PSO-SVR 算法在发酵过程控制中的应用

陈 树,徐保国,王海霞,吴晓鹏
CHEN Shu,XU Bao-guo,WANG Hai-xia,WU Xiao-peng

江南大学 通信与控制工程学院,江苏 无锡 214122
School of Communication and Control Engineering,Southern Yangtze University,Wuxi,Jiangsu 214122,China
E-mail:kjcuse@sytu.edu.cn

CHEN Shu,XU Bao-guo,WANG Hai-xia,et al.Study of fermentation process based on PSO-SVR.Computer Engineering and Applications,2007,43(19):214-216.

Abstract: In accordance with the hardship to get real-time and on-line biology parameters in fermentation process,a soft sensor model based on Support Vector Machines (SVM) is established for estimating the biology parameters.It is well known that the model complexity and generalization performance of this Support Vector Regression(SVR) model depend on a good setting of the three parameters(ε,C,γ).In this article,an algorithm called Particle Swarm Optimization(PSO) is applied to optimize the parameters (ε,C,γ) at the same time.Based on the proposed method,a PSO-SVR model is developed to estimate the products concentration of L-Asparaginase II.The results from fermenter control show that the PSO-SVR state estimation model has good learning accuracy and generalization performance so as to acquire the real-time and on-line estimation value of L-Asparaginase II products concentration.

Key words: Support Vector Regression(SVR);state estimation;Particle Swarm Optimization(PSO) algorithm;L-Asparaginase II

摘 要:针对发酵过程中生物参数难以实时在线测量的问题,建立了用于生物参数状态预估的支持向量机软测量模型。考虑到该支持向量回归模型的复杂性和推广能力的好坏很大程度上取决于其3个参数(ε,C,γ)能否取到最优值,采用粒子群算法实现对参数(ε,C,γ)的同时寻优。在此基础上,以L-天冬酰胺酶II为对象,建立其基于PSO-SVR的发酵过程产物浓度状态预估模型。发酵罐控制结果表明:该模型具有很好的学习精度和泛化能力,可实现对L-天冬酰胺酶II产物浓度的实时在线预估。

关键词:支持向量回归(SVR);状态预估;粒子群优化(PSO)算法;L-天冬酰胺酶II
文章编号:1002-8331(2007)19-0214-03 **文献标识码**:A **中图分类号**:TP391

1 引言

发酵过程控制有三个目的:提高过程稳定性、克服各类扰动、优化过程行为。前两个目的可通过发酵过程环境变量的智能控制来达到。后一个是过程性能优化,在补料优化中,先求取一些变量的优化轨线,然后在实际生产中加以控制,使它们按优化轨线变化。因此,发酵过程某些重要生物变量在线检测是发酵过程实施优化控制的基础。在开发生物传感器技术的同时,重点开发软测量状态变量估计技术具有重要的实际意义。SVM是基于统计学习理论的机器学习算法,SVR是建立在SVM思想上的回归算法,具有很强的理论优势,但是其理论优势得以实现的前提条件是要选取到合适的回归参数。对于ε-SVR,不敏感损失系数ε、惩罚系数C、核函数及其参数的优化选择对回归模型的学习精度和推广能力的好坏起着决定性作用。目前SVR的参数选择方法多是建立在经验和试凑的基础之上[1,2],以交叉验证法(CV)和留一法(LOO)为代表,耗时且缺乏通用性,模型的回归效果也得不到保证。对此,本文提出采用粒子群算法(Particle Swarm Optimization,PSO)实现对3个参数(ε,C,γ)的优化选择。并以L-天冬酰胺酶II发酵过程产物浓度为对象,建立其基于PSO-SVR的状态预估模型。通过实际发酵应用表明:该模型具有很好的学习精度和预测能力。

2 粒子群算法优化SVR参数对(ε,C,γ)

支持向量回归机(Support Vector Regression,SVR)是建立在支持向量机思想上的回归算法,实质是用非线性函数$f(x)=\omega\cdot\phi(x)+b$(其中$\phi(x)$为一非线性映射)拟合如下的样本数据集$\{x_i,y_i\}(i=1,2,\cdots,l;x_i\in R^n;y_i\in R)$,也就是在约束条件下寻找最优拟合超平面。最后得到的回归函数为:

$$f(x)=\omega\cdot\phi(x)+b=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)(\phi(x_i)\cdot\phi(x))+b^*=\sum_{i=1}^{n}(\hat{\alpha}_i-\alpha_i)K(x_i,x)+b^* \quad (1)$$

式中α_i、$\hat{\alpha}_i$为拉格朗日算子,b^*为阈值,$K(x_i,x)=\phi(x_i)\cdot\phi(x)$为核函数。在没有足够的先验知识时,很多研究和实验[3]表明,径

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2002AA217031-2)。
作者简介:陈树(1969-),博士生,研究方向:过程控制与模式识别;徐保国,教授,博士生导师,研究方向:过程控制与优化。

向基核函数(RBF,Radial Basis Function)是个不错的选择。本文就采用 RBF 核,即:

$$K(x_i,x_j)=\exp\left(-\frac{|x_i-x_j|^2}{2\gamma^2}\right) \tag{2}$$

现有 SVR 的研究表明[4]:SVR 模型的学习精度、泛化能力的好坏很大程度上取决于模型的不敏感损失系数 ε、惩罚系数 C、核函数及其参数是否同时取得最优。

PSO 算法是基于种群的并行全局搜索策略,采用简单的速度、位移模型实现对整个解空间的寻优,因而概念简单易于实现,且没有许多参数需要调整,具有更快的收敛速度,对处理高维优化问题也有一定的优势。粒子群可用下面 5 元素[5]形式描述:

$$PSO=(n,K_{iter},V,X,F_{fit}) \tag{3}$$

其中,n 为群体规模,K_{iter} 为迭代次数,V 和 X 分别表示所有粒子的速度空间和位置空间,F_{fit} 为适应度,从位置空间映射到实数空间。在每一次迭代中,粒子通过跟踪个体极值 $pbest$ 与全局极值 $gbest$ 来更新自己的位置和速度,公式如下:

$$v_i(t+1)=\omega*v_i(t)+c_1*rand()*(p_ibest-x_i(t))+ c_2*rand()*(gbest-x_i(t)) \tag{4}$$

$$x_i(t+1)=v_i(t)+v_i(t+1) \tag{5}$$

其中,$v(t)$ 是粒子的当前速度,$x(t)$ 是粒子的当前位置,$rand()$ 是[0,1]之间的随机函数,c_1 和 c_2 是学习因子,通常都取值为 2,ω 是加权系数,用来控制历史速度对当前速度的影响程度,一般在[0.1,0.9]之间取值。但若 ω 能随算法迭代的进行而线性减小,将显著改善算法的收敛性能。令 $\omega_{\max}$ 为最大加权系数,$\omega_{\min}$ 为最小加权系数,$iter$ 为当前迭代次数,$iter_{\max}$ 为算法总迭代次数,则有:

$$\omega=\omega_{\max}-iter*\frac{\omega_{\max}-\omega_{\min}}{iter_{\max}} \tag{6}$$

一般 $\omega_{\max}$ 取值为 0.9,$\omega_{\min}$ 取值为 0.4,更新过程中,粒子每一维的最大速率限制在 $v_{\max}$,粒子每一维的坐标也被限制在允许范围之内。同时,个体极值 $pbest$ 和全局极值 $gbest$ 在迭代过程中不断更新,最后输出的 $gbest$ 就是算法寻优到的最优解。

对于 PSO 优化 SVR 参数对(ε,C,γ),每个粒子由三维参数(ε,C,γ)决定其位置和速度,适应度函数就取为能直接反映 SVR 回归性能好坏的均方差(MSE):

$$MSE=\sqrt{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2/n} \tag{7}$$

那么,基于 PSO 的 ε-SVR 参数对(ε,C,γ)的优化选择步骤如下:

(1)粒子群初始化,确定群体规模 n,随机产生每个粒子的位置和速度;给定算法的最大加权系数 $\omega_{\max}$ 和最小加权系数 $\omega_{\min}$;设定算法的最大迭代次数 $G_{\max}$。

(2)将每个粒子的 $pbest$ 设置为当前位置,$gbest$ 设置为群体中最好粒子的当前位置。

(3)评价所有粒子的适应度。

(4)若某粒子当前适应度优于 $pbest$,$pbest$ 被当前位置替换;若所有粒子的当前最优适应值优于 $gbest$,$gbest$ 被当前最优位置替换。

(5)根据公式(4)、(5)、(6),更新每个粒子的位置与速度。

(6)检查终止条件,若达到最大迭代次数 $G_{\max}$ 或最优解停滞不再变化,就终止迭代;否则返回步骤(3)。

其中,文献[4]给出了(ε,C,γ)大致范围:$\varepsilon\in[0,0.2]$,$C\in[1,10^8]$,$\gamma\in[0.01,2.0]$,这样初始化参数对时可避免盲目选择。

对于 PSO-SVR 的实现,相当于把 SVR 模型的构造、预测算法嵌入到 PSO 计算适应值的步骤当中。这样,大家也许认为模型的回归精度应该有一定提高,但同时计算量也不会小。其实,用 PSO-SVR 时,只需随机选取少量样本进行模型训练,其余大量样本作为测试之用。这点不同于 CV、LOO 方法,它们是将大量样本作为训练,少量样本用作测试。所以,从理论上说,PSO-SVR 还是比 CV、LOO 节省时间的,特别是对噪声敏感的大样本数据集更能体现 SVR 和智能优化算法的优势。下面就用 PSO-SVR 建立 L-天冬酰胺酶 II 浓度的预估模型来验证这点。

3　基于 PSO-SVR 的 L-天冬酰胺酶 II 发酵过程状态预估

3.1　模型结构

L-天冬酰胺酶 II 发酵生产工艺采用典型的分批补料培养,在 25L 发酵罐中流加基质料液,基质浓度主要依靠补玉米浆混合物来维持。在其发酵过程中,最能体现生物发酵状态的测量值是补料速率、pH 值、溶解氧、菌体浓度、基质浓度、产物浓度和发酵时间等生化参数。因此,选择 PSO-SVR 的模型输入为:发酵当前时刻、当前时刻的补料速率、当前时刻的 pH 值、当前时刻的溶氧值、当前时刻的菌体浓度、当前时刻的基质浓度、当前时刻的产物浓度,模型的输出为下一个时刻的产物浓度,训练样本集主要来自实验室的发酵报表数据。

那么,对应于非线性 ε-SVR 的回归模型状态估计原理,可得到:所需拟合的样本数据集$\{x_i,y_i\}(i=1,2,\cdots,n;x_i\in R^d;y_i\in R)$中的 d=7,即回归模型输入为 7 维,输出为 1 维;另外,L-天冬酰胺酶 II 的最优发酵时间为 10 h 左右,每隔 0.5 h 采样一次,所以公式中的 n=20。

3.2　PSO 优化 SVR 参数对(ε,C,γ)

取 L-天冬酰胺酶 II 发酵过程中产物浓度较高的 4 组数据,利用 PSO 优化 SVR 参数对(ε,C,γ)的步骤如下:

步骤 1　以数据组 1 和 2 作为一对,用数据组 1 的输入、输出矩阵进行训练得到回归函数 $y=\omega\cdot x+b$,把数据组 2 的输入样本代入该回归函数求得其输出的估计值,取数据组 2 的输出估计值和测量值之间的均方差作为 PSO 算法适应度函数。由此,可以得到针对数据组 1 和 2 的一个最优参数对 $P1=(\varepsilon,C,\gamma)$=[0.208 0　100.082 0　2.002 1]。

步骤 2　再以数据组 1 和 3 作为一对,处理步骤同步骤 1,这样又可以得到第二组最优参数对 $P2=(\varepsilon,C,\gamma)$=[0.205 3　110.002 0　1.992 0]。

步骤 3　再以数据组 1 和 4 作为一对,处理步骤同步骤 1,这样又可以得到第三组最优参数对 $P3=(\varepsilon,C,\gamma)$=[0.210 1　115.002 1　1.939 7]。

步骤 4　对 $P1$、$P2$ 和 $P3$ 的对应参数元素分别取平均值,从而可以得到该发酵过程产物浓度 PSO-SVR 模型的最优参数对 $P=(P1+P2+P3)/3$=[0.207 8　108.362 0　1.995 9]。

3.3　PSO-SVR 模型预估

由上节内容得到的最优参数对 P=[0.207 8　108.362 0

1.995 9]就是 L-天冬酰胺酶 II 发酵过程产物浓度 PSO-SVR 预估模型的最佳参数选择,理论上说,由此所得的模型应该有较好的预测(状态预估)能力。模型应用于 5 批检验样本集,输出的产物浓度平均均方误差为 0.041,其中一批检验集的检验情况如表 1 所示,表中最大相对误差为 0.049,最小相对误差为 0.027,效果比较理想。

表 1 某批次检验集的检验情况

产物浓度(效价)/(万单位/L)							
时间/h	检验集测量值	模型输出值	相对误差	时间/h	检验集测量值	模型输出值	相对误差
2.5	0.14	0.147	+0.043	6.5	3.69	3.825	+0.036
3.0	0.16	0.168	+0.044	7.0	3.82	3.976	+0.041
3.5	0.17	0.161	−0.047	7.5	4.11	3.934	−0.043
4.0	0.28	0.268	−0.039	8.0	4.23	4.120	−0.027
4.5	0.54	0.518	−0.041	8.5	4.29	4.143	−0.034
5.0	1.02	1.048	+0.029	9.0	4.32	4.529	+0.049
5.5	1.78	1.702	−0.043	9.5	4.44	4.592	+0.034
6.0	2.63	2.526	−0.039	10.0	4.41	4.209	−0.046

为了进一步验证 PSO-SVR 算法的效果,将 PSO-SVR 算法、CV-SVR 算法、CMAC 算法分别对 5 批检验样本集进行测试。CV-SVR 算法输出的产物浓度平均均方误差为 0.061,CMAC 算法输出的产物浓度平均均方误差为 0.078,远高于 PSO-SVR 算法。

4 结论

本文以 L-天冬酰胺酶 II 发酵为对象,首先建立了旨在对其产物浓度状态预估的支持向量回归模型。鉴于该 SVR 模型回归性能(学习能力和推广能力)的好坏取决于模型参数(ε,C,γ)是否取到最优值,本文尝试着采用粒子群算法对这 3 个参数同时进行寻优。基于检验样本集发酵数据的仿真结果表明,该 PSO-SVR 模型具有很强的学习精度和良好的预测能力,基本可以实现对产物浓度的实时在线估计。接下来可以把该 PSO-SVR 模型用于发酵过程的先进控制当中,进一步验证该预估模型在实际生产中的价值。(收稿日期:2006 年 12 月)

参考文献:

[1] Cherkassky V,Mulier F.Learning from data:concepts,theory,and methods[M].New York:Wiley,1998.

[2] Cherkassky V,Ma Yun-qian.Practical selection of SVM parameters and noise estimation for SVM regression[J].Neural Networks,2004,17:113-126.

[3] 冯兴杰,魏新.基于支持向量机的旅客吞吐量预测研究[J].计算机工程,2005,31(14):172-174.

[4] Üstün B,Melssen W J.Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization[J].Analytical Chimica Acta,2005,544:292-305.

[5] 俞欢军,张丽平,陈德钊,等.复合粒子群优化算法在模型参数估计中的应用[J].高校化学工程学报,2005(5):284-287.

(上接 201 页)

[3] Shahabi C.Yoda:an accurate and scalable web-based recommendation system [C]//Proceedings of the Sixth International Conference on Cooperative Information Systems,Trento,Italy,September 2001:418-432.

[4] Mobasher B.Discovery and evaluation of aggregate usage profiles for web personalization[J].Data Mining and Knowledge Discovery,2002(6):61-82.

[5] Burke R,Hammond K,Young B C.The FindMe approach to assisted browsing[J].Journal of IEEE Expert,1997,12(4):32-40.

[6] Burke R.Hybrid recommender systems:survey and experiments[J].User Modeling and User-Adapted Interaction,2002,12(4):331-370.

[7] Middleton S E,Shadbolt N R,DE Roure D C.Ontological user profiling in recommender systems[J].ACM Transactions on Information Systems,2004,22(1):54-88.

[8] Hyvönen E,Saarela S,Viljanen K.Application of ontology techniques to view-based semantic search and browsing [C]//Davies J.LNCS 3053:ESWS 2004,2004:92-106.

[9] Hatala M,Wakkary R.Ontology-based user modeling in an augmented audio reality system for museums[J].User Modeling and User-Adapted Interaction,2005(15):339-380.

[10] Schafer J B,Konstan J A,Riedl J.Meta-recommendation systems:user-controlled integration of diverse recommendations[C]//CIKM'02,McLean,Virginia,USA,November 4-9,2002:43-51.

(上接 213 页)

系列二维切片的轮廓点数据进行综合处理,经过一些重采样、插值、平滑处理、滤波等工作,最终得出了一个三维显示矩阵,显示了心室几个方向的内窥曲面图。另外,又完成了心室切片面积及心室容积的测量工作,对心室切片面积的变化趋势做了直观的描述。本文中采用的改进 Snake 模型及三维显示、测量方法,可推广至特点相似的其它器官超声波图像处理中去。(收稿日期:2006 年 11 月)

参考文献:

[1] 罗军辉,冯平.MATLAB7.0 在图像处理中的应用[M].北京:机械工业出版社,2005:136-141.

[2] 李培华,张田文.主动轮廓线模型(蛇模型)综述[J].软件学报,2000,11(6):751-757.

[3] 李天庆,张毅,刘志,等.Snake 模型综述[J].计算机工程,2005,31(9):1.

[4] 杨杨,张田文.一种新的主动轮廓线跟踪算法[J].计算机学报,1998,21(8):297-302.

[5] 李熙莹,倪国强.一种自动提取目标的主动轮廓法[J].光子学报,2002,31(5):606-609.

[6] 赵暖,陈亚青,余建国,等.超声图像处理中 Snake 模型研究[J].上海生物医学工程,2004,25(4):3-10.

[7] 严加勇,庄天戈.医学超声图像分割技术的研究及发展趋势[J].北京生物医学工程,2003,22(1):67-71.

[8] Kass M,Witkin A,Terzopoulos D.Snakes:active contour models[J].International Journal of Computer Vision,1987:321-331.

[9] Williams D J,Shab M.A fast algorithm for active contours and curvature estimation[J].CVGIP:Image Understanding,1992,55(1):14-26.

[10] Zeng Li,Jansen C P,Marsch S,et al.Four-dimensional wavelet compression of arbitrarily sized echocardiographic data[J].IEEE Transactions on Medical Imaging,2002,21(9):1179-1187.