一种改进的最小二乘支持向量机软测量建模方法

An Improved LS-SVM Soft Sensing Modeling Method

毛晓娟　何小阳　温伟峰
（广西大学电气工程学院，广西 南宁　530004）

摘　要：针对最小二乘支持向量机（LS-SVM）缺少支持向量所具有的稀疏性和模型参数值难以选择的问题，提出利用马氏距离进行样本相似程度分析，去除集中部分样本，以恢复最小二乘支持向量机的稀疏性的方法。同时，采用 k-折交叉验证误差作为学习目标的粒子群优化算法来选取模型参数，并利用改进算法建立了精馏产品浓度的软测量模型。通过仿真验证了改进算法的有效性。结果表明模型精度较高，泛化能力强，满足工业测量要求。

关键词：最小二乘支持向量机　粒子群优化算法　软测量　建模
中图分类号：TP274　　**文献标志码**：A

Abstract: Aiming at the demerits of the least square support vector machine, such as losing the sparseness of support vector, and difficulty in selecting parameters of model, it is proposed that by using Mahalanobis distance to analyze the similarities among samples, for recover the sparseness of the least square vector machine. In addition, by adopting particle swarm optimization algorithm that with k-fold cross-validation error as learning object to select parameters of the model, and the improved algorithm is used to establish the soft sensing model for concentration of the product of distillation. The simulation verifies the effectiveness of the improved algorithm, and the result of research shows the model is accurate, and good for generalization to meet industrial measuring requirements.

Keywords: Least square support vector machine (LS-SVM)　Particle swarm optimization algorithm　Soft sensing　Modeling

0　引言

支持向量机（support vector machine，SVM）是由 Vapnik 等人在 20 世纪 90 年代中期提出的一种基于统计学习理论的机器学习方法[1]，它能较好地解决小样本、非线性等实际问题，已在许多领域取得了成功应用[2-4]。最小二乘支持向量机（least square support vector machine，LS-SVM）是标准支持向量机的一种扩展。它只求解线性方程，有效地简化了计算过程，提高了运算速度，在函数估计和逼近中得到了广泛的应用。但是，在 LS-SVM 中，所有的训练数据都成了支持向量，缺少标准支持向量机的“稀疏”特性，模型结构变得复杂。同时，实践表明，LS-SVM 的性能与核函数类型、核函数的参数和正规化参数 C 有很大关系。因此，如何合理地选取最优参数，对 LS-SVM 的模型精度和泛化能力有较大的影响，如安全仪表系统（safety instrumented system，SIS）

本文利用统计学中的马氏距离衡量各样本的相似程度，去除原始样本中的部分样本，使 LS-SVM 方法具有一定的“稀疏”性。同时，采用粒子群优化算法（particle swarm optimization，PSO）对 LS-SVM 的模型参数进行寻优。

1　改进的 LS-SVM 回归方法

1.1　LS-SVM 建模理论

l 组 n 维样本向量表示为 $(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_l,y_l)\in R^n\times R, i=1,2,\cdots,l$。LS-SVM 的基本思想是选择一个非线性映射 $z=\psi(\cdot)$，把输出样本向量从原空间映射到高维特征空间，在此高维特征空间构造最优线性回归函数 $f(x)=w\psi(x)+b$ 来进行数据拟合，这样非线性估计函数就转化为高维特征空间中的线性估计函数。LS-SVM 在优化目标中的损失函数为误差 ξ_i 的二次项[5]，则优化目标为：

$$\begin{cases}\min R(\boldsymbol{w})=\dfrac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\dfrac{C}{2}\sum\limits_{i=1}^{l}\xi_i^2\\ y_i=\boldsymbol{w}^{\mathrm{T}}\psi(x_i)+b+\xi_i\end{cases}\tag{1}$$

式中：$R(\boldsymbol{w})$ 为结构风险；C 为正规化参数。用拉格朗日法求解优化问题，有：

$$L(\boldsymbol{w},b,\xi_i,\alpha_i)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{C}{2}\sum_{i=1}^{l}\xi_i^2-\sum_{i=1}^{l}\alpha_i[\boldsymbol{w}\psi(x_i)+b+\xi_i-y_i]\tag{2}$$

修改稿收到日期：2011－01－11。

第一作者毛晓娟，女，1984 年生，现为广西大学控制理论与控制工程专业在读硕士研究生；主要从事综合自动化方面的研究。

式中:α_i 为拉格朗日乘子。定义核函数 $K(x_i,x_j)=\psi(x_i)\psi(x_j)$,$K(x_i,x_j)$是满足 Mercer 条件的对称函数。按照优化条件:$\frac{\partial L}{\partial \boldsymbol{w}}=0$、$\frac{\partial L}{\partial b}=0$、$\frac{\partial L}{\partial \xi_i}=0$、$\frac{\partial L}{\partial \alpha_i}=0$,则优化问题转化为如下求解线性方程:

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_i,x_j)+1/C & \cdots & K(x_i,x_j) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & K(x_i,x_j) & \cdots & K(x_i,x_j)+1/C \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \vdots \\ \alpha_l \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_l \end{bmatrix} \tag{3}$$

最后得到 LS-SVM 回归模型为:

$$f(x)=\sum_{i=1}^{l}\alpha_i K(x,x_i)+b \tag{4}$$

式中:核函数 $K(x,x_i)$ 采用径向基函数(radial basis function,RBF),即

$$K(x,x_i)=\exp(-\|x-x_i\|/2\sigma^2) \tag{5}$$

式中:σ 为核宽。

1.2　马氏距离

马氏距离是由印度统计学家马哈拉诺比斯于 1936 年引入的,具体定义如下[6]。

样本向量 $\boldsymbol{x}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_p]^{\mathrm{T}}$,其中,$\boldsymbol{x}_i=[x_{1i},\cdots,x_{ai},\cdots,x_{ni}]^{\mathrm{T}}$;$\boldsymbol{S}$ 为样本的协方差矩阵,且 $\boldsymbol{S}=(\sigma_{ij})_{p\times p}(i、j=1,2,\cdots,p;a=1,2,\cdots,n)$,即:

$$\sigma_{ij}=\frac{1}{n-1}\sum_{a=1}^{n}(x_{ai}-\bar{x}_i)(x_{aj}-\bar{x}_j) \qquad \bar{x}_i=\frac{1}{n}\sum_{a=1}^{n}x_{ai} \tag{6}$$

如果 $\boldsymbol{S}^{-1}$存在,则两个样本之间的马氏距离为:

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=[(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)]^{1/2} \tag{7}$$

2　PSO 算法

PSO 算法是群体智能的一个新的分支,由 Kennedy 和 Eberhart 于 1995 年首次提出[7],源于对鸟群捕食行为的研究。

PSO 算法初始化为一群随机粒子,粒子根据对个体和群体的飞行经验的综合分析来动态调整自己的速度,在解空间中进行搜索,通过迭代找到最优解。在每一次迭代中,粒子通过跟踪两个“极值”来更新自己:一个是粒子本身所找到的最优解 P_{ibest},即个体极值;另一个极值是整个种群目前找到的最优解 P_{gbest},即全局极值。在找到这两个最优值时,粒子根据如下两个公式来更新自己的速度和位置。

$$V_i(t+1)=w(t)V_i(t)+c_1r_1[P_{\mathrm{ibest}}(t)-X_i(t)]+c_2r_2[P_{\mathrm{gbest}}(t)-X_i(t)] \tag{8}$$

$$X_i(t+1)=X_i(t)+V_i(t) \tag{9}$$

式中:$X_i(t)$为粒子的当前位置;$V_i(t)$为粒子的当前速度;c_1 和 c_2 为学习因子,一般取 $c_1=c_2=2$;r_1 和 r_2 为介于(0,1)之间的随机数;$w(t)$为惯性权重,它随进化代数 t 从 0.9 线性递减至 0.4。

$$w(t)=0.9-(t/T_{\max})\times 0.5 \tag{10}$$

式中:$T_{\max}$为最大迭代次数。

3　改进的 PSO-LSSVM 算法

在利用 PSO 算法进行参数寻优时,适应度函数的大小一般采用 LS-SVM 模型的泛化能力估计值,交叉验证法可以用来评估模型的泛化能力。k-折交叉验证法是交叉验证法中的一种,具有估计无偏性。因此,采用 k-折交叉验证误差作为 PSO 算法中适应度函数的目标值。

k-折交叉验证需要把整个样本分为 k 个互不包含的子集,然后对数据进行 k 次训练与测试。经过 k 次测试后,将这 k 个泛化误差取平均值得到 k-折交叉验证误差,以此作为评价模型预测效果的依据,以调整模型参数[8]。

改进 PSO-LSSVM 建模方法描述如下。

① 将采样样本集经预处理后,根据样本数据,初始化参数 C 和 σ。

② 把整个样本分为 k 个互不包含的子集,将其中一个子集作为测试数据集,其余的子集合并为训练数据集 T,计算训练样本 T 两两之间的马氏距离 d。当 $d<\varepsilon$时,则认为是一对相似样本,删除相似样本中的一个样本,得到新的训练样本集 T'。利用 T'建立 LS-SVM 回归模型,再用回归模型对测试数据进行测试并计算泛化误差,然后对数据进行 k 次训练与测试。

③ 对 k 个泛化误差取平均值,该平均值作为 PSO 进化算法的个体适应度函数值。利用 PSO 算法对 C 和 σ 寻优,终止条件为达到最大迭代次数 $T_{\max}$。结束后将最优参数值 C 和 σ 赋予 LS-SVM,并重新建立回归模型。

4　仿真实例

为了验证上述改进 PSO-LSSVM 算法的有效性,本文采用一维测试函数 sinc 进行仿真测试。取 sinc 函数为:

$$y=\mathrm{sinc}(x)+v \qquad x\in[-5,5] \tag{11}$$

式中:v 是均值为 0、方差为 0.1 的高斯噪声。令 $x_i=-5+0.1i(i=1,\cdots,200)$,这样可在 $x\in[-5,5]$区间内取得 200 个样本点。将 x 的下标为 5 的倍数的样本点作为模型的测试样本,共 40 个样本,剩余的 160 个样本作为模型的训练样本。利用改进的 PSO-LSSVM

算法对该测试函数建立模型,其中,粒子数为20,最大迭代次数为100。经过PSO寻优,最优参数$C=453.77$,$\sigma=1.25$。图1为改进的PSO-LSSVM方法的sinc函数仿真结果,可见模型拟合效果较好。

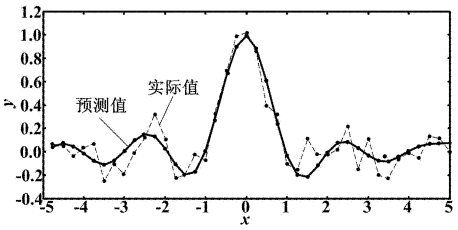

图1　sinc函数仿真结果

Fig. 1　Simulation result of sinc function

为了便于比较,分别利用RBF神经网络、基本的PSO-LSSVM和改进的PSO-LSSVM建立软测量模型,测试结果如表1所示。其中,性能指标分别为均方根误差MSE、平均泛化误差ABSE(误差绝对值的平均值)和最大绝对值误差。从表1中可以看出,利用改进的PSO-LSSVM方法所得的模型精度最高,泛化能力最强。

表1　三种软测量模型泛化结果比较

Tab. 1　Comparison of generalization results of three kinds of soft-sensing models

模　型	MSE	ABSE	最大绝对值误差
RBF神经网络	0.121 0	0.097 1	0.340 4
PSO-LSSVM	0.120 4	0.096 8	0.370 9
改进PSO-LSSVM	0.112 8	0.089 6	0.296 3

5　软测量建模的应用

精馏塔作为典型的化工单元,是大型石油化工企业生产的重要设备。为了对精馏塔实施先进控制和优化操作等任务,需要获取一些重要过程变量的实时数据,如精馏产品浓度。但是目前尚无理想工业化测量仪表能对其进行在线测量,而传统测量方法往往难以满足生产实时性要求。因此,精馏产品浓度在线软测量技术的研究意义重大。

根据精馏过程的工艺流程及现场经验,将精馏产品浓度作为软测量模型的主导变量,进料组成、进料流量、回流量和再沸器功率这四个变量作为辅助变量。

由于现场采集的数据过少,不能完全反映精馏过程的生产特性,所以,利用机理模型获取更多的数据。通过阿依达准则和平均值滤波方法筛选出200组数据,随机选择其中160组作为训练样本,剩余40组作为测试样本。

利用改进的PSO-LSSVM算法对精馏产品浓度进行软测量建模,模型参数设置为粒子群规模20、$T_{max}=80$、$\varepsilon=0.1$、$k=5$。独立运行30次,最优参数确定为$C=5\ 000$、$\sigma=38.1$。精馏产品浓度软测量模型泛化结果如图2所示。

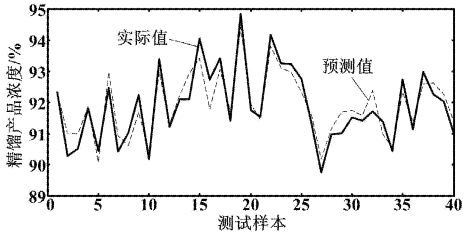

图2　泛化结果

Fig. 2　Generalization results

由图2可以看出,模型的预测值与实际值的拟合程度较好,模型测试过程满足要求,泛化性能较好,模型精度较高。

为了进一步验证改进PSO-LSSVM软测量方法的有效性,采用基本的PSO-LSSVM对精馏产品浓度进行软测量建模,泛化结果如表2所示。

表2　泛化结果比较

Tab. 2　Comparison of generalization results

模　型	MSE	ABSE	最大绝对值误差
PSO-LSSVM	0.470 9	0.396 7	0.951 6
改进PSO-LSSVM	0.422 1	0.360 5	0.910 0

从表2中可以看出,改进后的PSO-LSSVM模型的均方根误差MSE、平均泛化误差ABSE以及最大绝对值误差均小于基本的PSO-LSSVM模型,表明改进方法的泛化能力较好,改进方法有效。

6　结束语

针对LS-SVM方法由于支持向量个数多,模型结构复杂,失去了支持向量的稀疏性的问题,利用马氏距离进行样本相似程度分析,去除集中部分样本,以恢复LS-SVM的稀疏性,简化了模型结构。LS-SVM的参数选择一直是推广其应用的关键所在,本文在采用PSO算法进行寻优的基础上,以最小化k-折交叉验证误差作为PSO算法的个体适应度函数值。测试函数验证了改进PSO-LSSVM算法的有效性,并用改进算法建立了精馏产品浓度的软测量模型。仿真结果表明,该算法具有模型精度高、泛化能力强等特点,为精馏过程中难以在线软测量的产品浓度的实时检测提供了有效的手段。

(下转第45页)

制效果;基本没有迟滞现象,超调≤10%,稳态误差10 K左右,都在允许的范围内。

从响应曲线看,由于仿真时间设置得太短,系统并没有进入稳定状态,但跟踪效果已经有了很明显的改善。在仿真的末段出现了一个短时间的小尖峰,那是因为不小心触及了三段电阻炉上的温度传感器,导致采样温度与实际温度出现了巨大的偏差。

以上调试效果与使用 Matlab 仿真相比有着很大的差距,这可以从下四个方面进行分析:①仿真时的控制频率要比实验室调试下的控制频率大得多,即不在一个数量级上,控制频率与控制效果有很大的关系;②在对控制算法进行软件编程时,对模糊算法进行了简化处理,大大降低了控制算法的准确度;③在进行仿真时,控制对象的数学模型是不变的,但是在试验时,控制对象的数学模型是时变的;④温度越高,传感器的精度就越高,当温度较低时,传感器的精度就很不理想,而使用 Matlab 进行仿真就不存在这个问题。

5　结束语

针对电阻炉数学模型时变的特点,提出了用模糊控制器进行炉温控制系统设计的方法。本文设计了带积分输出的模糊控制器,用以解决基本模糊控制器难以兼顾系统动、静态性能的问题,并进行了仿真。最后借助组态王监控系统软件,以西门子 S7-200 可编程序控制器为平台对控制算法进行了调试。仿真和调试结果表明,该控制器可以改善基本模糊控制器的性能,并较好地运用到热处理电阻炉的炉温控制系统中。

参考文献

[1] 梁艳妮,黄运生. 软熔性能测试过程智能温度控制系统的设计[J]. 自动化仪表,2009,30(4):55 -60.

[2] Pivonka P. Comparative analysis of fuzzy PI/PD/PID controller based on classical PID controller approach[C]//Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, Honolulu, HI, USA: Pivonka P, 2002: 541 -546.

[3] 吴光英,吴光治,孙桂华. 新型热处理电炉[M]. 北京:国防工业出版社,1993.

[4] 潘健,刘斌,陈刚. 基于 Active-X 技术的电阻炉温度控制系统[J]. 控制工程,2008,15(1):61 -63.

[5] Chen Jiaxin, Li Wei. Application of fuzzy control PID algorithm in temperature controlling systems[C]//2003 International Conference on Machine Learning and Cybernetics, 2003: 2601 -2604.

[6] 粟梅,吴军,宋冬然. 多段电阻炉温度控制系统的设计与实现[J]. 自动化技术与应用,2007,26(3):58 -61.

[7] 马国华. 监控组态软件及其应用[M]. 北京:清华大学出版社,2001.

(上接第 41 页)

参考文献

[1] Vapnik V N. The nature of statistical learning theory [M]. New York: Springer-Verflag, 1995.

[2] Chen Wenjie, Wang Jing. Application of support vector machine in industrial process[J]. Computers and Applied Chemistry, 2005, 22 (3): 195 -200.

[3] 张健,李艳,朱学峰,等. 基于支持向量机的蒸煮过程卡伯值软测量[J]. 计算机测量与控制,2004,12(2):104 -106.

[4] 常玉庆,王福利,王小刚,等. 基于支持向量机的生物发酵过程软测量建模[J]. 东北大学学报:自然科学版,2005,26(11):1025 -1028.

[5] 方瑞明. 支持向量机理论极其应用分析[M]. 北京:中国电力出版社,2007:20 -21.

[6] 高惠旋. 应用多元统计分析[M]. 北京:北京大学出版社,2005:50 -53.

[7] Kennedy J, Eberhart R. Particle swarm optimization[C]//Proceedings of IEEE International Conference on Neural Networks, Perth, WA, Australia, 1995: 1942 -1948.

[8] 绍信光,杨慧中,陈刚. 基于粒子群优化算法的支持向量机参数选择及其应用[J]. 控制理论与控制工程,2008,23(5):740 -743.

行业信息

利德华福三次荣登“福布斯中国潜力企业榜”

近日,“2011‘福布斯中国潜力企业投资论坛’暨‘2011 福布斯中国最佳潜力企业颁奖典礼’”在上海隆重举行。北京利德华福电气技术有限公司已是第 3 次连续入榜的高压变频器企业。

此次评选,是福布斯中文版第七次对全国 2 万余家销售收入在 500 万元 ~10 亿元的中小企业进行全面、独立的调研,并根据增长性指标、回报率指标、盈利性指标(净利率)和销售及利润规模指标对候选企业进行综合排名,从中遴选出 200 家最具发展潜力的中小企业,其中包括 100 家上市公司和 100 家非上市公司。

对于自身的发展,利德华福制定了三条基本原则:把握本领域技术进步趋向、不断开辟新应用领域、树立良好企业形象。依靠这三条原则,利德华福才得以从名不见经传的企业发展成为如今颇具规模的企业,并力争在变频调速领域占有一席之地。