DOI:10.3969/j.issn.1673-4785.201012007
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120425.1647.001.html

基于GEP的最小二乘支持向量机模型参数选择

钱晓山[1,2],阳春华[1]
(1.中南大学 信息科学与工程学院,湖南 长沙 410083; 2.宜春学院 物理科学与工程技术学院,江西 宜春 336000)

摘 要:针对最小二乘支持向量机的多参数寻优问题,提出了一种基于基因表达式编程的最小二乘支持向量机参数优选方法.该算法将最小二乘支持向量机参数(C,σ)样本作为GEP的基因,按其变异算子随着进化代数和染色体所含基因数目动态变化的机制执行,其收敛速度和精确度大大提高.并与基于粒子群算法和遗传算法参数优选方法比较,通过标准测试函数验证了该算法的拟合误差最低.最后用其建立氧化铝生产蒸发过程参数预测模型,应用工业生产数据进行验证,实验结果表明该方法有效且获得了满意的效果.

关键词:基因表达式编程;最小二乘支持向量机;参数选择;粒子群算法;遗传算法
中图分类号:TP181 **文献标志码**:A **文章编号**:1673-4785(2012)03-0225-05

A parameter selection method of a least squares support vector machine based on gene expression programming

QIAN Xiaoshan[1,2], YANG Chunhua[1]
(1. School of Information Science and Engineering, Central South University, Changsha 410083, China; 2. Physical Science and Technology College, Yichun University, Yichun 336000, China)

Abstract: To solve the multi-parameter optimization problem of least squares support vector machines (LSSVM), a parameter optimization method based on gene expression programming (GEP) was proposed. The parameter (C, σ) samples of LSSVM were selected to be genes for GEP according to the mechanism of the dynamic change of the mutation operator with the gene number of the genome and the number of evolutionary generations. As a result, the convergence rate and accuracy were greatly increased. The new method was compared with other parameter optimization methods based on particle swarm optimization (PSO) and a genetic algorithm (GA) by several standard test functions, and the results show that the proposed method obtains the minimum fitting error. Finally, a parameter prediction model of the evaporation process of alumina production was established; the verification results using the industrial production data show that the method is effective and the result is satisfactory.

Keywords: gene expression programming (GEP); least squares support vector machine (LSSVM); parameter selection; particle swarm optimization (PSO); genetic algorithm (GA)

支持向量机(support vector machine, SVM)是Vapnik等[1-2]于1995年首先提出,它在解决小样本、非线性和高维模式识别问题中表现出许多优越特性,已成为智能科学技术研究领域的热点[3-4].最小二乘支持向量机(least squares support vector machine, LSSVM)[5-6]是标准支持向量机的一种扩展,是支持向量机在二次损失下的一种特殊形式,它采用最小二乘线性系统作为损失函数,将求解二次规划问题转化为求解一组线性方程;因而该方法求解速度较快,并广泛应用于非线性函数估计和逼近中,取得了较好的效果.

实践证明,最小二乘支持向量机的精度和泛化性能受核函数的参数以及惩罚系数的影响较大,因此,研究最小二乘支持向量机参数选择的方法对其发展有重要的实际意义.目前已经有一些最小二乘支持向量机参数优选方法,文献[7]针对LSSVM用交叉验证的方法进行核参数选择后应用于PCA的

收稿日期:2010-12-13. 网络出版日期:2012-04-25.
基金项目:国家自然科学基金资助项目(60874069);国家"863"计划资助项目(2009AA04Z124, 2009AA04Z137).
通信作者:钱晓山. E-mail: qianxiaoshan@126.com.

软测量建模;文献[8]将遗传算法用于核参数选择后对直流电机进行建模;文献[9]用粒子群算法进行核参数优选后用于软测量建模. 以上算法进行参数寻优时易陷入局部最优,从而影响了整个模型的精度及泛化性能. 文献[10]利用 GEP 和交叉验证法优选支持向量机的核参数,算法性能得到了大大改善. 基因表达式编程(gene expression programming, GEP)是由葡萄牙科学家 C. Ferreira 提出的一种基于基因型组(genome)和表现型组(phenome)的新型遗传算法,它继承和发展了遗传算法 GA 和遗传编程 GP,集成了它们的优点,因此该方法具有更强的解决问题的能力,在函数参数优化、演化建模、神经网络、分类和 TSP 问题等领域得到了广泛应用[11-12]. 本文提出了基于基因表达式编程的最小二乘支持向量机的参数寻优方法,在执行变异操作时,变异算子按照进化代数和染色体所含基因数目的不同而动态变化,这样优化了算法的收敛速度和精度. 同时通过与粒子群算法和遗传算法参数寻优方法比较,并用标准测试函数和实际工业过程生产数据进行验证,结果表明了该模型的预测精度较高.

1 基于 GEP 的支持向量机参数选择

1.1 基因表达式编程方法

GEP 沿袭了 GA 和 GP 中的复制、变异、交叉等遗传算子以及"物竞天择,适者生存"的自然选择思想,其解决问题的能力更强,比传统的 GA 和 GP 等遗传算法要快 100 ~60 000 倍[12].

在 GEP 中,个体采用固定长度的线性编码来表示. 个体染色体由 1 个或多个基因组成,每个基因由基因头 h 和基因尾 t 构成,h 中可以出现运算符或终结点,而 t 中只能出现终结点,并且 h 和 t 满足 $L(t)=L(h)\times(n-1)+1$,其中,n 为 h 中运算符、函数的最大参数个数. GEP 算法在对个体染色体进行适应度评价时,需要先将染色体按照自顶向下、自左至右的顺序将其编码为表达式树(expression tree, ET),再采用中根遍历 ET 的方法进行解码操作,计算其适应度[12].

基因表达式编程的实现技术主要包括编码方式、遗传算子、插串操作、重组算子、适应度函数选择、数值变量等[13],每个部分的具体实现可参考文献[13],这里不作详细叙述,变异算子动态变化机制可参考文献[14].

1.2 最小二乘支持向量机

在支持向量机回归法[15]中,设样本为 n 维向量,某区域的 m 个样本及其值表示为

$$(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\in\mathbf{R}^n\times\mathbf{R}.$$

首先通过非线性变换 $\boldsymbol{z}=\boldsymbol{\varphi}(\boldsymbol{x}_i)$ 将 m 维向量映射到 $l(l\gg m)$ 维这个高维特征空间中,之后采用线性函数 $f(\boldsymbol{x})=\boldsymbol{w}\boldsymbol{\varphi}(\boldsymbol{x})+b$ 来对其拟合,并容许出现拟合误差,目标是使回归模型在模型推广能力和经验风险之间找到最佳平衡点,即结构风险最小. LSSVM回归算法的优化目标为

$$\begin{cases}\min\limits_{\boldsymbol{w},b,\boldsymbol{\xi}} J(\boldsymbol{w},\boldsymbol{\xi})=\dfrac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\dfrac{C}{2}\sum\limits_{i=1}^{N}\xi_i^2;\\ \text{s.t. } y_i=\boldsymbol{w}\boldsymbol{\varphi}(\boldsymbol{x})+b+\xi_i,i=1,2,\cdots,m.\end{cases}\tag{1}$$

式中:$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ 为控制模型的复杂度,C 为误差惩罚参数,J 为误差控制函数. 利用拉格朗日法求解式(1)的优化问题,定义拉格朗日函数:

$$L(\boldsymbol{w},b,\boldsymbol{\alpha},\boldsymbol{\xi})=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{C}{2}\sum_{i=1}^{N}\xi_i^2-\sum\alpha_i(\boldsymbol{w}\boldsymbol{\varphi}(\boldsymbol{x}_i)+b+\xi_i-y_i).$$

式中:$\alpha_i(i=1,2,\cdots,m)$ 是拉格朗日乘子.

根据 KKT 优化条件:

$$\frac{\partial L}{\partial \boldsymbol{w}}=0,\frac{\partial L}{\partial b}=0,\frac{\partial L}{\partial \boldsymbol{\alpha}}=0,\frac{\partial L}{\partial \boldsymbol{\xi}}=0,$$

则有

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} J(\boldsymbol{w},\boldsymbol{\xi})=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{C}{2}\sum_{i=1}^{N}\xi_i^2,$$

$$\boldsymbol{w}=\sum_{i=1}^{m}\alpha_i\boldsymbol{\varphi}(\boldsymbol{x}_i),\sum_{i=1}^{m}\alpha_i=0,$$

$$\alpha_i=c\xi_i,\boldsymbol{w}\boldsymbol{\varphi}(\boldsymbol{x}_i)+b+\xi_i-y_i=0.\tag{2}$$

定义核函数 $K(\boldsymbol{x}_i,\boldsymbol{y}_i)=\boldsymbol{\varphi}(\boldsymbol{x}_i)\cdot\boldsymbol{\varphi}(\boldsymbol{y}_i)$,根据式(2),将求解优化问题转化为求解线性方程:

$$\begin{bmatrix}0 & 1 & \cdots & 1\\ 1 & K(\boldsymbol{x}_1,\boldsymbol{x}_1)+1/c & \cdots & K(\boldsymbol{x}_1,\boldsymbol{x}_m)\\ \vdots & \vdots & \vdots & \vdots\\ 1 & K(\boldsymbol{x}_m,\boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_m,\boldsymbol{x}_m)+1/c\end{bmatrix}\times\begin{bmatrix}b\\ \alpha_1\\ \vdots\\ \alpha_m\end{bmatrix}=\begin{bmatrix}0\\ y_1\\ \vdots\\ y_m\end{bmatrix}.$$

解上述线性方程组可得到拉格朗日乘子 α_i 和参数 b,由此确定 LSSVM 的输出为

$$y(\boldsymbol{x})=\sum_{i=1}^{m}\alpha_iK(\boldsymbol{x},\boldsymbol{x}_i)+b.$$

1.3 基于 GEP 的最小二乘支持向量机模型参数选择

由于最小二乘支持向量机的参数选择直接影响整个模型的收敛性、稳定性和精度,而 GEP 与 GA 和 GP 相比,具有更强的全局搜索能力[16-17];因此,将 GEP 算法引入到以径向基函数为核函数的 LSS-

VM 模型的参数优化中，形成基于 GEP 的 LSSVM 模型. 与 PSO 和 GA 优化算法比较，该算法可以得到更高的精度，其泛化性能和稳定性也大大提高.

染色体编码和适应度函数选择是进行惩罚系数 C 和核函数宽度 σ 参数优化的 2 个重要方面. 在 GEP 中，多基因结构可以用来进行有效的搜索以解决函数优化的问题，且最佳参数是在不停变化的随机数值常数上的数学运算中发现的. 为此，在染色体编码中采用处理随机数值常数的染色体组织结构. 随机数值常数集的选取十分容易，通常可以选择由 10 个随机常数构成的集合，如 $R=\{0,1,2,3,4,5,6,7,8,9\}$，这对大部分问题就能够达到很好的效果. 适应度函数选择如式(3)：

$$f(C,\sigma)=\sqrt{\frac{1}{n}\sum(y_i-y_i')^2}. \tag{3}$$

式中：y_i 为实际值，y_i' 为支持向量机的输出，n 为样本数. 利用 GEP 算法对惩罚系数 C 和核函数宽度 σ 进行寻优，具体算法步骤如下：

1)针对参数 C、σ 初始化种群，随机产生 60 组初始染色体，每个染色体由 5 个基因构成，每个基因头长度为 15(或更多)，初始化时采用 KARVA 编码；

2)读取样本数据，根据当前参数 C、σ 训练 LSSVM，得到支持向量机的输出 y_i'；

3)按式(3)计算适应值，并将适应值排序，保存适应值最高的个体；

4)执行变异，按照染色体所含基因的多少决定变异的基因位个数，本文选择每个基因变异 1 个基因位的方法；

5)执行 IS 插串、RIS 插串和 Gene 插串；

6)执行单点重组、两点重组和基因重组；

7)若运行到预先设定的最大代数或者适应度函数值收敛到设定精度，则执行 8)，否则执行 2)；

8)选择出最优染色体并保存记录；

9)对染色体解码，构建 LSSVM 模型.

2　算法性能验证

2.1　仿真测试

为了验证上述方法的有效性，选用标准测试函数进行仿真. 实验平台配置为2.8 GHz主频率，1 GB 内存，采用 Matlab 7.0 进行仿真实验.

1)取一维 sinc 函数：

$$f_1(x)=\mathrm{sinc}(x)+\phi.$$

式中：ϕ 是均值为 0、方差为 0.1 的高斯噪声. 输入变量取 150 个[-4, 4]之间的数据构成 LSSVM 的训练样本，以最小均方误差为目标，利用 GEP 算法对惩罚系数 C 和径向基核函数参数 σ 进行优选，其中 LSSVM 采用 $\xi=0.15$ 的一次不敏感损失函数. GEP 算法中选 60 组为初始染色体，最大迭代次数为 500. 为便于比较，采用同样大小的初始群体和最大迭代次数的 PSO 和 GA(交叉概率为 0.5，变异概率为 0.047)进行多次实验. 图 1 显示了 3 种算法的寻优过程对比结果，从中可以看出，GEP 和 PSO 的下降速度较快，而 GA 速度较慢，经过多次实验发现 GA 和 PSO 寻优的成功率低于 GEP，并且有时陷入局部最优，总的看来，GEP 算法的寻优能力和收敛速度都比 PSO 和 GA 算法好. 由图 2 可见，使用 3 种算法各自寻优的参数对 sinc 函数进行测试，发现 GEP 算法的拟合效果最好，且偏离实际值的幅度较小. 测试统计结果如表 1 所示，从中看出新方法的测试误差最小.

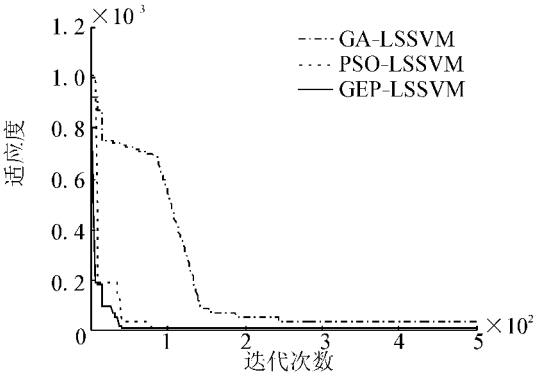


图 1　寻优过程比较

Fig. 1　Comparison of optimization process

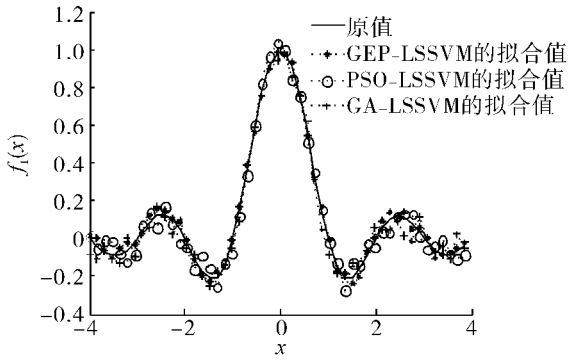


图 2　模型仿真结果比较

Fig. 2　Comparison of model simulation results

表 1　Sinc 函数测试结果比较

Table 1　Comparison of sinc function test results

寻优算法	C	σ	测试误差/10^{-4}
GA	800.16	3.87	6.010 8
PSO	756.36	3.95	3.696 0
GEP	735.29	4.12	0.910 8

2)取二维 Rosenbrock 函数:

$$f_2(x) = \sum_{i=1}^{N-1}[100 \times (x_i^2 - x_{i+1})^2 + (1 - x_i)^2],$$
$$x_i \in [-5.12, 5.12].$$

取1 000组数据样本按4:1的比例随机分组,训练样本数量为800,测试样本为200,其他设置如同 sinc 函数测试实验,得到的测试结果如表2所示.在二维函数的测试中,经过多次实验可以发现,与一维函数相比,二维函数的测试结果更能体现 GEP 的优越性,且相比于其他2个算法,基于 GEP 算法的 LSSVM 模型的拟合误差大大降低,进一步说明了该方法的有效性.

表2 Rosenbrock 函数测试结果比较

Table 2 Comparison of Rosenbrock function test results

寻优算法	C	σ	测试误差/10^{-4}
GA	903.16	4.035	8.621 3
PSO	656.36	2.915	7.026 4
GEP	451.08	0.893	2.518 2

2.2 工业生成过程验证

氧化铝蒸发过程是一类具有非线性、大滞后、多变量等特征的能量交换的复杂工业过程,在蒸发器内加热蒸汽,释放潜热,转移到料液中,使溶剂发生相变,溶液浓度得以提高.出料浓度是衡量产品质量的重要指标,由于技术、成本的限制难以实现在线检测,目前质量检测多以人工现场采集和实验室化验为主,检测结果严重滞后,不利于该过程的稳定控制.出液浓度的影响因素主要包括蒸发器的真空度、进料的流量、温度和浓度、加热蒸汽的流量和压力、蒸发器的料液位、不凝性气体和冷凝水的排除等[18].通过理论分析和生产经验选取影响较大的5个变量:进料温度 T_1、进料流量 F_1、进料浓度 L_{in}、新蒸汽温度 T_2、新蒸汽流量 F_2.以某氧化铝厂带闪蒸和强制循环的七效逆流降膜蒸发的蒸发过程为例,该厂实际生产1个月的数据作为训练数据和测试数据,建立基于 GEP 算法的 LSSVM 的蒸发过程出料浓度预测模型为

$$L_{out} = \text{GEP-LSSVM}(T_1, T_2, F_1, F_2, L_{in}).$$

式中:L_{out} 为预测模型输出,即出口料液浓度;GEP-LSSVM 为模型标示;T_1、F_1、L_{in}、T_2、F_2 为已知样本的输入.选用经过纠错、剔除和归一化处理后的400组工业数据中的300组用于建模,100组用于模型验证,选取 $\xi = 0.02$ 的一次不敏感损失函数和径向基核函数,通过 GEP 算法对最小二乘支持向量机建模参数进行优化,得到最优参数集 $\xi = 0.02$、$C = 428.56$、$\sigma = 0.072$.再用得到的最优参数训练 LSSVM,最终的出料浓度预测结果如图3所示.

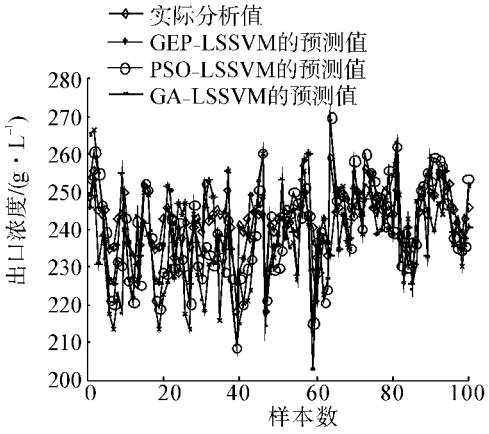


图3 模型泛化比较结果

Fig. 3 Comparison of model generalization results

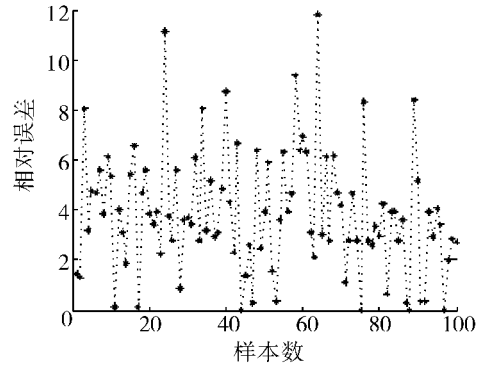


图4 GEP-LSSVM 模型预测相对误差

Fig. 4 Relative error of GEP-LSSVM model predict

图3显示了 GEP-LSSVM、PSO-LSSVM 和 GA-LSSVM 预测模型的泛化能力,从图中可知,GEP-LSSVM 模型的预测效果最好.通过进一步的数据分析,GEP-LSSVM 模型预测结果中相对误差(如图4所示)小于8%的样本达到92%,其最大相对误差小于12%,均方差 MSE(mean square error)为 $6.082\ 7 \times 10^{-5}$,具有较高的精度;另外,该模型相比于 PSO-LSSVM 的预测能力(MSE 为 8.959×10^{-5})和 GA-LSSVM 的预测能力(MSE 为 $1.618\ 5 \times 10^{-4}$),有了较大的提高.

3 结束语

最小二乘支持向量机的参数选择是支持向量机应用推广的一个重要方面,如何将各种算法应用于其中,一直以来是一个既有实际价值又有理论意义的研究课题.本文将 GEP 算法用于最小二乘支持向量机的参数优化,其中变异算子按照进化代数和染色体所含基因数目动态变化的机制进行变异操作,通过多个实验验证了该方法的有效性.然后将其应用于氧化铝蒸发过程出料浓度的预测模型的建立,

仿真结果表明了该预测模型预测精度高,完全满足实际工业生产的需要,同时也对其应用到其他生产过程有着一定的指导意义.另外,GEP算法本身的改进及其对支持向量机核参数的编码和解码方法也有待进一步研究.

参考文献:

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

[2] VAPNIK V,LEVIN E,CUN Y L. Measuring the VC-dimension of a learing machine[J]. Neural Computation, 1994,6(5): 851-876.

[3] SMOLA A J, SCHOLKOPF B. A tutorial on support vector egression[J]. Statistic and Computing,2004,14(3): 199-222.

[4] SANCHEZ A D. Advanced support vector machines and kernel methods[J]. Neurocomputing,2003,55(1): 5-20.

[5] SUYKENS J A K,VANDEWALL J. Least squares support vector machine classifiers [J]. Neural Processing Leters, 1999,9(3): 293-300.

[6] PELCKMANS K,SUYKENS J A K,DE MOOR B. Building sparse representations and structure determination on LS-SVM substrates[J]. Neurocomputing,2005,64: 137-159.

[7] 郑小霞,钱锋.基于PCA 和最小二乘支持向量机的软测量建模[J].系统仿真学报,2006,18(3): 739-741.
ZHENG Xiaoxia,QIAN Feng. Soft sensor modeling based on PCA and support vector machines[J].Journal of System Simulation,2006,18(3): 739-741.

[8] 赵吉文,刘永斌,孔凡让,等.核参数遗传选优的SVM在直线电机建模中的应用[J].系统仿真学报,2006,18(12): 3547-3549.
ZHAO Jiwen,LIU Yongbin,KONG Fanrang,et al. Application of SVM with genetic algorithms optimizing kernel parameters in linear motor model[J]. Journal of System Simulation,2006,18(12): 3547-3549.

[9] 刘瑞兰,牟盛静,苏宏业,等.基于支持向量机和粒子群算法的软测量建模[J].控制理论与应用,2006,23(6): 895-899,906.
LU Ruilan,MOU Shenging,SU Hongye,et al. Modeling soft sensor based on support vector machine and particle swarm optimization algorithms[J]. Control Theory and Applications,2006,23(6): 895-899,906.

[10] THADANI K,JAYARAMANVK,SUNDARARAJAN V. Evolutionary selection of kernels in support vector machines C]//International Conference on Advanced Computing and Communications. Mangalore,India,2006: 19-24.

[11] FERREIRA C. Gene expression programming in problem solving[C/OL].[2010-12-10]. http://w.gene-expression-programming. com/webpapers/GEPtutorial.pdf.

[12] FERREIRA C. Gene expression programming: a new adaptive algorithm for solving problems[J]. Complex Systems,2001,13(2): 87-129.

[13] MITCHEL M. An introduction to genetic algorithms[M]. Cambridge,UK: The MIT Press,1996: 143-164.

[14] 钱晓山,阳春华.改进基因表达式编程在股票中的研究与应用[J].智能系统学报,2010,5(4): 303-307.
QIAN Xiaoshan,YANG Chunhua. Improved gene expression programming algorithm-tested by predicting stock indexes[J]. CAAI Transactions on Intelligent Systems,2010, 5(4): 303-307.

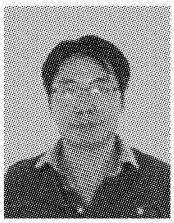
[15] 张春晓,张涛.基于最小二乘支持向量机和粒子群算法的两相流含油率软测量方法[J].中国电机工程学报,2010,30(2): 86-91.
ZHANG Chunxiao,ZHANG Tao. Soft measurement method for oil holdup of two phase flow based on least squares support vector machine and particle swarm optimization [J]. Proceedings of the CSEE,2010,30(2): 86-91.

[16] RIVERO D,DORADO J,RABUNAL J,et al. Using genetic programming for artificial neural network development and simplification[C]//Prceedings of the 5th WSEAS International Conference on Computational Intelligence,Man-Machine Systems and Cybernetics. Venice,Italy,2006: 65-71.

[17] XU Kaikuo,LIU Yintian,RONG Tang,et al. A novel method for real parameter optimization based on gene expression programming[J]. Applied Soft Computing, 2009,9(2): 725-737.

[18] 徐文熙,穆文俊.化工原理(上)[M].北京:中国石化出版社,1992.

作者简介:



钱晓山,男,1980年生,讲师,博士研究生,主要研究方向为复杂工业过程建模、优化控制.

阳春华,女1965年生,教授,博士生导师,博士,中国有色金属学会计算机学术委员会委员兼秘书长,中国自动化学会理事、应用专业委员会委员、技术过程故障诊断与安全性专业委员会委员,中国人工智能学会智能控制与智