文章编号: 1001-0920(2007)06-0693-04

基于支持向量机的纺纱质量预测模型研究

吕志军, 杨建国, 项　前, 王晓玲
(东华大学 机械工程学院, 上海 201620)

摘　要: 纱线的生产是一个多环节的复杂工业过程,其质量控制大多需要依赖领域专家的个人经验,为此提出一种基于支持向量机的纱线质量预测模型. 探讨了模型选择与验证问题,并利用"网格搜索"法对模型参数进行了优化. 试验结果表明,在小样本和"噪音"数据环境下,支持向量机模型仍能保持一定的预测精度,与人工神经网络模型相比,更适应于真实纺纱生产过程.

关键词: 支持向量机; 统计学习; 预测模型; 人工神经网络; 纺纱生产
中图分类号: TS131.8　　**文献标识码**: A

Research on support vector machines based predictive model for yarn quality

LV Zhi-jun, YANG Jian-guo, XIANG Qian, WANG Xiao-ling
(College of Mechanical Engineering, Donghua University, Shanghai 201620, China. Correspondent: LV Zhi-jun, E-mail: dhcims@tom.com)

Abstract: Yarn production is a multiple stage complex industrail process, and its quality control is heavly depended upon the domain expert's experience. An SVM model for predicting yarn properties is presented, and the model parameters are optimized with "grid-research" method. Experimental results show that under the real data sets and small population circumstances, SVM models are capable of maintaining the stability of predictive accuracy, and more suitable for noisy spinning process.

Key words: Support vector machines; Statistical learning; Predictive model; Artificial neural networks; Spinning process

1 引　言

纱线的生产需经过一个掺杂行为变量、多环节的复杂工艺过程,其间包括材料位移过程、流体动力学过程、物质热交换过程、化学过程,以及借助于工艺设备顺序或并列完成的工艺操作过程[1]. 纱线生产的特点是工艺过程长,各类工艺过程既有纤维材料成形方式的不同,也有为得到不同性能和质量指标的最终制品所采用加工方法的差别. 纱线的性能指标与工艺参数之间的因果关系十分复杂,难以建立精确的数学解析模型[2]. 因此,工艺过程的控制通常还要依赖技术人员的个人经验,产品质量的波动性比较大.

针对这种情况,国内外学者进行了大量的研究,其中人工神经网络(ANNs)技术是当前普遍采用的一种建模方法[3-5]. 其基本原理是通过大样本的自我学习来映射输入和输出关系,从而推测纱线的质量指标,并以此来调节相应的工艺参数(如原料、上机参数等),达到预测控制的目的. 然而,ANNs 模型还存在某些难以解决的问题,例如需要大量的数据样本以及模型的过拟合现象等,这在很大程度上制约着该技术在工业纺纱过程中的应用[4].

近几年,一种基于统计学习理论的支持向量机(SVM)算法正受到越来越多的关注,其卓越性能体现在: 1) 与人工神经网络类似,SVM 也是一个完全基于数据的非线性建模工具; 2) SVM 模型是基于结构风险最小化原则(SRM)的,泛化性能潜力巨大; 3) SVM 的目标函数是一个凸优化问题,其最优解具有唯一性; 4) 在 SVM 模型中,应用核技术,将输入空间中的非线性问题通过非线性函数映射到高维特征空间中,在高维空间中构造线性判别函数;

收稿日期: 2006-02-19; **修回日期**: 2006-04-01.
基金项目: 国家自然科学基金项目(70371040); 国家经贸委技术创新项目(02LJ-14-05-01).
作者简介: 吕志军(1967—),男,太原人,高级工程师,博士生,从事数据挖掘、知识工程等研究; 杨建国(1951—),男,河北保定人,教授,博士生导师,从事先进制造技术、机电一体化等研究.

5) SVM 专门针对小样本情况,其最优解基于已有样本信息,而不是样本数趋于无穷大时的最优解. 正因为如此,SVM 算法自创立以来,已在许多领域取得令人鼓舞的研究成果[6,7]. 然而现有的文献中较少发现有 SVM 模型用于纺纱过程质量预测的报道,本文拟就此展开讨论.

2 基于 SVM 的回归算法

由统计学习理论[8]可知,对于回归估计 f,实际风险 $R_s(f)$ 和经验风险 $R_{emp}(f)$ 之间以至少 $(1-\eta)$ 的概率 $(\eta>0)$ 满足

$$R_s(f) \leq R_{emp}(f) + \Phi(h/n). \tag{1}$$

其中:h 为 VC 维,n 为样本数. 这一结论从理论上说明了学习机器的实际风险是由两部分组成的:一部分为经验风险(训练误差);另一部分称作置信范围,它与学习机器的复杂性和训练样本数有关. 这表明,在有限训练样本下,学习机器的 VC 维越高(复杂性越高)则置信范围越大,导致真实风险与经验风险之间可能的差别越大,这就是出现过学习现象的原因. 机器学习过程不但要使经验风险最小,还要使 VC 维尽量小以缩小置信范围,才能取得较小的实际风险,即满足 SRM 准则.

假设训练样本集

$$G = \{(x_i, y_i)\}_{i=1}^{N}.$$

其中:$x_i \in R^m$ 为输入值,$y_i \in R$ 为输出值. SVM 回归模型的基本思想就是将 m 维输入向量 x 通过某种非线性关系 ϕ 映射到高维特征空间 F 中,从而在特征空间 F 中实现线性回归

$$f(x) = \sum_{i=1}^{D} w_i \phi_i(x) + b. \tag{2}$$

其中:$\{\phi_i(x)\}_{i=1}^{D}$ 表示特征空间;未知参数 $\{(w_i)\}_{i=1}^{D}$ 和 b 分别表示权重和偏差系数,可以在样本集训练过程中获得. 为了避免出现过拟合现象进而提高模型的泛化能力,需要考虑结构风险原则并使下列函数极小化:

$$R[w] = \frac{1}{N}\sum_{i=1}^{N} |f(x_i) - y_i|_{\varepsilon} + \lambda \|w\|^2. \tag{3}$$

其中:λ 为调节因子,$|f(x_i)-y_i|_{\varepsilon}$ 定义为 Vapnik-ε 敏感损失函数[8]

$$|f(x)-y|_{\varepsilon} = \begin{cases} |f(x)-y| - \varepsilon, & |f(x)-y| \geq \varepsilon; \\ 0, & |f(x)-y| < \varepsilon. \end{cases} \tag{4}$$

式(4)中 $f(x)$ 为通过对样本集的学习而构造的回归估计函数,y 为 x 对应的目标值,$\varepsilon>0$ 为与函数估计精度直接相关的设计参数,将 ε- 不敏感损失函数形象地比喻为 ε 管道,见图 1.

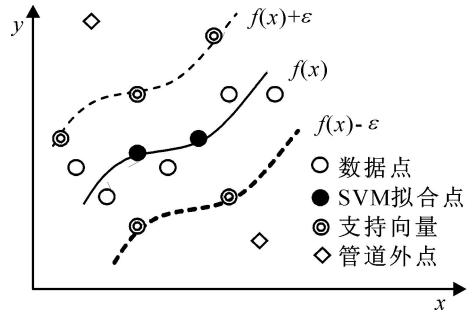


图 1 ε- 敏感损失函数下 SVM 回归模型

图 1 中,ξ_i 和 ξ_i^* 分别表示训练样本超出 ε 管道的正负偏差值,即所谓的松弛变量. SVM 通过下列条件拟合回归曲线 $f(x)$:1) 极小化松弛变量 (ξ, ξ^*) 以使训练误差极小化;2) 极小化 $\|w\|^2$ 以平滑回归曲线 $f(x)$. Vapnik[8] 认为式(3)极小化后可得

$$f(x, \alpha, \alpha^*) = \sum_{i}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b. \tag{5}$$

其中:$\alpha_i, \alpha_i^* \geq 0$ 且 $\alpha_i \alpha_i^* = 0$ 为拉格朗日算子,$K(x_i, x)$ 为核函数且有

$$K(x_i, x_j) = \sum_{i}^{N} \phi(x_i)\phi(x_j). \tag{6}$$

式(5)和(6)的一个重要特点是对于特征 $\{\phi_i(x)\}$,核函数 K 都可以被解析表达且形式相对简单. 因此,无需将矢量 x_i, x_j 直接映射到特征空间 F 中(即计算 $\phi(x_i), \phi(x_j)$)就可以计算特征空间 F 的内积,但前提是该核函数须满足 Mercer 条件. 常见的核函数有多项式、径向基以及 sigmoidal 函数等. 根据凸优化的充要条件,拉格朗日乘子 α_i, α_i^* 可由下式获得:

$$\max: R(\alpha^*, \alpha) = -\frac{1}{2}\sum_{i,j}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(x_i, x_j) - \varepsilon \sum_{j}^{N} (\alpha_j^* + \alpha_j) + \sum_{j}^{N} y_j (\alpha_j^* - \alpha_j). \tag{7}$$

$$\text{s.t.} \sum_{i}^{N} (\alpha_i^* - \alpha_i), \quad 0 \leq \alpha_i^* \leq C, 0 \leq \alpha_i \leq C. \tag{8}$$

注意到对于 Vapnik-ε 敏感损失函数而言,拉格朗日乘子 α_i, α_i^* 具有稀疏性,即只有 α_i, α_i^* 不为零所对应的向量 x 被称为支持向量. 参数 (ε, C) 需要根据实际问题由用户确定,SVM 回归算法详细的描述可以参考文献[6,8,9].

3 纱线质量预测模型

3.1 模型设计

在预测型学习任务里,样本的属性结构、模型及其参数选择是有效预测的前提,这里采用 v- 支持

向量回归机[9] 作为纱线的质量预测模型. 在该模型中,需要确定的结构参数有:

1) 惩罚因子 C,表示在决策函数的复杂性与错误决策之间的折衷程度.

2) 稀疏参数 v,近似于整个样本中所含"噪音"数据的比例. 当 $v \in [0.3, 0.6]$ 时模型有较为理想的泛化性能[10],建议取 $v = 0.54$.

3) 核函数,这里选择较为常用的径向基核

$$K(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2), \quad (9)$$

其中 σ 是核函数的带宽.

3.2　参数优化

模型选定之后,需要人为确定的参数包括 (C, σ). 可惜的是,这些参数一般无法直接获得. 这里采用网格搜索并结合交叉验证的方法[11] 对 (C, σ) 进行优化,基本算法如下:

Step1: 系统初始化,给出 (C, σ) 可能存在的取值范围.

Step2: 将参数 (C, σ) 各自分成 q 等份,两两配对并构成 $q \times q$ 个参数网格.

Step3: 依次按网格节点将相关参数代入模型中,并对其进行交叉验证,计算预测值与真值之间的差值均平方(MSE)

$$\mathrm{MSE}(\lambda) = \sum_{i}^{m} (y_{ti} - y_{pi})^2 / m. \quad (10)$$

其中: y_{ti} 和 y_{pi} 分别为第 i 组实测值和观测值, m 为交叉验证分组数,这里 $m = 5$.

Step4: 取

$$E = \min\{\mathrm{MSE}(\lambda) \mid \lambda = 1, \cdots, q \times q\}$$

时相对应的 (C, σ) 值.

Step5: 判定 E 是否可以满足精度要求 e,如果不满足则转向 Step1.

Step6: 获取优化的 (C, σ) 值,结束.

4　试验和分析

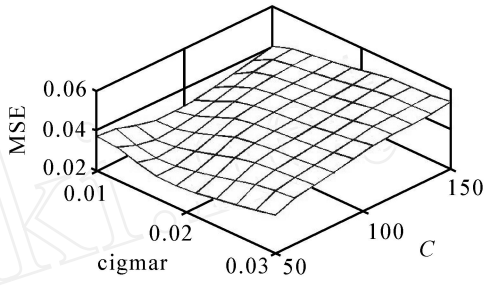
纱线生产过程中质量指标 Q 定义为

$$Q = f(A, B). \quad (11)$$

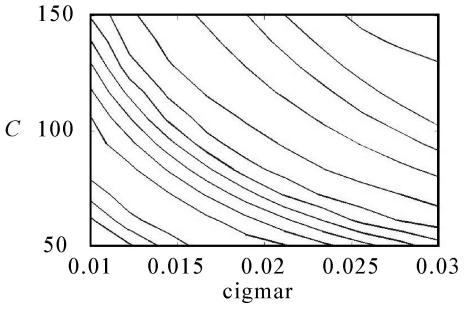
其中: A 为原料指标,包括 5 个毛纤维指标(平均细度、细度离散、豪特长度、长度离散、短毛率),纱线设计规格(纱线支数、纱线捻度); B 为纺纱设备(细纱机) 上机参数,包括牵伸倍数、锭子转速、钢丝圈号数;取纱线的两个重要质量指标断裂强力(BF) 和断头率(ED) 作为函数的输出向量 Q.

从国内某毛纺厂采集真实纺纱生产过程中的小样本数据 26 组,其中 20 组用于模型训练,其余数据用于模型测试. 确定模型的输入输出关系后,对模型进行训练,并依据网格搜索法对模型参数进行优化. 目前基于 SVM 的算法软件已相对成熟,本文模型训练采用 LIBSVM 2.8 软件包,该软件包主要应用 SMO 算法求解凸优化问题,具有快速高效的特点[11,12].

表 1 给出了可以满足精度要求的参数 (C, σ) 对以及当时的搜索范围,图 2 和图 3 给出了对应的误差风险曲面以及云图,表2给出了训练完成后模型

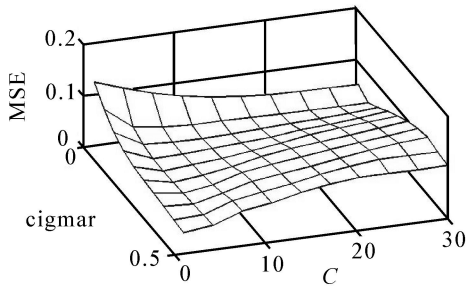


(a) 误差风险曲面

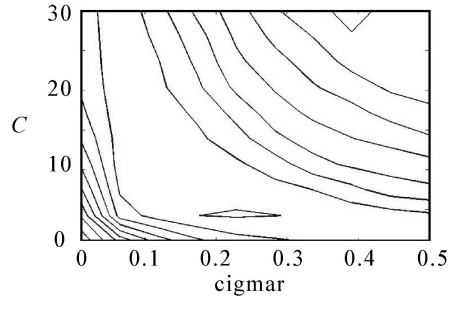


(b) 云 图

图 2　纱线断裂强力预测模型的误差风险曲面与云图



(a) 误差风险曲面



(b) 云 图

图 3　纱线断头预测模型的误差风险曲面与云图

表 1 优化的(C,σ) 参数对及其参考搜索范围

输出参数	搜索区间		优化值
	σ	C	
BF	[0.01,0.03]	[50,150]	σ = 0.012, C = 101.2
ED	[0.1,0.5]	[1,30]	σ = 0.287, C = 2.975

表 2 纺纱过程质量预测模型性能对比

试验编号	真实值		ANN 模型		SVM 模型	
	BF	ED	BF	ED	BF	ED
1	103	60	113.89	70.41	116.24	72.06
2	75	60	61.91	75.78	76.87	72.40
3	152	30	153.46	39.40	156.57	42.22
4	75	65	61.91	75.78	76.87	72.40
5	70	60	47.00	69.84	76.86	59.31
6	65	90	66.76	79.22	66.62	81.27
R			0.96	0.88	0.99	0.91
MSE			165.92	129.16	42.14	96.66
e%			13.67	19.99	5.52	17.29
误差大于10%的个数			4	6	1	3

测试试验的结果,通过相关系数(*R*)、差值均平方(MSE)和平均相对误差(*e*%)统计指标与对应的人工神经网络模型进行了比对.

可以看到,就 BF 而言,SVM 模型的预测精度与可靠性均比ANN模型有明显提高;受设备操作者与观测者的局限[1],ED 一直是一个比较难以预测的主观指标.但即使这样,SVM 模型的预测能力同 ANN 模型相比较也得到了一定程度的改善.

5 结 语

SVM 是一个新兴的基于统计学习理论的数据挖掘方法,其优势在于小样本学习、解的唯一性和稀疏性、良好的泛化能力等.这里建立了基于 SVM 的纺纱过程质量预测模型,探讨了模型的选择与参数优化问题.试验数据显示,SVM 模型在小样本以及"噪音"数据环境下仍能保持一定的预测精度,同人工神经网络模型相比,其泛化性能有所提高,更适合于真实的纺纱生产过程.当然,与其他任何新兴技术一样,SVM 的工业应用仍需深入和拓展,如在核函数的设计方面尚待进一步优化.研究表明,SVM 模型可以为纺纱企业进行质量控制与决策提供一个新的有效工具.

参考文献(References)

[1] Peter R. Lord handbook of yarn production [M]. Abinhton: Woodhead Publishing Limited, 2003.

[2] Les M Sztanera. Soft computing in textile sciences[M]. New York: Physica-Verlag Heidelberg, 2003.

[3] Chattonpadhyay R, Guha A. Artificial neural networks: Applications to textiles [J]. Textile Progress, 2004, 35(1): 1-42.

[4] Rafael Beltran, Wang L J, Wang X G. Predicting worsted spinning performance with an artificial neural model[J]. Textile Research J, 2004, 74(9): 757-763.

[5] Sette S, Boullart L, Langenhove Van. Using genetic algorithms to design a control strategy of an industrial process[J]. Control Engineering Practice, 1998, 7(6): 523-527.

[6] Sanchez David V. Advanced support vector machines and kernel methods[J]. Neurocomputing, 2003, 55(3): 5-20.

[7] 常玉清, 邹伟, 王福利, 等. 基于支持向量机的软测量方法研究[J]. 控制与决策, 2005, 20(11): 1307-1310.
(Chang Y Q, Zou W, Wang F L, et al. Research on soft sensing method based on support vector machines [J]. Control and Decision, 2005, 20(11): 1307-1310.)

[8] Vapnik V N. Statistical learning theory [M]. New York: Wiley, 1998.

[9] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京: 科学出版社, 2004.
(Deng N Y, Tian Y J. A novel method in data mining: Support vector machines[M]. Beijing: Science Press, 2004.)

[10] Athanassia Chalimourda, Scholkopt B, Smola A. Experimentally optimal *v* in support vector regression for different Noise models and parameter settings[J]. IEEE Trans on Neural Networks, 2004, 15(2): 127-141.

[11] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification [DB/OL]. http://www.csie.ntu.edu.tw/~cjlin/papers/guide, 2003-06.

[12] Lin C J. Asymptotic convergence of an SMO algorithm without any assumptions[J]. IEEE Trans on Neural Networks, 2002, 13(2): 248-250.