一种稀疏最小二乘支持向量机

赵 会,黄景涛
ZHAO Hui,HUANG Jing-tao

河南科技大学 电子信息工程学院,河南 洛阳 471003
College of Electronic and Information Engineering,Henan University of Science and Technology,Luoyang,Henan 471003,China
E-mail:zhaohuihaha@126.com

ZHAO Hui,HUANG Jing-tao.Sparse least squares support vector machine.Computer Engineering and Applications,2009,45(26):40-42.

Abstract: To solve the problem of sparseness lacking in the Least Squares Support Vector Machine(LS-SVM),a new least squares support vector machine based on the boundary samples is proposed,which uses center distance ratio to select bigger support value boundary samples and making them as training samples.Thus,the number of support vector is reduced and the speed of computing is improved.Finally,the new algorithm is tested on the four benchmarking UCI datasets.The result shows that the proposed algorithm can adaptively obtain the sparse solutions almost not losing generalization performance,and the speed of classifiers is also improved.

Key words: sparseness;least squares support vector machine;center distance ratio;boundary sample

摘 要:针对最小二乘支持向量机缺乏稀疏性的问题,提出了一种基于边界样本的最小二乘支持向量机算法。该算法利用中心距离比来选取支持度较大的边界样本作为训练样本,从而减少了支持向量的数目,提高了算法的速度。最后将该算法在4个UCI数据集上进行实验,结果表明:在几乎不损失精度的情况下,可以得到稀疏解,且算法的识别速度有了一定的提高。

关键词:稀疏性;最小二乘支持向量机;中心距离比;边界样本
DOI:10.3778/j.issn.1002-8331.2009.26.011 **文章编号:**1002-8331(2009)26-0040-03 **文献标识码:**A **中图分类号:**TP181

1 引言

支持向量机(Support Vector Machine,SVM)是建立在统计学习理论[1]之上的一种新的机器学习方法,它具有出色的学习泛化能力,近年来已被成功地应用在模式识别、函数回归、预测模型、图像处理等领域。经典的SVM的训练算法需要求解一个凸二次规划问题,计算的复杂性较大,为减小支持向量机的计算复杂度,Suykens等人[2]1999年提出了最小二乘支持向量(Least Squares Support Vector Machine,LS-SVM),该算法只需通过求解一组线性方程而获得最优分类面,学习速度较快,而且内存需求少。但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[3-5],主要是通过修剪样本,去掉拉格朗日乘子相对较小的支持向量,从而提升算法的稀疏性。但这些方法都没有对边界近邻样本给予更多的关注。对于分类问题,边界样本包含了分类的重要信息,最难分类和最易引起分类错误的样本也集中在边界附近,因此如果失去了训练样本的边界信息,会使SVM不能更好地体现样本的实际分布信息。

该文是在文献[2]的基础上提出了一种基于边界样本的最小二乘支持向量机(Boundary samples Least Squares SVM,BLS-SVM)分类算法,该算法利用高维空间中样本点的中心距离比求出每类的支持度较大的边界样本,而忽略那些支持度较小的样本,并利用求出的边界样本训练最小二乘支持向量机,从而提高训练速度。

2 最小二乘支持向量机

和传统的SVM方法类似,LS-SVM也是通过构造最优分类超平面实现分类的。对于一个给定的训练样本集:(x_k,y_k),$k=1,\cdots,N$,$x_k\in R^m$,$y_k\in R$,x_k是第k个输入样本,y_k是第k个输出样本,构造的最优分类超平面为:

$$y(x)=\text{sign}\left[\sum_{k=1}^{N}\alpha_k y_k K(x,x_k)+b\right] \tag{1}$$

其中,$\alpha_k\in R$,$K(x,x_k)$为核函数,核函数形式有很多,均采用Gaussian核,即:

$$K(x,x_k)=\exp\left(-\frac{\|(x-x_k)^2\|}{2\sigma^2}\right) \tag{2}$$

根据结构风险最小化原则,最小二乘支持向量机问题可以表示为约束优化问题:

基金项目:河南省重点科技攻关项目(No.082102210015);河南科技大学青年基金(No.2007QN041)。
作者简介:赵会(1982-),女,硕士研究生,主要研究方向:模式识别与智能系统;黄景涛(1977-),男,博士,副教授,主要研究方向:数据挖掘和智能优化算法及应用。
收稿日期:2009-01-07 **修回日期:**2009-03-11

$$\min_{w,b,e} J_p(w,e)=\frac{1}{2}w^{\mathrm{T}}w+\frac{1}{2}\gamma\sum_{k=1}^{N}e_k^2 \tag{3}$$

$$\text{s.t. } y_k[w^{\mathrm{T}}\varphi(x_k)+b]=1-e_k,k=1,\cdots,N \tag{4}$$

其中,w 为权系数向量,γ 为错分惩罚参数,e_k 为误差。$\phi(\cdot)$为非线性函数,把数据映射到高维空间。

定义 Lagrangian 函数:

$$L(w,b,e,\alpha)=J_p(w,e)-\sum_{k=1}^{N}\alpha_k\{y_k[w^{\mathrm{T}}\varphi(x_k)+b]-1+e_k\} \tag{5}$$

根据 KKT(Karush-Kuhn-Tucker)条件有:

$$\frac{\partial L}{\partial w}=0\rightarrow w=\sum_{k=1}^{N}\alpha_k y_k\varphi(x_k) \tag{6}$$

$$\frac{\partial L}{\partial b}=0\rightarrow\sum_{k=1}^{N}\alpha_k y_k=0 \tag{7}$$

$$\frac{\partial L}{\partial e_k}=0\rightarrow\alpha_k=\gamma e_k \tag{8}$$

$$\frac{\partial L}{\partial \alpha_k}=0\rightarrow y_k[w^{\mathrm{T}}\varphi(x_k)+b]-1+e_k=0 \tag{9}$$

从上面可知,α 的每个分量均与样本的误差成正比,因此支持向量失去了稀疏性。

3　边界样本最小二乘支持向量机

通过研究发现,LS-SVM 的最优解中所有支持度 α_k 都不为零,每个训练样本点的支持度对最优解的贡献不同[5],支持度 $|\alpha_k|$ 大的训练样本点对最优解的贡献大,支持度 $|\alpha_k|$ 小的训练样本点对最优解的贡献小,而每个训练样本点支持度 $|\alpha_k|$ 的大小与该点到最优分类面的距离有关,即:

$$d_k=\frac{(w^{\mathrm{T}}x_k+b)}{\|w\|}=\frac{(1-e_k)}{y_k\|w\|} \tag{10}$$

而在最小二乘支持向量机中,$\alpha_k=\gamma e_k$

$$d_k=\frac{(w^{\mathrm{T}}x_k+b)}{\|w\|}=\frac{(1-e_k)}{y_k\|w\|}=\frac{(1-\alpha_k/\gamma)}{y_k\|w\|} \tag{11}$$

$$\Rightarrow\alpha_k=\gamma(1-d_k y_k\|w\|) \tag{12}$$

由式(11)、式(12)可知,y_k 只能等于±1、$\|w\|$ 近似不变。在 r 选定后,当 $d_k\ll1$ 或 $d_k\gg1$ 时,$|\alpha_k|$ 较大。即当样本点离最优分类面很近或很远时,$|\alpha_k|$ 都较大,当样本点离最优分类面距离中等时,$|\alpha_k|$ 较小。

基于该理论,提出了一种基于边界样本的最小二乘支持向量机(BLS-SVM)算法,BLS-SVM 算法分为两步:(1)从所有训练样本中挑选出那些中心距离比较大的样本作为支持度 $|\alpha_k|$ 可能大的边界训练样本,以减少训练样本数量,提高算法速度;(2)利用所选择的样本进行一次最小二乘支持向量机训练,求出近似最优解。BLS-SVM 算法根据各类中样本点的中心距离比来选择边界训练样本,减少了支持向量的数量,因而具有稀疏性,而且只需要进行一次 LS-SVM 训练,就可以获得近似最优解,因此算法的识别速度较快。

依据上面的分析,从支持向量几何角度出发,选择那些对最优分类面贡献大的边界样本进行训练,而忽略那些贡献小的训练样本,在保证算法精度的同时减少参与训练的样本数,从而提高算法的速度。

定义 1(距离)给定两个训练样本 x_i、x_j,则两样本点在再生核 Hilbert 空间的距离可表示为:

$$d(\phi(x_i),\phi(x_j))=\sqrt{K(x_i,x_i)-2K(x_i,x_j)+K(x_j,x_j)}=\|\phi(x_i)-\phi(x_j)\|^2 \tag{13}$$

定义 2(中心距离)就是样本到中心的距离。假设两类训练样本数分别为 N_+、N_-。则两类的类中心[6]分别为:$\phi(x_c^+)=\frac{1}{N_+}\sum_{i=1}^{N_+}\phi(x_i^+)$、$\phi(x_c^-)=\frac{1}{N_-}\sum_{i=1}^{N_-}\phi(x_i^-)$,且对每个正样本来说,都有两个中心距离:自中心距离 $D_{si}^+=d(\phi(x_i^+),\phi(x_c^+))$和互中心距离 $D_{mi}^+=d(\phi(x_i^+),\phi(x_c^-))$,同理对于每个负类样本点的自中心和互中心距离分别为:$D_{si}^-=d(\phi(x_i^-),\phi(x_c^-))$、$D_{mi}^-=d(\phi(x_i^-),\phi(x_c^+))$。

定义 3(中心距离比)已知两类样本,某一类中样本点 x_i 的自中心距离和互中心距离的值分别为 D_{si} 和 D_{mi},则该样本点的中心距离比:$R_i=D_{si}/D_{mi}$。

算法具体实现步骤如下:

(1)根据定义 2 求出两类样本的类中心,并计算两类类中心之间的距离 $D=d(\phi(x_c^+),\phi(x_c^-))$。

(2)分别计算各类训练样本所构成的近似平均半径。在计算之前,首先去除类中的奇异点,即那些离类中心的距离值非常大(约占该类样本点数量的 5%左右)的样本点,其中 $\bar{R}_+=\frac{\sum_{k=1}^{N_+}D_{si}^+}{N_+}$,$\bar{R}_-=\frac{\sum_{k=1}^{N_-}D_{si}^-}{N_-}$。

(3)利用定义 3 分别求出高维空间中每类样本点的中心距离比集合 R。若已知正类训练样集合 $X^+=\{x_i|x_i\in R^m,y_i=1,i=1,2,\cdots,N^+\}$,则 $R^+=\{R_i^+|R_i^+\in D_{si}^+/D_{mi}^+,i=1,2,\cdots,N^+\}$,同理求 R^-。

(4)设置阈值,选取支持度较大的边界样本。因为最优分类面总是位于两类训练样本所构成的凸包之间。因此根据各类训练样本中心距离比 R_i,结合两类类中心之间的距离 D 以及近似平均半径 $\bar{R}_+$、$\bar{R}_-$,来确定用于选取支持度较大的边界样本的阈值。

支持度较大样本的选取方法:当 $R_i^+>T_+$ 和 $R_i^->T_-$ 时,则该训练样本选定为可能支持度较大的样本。这里 T_+、T_- 分别是两类样本进行选择的阈值,可以不相等。

当 $D>\bar{R}_+-\bar{R}_-$,T_+、T_- 可以取较大值,如 $T_+=\bar{R}_+/D-\mu$,$T_-=\bar{R}_-/D-\mu$($\mu=\bar{R}/4$)。

当 $D<\bar{R}_+-\bar{R}_-$,T_+、T_- 可以取较小值,如 $T_+=\bar{R}_+/D-\mu$,$T_-=\bar{R}_-/D-\mu$($\mu=\bar{R}/2$)。对均衡数据,边界样本数可取同一个比例初值,对于非均衡数据,选取边界样本的比率和原始训练样本数成反比(改变 μ 的值可以改变所选取边界样本的数量)。

4　仿真实验

为了验证算法的有效性,进行了一系列实验:即先按照上述方法根据样本的中心距离比选择边界样本,然后采用 LS-SVM 算法包进行训练[7]。

(1)对两类分类的实验分析

在实数空间上随机生成两类正态分布的样本集,如图 1 所

示。它们的中心分别位于μ_1=[-0.5;-0.5]和μ_2=[-0.5;0.5],协方差矩阵$\boldsymbol{C}_1$=$\boldsymbol{C}_2$=0.062 5I。实验分别随机产生200,400,800个数据点,每个数据集随机等分成5组,其中一组作为测试集,其余4组的并集作为初始训练集。对分类器集成的性能测试,采用5倍交叉验证方法,实验结果如表1、表2所示(表中为5次测试结果的平均值)。

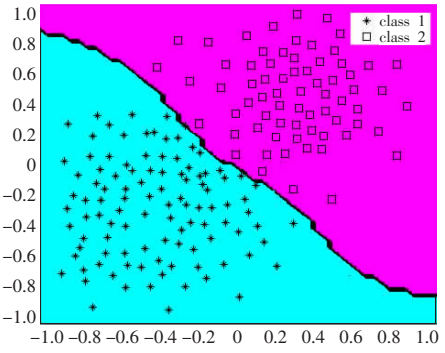

图1 二维平面实验数据

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	σ^2=2	160	94.01	0.247	96.83	0.092
400	r=10	320	99.36	0.978	98.75	0.062
800		640	99.38	3.469	98.65	0.079

表2 采用边界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	σ^2=2	36	97.95	0.125	96.47	0.016
400	r=10	76	98.82	0.641	98.52	0.035
800		92	98.98	3.172	97.86	0.047

由表1和表2得知,对于两类分类情况,在采用相同的核函数和相同的γ时,与LS-SVM相比,BLS-SVM算法具有如下优点:弥补了稀疏性,减少了存储空间和计算量,提高了训练和预测的速度,且训练样本集及测试样本集的准确率基本不受影响。

(2)对多类分类的实验分析

本实验分别使用BLS-SVM分类方法和LS-SVM方法对UCI数据集[8](如表3所示)中几种多类数据进行实验,在进行两种方法训练时,γ=10,核参数σ^2=0.8,因为边界样本一般占训练样本总数的20%~30%,故在BLS-SVM算法中,以选取1/3左右的训练样本为标准,来决定阈值。对于多类数据分类,采用o-a-r(one-against-rest)[9]方法进行分类。每个数据集的测试实验共进行5次,表4、表5给出的是5次总的平均值。实验平台为PC机(Intel 1.6G/1G DDR)。

从实验结果可以看出运用BLS-SVM和LS-SVM方法处理同一个训练样本集时,前者只使用了支持度较大的边界样本作为训练样本,减少了训练样本的数量,因此算法的速度得到了一定的提高。尤其当训练样本数量较大时,BLS-SVM算法比LS-SVM的速度快的多,且分类器在训练样本集和测试样本集上的准确率基本没有改变。

表3 实验数据

数据集	类别数	训练样本集	测试样本集	属性
IRIS	3	75	75	4
WINE	3	118	59	13
Stimage	7	4 435	2 000	36
Shuttle	7	43 500	14 500	9

表4 采用最小二乘支持向量机实验结果

数据集	训练样本集		测试样本集	
	分类正确/(%)	学习时间/s	分类正确/(%)	花费时间/s
IRIS	97.31	0.172	94.06	0.172
WINE	99.23	0.343	96.83	0.265
Stimage	98.70	35.290	92.71	0.062
Shuttle	95.14	274.580	92.76	72.080

表5 采用边界最小二乘支持向量机实验结果

数据集	训练样本集		测试样本集	
	分类正确/(%)	学习时间/s	分类正确/(%)	花费时间/s
IRIS	97.10	0.128	92.12	0.097
WINE	98.75	0.310	98.57	0.133
Stimage	99.19	23.250	90.88	0.042
Shuttle	93.87	184.780	94.53	53.210

对非均衡数据集(WINE),选取边界样本时,选取的数量和原始训练集两类数据量的大小约成反比,这样就能使得两类训练样本的数量基本上达到均衡,有效地解决了传统支持向量机构造的多分类器方法中因样本之间的不均衡对精度产生影响的问题。

通过以上实验分析验证了边界样本最小二乘支持向量机算法的有效性和可行性:即该方法有效地利用了中心距离比值,通过选择支持度较大的边界样本的方法来减少训练样本的数量,从而在保证算法推广能力的同时也提高了支持向量机的训练速度。

5 结束语

根据支持向量、中心距离比以及边界向量之间的关系,提出了一种基于边界样本的最小二乘支持向量机算法。该算法利用高维空间中样本点的中心距离比选择支持度较大的边界样本进行LS-SVM训练,在保证算法推广能力的同时,减少了训练样本的数量,提高了算法的训练速度。另外,有效地解决了传统支持向量机构造的多类分类器方法中存在的因训练样本之间的不均衡对精度产生影响的问题。文中的算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地探索研究。

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] Hoegaerts L,Suykens J A K,Vandewalle J,et al.A comparison of pruning algorithms for spares least squares support vector machines[C]// Proceeding of International Conference on Neural Information,Calcutta,India,2004:1247-1253.

(下转77页)