维普资讯 http://www.cqvip.com

支持向量机泛化能力估计若干方法*)

宋晓峰[1] 陈德钊[2] 胡上序[2]

(南京航空航天大学生物医学工程系 南京210016)[1]

(浙江大学智能信息工程研究所 杭州310027)[2]

摘 要 支持向量机性能很大程度上依赖于其参数,为了确定最优参数,往往需要估计所建模型的泛化能力。本文详细介绍了目前国外支持向量机的各种泛化能力估计方法,分析了各种泛化能力估计方法的特点及其关系,并指出了今后的发展方向。

关键词 支持向量机,泛化能力估计

Survey of Generalized Performance Estimation in Support Vector Machine

SONG Xiao-Feng[1] CHEN De-Zhao[2] HU Shang-Xu[2]

(Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)[1]

(Laboratory of Intelligent Information Engineering, Zhejiang University, Hangzhou 310027)[2]

Abstract The performance of support vector machine is dependent on its parameters to a great extent. In order to determine the parameters of SVM, we should estimate the SVM's generalization performance firstly. Some kinds of generalization performance estimators in overseas are introduced. We also analyze the characteristics of various generalization performance estimator, and indicate the trend of research on the generalization performance estimator.

Keywords Support vector machine, Generalization performance estimation

1 引言

自1995年Vapnik等提出统计学习理论(Statistical Learning theory或SLT)后,基于SLT理论的支持向量机(Support Vector Machine或SVM)技术,日渐成为国际学术界的研究热点。SVM用于模式分类和回归估计时显示出前所未有的良好性能,这主要归因于SVM摒弃了经验风险最小化(ERM)原则,而采用结构风险最小化(SRM)原则[1,2]。Vapnik等分析了模型复杂性与预测能力之间的矛盾,进而提出SRM原则,均衡地考虑对数据的拟合精度和拟合函数的复杂性,从而使模型的拟合偏差较小,且有较强的泛化推广能力,而这正是建模所追求的重要目标。

支持向量机不存在结构设计的问题,其结构可由算法自动确定,因此支持向量机的主要参数对其回归估计性能影响很大,一般通过交叉验证或bootstrapping的方法选定最佳参数,但步骤繁琐,计算量大,为此,人们希望能通过更为简便的方法选定最佳参数,在将全部样本作为训练集来构造支持向量机时,就需要估计支持向量机的泛化误差,以此来确定其最佳参数。目前,人们已提出了多种估计支持向量机泛化能力的方法。

2 支持向量机简介

SVM最初来自对模式分类问题的处理,假定已知样本,

$(x_1,y_1),\cdots,(x_l,y_l),x\in R^n,y\in\{+1,-1\}$

可以被一个超平面$(w\cdot x)-b=0$分开,SVM寻求的是这样一个分割平面:使得训练数据集中的点距离这一分类面尽量的远,也即使其两侧的空白区域尽量的大。数学描述如下,

在满足式:$y_i[(w\cdot x_i)-b]\geqslant 1,\ i=1,\cdots,l.$ (1)

的条件下,求解下面的二次规划问题,

$$\Phi(w)=\frac{1}{2}(w\cdot w) \tag{2}$$

这个优化问题的解可通过求解拉格朗日函数的鞍点得出,利用对偶性原理,经过推导转化为求解如下泛函的优化问题,

$$\begin{aligned}&\max\quad W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)\\&\text{s.t.}\quad \alpha_i\geqslant 0,i=1,\cdots,l;\\&\qquad \sum_{i=1}^{l}\alpha_i y_i=0;\end{aligned} \tag{3}$$

设$\alpha_0=(\alpha_1^0,\cdots,\alpha_l^0)$为上面二次优化问题的解,则最优超平面中的向量$w_0$的模为,

$$\|w_0\|^2=2W(\alpha_0)=\sum_{S.V.}\alpha_i^0\alpha_j^0(x_i\cdot x_j)y_i y_j \tag{4}$$

最后得到的分类规则函数为,

$$f(x)=\operatorname{sgn}(\sum_{S.V.}y_i\alpha_i^0(x_i\cdot x)-b_0), \tag{5}$$

其中x_i是支持向量,α_i^0是对应的拉格朗日系数,b_0是常数(可由w_0和两类中任一对支持向量求得)。

在线性不可分的情况下,在式(1)中增加一松弛变量$\zeta_i\geqslant 0$,成为,

$$y_i[(w\cdot x_i)-b]-1+\zeta_i\geqslant 0,i=1,\cdots,l. \tag{6}$$

综合考虑了最小化错分样本数和最优推广能力,则目标函数改为,

$$\Phi(w,\zeta)=\frac{1}{2}(w\cdot w)+c(\sum_{i=1}^{l}\zeta_i) \tag{7}$$

其中$c>0$为一常数,它控制对错分样本的惩罚程度,则求解在两值线性不可分情况下的广义最优分类面,就是求解在(6)

*)基金项目:国家自然科学基金资助项目(编号:20076041)。**宋晓峰** 讲师,博士,研究方向为计算智能,神经网络等;**陈德钊** 教授,博导,研究方向为建模与优化,信息智能处理等;**胡上序** 教授,博导,研究方向为计算机仿真与智能信息处理。

维普资讯 http://www.cqvip.com

式约束条件下，式(7)的二次优化问题。最终转化为与两值线性可分情况下相同的问题，即根据求拉格朗日函数的鞍点转化为求解原问题的对偶问题(3)式，只是约束条件 $\alpha_i \geqslant 0, i=1,\cdots,l$；改为 $0 \leqslant \alpha_i \leqslant c, i=1,\cdots,l$。

对于非线性的分类问题，通过引入核函数，将原空间的数据通过非线性变换映射到一高维特征空间，在这一空间求解广义线性最优分类面，定义不同的核函数，则对应不同的算法，较为常用的核函数为：多项式核函数 $K(x,x_i)=[(x\cdot x_i)+1]^d$，径向基核函数 $k(x,x_i)=\exp(-\frac{\|x-x_i\|^2}{\sigma^2})$，Sigmoid 核函数 $K(x,x_i)=S[v(x\cdot x_i)+c]$。

因此，在广义最优分类面求解中引入核函数 $K(x,x_i)$ 就可以实现某一非线性变换后的线性分类，上面几种情况可统一为求解如下的二次优化问题，

$$\begin{aligned} &\max \quad W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k(x_i\cdot x_j) \\ &\text{s.t} \quad 0\leqslant\alpha_i\leqslant c, i=1,\cdots l; \\ &\qquad \sum_{i=1}^{l}\alpha_i y_i=0 \end{aligned} \tag{8}$$

α_i 为与每个样本对应的拉格朗日乘子，式(8)存在唯一解，其解中只有一少部分的 α_i 不为零，其对应的样本就是支持向量。则其广义最优分类面决策函数为，

$$f(x)=\operatorname{sgn}(\sum_{S.V.} y_i\alpha_i K(x_i,x)+b) \tag{9}$$

3 SVM 泛化能力估计的理论基础

定理1 给定分类训练集 $D=\{(x_1,y_1),\cdots,(x_l,y_l)\}$、特征空间 H 和超平面 (w,b)，则分类间隔 $\gamma(w,b,D)$ 和数据集半径 $R(D)$ 定义为，

$$\gamma(w,b,D)=\min_{(x_i\cdot y_i)\in D}\frac{y_i(w\cdot\Phi(x_i)+b)}{\|w\|}$$

$$R(D)=\min_{a,x_i}\|\Phi(x_i)+a\|$$

假设 $\gamma>0$，则错分概率和期望值为，

$$p_{err}(w,b)=P(sign(w\cdot\Phi(X)+b)\neq Y)$$

$$E(p_{err})\leqslant\frac{1}{l}E\{\frac{R^2(D)}{\gamma^2(L(D),Z)}\}$$

定义1 阶跃函数 $\Psi(x)=\begin{cases}1 & x>0\\ 0 & x\leqslant 0\end{cases}$

定义2 已知有 l 组样本数据用于建立分类器模型，p 组样本数据用于验证分类模型误差，则分类模型误差由下式确定，

$$T=\frac{1}{p}\sum_{i=1}^{p}\Psi(y_i' f(x_i'))$$

4 SVM 泛化能力估计方法

支持向量机在获得分类和回归模型后，往往需要对 SVM 模型的泛化能力进行估计，得到模型误差的一个上限值，以便对所获模型的精度进行评价，并进而优化模型参数，以获得最优 SVM 模型。

支持向量机中参数众多，关系复杂，参数确定比较困难，是一个制约支持向量机广泛应用的关键因素。因此，目前国际上对支持向量机中优化参数的研究方兴未艾。在此领域中，SVM 泛化能力的估计是一个比较重要的问题。传统的泛化能力估计方法是交叉验证和 bootstrapping 方法，这些方法比较成熟且已经被广泛应用。但其固有的缺点是计算繁琐，效率低下。为了有效估计 SVM 的泛化能力，最近，一些研究者相继提出了一些比较有效的方法，下面对其逐一进行介绍[3~8]。

4.1 Leave-One-Out bound

对于容量为 l 的样本数据集，每次从中选出一个样本数据作为测试集，剩余 $l-1$ 个样本数据作为训练集，这样共需构造 $l-1$ 个分类器，分别用于验证其余下的测试数据。验证误差记为 $L(x_1,y_1;\cdots;x_l,y_l)$，则可给出此种分类器的泛化误差估计为，

$$E(p_{err}^{l-1})=\frac{1}{l}E(L(x_1,y_1;\cdots;x_l,y_l))$$

Leave-One-Out 方法是一个估计泛化能力比较好的方法，但其需要训练 l 次，计算步骤冗长繁琐，效率不高。

将所有样本数据用于建立分类器模型时，所得的分类器模型记为 f^0，将除第 i 个以外样本数据用于建立分类器模型，第 i 个样本数据用于验证模型，此时所得的分类器模型记为 f^i，则可得到验证误差为，

$$L(x_1,y_1;\cdots;x_l,y_l)=\sum_{p=1}^{l}\Psi(-y_p f^p(x_p)) \tag{10}$$

式(10)可进一步写为，

$$L(x_1,y_1;\cdots;x_l,y_l)=\sum_{p=1}^{l}\Psi(-y_p f^0(x_p)+y_p(f^0(x_p)-f^p(x_p)))$$

设 $U_p=\max\{y_p(f^0(x_p)-f^p(x_p))\}$，对于硬间隔分类问题，有 $y_p f^0(x_p)\geqslant 1$，所以，

$$L(x_1,y_1;\cdots;x_l,y_l)=\sum_{p=1}^{l}\Psi(U_p-1)$$

4.2 Jaakkola-Haussler bound

在1999年，Jaakkola 和 Haussler 曾证明了如下的不等式，

$$y_p(f^0(x_p)-f^p(x_p))\leqslant\alpha_p^0 K(x_p,x_p)=U_p$$

因此，就有如下的验证误差估计，

$$T=\frac{1}{l}\sum_{p=1}^{l}\Psi(\alpha_p^0 K(x_p,x_p)-1)$$

对于 $x\geqslant 0$，有 $\Psi(x-1)\leqslant x$，故分类器模型的验证误差：Jakkola-Haussler 上限为，

$$T=\sum_{p=1}^{l}\alpha_p^0 k(x_p,x_p)$$

4.3 Opper-Winther bound

Opper 和 Winther 于2000年证明了如下等式，

$$y_p(f^0(x_p)-f^p(x_p))=\alpha_p^0/(K_{SV}^{-1})_p$$

其中 K_{SV} 为支持向量点积矩阵，则有分类器的验证误差估计：Opper-Winther 上限估计为，

$$T=\frac{1}{l}\sum_{p=1}^{l}\Psi(\frac{\alpha_p^0}{(K_{SV}^{-1})_p}-1)$$

4.4 Radius-margin bound

Vapnik 等人在1998年提出了如下的泛化误差估计式，

$$T=\frac{1}{l}\frac{R^2}{\gamma^2}$$

其中 R 和 γ 的定义见定理1。

4.5 Span bound

Vapnik 和 Chapelle 在2000年定义了支持向量范围(span of support vector)，首先设 $\alpha^0=(\alpha_1^0,\alpha_2^0,\cdots,\alpha_n^0)$ 为式(8)优化问题的解，那么对任一支持向量 x_p 而言，定义集合 Λ_p 为，

$$\Lambda_p=\{\sum_{i\neq p,0<\alpha_i^0<C}^{l}\lambda_i\Phi(x_i),\ \sum_{i=1,i\neq p}^{l}\lambda_i=1, 0\leqslant\alpha_i^0+y_i y_p\alpha_p^0\lambda_i\leqslant C\}$$

那么，定义支持向量 x_p 的范围 S_p 为，$S_p^2=d^2(x_p,\Lambda_p)=\min_{x\in\Lambda_p}(x_p-x)^2$

Vapnik 等人又证明了 $S_p=d(x_p,\Lambda_p)\leqslant D_{SV}$，$D_{SV}$ 为包含所有支持向量的超球直径。

(下转封四)

维普资讯 http://www.cqvip.com

(上接第 126 页)

同时又有，$y_p(f^0(x_p)-f^p(x_p))=\alpha_p^0 S_p^2$

那么就有基于支持向量范围 S_p 的分类模型泛化误差估计公式(span bound)为，

$$T=\frac{1}{l}L(x_1,y_1,\cdots,x_l,y_l)=\frac{1}{l}\sum\{p:\alpha_p^0 S_p^2\geqslant y_p f^0(x_p)$$

4.6 ξα-estimator

留一法的缺点是计算效率低，对一个有 n 个元素的训练集，学习器需要被调用 n 次。对大多数实际问题，这个代价通常很大。下面提出的泛化误差估计器是基于留一法的，但是充分利用 SVM 的特性，它只需要较少的计算次数，它不需要重复取样和重复训练，而是在用全部样本数据作为训练数据建模后直接应用。这种方法需要用到 SVM 训练结果的拉格朗日乘子向量 α 和训练损失向量 ξ，考虑到这一点，这种估计器被称为 $\xi\alpha$-estimators[9]。α 和 ξ 在训练 SVM 后，不需要额外的代价就能获得。

对于稳定的软间隔 SVM，$\xi\alpha$-estimators 估计的模型误差率为 $Err_{\xi\alpha}=\frac{d}{n}$，其中，$d=\sum\{i:(p\alpha_i R_\Delta^2+\xi_i)>1\}$ ($\rho=2$，α 和 ξ 可由 SVM 训练结果得到，R_Δ^2 为 $K(x,x)-K(x,x')$的上界)。

$\xi\alpha$-estimators 表明了对于满足 $\rho\alpha_i R_\Delta^2+\xi_i>1$ 关系式的样本点在用留一法预测时会出现分类错误，而且；$\xi\alpha$-estimators 给出的是留一法模型误差估计的一个上限。

5 各种方法的分析比较

上述各种估计分类模型的泛化能力估计式最基本的是 leave-one-out bound，也就是通常所说的留一法估计界限，Jaakkola-Haussler bound 和 Opper-Winther bound 均可由留一法估计界限导出，但无需多次的求解分类器，计算上简单，但只是给出了泛化误差的宽的估计上限，$\xi\alpha$-estimators 是新近提出的一种泛化能力估计方法，应用起来方便、简单。但其给出的误差界限较为宽松，且与惩罚系数 C 还有很大关系，在惩罚系数较大时，计算结果会不甚准确，这使得这一估计方法应用起来不甚方便；Support vector count bound 给出的误差界限最为宽松，是计算最为方便简单的一种估计方法，但其给出的误差界限是这几种估计方法最大的一种，可信度最低。Span bound 通过定义一个超平面集合 Λ_p，以样本数据 x_p 到这一超平面的距离来估计模型的泛化误差，思路新颖，计算不复杂，是一个比较好的估计方法。Radius-margin bound 是通过分类器的最大分类间隔和包含样本数据集的超球半径来估计模型的泛化误差，是目前应用比较多的一种方法。

表 1 为用 SVM 来建立留兰香模式分类器时采用上述各种估计方法所得的训练误差结果[10]，从中可看出，在上述几种泛化误差估计器中，除了 Leave-one-out 方法估计的测试误差比较小外，Radius-margin bound 方法估计的测试误差比其它几种方法要小。用于留兰香模式分类器的泛化误差估计比较好。

表 1 泛化误差估计值比较

Methods	Estimated testing error(%)
Leave-one-out	18.2%
Support vector count bound	47.7%
Jaakkola-Haussler bound	20.5%
Opper-Winther bound	22.7%
Radius-margin bound	20.1%
Span bound	27.3%
ξα-estimator	31.8%

P-2817
TP
2004.31(8)

结论 支持向量机用于模式分类或回归估计时的性能好坏很大程度上依赖于其参数的确定。为了确定最优参数，往往要估计所建模型的泛化能力，依据泛化误差来确定支持向量机所含参数。确定参数比较有效的方法就是交叉验证，但其计算繁琐，效率低下，应用不甚方便。近来许多学者相继提出了一些估计泛化误差的方法，但其应用都有各自的局限性，需针对不同的研究对象采用不同的泛化误差估计方法。今后，计算简单，效率高，通用的泛化误差估计方法是人们研究的一个重要方向。

参 考 文 献

1 Vapnik V N. The Nature of Statistical Learning Theory. Springer, New York, 1995
2 张学工．关于统计学习理论与支持向量机[J]．自动化学报，2000，26(1)：32～42
3 Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing Multiple Parameters for Support Vector Machines. Machine Learning, 2002,46(1):131～159
4 Chapelle O, Vapnik V. Model Selection for Support Vector Machines. In Advances in Neural Information Processing System, 1999
5 Jaakkola T S, Haussler D. Probabilistic kernel regression models. In: proc. of the 1999 Conf. on AI and Statistics, 1999
6 Vapnik V, Chapelle O. bounds on error expection for support vector machines. Neural Computing, 2000, 12(9)
7 Lee J-H, Lin C-J. Automatic model selection for support vector machines. Neural Computation, 2002
8 Chung K-M, et al. Radius Margin Bounds for Support Vector Machines with the RBF Kernel. To appear in Neural Computation, 2003
9 Joachims T. Estimating the Generalization Performance of a SVM Efficiently. In: Proc. of ICML-00, 17th Intl. Conf. on Machine Learning. Morgan Kaufman, 2000
10 颜学峰．高维复杂模式识别的新方法：[浙江大学博士学位论文]．2002

计算机科学

(1974 年 1 月创刊)
第 31 卷第 8 期 (月刊)
2004 年 8 月 25 日出版

ISSN 1002－137X
CN50－1075/TP

定价：20.00 元　国外定价：5 美元
邮发代号：78－68
发行范围：国内外公开

主管单位：国家科学技术部
主办单位：国家科技部西南信息中心
编辑出版：《计算机科学》杂志社
重庆市渝中区胜利路 132 号　邮政编码：400013
电话：(023) 63500828　E-mail：jsjkx@swic.ac.cn
社　长：牟炳林
主　编：彭　丹
副主编：朱宗元
印刷者：重庆科情印务有限公司
总发行处：重庆市邮政局
订购处：全国各地邮政局
国外总发行：中国国际图书贸易总公司（北京399 信箱）
国外代号：6210—MO