基于结构风险上界的 SVM 参数选择

宋小杉[1],蒋晓瑜[1],罗建华[2],汪 熙[1]

1. 装甲兵工程学院控制工程系,北京 100072
2. 装甲兵工程学院科研部,北京 100072

摘要 提出了基于结构风险上界的 SVM 参数选择方法。首先,从理论上分析了 SVM 结构风险上界的计算方法,给出了结构风险上界的算法步骤;其次,以结构风险上界作为 SVM 泛化性评价准则对 5 个 UCI 公开数据库和经过实测建立的两个特征库(包括二类和多类数据)进行了参数选择仿真实验,并与 5-折交叉验证的实验结果进行了比较,结果表明,基于结构风险上界的 SVM 参数选择方法有效、省时。

关键词 支持向量机;结构风险上界;参数选择

中图分类号 TP18 **文献标识码** A **doi** 10.3981/j.issn.1000-7857.2011.08.012

SVM Parameter Selection Based on the Bound of Structure Risk

SONG Xiaoshan[1], JIANG Xiaoyu[1], LUO Jianhua[2], WANG Xi[1]

1. *Department of Control Engineering, Academy of Armored Force Engineering, Beijing 100072, China*
2. *Department of Science Research, Academy of Armored Force Engineering, Beijing 100072, China*

Abstract Support Vector Machine (SVM) is an intelligent technology for classification problems. Because of its flexibility, computational efficiency and capacity to handle high dimensional data, SVM has become a popular research issue in recent years. Selection of optimal parameters is important for an SVM. The traditional methods, such as the k-fold cross validation, can select optimal parameters, but would take too much time. In this paper, a method of SVM parameter selection based on the bound of structure risk is proposed. First, the bound of the structure risk is theoretically analyzed. Then, the simulated experiments with several datasets are designed. Comparisons are made between the proposed method and the method based on the 5-folds cross validation. The results show that the proposed method is effective and takes less time, and it would be very suitable for target recognition problems.

Keywords support vector machine; bound of structure risk; parameter selection

0 引言

支持向量机(Support Vector Machine,SVM)建立在统计学习理论(Statistic Learning Theory,SLT)中的 VC 维理论和结构风险最小化原理的基础上[1],根据有限样本信息在模型的复杂性和学习能力之间寻求最佳折中,以期获得最好的泛化能力。SVM 已经被看作是对传统学习方法的一个好的替代,特别在小样本、高维非线性情况下,具有较好的泛化性能[1]。目前,SVM 正在成为继神经网络研究后新的研究热点。

参数选择直接影响 SVM 目标识别性能的优劣。如何确定 SVM 参数是研究 SVM 的重要内容,也是 SVM 的研究热点[1-9]。SVM 参数选择的传统方法是 k-折交叉验证[1-2],张学工等[3]最早引进 SVM,于 2002 年提出基于变异函数的径向基函数(Radial Basis Function,RBF)参数估计方法。2005 年,Ayat[4]提出了一种基于 SVM 分类错误概率估计的 SVM 性能评价准则,并与 GACV 和 VC 维准则进行比较,显示了该方法选择参数的有效性,但该方法实现较为复杂。近几年,人们将遗传算法和粒子群算法应用到 SVM 参数选择中[5-8],Mu 等[9]将参数选择看作非线性动态系统, 利用扩展卡尔曼滤波算法进行选择。这些算法均得到了较好的结果,但他们没有从 SVM 的本质作用入手研究,所采用的 SVM 评价方法基于经验风险或交叉验证,均存在一定局限。

SVM 的本质是要得到期望风险最小的判别函数,但期望风险往往不能直接计算出来,故采用结构风险对期望风险进行近似。结构风险反映了 SVM 的经验风险和泛化性的总体要

收稿日期:2010-08-06;修回日期:2011-02-16
基金项目:"十一五"装备预研项目(2009YY02)
作者简介:宋小杉,博士研究生,研究方向为目标探测与识别,电子信箱:sxsh029@yahoo.com.cn;蒋晓瑜(通信作者,中国科协所属全国学会个人会员登记号:S040420880S),教授,研究方向为战场多传感器信息融合、电子稳像、目标探测与识别等,电子信箱:jiangxiaoyu2007@gmail.com

求,但目前并没有普适的计算公式,而结构风险上界则可以通过公式计算出来。本文对 SVM 结构风险上界进行了理论分析,给出了结构风险上界的算法步骤,并提出了基于结构风险上界的参数选择方法。

1　结构风险上界的计算

下面定理给出了结构风险与经验风险之间的关系,它是分析 SVM 性能的重要基础。

定理 1 (结构风险上界)[10]　对于判别函数集中的所有函数,结构风险和经验风险之间至少以概率 $1-\eta$ 满足如下关系:

$$R(\alpha)\leqslant R_{\text{emp}}(\alpha)+\sqrt{\frac{h[\ln(2l/h)+1]-\ln(\eta/4)}{l}} \tag{1}$$

其中,$R(\alpha)$为结构风险,$R_{\text{emp}}(\alpha)$为经验风险,l 为样本数,h 为函数集的 VC 维。

由定理 1 可知,SVM 的风险由两部分组成:一部分为经验风险(训练误差);另一部分称做置信范围,它与 SVM 的 VC 维 h 和训练样本数 l 有关。

对于样本集 $D=\{(x_i,y_i),i=1,2,\cdots,l\}$,$x_i$ 为样本向量,$y_i=\{1,-1\}$为样本所属类别,SVM 的经验风险定义为

$$R_{\text{emp}}(\alpha)=\frac{1}{l}\sum_{i=1}^{l}L(y,f(x_i,\alpha)) \tag{2}$$

式中 $f(x_i,\alpha)$为 SVM 对样本 x_i 的预测类别。$L(y,f(x,\alpha))$为 SVM 用于模式识别的损失函数:

$$L(y,f(x,\alpha))=\begin{cases}1 & y=f(x,\alpha)\\ 0 & y\neq f(x,\alpha)\end{cases} \tag{3}$$

因此,模式识别的经验风险实际上是对训练数据集进行识别的错误率,即

$$R_{\text{emp}}(\alpha)=\frac{N_{\text{err}}}{l} \tag{4}$$

其中,N_{err} 为 SVM 对测试集识别错误的样本个数。

定理 2 (VC 维上界)[10]　设向量 $\boldsymbol{x}\in X$ 属于一个半径为 r 的球中,那么 Δ-间隔分类超平面集合的 VC 维 h 以下面不等式为界:

$$h\leqslant\min\left(\left[\frac{r^2}{\Delta^2}\right],n\right)+1 \tag{5}$$

其中,r 为特征空间中覆盖所有样本的最小超球半径,Δ 为特征空间中分类超平面的分类间隔,n 为特征空间的维数。

式(5)中

$$\frac{1}{\Delta^2}=\|w\|^2=\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(\varphi(x_i),\varphi(x_j))$$
$$=\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{6}$$

其中,$K(x_i,x_j)$为核函数,表示特征空间中两个向量之间的内积。

设

$$m=\frac{1}{l}\sum_{i=1}^{l}\Phi(x_i)$$

是特征空间中所有特征向量的中心,则超球半径 r^2 可表示为

$$r^2=\max(\|\Phi(x_i)-m\|^2)=\max(\|\Phi(x_i)\|^2+\|m\|^2-$$
$$2(\Phi(x_i),m))=\max\left(K(x_i,x_j)+\right.$$
$$\left.\frac{1}{l^2}\sum_{i=1}^{l}\sum_{j=1}^{l}K(x_i,x_j)-\frac{2}{l}\sum_{i=1}^{l}K(x_i,x_j)\right) \tag{7}$$

由定理 1 知,置信区间可写为

$$\phi\left(\frac{l}{h}\right)=\sqrt{\frac{h[\ln(2l/h)+1]-\ln(\eta/4)}{l}} \tag{8}$$

随着样本数 l 的增大,置信区间单调减少,而式(1)成立的概率单调增大,故有[10]

$$\eta=\frac{1}{\sqrt{l}} \tag{9}$$

将式(9)代入式(8)可得置信区间为

$$\phi\left(\frac{l}{h}\right)=\sqrt{\frac{h[\ln(2l/h)+1]+\ln 4+\ln(\sqrt{l})}{l}} \tag{10}$$

综上,结构风险上界的计算步骤如下。

(1) 用 SVM 对样本集进行训练得到判别函数。

(2) 根据式(3)计算经验风险 R_{emp}。

(3) 将式(6)、式(7)代入式(5)计算 VC 维上界 h_{bound}:如果 $r^2/\Delta^2>n$,$h_{\text{bound}}=n+1$;如果 $r^2/\Delta^2<n$,$h_{\text{bound}}=r^2/\Delta^2+1$。

(4) 根据式(10)计算置信区间 Φ。

(5) 将 R_{emp}、h、Φ 代入式(1)得到结构风险上界 R_{bound}。

2　基于结构风险上界的参数选择

SVM 的参数包括惩罚因子和核参数。惩罚因子影响着 SVM 决策函数对特征向量的错分程度,一般用 C 表示,C 越大,越不允许错分。而核参数则因核函数的不同而不同。SVM 常用的核函数有多种,其中 RBF 具有较少的参数和广泛的普适性,在 SVM 的应用中是首要选择[2],其表达式为

$$K(x,x')=\exp(-\gamma\|x-x'\|^2) \tag{11}$$

其中 γ 为核参数。

2.1　算法

基于结构风险上界,采取网格法进行参数选择,为穷极搜索范围,C 的取值集合为 $\{2^{-5},2^{-4},\cdots,2^{10}\}$,$\gamma$ 为 $\{2^{-10},2^{-9},\cdots,2^{5}\}$,则参数选择具体算法如下。

(1) 从参数取值集合中依次取出一对参数(C,γ)。

(2) 设定一个长度为 16×16 的数组 $R_{\text{bound}}[\,]$。

(3) 根据第 1 节计算结构风险上界 $R_{\text{bound}}(i,j)$,并将其值装入数组 $R_{\text{bound}}[\,]$中。

(4) 比较数组 $R_{\text{bound}}[\,]$中所有值的大小,其最小值对应的参数对即为最优参数对。

2.2　仿真实验

由于高斯核函数的普适性[2],本文主要以高斯型支持向量机进行实验,用基于结构风险上界的方法对各数据库进行 SVM 参数选择,并与基于 5-折交叉验证的方法进行比较。5-折交叉验证是一种经典的 SVM 性能评价准则,研究证明,5-

折交叉验证能够较好地评价 SVM 性能，但其时间开销较大，在 l 个样本中完成一次参数选择需要 $16\times16\times5\times l^2$ 次运算，而基于结构风险上界的方法则需要 $16\times16\times2\times l^2$ 次运算。

实验采用台湾林智仁的 Libsvm 软件为核心代码[2]。实验数据一部分来自加州大学欧文分校网站的公开实测特征数据库[11] Heart、Australian、German、Vehicle 和 Satimage，另一部分为经过实测建立的小物件特征库和装甲车辆特征库。其中小物件特征库中包含 4 类目标，分别是充电器、打火机、瓶盖和钥匙；装甲车辆特征库包含 2 类目标，分别是国产某型坦克和某型步战车。每个特征库包含一个训练集和一个测试集，它们是独立同分布的两个数据集。SVM 基于训练集进行训练得到 SVM 判别函数，然后在测试集上进行识别测试。7 组实验都是在 CPU 主频为 1.85GHz、内存为 512MB 的 PC 机上用 C 语言编程完成的。结果如表 1 所示。

表 1　两种参数选择方法的实验比较

Table 1　Comparison between results of the two methods

数据	训练集样本数/测试集样本数	结构风险上界			5-折交叉验证		
		最优参数	测试准确率/%	所用时间/s	最优参数	测试准确率/%	所用时间/s
Heart	100/170	C=1,γ=0.0625	83.5294	2.812	C=1,γ=0.0909	82.9412	5.906
Australian	200/490	C=1,γ=0.0625	87.1166	7.968	C=1,γ=0.3333	84.8671	12.656
German	600/400	C=1,γ=0.0625	75.4386	91.876	C=1,γ=0.1000	72.6817	187.425
Vehicle	400/445	C=512,γ=0.5000	77.8400	50.183	C=16384,γ=0.0624	77.3000	121.560
Satimage	800/3635	C=64,γ=0.1250	82.1900	332.512	C=8,γ=1.0000	81.8800	857.987
小物件	144/68	C=64,γ=0.5000	94.1200	0.856	C=64,γ=0.5000	94.1200	1.987
装甲车辆	720/100	C=512,γ=2.000	92.0000	27.920	C=512,γ=2.0000	92.0000	50.796

表 1 中，Heart、German、Australian 和装甲车辆均为 2 类数据，Vehicle 和小物件均为 4 类数据库，而 Satimage 则是包含了 6 类目标的多类数据，实验对多类分类采取一对一策略[1-2]。可以看出，① 对所有 7 组实验，基于结构风险的 SVM 参数选择所用的运算时间更短、速度更快；② 对前 5 组实验，基于结构风险上界的 SVM 参数选择结果，对测试集的识别率明显高于基于 5-折交叉验证参数选择结果，而后两组实验中两种方法所得到的参数相同，但本文提出的方法用时较少；③ 本文提出的方法对 2 类和多类均适用。

3　结论

SVM 参数选择直接影响 SVM 泛化性的大小，从而影响基于 SVM 的目标识别准确率。如何进行参数选择从而省时有效地得到最优参数，是目前的研究热点。结构风险上界可以较好地反映 SVM 期望风险的大小，是一种较为理想的 SVM 泛化性评价标准，以其为基础进行 SVM 参数选择，可以得到使 SVM 泛化性较强的参数。本文给出了结构风险上界的算法，提出了一种基于结构风险上界的 SVM 参数选择方法，对 5 个 UCI 公开实测数据库和经过实测建立的 2 个数据库进行 SVM 参数选择仿真实验，并与 5-折交叉验证进行了比较。结果表明，基于结构风险上界的方法对 2 类分类 SVM 模型和多类分类 SVM 模型均可适用，是一种有效、省时的 SVM 参数选择方法。

参考文献(References)

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

[2] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification [R/OL]. [2008-06-16]. http://www.csie.ntu.edu.tw/cjlin/papers/guide/guide.pdf.

[3] 阎辉, 张学工, 马云潜, 等. 基于变异函数的径向基核函数参数估计[J]. 自动化学报, 2002, 28(3): 450-455.
Yan Hui, Zhang Xuegong, Ma Yunqian, *et al*. *Acta Automatica Sinica*, 2002, 28(3): 450-455.

[4] Ayat N E, Cheriet M, Suen C Y. Automatic model selection for the optimization of SVM kernels[J]. *Pattern Recognition*, 2005, 38(10): 1733-1745.

[5] Huang C L, Wang C J. A GA-based feature selection and parameters optimization for support vector machines [J]. *Expert Systems with Applications*, 2006, 31(2): 231-240.

[6] Yu Q, Zhang B H, Wang J L. Parameter optimization of e-SVM by genetic algorithm[C]. 5th International Conference on Natural Computation, Tianjin, China, Aug 14-16, 2009.

[7] Guo X C, Yang J H, Wu C G, *et al*. A novel LS-SVM hyper-parameter selection based on particle swarm optimization[J]. *Neurocomputing*, 2008, 71(16-18): 3211-3215.

[8] Zhang X Y, Guo Y L. Optimization of SVM parameters based on PSO algorithm [C]. 5th International Conference on Natural Computation, Tianjin, China, Aug 14-16, 2009.

[9] Mu T T, Nandi A K. Automatic tuning of L2-SVM parameters employing the extended Kalman filter[J]. *Expert Systems*, 2009, 26(2): 160-175.

[10] Vapnik V. Statistical learning theory[M]. New York: J Wiley, 1998.

[11] Blake C L, Merz C J. UCI repository of machine learning databases[EB/OL]. [2008-08-28]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

(责任编辑　代丽)