

自动化学院

034 系

035 系



南京航空航天大学学报

(社会科学版)(季刊)

NANJING HANGKONG HANGTIAN DAXUE XUEBAO

(SHEHUI KEXUE BAN)

第6卷(总第30期)

2004年12月

主 编 伍贻兆

执行主编 张 辉

JOURNAL OF NANJING

UNIVERSITY OF AERONAUTICS

& ASTRONAUTICS (Quarterly)

(Social Sciences)

Vol.6 Suppl., (Serial 30)

Dec. 2004

Editor in chief

Wu Yizhao

Executive Editor in Chief

Zhang Hui

主 办 南京航空航天大学
编辑出版 南京航空航天大学学报
(社会科学版)编辑部
印 刷 华东地研所印刷厂
地 址 南京市御道街29号
邮 编 210016
电 话 (025)84890772

Sponsored by Nanjing University of
Aeronautics & Astronautics
Edited & Published by Editorial Department
of Journal of NUAA(Social Sciences)
Address 29 Yudao Street Nanjing, China
Tel (025)84890772 E-mail hec01 @ nuaa.edu.cn

公开发行人 刊号: ISSN 1671-2129(国际标准刊号) 增刊号: (2004)084号 定价: 4.50元
CN32-1548/C(国内统一刊号)

序号	姓名	职称	单位	论文题目	刊物会议名称	年卷期	类别
1	宋晓峰 韩平 邹丽	讲师	034	基于支持向量机的足月胎儿体重预测新方法	中国生物医学工程学报	2004, 23, 6	
2	宋晓峰 陈德钊 胡上序	讲师	034	支持向量机泛化能力估计若干方法	计算机科学	2004, 31, 8	
3	宋晓峰 胡志忠 王惠南	讲师	034	我校生物医学工程学科发展的整体思路分析	南京航空航天大学(社会科学版)	2004, 6, 30	
4	陈 春 晓	讲师	034	基于小波变换的心电信号去噪处理	生物医学工程学杂志	2004, 21, 1	
5	钱 志 余	讲师	034	生物组织光学参数:优化散射系数(u)的实时在位测定	南京航空航天大学学报	2004, 36, 6	
6	吴了泥 黄一敏	硕士 副高	035	基于Stateflow技术的多模态飞行控制律仿真	中国自动化学会华东地区第十七届学术会议	2004	
7	马启鑫 黄一敏	硕士 副高	035	输出反馈在飞行控制律设计种的应用研究	中国自动化学会华东地区第十七届学术会议	2004	
8	李立早 黄一敏	硕士 副高	035	无人机上升段飞行轨迹设计	中国自动化学会华东地区第十七届学术会议	2004	
9	汪 元 黄一敏	硕士 副高	035	可重复运载器自动着陆段内回路的设计研究	中国自动化学会华东地区第十七届学术会议	2004	
10	王春峰 黄一敏	硕士 副高	035	一种实现舵机带宽测试的系统辨识方法	南京航空航天大学第六届学术年会	2004	
11	孙春贞 黄一敏	博士 副高	035	基于输出反馈 H_∞ 回路成形的飞行控制律设计	中国自动化学会华东地区第十七届学术会议	2004	
12	黄一敏 孙春贞	副高 博士	035	无动力弹体的纵向弹道控制	南京理工大学学报 2004年28卷1期	2004	
13	耿通奋 黄一敏	硕士 副高	035	小型无人机实时仿真系统研究	计算机仿真 2004年21卷9期	2004	
14	邱浩 王道波 张焕春	博士 教授 教授	035 035 033	一种改进的反向传播神经网络算法	应用科学学报	2004 年 22 卷 3 期	
15	邱浩 王道波 张焕春	博士 教授 教授	035 035 033	控制系统的故障诊断方法综述	航天控制	2004 年 22 卷 2 期	
16	曾丽兰 王道波 樊春霞	博士 教授	035 035	基于模糊 PID 的无人直升机解耦控制研究	应用科学学报	2004 年 22 卷 4 期	
17	柳玉甜 陈欣	博士 研究员	035	无人飞行器航向静稳定性研究	飞行力学	2004, 22, 2	
18	柳玉甜 陈欣	博士 研究员	035	无人飞行器静稳定性问题的研究	现代防御技术	2004, 32, 5	
19	潘常春 陈欣	博士 研究员	035	基于鲁棒特征结构配置的无人机直接侧力控制	飞行力学	2004, 22, 3	

20	步红 余德义	硕士 研究员	035	无人机远程控制系统仿真 及可视化	中国航空学会青年科技 论坛	2004, 9	
21	郝红雨 余德义	硕士 研究员	035	无人机远程控制系统 中的三维可视化	2004系统仿真技术及其 应用学术交流会	2004, 8	
22	段海滨 王道波 朱家强 黄向华	博士 教授 博士后 副教授	031	Parameter identification of LuGre friction model for flight simulation servo system based on ant colony algorithm	Transactions of Nanjing University of Aeronautics & Astronautics	2004, 21, 3.	
23	段海滨 王道波 黄向华 朱家强	博士 教授 副教授博 士后	031	Research on non-linear angular potentiometer and its application in simulation servo system	Journal of Sichuan University	2004, 41,3	
24	段海滨 王道波 于秀芬	博士 教授 研实	031	Design of angular measuring servo system based on AT89C52 microcontroller	Proceedings of the 3rd International Symposium of Instrument Science & Technology	ISBN: 7-5603-2056-2/ TH.122, 2004, Vol. 3	
25	段海滨 王道波 朱家强 黄向华	博士 教授 博士后 副教授	031	蚁群算法理论及应用研究 的进展	控制与决策	2004,19, 12	
26	段海滨 王道波 黄向华 韩朔眺	博士 教授 副教授教 授	031	基于粗集理论的高精度伺 服仿真转台故障诊断研究	中国机械工程	2004,15, 21	
27	段海滨 王道波	博士 教授	031	一种快速全局优化的改进 蚁群算法及仿真	信息与控制	2004,33, 2	
28	段海滨 王道波 黄向华 朱家强	博士 教授 副教授博 士后	031	基于蚁群算法的PID参数优 化	武汉大学学报	2004,37, 5	
29	段海滨 王道波	博士 教授	031	蚁群算法的全局收敛性研 究及改进	系统工程与电子技术	2004,26, 10	
30	段海滨 王道波 于秀芬 朱家强	博士 教授 研实 博士后	031	一种改进的蚁群算法用于 约束非线性规划问题求解	.四川大学学报	2004,41, 5	
31	段海滨 王道波 尧放哉	博士 教授高工	031	一种大负载三轴电动仿真 测试转台设计与实现	南京航空航天大学学报	2004,36, 3	
32	段海滨 王道波 张天宏 黄向华	博士 教授 副教授副 教授	031	飞行仿真电液伺服转台功 能模拟器设计研究	应用基础与工程科学学 报	2004,12, 2	

33	段海滨 王道波 于秀芬 尧放哉	博士 教授 研实 高工	031	武装直升机火控系统仿真 测试转台的综合性能指标 调试策略研究	应用基础与工程科学学 报	2004,12,3	
34	段海滨 盛守照 王道波	博士 博士 教授	031	基于 Windows VxD 技术的 高精度仿真伺服控制系统	吉林大学学报	2004,22, 6	
35	段海滨 王道波 盛守照 黄向华	博士 教授 博士 副教授	031	一种高精度飞行仿真模拟 器的软件设计	华中科技大学学报	2004,32, 10	
36	段海滨 王道波 尧放哉	博士 教授高工	031	武装直升机火控系统仿真 转台的控制系统设计研究	系统仿真学报	2004,16, 7	
37	段海滨 王道波 黄向华	博士 教授 副教授	031	光电轴角编码器在飞行仿 真伺服系统中的应用	传感器技术	2004,23, 5	
38	段海滨 王道波 黄向华	博士 教授 副教授	031	基于光电轴角编码器的测 试转台鉴频技术	传感器技术	2004,23, 10	
39	段海滨 王道波 黄向华	博士 教授 副教授	031	一种新型高精度大功率电 动飞行仿真转台的驱动方 案设计	中国惯性技术学报	2004,12, 3	
40	段海滨 于秀芬 王道波 黄向华	博士 研实 教授 副教授	031	基于内模PID鲁棒控制的飞 行仿真伺服系统设计	中国空间科学技术	2004,24, 6	
41	于秀芬 段海滨 龚华军	研实 博士 副教授	031	移动机器人视觉定位方法 的研究与实现	数据采集与处理	2004,19, 4	
42	段海滨 王道波	博士 教授	031	Novel type of nonlinear backstepping controlling scheme for flight simulation servo system	<i>Proceedings of the 6th Asia-Pacific International Conference on Control and Measurement</i>	ISBN: 7-80183-428-3/ T.011, Chengdu, 2004	
43	段海滨 王道波	博士 教授	031	Nonlinear friction compensation controller design for high precision flight simulator	<i>Proceedings of the 23rd Chinese Control Conference (International)</i>	ISBN:7-5628-1 583-6/TP.129, Wuxi, 2004, Vol. 1	
44	段海滨 王道波 黄向华	博士 教授 副教授	031	Study of the implementation of a novel type of adaptive-PID-based electric-driven high precision turntable	<i>Proceedings of the PRC Conference' 2004 Electronics & Information Technology</i>	ISBN:7-80128- 591-3, Huangshan, 2004, Vol. 1	
45	于秀芬 段海滨	硕士 博士	031	基于时滞卡尔曼滤波的机 器人视觉定位	电光与控制	2004, 11, 2	

	龚华军	副教授					
46	王志胜 王道波 蔡宗琰	博士后 教授 教授	031	用于软故障检测的残差检验法	西安电子科技大学学报	2004, 31, 3	
47	王志胜 王道波 蔡宗琰	博士后 教授 教授	031	传感器标定的统一数据处理方法	传感器技术	2004, 23, 3	
48	王志胜 王道波 蔡宗琰	博士后 教授 教授	031	双余度传感器缓变故障检测的改进残差检验法	传感器技术	2004, 23, 4	
49	王志胜 刘元祥 王道波	博士后 高工 教授	031	一种无矩阵求逆的最优滤波计算方法	计算技术与自动化	2004, 23, 3	
50	王志胜 刘建业 周军	博士后 教授 教授	031	推广联合滤波算法在卫星组合定姿系统中的应用	宇航学报	2004, 25, 5	
51	王志胜 覃媛媛 王道波	博士后 硕士 教授	031	A method of reliability allocation of complicated large system	International Journal of PEM	2004, 9, 1	
52	王志胜	博士后	031	输入和干扰均可预见的最优控制器设计	第23届中国控制会议	2004	
53	李春涛	中级博士	033	Neural Model-Based Adaptive Control for Systems with Unknown Preisach-type Hysteresis	控制理论与应用	2004, 1, 1	
54	李春涛	中级博士	035	Observer-based Adaptive Output Feedback Control of Nonlinear Systems Using Neural Network	系统仿真学报	2004, 16, 10	
55	李春涛	中级博士	035	Adaptive Control Based on Neural Estimation for Systems with Unknown Hysteresis	IEEE Conference on Control Applications/International Symposium on Intelligent Control/Computer Aided Control Systems Design	2004, 9, 2-4	
56	李春涛	中级博士	035	Adaptive Neural Sliding Mode Control for Systems with Unknown Hysteresis Using Neural Model Based Prediction	Eighth International Conference on Control, Automation, Robotics and Vision	2004, 12, 6-9	
57	祖家奎 赵淳生	博士后 教授	035 01	Research on Resonance and Antiresonance States of Free Stator of Ultrasonic Motors	Chinese Journal of Acoustics	2004, 23, 04	
58	祖家奎 赵淳生 戴冠中	博士后 教授 教授	035 01 外校	模糊逻辑系统的一种自适应参数学习算法	系统仿真学报	2004, 16, 05	

59	祖家奎 赵淳生 戴冠中	博士后 教授 教授	035 01 外校	基于聚类算法的模糊逻辑 结构与系统性能分析	南京航空航天大学学报	2004, 36, 01	
60	祖家奎 赵淳生 戴冠中	博士后 教授 教授	035 01 外校	非单点模糊逻辑系统在非 线性信道均衡器中的应用	控制与决策	2004, 19, 04	
61	祖家奎 赵淳生	博士后 教授	035 01	行波型超声电机频率跟踪 控制技术评述	微电机	2004, 37, 06	
62	祖家奎 赵淳生	博士后 教授	035 01	行波型超声电机驱动和控 制技术的现状与发展	微特电机	2004, 32, 06	
63	李志宇 肖前贵	助教 高工	035 035	基于LM35温度传感器的温 控系统设计	微计算机信息	2004, 20, 09	
64	李志宇 肖前贵	助教 高工	035 035	应用PC/104模块实现多串 口通讯扩展设计	工业控制计算机	2004, 17, 09	

基于支持向量机的足月胎儿体重预测新方法

宋晓峰¹, 韩平², 邹丽², 陈德钊³, 胡上序³

(1. 南京航空航天大学 生物医学工程系, 南京 210016; 2. 华中科技大学同济医学院附属协和医院, 武汉 430022;
3. 浙江大学 智能信息工程研究所, 杭州 310027)

摘要: 支持向量机回归估计的性能往往依赖于核函数及其参数、不敏感系数和惩罚因子的确定。支持向量机中参数的确定是一个较为困难的问题, 以往通常采用交叉验证的方法确定参数。本研究提出了自适应参数调整支持向量回归估计方法(A-SVM), 给出了自适应参数调整算法。并已成功应用于足月胎儿体重的建模, 且经与径向基网-偏最小二乘回归方法(RBFN-PLSR)建立的模型比较, A-SVM方法拟合精度和预测能力均比RBFN-PLSR方法好。

关键词: 支持向量机; 胎儿体重; 建模; 回归估计

中图分类号: R319

文献标识码: A

引言

根据孕妇的宫高、腹围等参数以及通过B超测量胎儿的双顶径、股骨长度、腹经线等指标, 采用一定数学方法建立胎儿体重模型, 来对胎儿体重进行预测, 对产科医师具有重要的实际意义。根据所得到的胎儿体重, 医师可合理选择分娩方式。若所建立的胎儿体重模型不能真实反映实际情况, 所预测得到的胎儿体重误差较大, 将影响产科医生对分娩方式的选择。目前, 用于预测胎儿体重的方法和公式有多种^[1,2], 都是一些传统的经验性回归分析方法, 预测结果误差较大。也有一些研究者采用人工神经网络等人工智能方法^[3,4], 取得一定效果。但人工神经网络还存在网络结构复杂, 训练时间长, 易于过学习等缺点。

自1995年Vapnik等提出统计学习理论(Statistical Learning Theory, SLT)后, 基于SLT理论的支持向量机(Support Vector Machine, SVM)技术, 日渐成为国际学术界的研究热点。SVM较好地解决了上述缺陷, 用于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^[5,6]。回归分析和人工神经网络所基于的ERM原则过于寻求样本数据的偏差最小, 从而导致模型预测性能的下降。Vapnik等分析了模型复杂性与预测能力间的矛盾, 进而提出SRM原则, 均衡地考虑对数据的训练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主要目标。

本研究将提出一种参数自适应调整的算法, 以避免通过交叉验证的冗长步骤选定最佳参数^[5], 由此构成自适应支持向量机(A-SVM), 并将其应用于足月胎儿体重模型的建立, 效果令人满意。与径向基网-偏最小二乘(RBFN-PLSR)方法对比, 其训练精度及预测性能都比较好。

1 支持向量机回归建模方法

基于统计学习理论的支持向量机原先是为求解小样本模式识别问题而提出的, 后来推广应用于函数拟合(回归估计)。SVM的核心思想是在降低经验风险的同时, 尽量减小期望风险的上界, 从而在已有样本集的基础上, 确保模型有足够的预测精度, 即提高模型的推广能力。

基金项目: 国家自然科学基金资助项目(20076041)

收稿日期: 2004-01-08; 修回日期: 2004-06-08

先讨论用线性回归函数 $f(x) = w \cdot x + b$ 拟合一样本集 $\{x_i, y_i\} (i = 1, 2, \dots, n; x_i \in R^d; y_i \in R)$, 设所有训练样本的精度为 ε (又称其为不敏感系数), 即满足(1)式,

$$\begin{aligned} y_i - w \cdot x_i - b &\leq \varepsilon \\ w \cdot x_i + b - y_i &\leq \varepsilon \end{aligned} \quad (1)$$

为控制函数的复杂性应使线性回归函数尽量平坦, 亦即使 w 的欧拉范数 $\|w\|^2$ 最小^[5]。为考虑超出精度要求的拟合误差, 引入非负松弛变量 ζ_i, ζ_i^* , 以处理未满足(1)式的样本点。这样函数拟合问题演变为求解向量 w 和 b 的二次优化问题, 如式(2)。

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\zeta_i + \zeta_i^*), \\ \text{S. T.} \quad & y_i - w \cdot x_i - b \geq \varepsilon + \zeta_i, \\ & w \cdot x_i + b - y_i \geq \varepsilon + \zeta_i^*, \\ & \zeta_i \geq 0, \\ & \zeta_i^* \geq 0, \\ & i = 1, 2, \dots, n \end{aligned} \quad (2)$$

其中常数 C 为一惩罚系数, 以控制对误差超出 ε 的样本点的惩罚度, 为便于计算, 常将(2)式转换为(3)式所示的对偶形式,

$$\begin{aligned} \max \quad & W(\alpha, \alpha^*) = \varepsilon \sum_{i=1}^n (\alpha_i^* + \alpha_i) + \sum_{i=1}^n y_i (\alpha_i^* - \alpha_i) - \frac{1}{2} \sum_{i=1}^n (\alpha_i^* - \alpha_i) (\alpha_j^* - \alpha_j) (x_i \cdot x_j) \\ \text{S. T.} \quad & \sum_{i=1}^n (\alpha_i^* - \alpha_i) = 0 \\ & 0 \leq \alpha_i^* \leq C \quad i = 1, \dots, n \\ & 0 \leq \alpha_i \leq C \quad i = 1, \dots, n \end{aligned} \quad (3)$$

由此可求得拉各朗日 (Lagrange) 乘子 α_i, α_i^* 以及向量 w 与 b 为

$$w = \sum_{i=1}^n (\alpha_i^* - \alpha_i) x_i \quad (4)$$

$$b^* = y_i - (w \cdot x_i) - \varepsilon \quad \text{当 } \alpha_i - \alpha_i^* > 0 \text{ 时} \quad (5)$$

$$b^* = y_i - (w \cdot x_i) + \varepsilon \quad \text{当 } \alpha_i - \alpha_i^* < 0 \text{ 时}$$

进而解得线性回归函数,

$$f(x) = w \cdot x + b = \sum_{i=1}^n (\alpha_i^* - \alpha_i) (x_i \cdot x) + b^* \quad (6)$$

对于回归函数为非线性的问题, 仅须将(3)、(5)、(6)式的内积运算 $x_i \cdot x$ 替换为核函数 $K(x_i, x)$ ^[5,6], 这相当于通过核函数, 将自变向量非线性地转换到高维特征空间, 然后在高维特征空间内进行线性函数拟合, 以实现原问题的非线性回归估计。常用的核函数有三种, 多项式函数 $[(x \cdot x_i) + 1]^d$; 径向基函数 $\exp(-\frac{\|x - x_i\|^2}{\sigma^2})$; 和 Sigmoid 函数 $\tanh(v(x \cdot x_i) + c)$, 本研究将应用径向基核函数。

通常在求解(3)式二次优化问题时所得到的拉各朗日乘子 α_i, α_i^* 中只有一小部分不为零, 它们对应的样本即为支持向量, 因此支持向量机就是在样本集中挑选具有代表性的特征向量子集 (即支持向量) 进行回归估计。在样本数较多时, 实际上只有其中一部分样本参与回归, 估计输入输出之间的函数关系。

也可将支持向量机类比于人工神经网络, 如图1所示, 与人工神经网络所不同的是, 它无需选定结构, 而

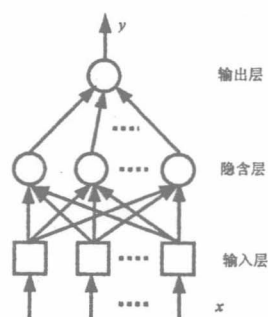


图1 支持向量机网络示意图

由算法自动确定,隐单元数为支持向量的个数,其处理核函数的中心向量就是对应的支持向量。又因归结为二次优化问题,容易实现全局最优,可避免陷入局部极小。

2 支持向量机参数的自适应调整

为选定支持向量机的最佳参数,通常采用交叉验证的方法,这需要冗长的试验计算。为此,本研究拟提出参数自适应调整的算法,由此构成自适应支持向量机(A-SVM)。在训练计算中分析拟合误差的演变情况,自动地为支持向量机选定较佳的参数,以获取较优的回归估计函数。为算法所需,建立如下定义和规则。

定义1:有序样本 将所有样本按如下原则排序,任选一样本记为 x_1 ,除 x_1 外与 x_1 距离最小的样本记为 x_2 ,以此类推,对于样本 $x_i (i=1, \dots, n-1)$,除 $\{x_1, x_2, \dots, x_i\}$ 外与 x_i 距离最小的样本记为 x_{i+1} ,这样可将所有样本排序为 $x_1, x_2, \dots, x_n$ 。称已经排序的样本为有序样本,约定下文中的样本已被排序,称 x_{i+1} 为样本点 $x_i (i=1, \dots, n-1)$ 的相邻点。

定义2:相邻中值点 对于样本点 $x_i (i=1, \dots, n-1)$,定义它的相邻中值点为:

$$x_{mi} = (x_i + 0.5(x_{i+1} - x_i))。$$

计算相邻点距离为 $d_i = \|x_{i+1} - x_i\|$,距离均值为 $\bar{d} = \frac{1}{n-1} \sum_{i=1}^{n-1} d_i$ 。

规则1:有样本集 $\{x_i, y_i\} (i=1, 2, \dots, n; x_i \in R^d; y_i \in R)$,设精度为 $\delta_y > 0$,回归函数为 $f(x)$ 。对于样本点 $x_i (i=1, \dots, n-1)$,若与相邻点的距离不是很大,即有 $d_i < \bar{d}$,且两相邻点的因变量取值也不是太靠近,即有 $|y_{i+1} - y_i| > \delta_y$,检查回归函数在相邻中值点的取值 $f(x_{mi})$,如果它未落在 y_i 和 y_{i+1} 间,则认为回归函数不够“平坦”,或过于复杂。此时可能导致过拟合。

对于采用径向基核函数的 A-SVM,自适应调整参数的算法,其主要步骤,见图2。

对于不敏感系数,本算法将取值为

$$\epsilon = 0.5s_d \quad (7)$$

其中 s_d 为噪声分布密度函数的标准差,且在整个迭代过程中不敏感系数将保持不变。设迭代轮次计数器为 t ,在训练迭代的第 t 轮,惩罚系数和核函数的宽度系数分别记为 $C^{(t)}$ 和 $\sigma^{2(t)}$ 。采用这些参数,调用 SVM 的一种算法,对训练样本集作回归估计,并进而计算各训练样本点的拟合相对误差为 $E_i^{(t)} = \frac{|y_i - f(x_i)|}{y_i}$,以

及平均拟合相对误差为 $MRE^{(t)} = \frac{1}{n} \sum_{i=1}^n E_i^{(t)}$ 。又记回归的相对精度为 δ ,惩罚系数和宽度系数的调整步长为 ΔC 和 $\Delta \sigma^2$,迭代的最大轮次为 T_m 。

计算宽度系数的初值为

$$\sigma^{2(0)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^T (x_i - \bar{x}) \quad (8)$$

其中 $\bar{x}$ 为所有自变向量的均值, n 为样本个数。

计算惩罚系数的初值为

$$C^{(0)} = 2[\max(y_i) - \min(y_i)] \quad (9)$$

在上述算法中, SVM 算法可采用文献^[7,8]介绍的 Lin 算法,其基本思想是在 SMO 算法中嵌入 SVM^{light}算法的工作集选择策略。

3 足月胎儿体重的建模

在产前如何做好胎儿体重的准确预测,判断其成熟度,及时掌握胎儿在宫内的生存状态,提高围产保健

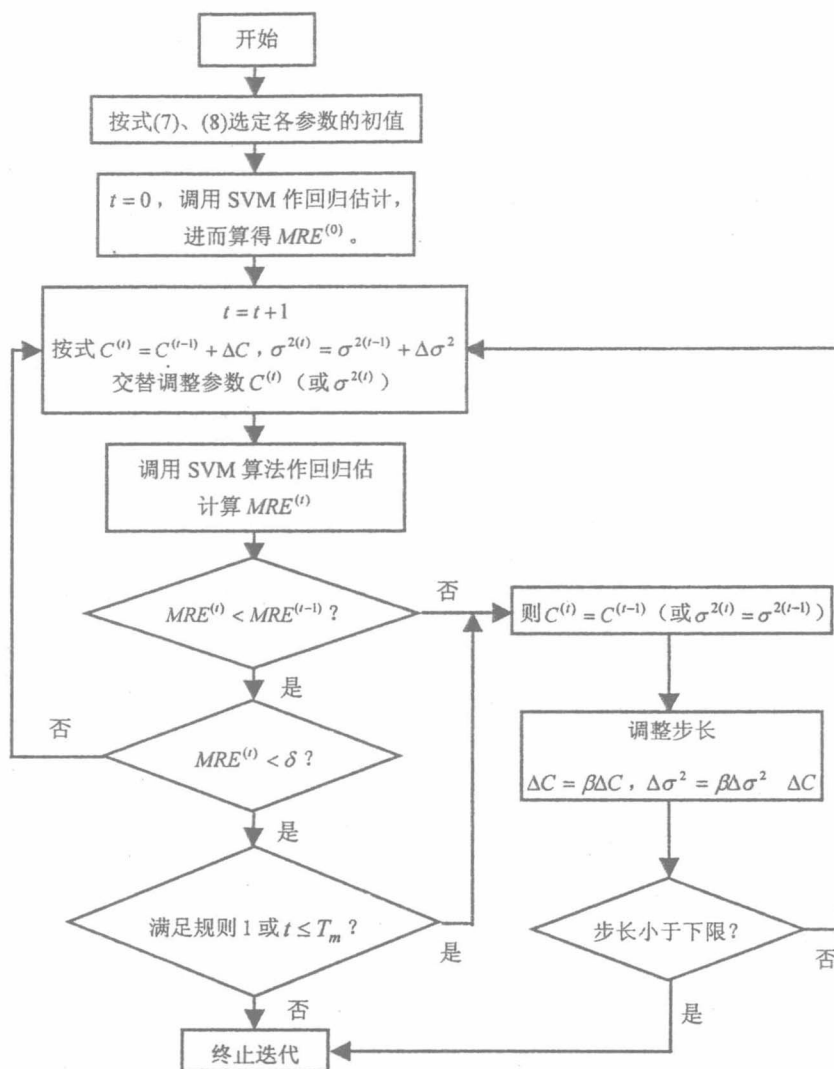


图2 A-SVM 算法流程框图

质量,这是产科医生的一项重要工作。本研究收集了武汉某医院妇产科近两年的病例资料 150 例样品作为样本,其中有 9 例样品为巨大儿。拟用超声检查的参数(双顶径 BPD、枕额径 OFD、股骨长 FL、腹经线 AC(D1)、腹经线 AC(D2))以及孕妇的宫高、腹围等七个指标作为自变量,分娩时的实际体重作为因变量。采集这些数据时不可避免地带有噪声。自变量与因变量间存在的定量关系,尚难按机理确定,也无法给出明确的函数形式。这些都将影响用常规方法建立足月胎儿预期体重经验模型的适用性^[1]。鉴于支持向量机的优势,并期望在实际采集的样本数据基础上建立一个能预测足月胎儿预期体重的有效模型。本研究采用 A-SVM 算法来建立模型以提高模型的精度和预测能力。

还将采用径向基函数网-偏最小二乘回归方法(RBFN-PLSR)建模^[9],以作比较。该法较好地将人工神经网络与多元统计分析相结合,以避免径向基网隐含层设计及其参数选用的困难。它将隐结点数 m 取为训练样本数据个数 n ,每个隐结点径向基函数的中心向量取为相应样本数据点的值。这相当于将每个样本数据点视为一个聚类,样本数据点即为该聚类的中心。然后应用偏最小二乘回归,有效地消除由于隐结点数过多所造成的自变量间的复共线性,稳定地求解隐层与输出层间的连接权。

3.1 SVM 参数对建模结果的影响

先以足月胎儿体重建模为例,检验 SVM 所用参数的影响。为此,采用不同的惩罚系数 C ,且将不敏感系数 ϵ 和宽度系数 σ^2 保持不变(ϵ 取 0.01, σ^2 取 100),应用 SVM 进行回归估计。为避免偶然性,试验将以交叉验证的方式进行,并计算回归函数的训练、测试相对误差。现将预测相对误差的平均值及其误差范围随惩罚系数 C 演变的情况示于图 3(训练相对误差均值随惩罚系数 C 有类似的变化情况,图示略)。从中可看出模型的测试误差将随惩罚系数 C 的增大而降低,达到一定值时不再有变化,但运行时间会增加很多。同时,误差范围的表示是以标准差的 2 倍画出的,从图中误差范围的变化情况可看出,随着惩罚系数 C 的增大,模型的稳定性也趋好。因此存在惩罚系数 C 的合理选择问题。

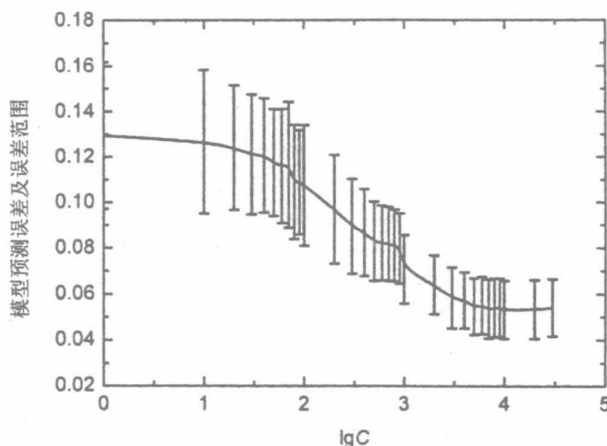


图3 模型测试误差及误差范围随惩罚系数的变化情况

同样地,将不敏感系数 ϵ 和惩罚系数 C 保持不变(ϵ 取 0.01, C 取 4000),采用不同的宽度系数 σ^2 ,应用 SVM 建模,来检验宽度系数 σ^2 对建模结果的影响。并计算回归函数的训练、测试相对误差。现将预测相对误差的平均值及其误差范围随宽度系数 σ^2 演变的情况示于图 4(训练相对误差均值随宽度系数 σ^2 的有类似的变化情况,图示略)。从中可看出模型的测试误差随宽度系数 σ^2 增加有所降低,在 $\sigma^2 = 100$ 时,测试误差最小,随后将随着宽度系数 σ^2 的变大而增大。同时,误差范围在 $\sigma^2 = 100$ 附近区域较短,说明模型在此范围内稳定性较好。在其他区间,模型的稳定性则较差。

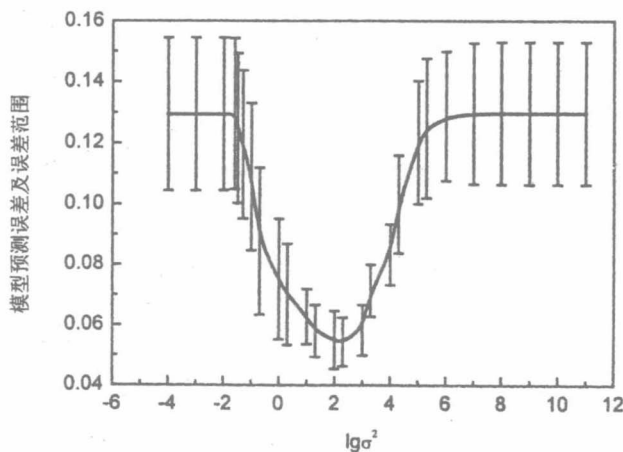


图4 模型测试误差及误差范围随宽度系数的变化情况

依上述分析,显见参数调整对支持向量机回归估计性能的重要性。为选用较佳的参数,以建立较优的

足月胎儿体重模型,拟用3节的支持向量机自适应算法A-SVM进行非线性建模。先由样本数据和算式(7)、(8)、(9)计算确定不敏感系数 ϵ ,初始惩罚系数C和径向基核函数的初始宽度系数 σ^2 ,列于表1。此后,随着训练迭代的进行,回归估计函数的拟合误差将逐步减少,由此自适应算法将调整惩罚系数C和核函数宽度系数 σ^2 ,进一步改善回归函数的拟合性能。完成训练时,调整惩罚系数C和核函数宽度系数 σ^2 所达到的最优值也列于表1,它们与初值相比已有较大改变。该表还给出了SVM所用的支持向量的个数,在训练迭代的初始为150,迭代结束时降为131,简化了SVM结构,降低了结构风险。

表1 胎儿体重模型的支持向量机参数

参数	不敏感系数 ϵ	核函数宽度系数 σ^2	惩罚因子 C	支持向量数	迭代次数
初始值	0.01	0.1	10	150	17
调整值		100	4000	131	

3.2 建模结果的比较

为避免偶然性,采用交叉验证的方式,应用A-SVM和RBFN-PLSR方法为胎儿体重建模。将150个样本数据,随机均分成15组,每次取出一组(10个样本数据)作为预测样本,其余的作为训练样本。两种方法各作15次试验。A-SVM的运行方式与3节相同。应用RBFN-PLSR方法时,RBF的宽度系数采用3.1节所求得的较优值 $\sigma^2 = 100$,而回归所用的PLS成分数为5,此时因变量方差的95.1%已被解释。两种方法的拟合相对误差绝对值(用 ASE_r 表示)、预报相对误差绝对值(用 APE_r 表示),以及这两种相对误差的均值及标准差、两种算法的运行时间列于表2,从中可以清楚地看出A-SVM明显优于RBFN-PLSR。RBFN-PLSR方法较好地解决了径向基网络结构设计和训练的困难,但仍依据ERM原则,影响了预报精度的进一步提高,而A-SVM采用了SRM原则,使预报精度十分接近于拟合精度,模型的推广泛化能力有了显著的提高。A-SVM算法的运行时间虽高于RBFN-PLSR方法,但其所建模型的预测精度和稳定性要高于RBFN-PLSR方法,这种时间开销还是值得的。

表2 两种模型的训练与测试相对误差

NO.	A-SVM		RBFN-PLSR		Num. of PLS
	$ASE_r(\%)$	$APE_r(\%)$	$ASE_r(\%)$	$APE_r(\%)$	
1	4.54	5.11	7.63	8.12	5
2	5.79	6.80	8.89	11.45	5
3	4.97	4.57	7.94	9.04	5
4	5.16	5.38	5.39	4.67	5
5	2.84	4.62	6.14	9.66	5
6	5.53	6.41	4.28	7.60	4
7	2.71	4.08	3.42	4.08	5
8	4.87	6.72	8.43	11.43	5
9	5.73	5.45	6.47	10.49	5
10	4.96	5.04	5.73	5.78	5
11	5.47	5.84	5.97	10.11	5
12	4.03	5.19	7.41	9.12	5
13	4.37	5.19	7.82	9.44	5
14	5.01	5.58	5.67	6.01	4
15	4.76	5.35	5.27	5.84	5
平均值	4.72	5.42	6.43	8.19	/
标准差	0.93	0.80	1.56	2.41	/
算法运行时间(s)	36.1		8.3		/

4 结论

支持向量机技术(SVM)摒弃了传统方法中采用的经验风险最小原则,而采用结构风险最小原则,可避免经典方法的弊端,其回归估计函数有很强的推广泛化能力。但支持向量机的参数对其回归估计性能影响很大。为了获得支持向量机最佳的回归估计性能,往往通过交叉验证的冗长步骤选定最佳参数。本研究通过定义有序样本、相邻中值点等概念以及判断过拟合等规则,提出一种自适应支持向量机方法 A-SVM,其应用自适应方法高效选用支持向量机的最佳参数,可使回归估计函数的拟合和预报精度均得到显著的提高。将其应用于足月胎儿体重的建模,效果明显优于经典方法,所建模型可以有效地应用于 B 超软件系统中。实例表明这种新技术也适用于生理系统的建模。

5 参考文献

- [1] 朱洁萍,戴钟英. B 型超声预测胎儿体重的研究进展[J]. 中华妇产科杂志, 1997,32(6):380-382.
- [2] 康岚,桂占吉,腾丽. B 型超声测量胎儿生长参数预测胎儿体重[J]. 中华超声影像学杂志, 2003,12(3):185-186.
- [3] 刁晓娣,江志斌,刘瑾. 根据孕妇参数预测胎儿体重的神经网络方法[J]. 中国生物医学工程学报, 1999,18(2):155-158.
- [4] 李笑天,庄依亮,常才,张玉华. 神经网络在预测胎儿体重中的作用[J]. 中国医学影像技术, 1999,15(5):400-403.
- [5] Vapnik VN. The Nature of Statistical Learning Theory[M]. Springer, New York, 1995.
- [6] 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.
- [7] Lin CJ. On the convergence of the decomposition method for support vector machines. IEEE Transactions on Neural Networks[J]. 2001,12(6):1288-1298.
- [8] Liao SP, Lin HT, Lin CJ. A Note on the Decomposition Methods for Support Vector Regression[J]. Neural Computation, 2002(14): 1267-1281.
- [9] Walczak B, Massart DL. The radial basis function - partial least squares approach as a flexible non - linear regression technique[J]. Analytica Chimica Acta, 1996,331:177-185.

A NEW METHOD FOR ESTIMATION OF FETAL WEIGHT USING SUPPORT VECTOR MACHINE

SONG Xiao-feng¹, HAN Ping², ZOU Li², CHEN De-zhao³, HU Shang-xu³

(1. Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016;

2. Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022;

3. Laboratory of Intelligent Information Engineering, Zhejiang University, Hangzhou 310027)

Abstract: The performance of support vector machine for regression estimation depends on the kernel function, its parameter, insensitive factor ϵ and penalty factor C . The determination of the parameters of SVM remains a difficult problem. Cross-validation method is commonly used in practice to decide the parameters of SVM, however, it is not efficient enough. A novel adaptive support vector machine (A-SVM) was proposed in this work to determine the optimal parameters automatically. The algorithms for adaptively tuning parameters of SVM established. A-SVM was successfully applied in estimating the fetal weight. Compared with RBFN-PLSR methods, A-SVM was superior in both fitting and testing accuracy.

Key words: Support vector machine; Fetal weight; Modeling; Regression estimation

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-Feng¹ CHEN De-Zhao² HU Shang-Xu²

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

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远,也即使其两侧的空白区域尽量大。数学描述如下,

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

的条件下,求解下面的二次规划问题,

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

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,

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

设 $\alpha_0 = (\alpha_1^0, \dots, \alpha_l^0)$ 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,

$$\|w_0\|^2 = 2W(\alpha_0) = \sum_{i,j=1}^l \alpha_i^0 \alpha_j^0 (x_i \cdot x_j) y_i y_j \quad (4)$$

最后得到的分类规则函数为,

$$f(x) = \text{sgn} \left(\sum_{i=1}^l y_i \alpha_i^0 (x_i \cdot x) - b_0 \right), \quad (5)$$

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 $\zeta_i \geq 0$,成为,

$$y_i[(w \cdot x_i) - b] - 1 + \zeta_i \geq 0, i = 1, \dots, l. \quad (6)$$

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,

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

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。

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 $\alpha_i \geq 0, i=1, \dots, l$; 改为 $0 \leq \alpha_i \leq c, i=1, \dots, l$ 。

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,

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

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,

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

3 SVM 泛化能力估计的理论基础

定理1 给定分类训练集 $D = \{(x_1, y_1), \dots, (x_l, y_l)\}$ 、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,

$$\begin{aligned} \gamma(w, b, D) &= \min_{(x_i, y_i) \in D} \frac{y_i (w \cdot \Phi(x_i) + b)}{\|w\|} \\ R(D) &= \min_{a, x_i} \|\Phi(x_i) + a\| \end{aligned}$$

假设 $\gamma > 0$, 则错分概率和期望值为,

$$\begin{aligned} p_{err}(w, b) &= P(\text{sign}(w \cdot \Phi(X) + b) \neq Y) \\ E(p_{err}) &\leq \frac{1}{l} E\left\{\frac{R^2(D)}{\gamma^2(L(D), Z)}\right\} \end{aligned}$$

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

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 $L(x_1, y_1; \dots; x_l, y_l)$, 则可给出此种分类器的泛化误差估计为,

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

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,

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

式(10)可进一步写为,

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

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

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

4.2 Jaakkola-Haussler bound

在1999年,Jaakkola 和 Haussler 曾证明了如下的不等式, $y_p (f^0(x_p) - f^p(x_p)) \leq \alpha_p^0 K(x_p, x_p) = U_p$

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 $x \geq 0$, 有 $\Psi(x-1) \leq x$, 故分类器模型的验证误差: Jaakkola-Haussler 上限为,

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

4.3 Oppen-Winther bound

Oppen 和 Winther 于2000年证明了如下等式,

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

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: Oppen-Winther 上限估计为,

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

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 $\alpha^0 = (\alpha_1^0, \alpha_2^0, \dots, \alpha_n^0)$ 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,

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

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 $S_p = d(x_p, \Lambda_p) \leq D_{SV}$, D_{SV} 为包含所有支持向量的超球直径。

(下转封四)