

# 基于边界近邻的最小二乘支持向量机实现

马 波,王正群,侯艳平,邹 军

(扬州大学 信息工程学院,江苏 扬州 225009)

**摘 要:**最小二乘支持向量机采用最小二乘线性系统代替传统的支持向量即采用二次规划方法解决模式识别问题,能够有效地减少计算的复杂性。但最小二乘支持向量机失去了对支持向量的稀疏性。文中提出了一种基于边界近邻的最小二乘支持向量机,采用寻找边界近邻的方法对训练样本进行修剪,以减少了支持向量的数目。将边界近邻最小二乘支持向量机用来解决由  $1-a-r$  (one-against-rest) 方法构造的支持向量机分类问题,有效地克服了用  $1-a-r$  (one-against-rest) 方法构造的支持向量机分类器训练速度慢、计算资源需求比较大、存在拒分区域等缺点。实验结果表明,采用边界近邻最小二乘支持向量机分类器,识别精度和识别速度都得到了提高。

**关键词:**最小二乘支持向量机;一对多方法;边界近邻

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2008)05-0108-04

## Least Squares Support Vector Machine Based on Boundary Nearest

MA Bo, WANG Zheng-qun, HOU Yan-ping, ZOU Jun

(School of Information Engineering, Yangzhou University, Yangzhou 225009, China)

**Abstract:** Least squares support vector machines can reduce the high computational complexity. Duo to equality type constraints in the formulation, the solution follows from solving a set of linear equations, instead of quadratic programming for classical SVM. But unfortunately a very attractive feature of SVM, namely its sparseness, was lost. A new least squares support vector machines based on boundary nearest was proposed, which reduced the number of support vector by using boundary nearest methods pruning the training sample. Meanwhile, uses least square support vector machine for solving multiclass problems constructed by  $1-a-r$  (one-against-rest) method, which overcame the demerits of low training speed, high computational requirements and existing reject area in  $1-a-r$  (one-against-rest) method effectively. Experimental result shows that the performance of accuracy and speed of classifiers are improved after using least squares support vector machines based on boundary nearest.

**Key words:** least squares support vector machine; one-against-rest; boundary nearest

### 0 引 言

支持向量机(SVM)是 Vapnik<sup>[1]</sup>等人提出的一种全新的机器学习方法,由于其优越的学习能力,在国内外学术界受到广泛重视,并在模式识别和函数估计方面取得了越来越多的应用。由于标准的支持向量机需要求解一个受约束的二次规划问题,计算的复杂性较大,为了减少支持向量机的计算复杂度, Suykens<sup>[2]</sup>提出了最小二乘支持向量机(LSSVM),它把支持向量机学习问题转化为线性方程组问题,因此具有较快的运算速度。但最小二乘支持向量机同样存在不足,最小

二乘支持向量机对支持向量失去了稀疏性。针对这一问题许多学者提出了不同的方法<sup>[3~4]</sup>,主要是对样本进行修剪,去掉对应拉格朗日乘子较小支持向量,同时这样会带来分类器性能的下降。

文中提出了一种改进算法——边界近邻最小二乘支持向量机,首先将样本映射到高维空间,利用高维空间的距离公式求出每类的边界样本,最后利用求出的边界样本训练最小二乘支持向量机。考虑到传统的支持向量机是针对两类问题提出的,如何将其有效地推广到多类分类是当前研究的热点。传统的方法是将多类问题转化为多个两类问题,其中最典型的是  $1-a-r$  (one-against-rest)<sup>[5]</sup> 和  $1-a-1$  (one-against-one)<sup>[6]</sup> 方法。这两种多类分类方法对于  $N$  类分类问题分别需要构造  $N$  和  $N(N-1)/2$  个分类器,当类别数目比较多时,这两种算法的速度都比较慢,而且存在不可识别区域。文中将上述的边界近邻最小二乘支持向

收稿日期:2007-08-27

基金项目:江苏省自然科学基金项目(05KJB5201);扬州大学自然科学基金(KK0413160)

作者简介:马 波(1981-),男,硕士研究生,研究领域为模式识别与神经网络;王正群,博士,副教授,研究领域为模式识别、机器学习。

量机运用到由  $1 - a - r(\text{one-against-rest})$  方法构造的分类器中,有效地解决了训练样本之间的不平衡对精度产生的影响和存在拒分区域的缺点。

## 1 最小二乘支持向量机

最小二乘支持向量机与标准的支持向量机的不同在于它把不等式约束改成了等式约束,并把经验风险的一次方改为二次方。最小二乘支持向量机的实现如下:

对于一个给定的训练数据集:

$$(x_i, y_i), \quad i = 1, 2, \dots, l, \quad x_i \in R^n, \quad y_i \in R$$

利用高维特征空间中的线性函数:

$$y(x) = w^T \phi(x) + b \quad (1)$$

来拟合同本集,其中非线性映射  $\phi(\cdot)$  把数据从输入空间映射到高维特征空间,以便使输入空间中的非线性拟合问题变成高维特征空间中的线性拟合问题。根据结构风险最小化原理,支持向量机问题可以表示为约束优化问题(下式中  $d_k$  表示类标签):

$$\min_{w, b, e} J_p(w, e) = \frac{1}{2} w^T w + \frac{C}{2} \sum_{k=1}^N e_k^2 \quad (2)$$

$$\text{s. t.} \quad d_k [w^T \phi(x_k) + b] = 1 - e_k$$

为了解上述优化问题,把约束优化问题转化成无约束优化问题,建立 Lagrange 函数:

$$L(w, b, e, a) = J_p(w, e) - \sum_{k=1}^N a_k \{d_k [w^T \phi(x_k) + b] - 1 + e_k\} \quad (3)$$

根据 KKT(Karush - Kuhn - Tucker) 条件有:

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^N a_k d_k \phi(x_k) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^N a_k d_k = 0 \\ \frac{\partial L}{\partial e_k} = 0 \rightarrow a_k = C e_k \\ \frac{\partial L}{\partial a_k} = 0 \rightarrow d_k [w^T \phi(x_k) + b] - 1 + e_k = 0 \end{cases} \quad (4)$$

从方程组(4)中消去  $e_k, w$  后,可以得到:

$$\begin{bmatrix} 0 & d^T \\ d & \Omega + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{1} \end{bmatrix} \quad (5)$$

其中  $d = [d_1, d_2, \dots, d_N]^T, a = [a_1, a_2, \dots, a_N]^T,$

$$\Omega_{i,j} = d_i d_j K(x_i, x_j), \vec{1} = [1, 1, \dots, 1]^T$$

支持向量机的输出为:

$$y(x) = \sum_{k=1}^N a_k d_k(x, x_k) + b \quad (6)$$

从上面可以看出,最小二乘支持向量机的训练问题归结为一个线性方程组的求解问题,而不像标准支持向量机那样求解一个二次规划问题,这样的转化使

求解更为简单快速。但是最小二乘支持向量机将不等式约束条件转变成了等式的约束条件,这样使得最小二乘支持向量机算法对支持向量失去了稀疏性。

对于监督学习而言,不可否认的是训练样本越多越好,训练样本越多越能够得到更多的样本分布的结构信息,经过学习可以得到泛化能力更强的支持向量机。但是对于大样本来说,参数的搜寻是比较困难的,因此,如果能够减少训练样本的数量,则必然可以增加搜寻的速度。随着训练样本的减少,得到的支持向量的个数也在减少,从而提高了分类器的速度。现在已有一些精简训练样本的方法,如利用自组织理论方法,利用 K 均值原理方法等。利用这些方法对训练样本进行精简然后训练支持向量机取得了一定的效果,但是这些方法共同的问题是没有对边界近邻样本给予更多的关注。对于分类问题,邻界样本包含了分类的重要信息,最难分类的最容易引起分类错误的样本也集中在邻界附近,如果失去了训练样本的邻界信息,必然使得 SVM 不能更好地体现样本的分布信息。

## 2 边界近邻最小二乘支持向量机

文中针对这些问题,提出了一种改进的最小二乘支持向量机——边界近邻最小二乘支持向量机。首先将训练样本变换到高维空间,然后在高维空间中求出每类的边界近邻样本,最后用边界样本来训练最小二乘支持向量机。具体实现方法如下:

选择任意一类中的一个样本  $x$ , 利用公式(7)求高维空间距离所选择类最近的另一类样本点,其中  $k(x_1 \cdot x_1) = (\phi(x) \cdot \phi(y))$  是核函数,即为边界近邻样本。

$$D(x, y) = \|\phi(x) - \phi(y)\|_2 = \sqrt{k(x \cdot x) - 2k(x \cdot y) + k(y \cdot y)} \quad (7)$$

这里选择出来的边界样本为最小二乘支持向量机训练的负样本,同理运用公式(7)将另一类样本作为所选择类的样本,求出边界近邻样本,这些边界样本作为支持向量机训练的正样本。然后利用所选择出来的训练样本对支持向量机进行训练。由于边界样本的数目远远小于原训练样本的数目,这样训练时间可以明显减少。

## 3 基于边界近邻最小二乘支持向量机的多分类问题解决

传统的支持向量机是针对两类问题提出来的,而实际应用中常常出现的是多类问题。因此将 SVM 应用于多类问题对挖掘 SVM 的应用潜力具有非常重要

的意义。其中  $1-a-r(\text{one-against-rest})$  方法是应用比较广泛的一种。在该分类方法中对  $n$  个类别仅需要构造  $n$  个支持向量机, 每一个支持向量机分别将某一类与其它类别中分离开来。在分类时, 取决策函数输出值最大的类别为测试样本的类别。

根据  $1-a-r(\text{one-against-rest})$  方法设计支持向量机简单、有效、训练时间较短, 可用于大规模数据, 但其缺点在于:

1) 当类别数较大时, 某一类的训练样本将大大少于其它类训练样本的总和, 这种训练样本之间的不平衡将对精度产生影响;

2) 存在误分、拒分区域。

文中将上述的边界近邻最小二乘支持向量机与  $1-a-r(\text{one-against-rest})$  方法结合运用到多类情况。由于边界近邻最小二乘支持向量机训练样本的选择采用了将训练样本映射到高维空间然后求取边界样本的方法, 对于某一类训练样本其边界的训练样本正负样本的数目应很接近, 这样就有效地解决了  $1-a-r(\text{one-against-rest})$  方法中由于样本之间的不平衡对精度产生的影响。同时最小二乘支持向量机的输出为测试样本与分类超平面之间的距离, 这样对于  $N$  类分类问题, 将产生  $N$  个分类超平面, 利用样本距分类超平面正向距离最远的分类器作为测试样本的输出类别。这样也有有效地解决了  $1-a-r(\text{one-against-rest})$  方法中存在拒分区域不足的问题。

算法具体描述如下:

Input: data set  $S$ , Regularization factor  $C$ , Kernel function parameter  $\sigma^2$ . Output: Class of input sample  $k$ .

(1) Produce the trainset  $s_T$  and valifiset  $s_v$  from the date set  $S$ .

(2) Choose the parameter  $C$  and  $\sigma^2$ , and using equation (7) get the boundary samples  $x_{ij}$  of each class.

(3) Train the LS-SVM based on the boundary samples train set.

(4) Test the LS-SVM performance, if the performance worse, change the parameter  $C$  and  $\sigma^2$ , go to step (1); else go to step(5).

(5) Using LS-SVM test the class of input sample  $k$ .

## 4 实验结果与分析

为了验证文中所提出的边界近邻最小二乘支持向量机的有效性, 实验采用了将边界近邻最小二乘支持向量机与最小二乘支持向量机比较的方法。实验是在 PC 机上 (256M 内存, 40G 硬盘) 完成的, 采用 Matlab 语言实现。

实验分两部分:

(1) 利用二维平面的模拟数据验证边界最小二乘支持向量机对于两类情况的有效性;

(2) 将边界最小二乘支持向量机与  $1-a-r(\text{one-}$

$\text{against-rest})$  方法相结合对标准数据集中的多类情况进行分类。

### 4.1 实验一

在归一化的二维平面  $X-Y$  上用随机函数产生两类数据点, 以式(8)所示的双曲线作为分类的边界 (如图(1)所示), 实验分别随机产生 200, 400, 800 个数据点, 每个数据集随机分成 5 组, 其中 1 组作为测试集, 其余 4 组的并集作为初始训练集。因此, 对分类器集成的性能测试, 采用 5 倍交叉验证方法, 5 组示例中, 每 1 组都有一次作为测试集。实验采用将文中所提出的边界近邻最小二乘支持向量机与最小二乘支持向量机进行比较的方法。实验结果如表(1)、表(2)所示 (表中为 5 次测试结果的平均值)。本实验中选取高斯核作为核函数, 具体形式如式(9)。

$$(X-0.5)^2 - (Y-0.5)^2 = 0.01 \quad (8)$$

$$K(x, x_i) = \exp\left[-\frac{|x - x_i|^2}{2\sigma^2}\right] \quad (9)$$

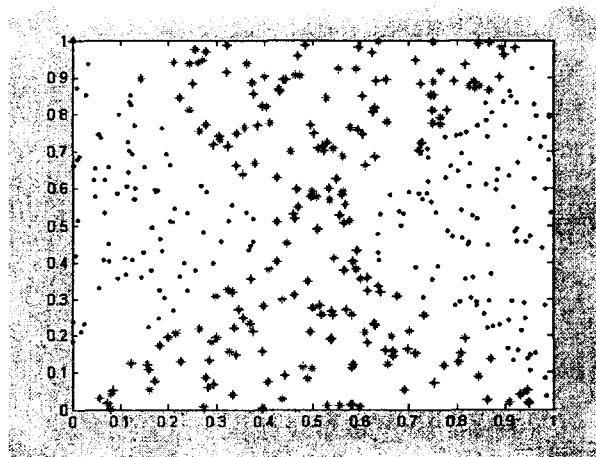


图 1 二维平面实验数据

表 1 采用最小二乘支持向量机实验结果

数据集	参数选择	支持向量	训练样本集		测试样本集	
			分类正确率 (%)	学习时间 (s)	分类正确率 (%)	花费时间 (s)
200	$\sigma = 0.2$ $C = 100$	160	96.3	0.45	95.0	0.094
400		320	96.5	2.03	94.5	0.422
800		640	96.1	11.5	95.1	1.656

表 2 采用边界近邻最小二乘支持向量机实验结果

数据集	参数选择	支持向量	训练样本集		测试样本集	
			分类正确率 (%)	学习时间 (s)	分类正确率 (%)	花费时间 (s)
200	$\sigma = 0.2$ $C = 100$	35	98.9	0.688	96.5	0.031
400		60	99.8	2.860	98.5	0.079
800		79	99.7	11.04	99.3	0.219

从实验结果可以看出由于采用了边界近邻方法, 可以明显减少支持向量的数量, 从而减少了样本的测试时间, 提高了样本的识别率。因刚开始由于样本的数

目较少,采用边界近邻最小二乘支持向量机由于需要寻找边界近邻样本,在训练时间上高于最小二乘支持向量机,但随着样本数量的增加,边界近邻最小二乘支持向量机训练时间将小于最小二乘支持向量机。所以从总体上来看,边界近邻最小二乘支持向量机在提高识别率的同时也减少了样本的训练时间。

#### 4.2 实验二

本实验的数据均来自 UCI 数据库<sup>[7]</sup>,实验数据如表 2 所示。在实验过程中,Iris 和 Letter 均采用原始数据;对于 Wine 和 Segment,由于属性值之间差异太大,故利用式(10)作了预处理。其中  $x_i$  是属性值, $\bar{x}$  是均值, $S$  是标准差。

$$x'_i = (x_i - \bar{x})/S \quad (10)$$

实验数据和实验结果分别如表 3 和表 4 所示,从实验结果可以看出运用边界最小二乘支持向量机来处理多类问题从总体上提高了样本的识别精度和识别速度。但对于大样本问题,由于样本的数量比较多,边界近邻最小二乘支持向量机产生的支持向量也比较多,这样增加了样本的学习时间。实验中可以采用聚类的方法减少类中心样本,以减少边界样本的搜寻时间来减少样本的学习时间。

表 3 实验数据列表

数据名称	训练集	测试集	类别数	属性值
Iris	75	75	3	4
Wine	118	59	3	13
Segment	1540	770	7	19
Letter	16000	4000	26	16

表 4 实验结果

数据名称	训练集		测试集	
	分类正确率(%)	学习时间(s)	分类正确率(%)	花费时间(s)
Iris	97.33	1.860	96.00	0.016
Wine	99.85	1.375	98.33	0.031
Segment	99.74	92.35	95.58	0.042
Letter	98.93	2653.4	94.30	56.38

(上接第 107 页)

都是利用商业化的桥接产品。这里对 CORBA 与 DCOM 桥接方面的研究,是为了探讨更好的桥接模式,开发更好的桥接产品,推动分布式应用的发展。

#### 参考文献:

- [1] Kraus E. Interworking Methodologies for DCOM and CORBA[D]. Tennessee: Faculty of Computer and Information Science, East Tennessee State University, 2003.

## 5 结束语

针对多分类方法中的 1-a-r(one-against-rest)方法存在由于训练样本之间的不平衡将对精度产生影响,存在不可识别区域等不足,提出了利用边界近邻最小二乘支持向量机来解决多分类问题。有效地解决了 1-a-r 方法中存在的由于训练样本之间的不平衡将对精度产生影响及存在拒分区域的问题,实验结果表明该方法的有效性。然而,文中的方法还存在一定的局限性。如在训练过程中的参数如何确定,如何减小噪声对训练精度的影响等,这些问题都有待于进一步的探索研究。

#### 参考文献:

- [1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995.
- [2] Suykens J A K, Vandewalle J. Least Squares Support Vector Machine Classifiers[J]. Neural Processing Letter, 1999, 9(3): 293-300.
- [3] De Kruif B J, De Vries T J A. Pruning error minimization in least squares support vector machines[J]. IEEE Trans. Neural Networks, 2003, 14(3): 696-702.
- [4] Hoegaerts L, Suykens J A K, Vandewalle J, et al. A comparison of pruning algorithms for sparse least squares support vector machines[C]//in: Proceeding of International Conference on Neural Information Processing 2004. Calcutta, India: [s. n.], 2004: 1247-1253.
- [5] Botton L, Cortes C, Denker J, et al. Comparison of Classifier Methods: A Case Study in Hand-writing Digit Recognition [C]//International Conference on Pattern Recognition. [s. l.]: IEEE Computer Society Press, 1994: 77-87.
- [6] Krebel U. Pairwise Classification and Support Vector Machines [C]//Scholkopf B, Burges C J C, Smola A J. Advances in Kernel Methods; Support Vector Learning. MA: MIT Press, 1999: 255-268.
- [7] UCL Machine Learning Group. Elena database[EB/OL]. 2003. <http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/elena.htm#stuff>.

- [2] OMG. Common Object Request Broker Architecture: Core Specification[M]. [s. l.]: OMG, 2004.
- [3] Norman R J. CORBA and DCOM: Side by Side[EB/OL]. 1998-05. [www.DistributedComputing.com](http://www.DistributedComputing.com).
- [4] 钟灿, 钟本善, 周熙襄. COM 和 CORBA 的桥接应用[J]. 电子科技大学学报, 2003, 32(2): 188-191.
- [5] Tolba M F, Fathy S K, Ismail H M. Design and Implementation of Interworking Architecture[J]. IJICIS, 2003, 3(1): 10-21.