on Fisher discriminant analysis method and uses it to design FSVM multi-classification. Test results of three data sets show that the recognition rate of FSVM using new offset is higher than the one using normal offset.

【Key words】 offset; Support Vector Machine(SVM); Fuzzy Support Vector Machine(FSVM); machine learning

1 概述

支持向量机(Support Vector Machine, SVM)是基于统计学习的机器学习新方法[1]，它具有良好的数学形式、直观的几何解释和良好的泛化能力，解决了模型选择与欠学习、过学习问题以及非线性问题，避免局部最优解，有效克服了“维数灾难”。由于SVM中人为设定的参数较少，因此便于使用，已被成功用于各种分类问题。

SVM方法可以解决两类问题，而处理多类问题时，如果一些输入样本不能被确切归为某一类，则普通SVM将无法顺利运行。针对该问题，文献[2]提出模糊支持向量机(Fuzzy Support Vector Machine, FSVM)，并引入模糊隶属度函数，在一定程度上避免了传统SVM的局限性。SVM和FSVM的分类性能由求解二次规化问题得到的权重和偏移量确定，偏移量过大或过小将直接影响最优分类面位置和分类性能。因此，本文采用一种新的偏移量求解方法。

2 两类支持向量机

支持向量机训练时，多类问题被转化为 n 个两类问题进行训练。在一个两类问题中，m(m<1)维训练集被映射到一维特征空间 Z 中。两类的最优分类超平面基于特征空间中的二次优化问题进行求解。

2.1 最优超平面

SVM由线性可分情况下的最优分类面发展而来，其基本思想如图1所示。在图1中，空心点和方格点代表2类样本；H 为分类线，H_1, H_2 分别为经过各个类中离分类线最近的样本且平行于分类线的直线，它们之间的距离称为分类间隔 $margin$。最优分类线能将2个类正确分开，并使分类间隔最大[3]。

设样本集为(x_i, y_i), $i=1,2,\cdots,n$, $x\in R^d$, $y\in\{+1,-1\}$ ，且满足：

$$y_i[(w\cdot x_i)-b]-1\geq 0 \tag{1}$$

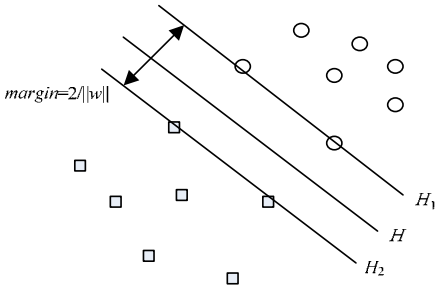

图1 线性可分情况下的最优分类面

此时分类间隔等于 $\frac{2}{\|w\|}$ ，使间隔最大等价于使 $\|w\|^2$ 最小。满足式(1)且 $\frac{1}{2}\|w\|^2$ 最小的分类面即最优分类面，H_1, H_2 上的训练样本点称为支持向量。

利用Lagrange优化方法可以把上述最优分类面问题转化为其对偶问题，即在约束条件 $\sum_{i=1}^{n}a_i y_i=0, a_i\geq 0$, $i=1,2,\cdots,n$ 下对 a_i 求解下列函数的最大值：

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_i a_j y_i y_j (x_i,x_j) \tag{2}$$

a_i 为原问题中与式(1)对应的Lagrange乘子。上述问题是一个不等式约束下二次函数寻优的问题，存在唯一解。容易

基金项目：2006年教育部新世纪优秀人才计划基金资助项目(NCEG-06-0487)；国家自然科学基金资助项目(60472060, 60572034)；江苏省自然科学基金资助项目(BK2006081)

作者简介：陈家德(1980－)，男，硕士，主研方向：模式识别，模糊支持向量机；吴小俊，教授、博士生导师

收稿日期：2009-05-09 **E-mail：**chjd81@163.com

证明，解中只有一部分 a_i 不为零，它们对应的样本即支持向量。求解上述问题得到最优超平面的 w^* 和 b^*，此时最优分类函数为

$$D(x)=\mathrm{sgn}\{(w^*\cdot x)-b^*\}=\mathrm{sgn}(\sum_{i=1}^{n}a_i^*y_i(x_i\cdot x)-b^*) \tag{3}$$

其中，求和部分只对支持向量进行；b^* 是偏移量。

2.2 核函数

对于非线性问题，可以通过非线性变换将其转化为某个高维空间中的线性问题，并在变换空间中求最优分类面。在高维空间中只要进行内积运算，此类内积运算可以用原空间中的函数实现[3]。根据泛函分析的相关理论，只要一种核函数 $K(x_i,y_j)$ 满足 Mercer 条件，它就对应某一变换空间中的内积。

可以用特征空间的 $\varphi(x)$ 代替 x，则式(3)转化为

$$Q(a)=\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_j\varphi(x_i)\cdot\varphi(x_j) \tag{4}$$

若 $K(x,y)$ 满足 Mercer 条件，则令 $K(x,\ y)=\varphi(x)\cdot\varphi(y)$，式(3)转化为

$$D(x)=\mathrm{sgn}(\sum_{\text{支持向量}}a_iy_iK(x_i,x)-b^*) \tag{5}$$

3 “一对一”多类 FSVM

3.1 多类 SVM

在 k 类问题中，该方法共构建 $k(k\text{-}1)/2$ 个 SVM，其中每个 SVM 分开 2 个类别，分类函数如下：

$$D_{ij}=w_{ij}\boldsymbol{x}-b_{ij} \tag{6}$$

其中，w_{ij} 为权重；b_{ij} 为偏移量；$D_{ij}(\boldsymbol{x})=-D_{ji}(\boldsymbol{x})$。对于输入向量 $\boldsymbol{x}$，计算 $D_i(\boldsymbol{x})=\sum^{n}\mathrm{sgn}(D_{ij}(\boldsymbol{x}))$。

$$sign(\boldsymbol{x})=\begin{cases}1 & \boldsymbol{x}>0\\ 0 & \boldsymbol{x}\le 0\end{cases}\pi \tag{7}$$

将 $\boldsymbol{x}$ 划入如下类别：

$$\arg\max_{i=1,2,\cdots,k}D_i(\boldsymbol{x}) \tag{8}$$

采用上述方法时，会出现不确定区域。

3.2 多类 FSVM

为了解决多类 SVM 中的不确定区域问题，文献[4-5]提出 FSVM 的多类算法，在多个类别满足式(8)而造成相同分类结果的情况下，引入模糊隶属度函数。在“一对一”分类模式下，假设最优决策面方程为 $D_{ij}(\boldsymbol{x})=0(j\ne i,i,\ j=1,2,\cdots,k)$，在垂直于 $D_{ij}(\boldsymbol{x})=0$ 方向上定义如下一维隶属函数：

$$m_{ij}(\boldsymbol{x})=\begin{cases}1 & \text{if } D_{ij}(\boldsymbol{x})\ge 1\\ D_{ij}(\boldsymbol{x}) & \text{else}\end{cases} \tag{9}$$

其中，$j\ne i,i,\ j=1,2,\cdots,k$。利用 $m_{ij}(\boldsymbol{x})$ 定义第 i 类隶属函数为

$$m_i(\boldsymbol{x})=\min_{j\ne i,\ j=1,2,\cdots,k}m_{ij}(\boldsymbol{x}) \tag{10}$$

一个未知的待识样本 $\boldsymbol{x}$ 隶属于以下类别：

$$\arg\max_{i=1,2,\cdots,k}m_i(\boldsymbol{x}) \tag{11}$$

图 2 描述了“一对一”方法对不可分区域的划分结果。

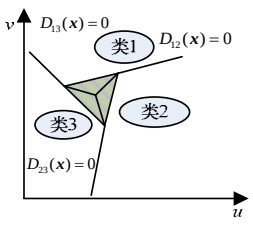

图 2 “一对一”方法对不可分区域的划分结果

4 偏移量

模糊支持向量机的最优分类面由权重 w 和偏移量 b 同时决定，权重确定其形态，偏移量确定最优分类面的位置。通过适当改变偏移量 b，可以平滑地移动分类面而不改变其形状，但对分类结果会产生一定影响。

在特征空间的线性可分两类问题中，令

$$\beta=\min_{z_i}Z_1(w,x_i)$$

$$\alpha=\max_{z_i}Z_2(w,x_i)$$

其中，Z_1 和 Z_2 分别为标记“+1”和“-1”的两类；w 为权重；$x_i\in(Z_1\cup Z_2)$；$\beta>\alpha$。此时，分类间隔转换为 $\pi_\alpha:(w,x_i)-\alpha=0$ 和 $\pi_\beta:(w,x_i)-\beta=0$ 两面之间的距离。在多类问题中，可以利用 α 和 β 来调整偏移量大小。对多类问题中的每个类分别用 $\theta_1.\theta_2,\cdots,\theta_k\ (k\ge 2)$ 标记。本文采用 FSVM 多类问题中的一对一算法，k 类问题分为 $k(k-1)/2$ 个两类问题。将分别标记为 $\{+1,-1\}$ 的两类中的第 1 类训练数据的个数记为 N_1，另一类训练数据的个数记为 N_2。

FSVM 的标准偏移量为

$$b_S=\frac{\alpha+\beta}{2} \tag{12}$$

它与上文用二次规化问题求解的偏移量 b^* 一致，且与每对两类问题中的数据个数相独立。该偏移量不一定是最佳选择。根据 Fisher 判别分析方法[2]可以得到如下偏移量：

$$b_N=d\frac{N_1\alpha+N_2\beta}{N_1+N_2} \tag{13}$$

此偏移量是 α，β 以及每类 N_1 和 N_2 的凸组合。如果 $N_2>N_1$，则偏移量离数据为 N_2 的一类较远，反之则离数据为 N_1 的一类较远。当 $N_2=N_1$ 时，式(13)可以简化为

$$b_N=d\frac{\alpha+\beta}{2}=db_S \tag{14}$$

其中，$0.5<d<1.5$，是新偏项的参数，一般根据数据和先验知识给定。最优分类面中不同偏移量的最优分类线见图 3。

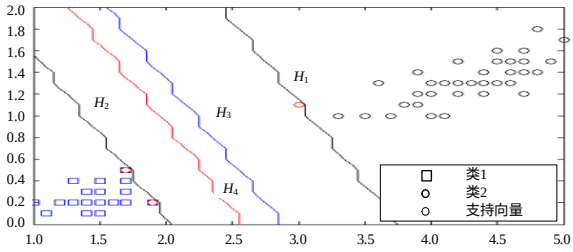

图 3 最优分类面中不同偏移量的最优分类线

图 3 以 Iirs 二维数据为例，其中，空心圆点代表一个类；方框空心点代表另一个类，空心圆点类数据多于方框空心点类数据；H_1,H_2 分别为经过各类中离分类线最近的样本且平行与分类线的直线，线上的点为支持向量点；H_3 为标准偏移量的分类线；H_4 为新偏移量的分类线。图 3 描述了 $N_2>N_1$ 的情形，可以看到分类线由于偏移量不同而发生的位置变化。

由权重和偏移量构成的最优分类面函数在对测试集进行测试前已经确定，因此，无论标准偏移量或新偏移量，在 SVM 和 FSVM 中的求法相同，且参数的选取一致。

5 实验与结果分析

为了验证本文方法的性能，利用 PCA 方法提取 ORL 人脸库数据与水仙花数据进行实验比较。取压缩后维数为 46×56 的每类人脸中的前 5 幅作为训练集，后 5 幅作为测试数据，

并用 PCA 方法提取人脸特征数据。以水仙花数据每类中的前 25 个作为训练集，后 25 个作为测试数据。Wine 数据集的每类数据各不相同。各种数据特征如表 1 所示。

表 1　实验数据的特征

数据集	类别数	维数	每类数据个数	总个数
人脸	40	10	10	400
水仙花	3	4	50	150
Wine	3	13	59, 71, 48	178

用上述“一对一”多类 FSVM 算法，选用多项式核函数，惩罚因子 $C=100$。实验计算机配置为奔腾双核 1.86 GHz CPU, 1 GB RAM，仿真程序用 Matlab 7.0 编写。由图 4~图 6 可知，对于人脸数据、水仙花数据和 Wine 数据，最高识别率分别在新偏移量参数 d 为 0.9, 1.1, 1.0 时取得。

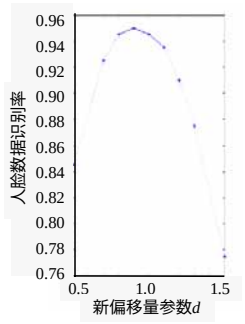

图 4　人脸数据识别率

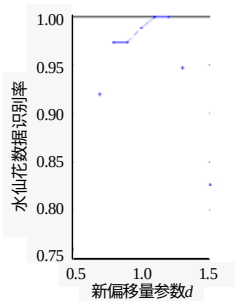

图 5　水仙花数据识别率

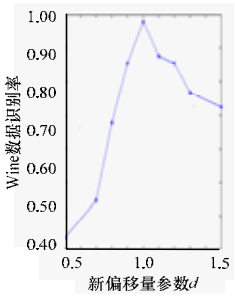

图 6　Wine 数据识别率

由表 2 可以看出，对于人脸数据和水仙花数据，因为每类数据都相同，即 $N_1 = N_2$，所以新偏移量由标准偏移量通过适当调整参数 d 得到。对于 Wine 数据，虽然其每类数据各不同，但通过改变偏移量的求法，即使用新偏移量，其识别率得到了一定提高。

表 2　标准偏移量和新偏移量

数据集	d	b_S/(%)	b_N/(%)
人脸	0.9	94.50	95.00
水仙花	1.1	98.67	100.00
Wine	1.0	94.44	98.89

在水仙花数据中虽然存在交叉点，但由于训练时选取的每类中的前 25 个数据已经包括了交叉点，即指定了其类别，因此识别率可以达到 100%。

由图 4~图 6 可以看出，不同参数求出的偏移量对识别率有一定影响。通过参数调整，对每种数据分类时可以取得一个峰值，其值高于采用原标准偏移量进行分类时的峰值。

6　结束语

本文使用一种新的偏移量计算方法，得到的偏移量对最优分类面的位置进行了适当调整，不会影响 FSVM 的学习能力和外推能力。该方法为偏移量的获取提供了新思路，但仍然需要人工设置参数值而没有实现自适应，有待进一步完善。

参考文献

[1] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2000.

[2] Tsujinishi D, Shigeo A. Fuzzy Lest Squares Support Vector Machines for Multiclass Problem[J]. Neural Networks, 2003, 16(5/6): 785-792.

[3] Vapnik V N. 统计学习理论的本质[M]. 2 版. 北京: 清华大学出版社, 2000.

[4] Shigeo A, Inoue T. Fuzzy Support Vector Machines for Multiclass Problems[C]//Proc. of ESANN'02. Bruges, Belgium: [s. n.], 2002.

[5] Inoue T, Shigeo A. Fuzzy Support Vector Machines for Pattern Classification[C]//Proc. of IJCNN'01. Washington D. C., USA: [s. n.], 2001.

编辑　陈　晖

（上接第 131 页）

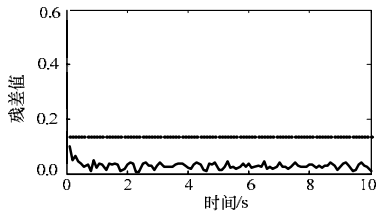

图 4　正常情况下的检测结果

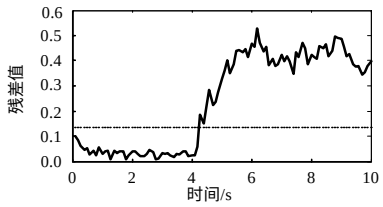

图 5　故障情况下的检测结果

7　结束语

本文研究具有输出时延的网络控制系统，通过将传感器采样周期等分，可将网络时延近似看成控制器端采样周期的整数倍，从而减小时间驱动方式对时延的放大作用。考虑到系统具有外部干扰，设计基于切换系统的鲁棒故障观测器，将故障检测阈值的选取归结为具有线性矩阵不等式约束的最小化问题。下一步研究方向是考虑系统同时具有输出时延和控制时延的情况。

参考文献

[1] Yong Bao, Dai Qiuqiu, Cui Yingliu, et al. Fault Detection Based on Robust States Observer on Networked Control Systems[C]//Proc. of ICCA'05. Budapest, Hungary: [s. n.], 2005.

[2] 于之训, 蒋　平, 陈辉堂, 等. 具有传输延迟的网络控制系统中状态观测器的设计[J]. 信息与控制, 2000, 29(2): 125-130.

[3] 陈雪丽, 张建明, 陈　良. 基于神经元控制的网络时延补偿策略研究[J]. 自动化仪表, 2006, 27(6): 22-25.

[4] Zhang Mingjun, Tarn T J. A Switching Control Strategy for Nonlinear Dynamic Systems[C]//Proc. of ICRA'03. Taipei, Taiwan, China: IEEE Press, 2003.

[5] 俞　立. 鲁棒控制——线性矩阵不等式处理方法[M]. 北京: 清华大学出版社, 2002.

编辑　金胡考