浙江理工大学学报(自然科学版),第 39 卷,第 4 期,2018 年 7 月
Journal of Zhejiang Sci-Tech University (Natural Sciences)
Vol. 39, No. 4, Jul. 2018

DOI:10.3969/j.issn.1673-3851(n).2018.04.012

基于移动窗口和粒子群寻优的集成偏最小二乘改进算法

马仕强,任　佳,赵梦恩
(浙江理工大学机械与自动控制学院,杭州 310018)

摘　要:为了克服传统偏最小二乘算法对时序非线性数据拟合效果差的问题,提出了结合移动窗口技术和粒子群算法的集成偏最小二乘算法(EMWPLS_PSO)。该算法通过移动窗口判定时序数据状态突变时刻以对原始数据集进行数据划分,添加冗余检查步骤,简化模型结构,同时引入粒子群算法对关键参数寻优,提高了模型性能。采用工业数据集 Debutanizer_data 验证算法,结果表明:该算法在处理时序、非线性数据时具有更高的精准度和更好的稳定性。这表明基于 EMWPLS_PSO 的软测量建模算法在工业领域的实用性和可操作性。

关键词:软测量;偏最小二乘;局部加权;移动窗口;粒子群算法
中图分类号:TP183　　**文献标志码**:A　　**文章编号**:1673-3851 (2018) 07-0453-08

0　引　言

随着社会经济的快速发展,工业过程更加注重生产的安全、效益和环保,对过程关键参数的监测要求也愈加严格,仅仅依靠传统的检测技术已无法满足工业生产的多样化需求,在此背景下人们提出了软测量技术(软测量)。软测量通过建立过程辅助变量与主导变量之间的数学模型[1]实现对主导变量的预测,其中辅助变量是指那些易测且与主导变量有直接或间接联系的过程变量。软测量模型一般可分为模型驱动型和数据驱动型两大类[2-3],在动态变化的工业生产过程中,数据驱动软测量建模优势明显[4]。数据驱动型模型又可细分为线性模型和非线性模型两类,在软测量中常采用线性模型[5],偏最小二乘法(Partial least square,PLS)[6]是其中一种应用广泛的建模方法。相比于最小二乘法,PLS 在处理多输入多输出数据时更具有优势,原因在于 PLS 中包含主成分提取的思想,可以在减小噪声干扰的同时,使有效信息更加集中,从而避免"维数灾难"的问题。

在实际工业生产中,时序数据会往往会呈现较强的非线性,将 PLS 非线性优化是当今研究的一个热点。根据非线性系统理论,如果系统的输出面光滑,那么任何非线性系统都可以通过多个线性模型进行逼近[7],Kaneko 等[8]由此提出了基于局部加权的偏最小二乘法算法(Locally weighted partial least square,LWPLS)。该算法首先将数据集划分成若干个子数据集并建立对应的局部子模型,然后利用集成学习的思想[9]给每个子模型匹配适当权值从而构成全局模型。与单个学习器相比,集成学习可以显著提高模型的泛化能力[10]。

本文针对工业过程中具有时序性、非线性、易突变等特性的数据,提出了以局部加权为主体框架的 PLS 建模方法。首先通过移动窗口(Moving-windows,MW)[11]对数据集进行划分,建立子模型;然后引入模型剪枝技术减少子模型间的冗余,进一步提高模型的效率;最后根据贝叶斯定理实现模型集成预测。由于模型性能受算法内部自定参数影响,本文引入粒子群算法(Particle swarm optimization,PSO)[12]进行参数自动寻优,确保模型性能最优,由此本文提出基于移动窗口和粒子群寻优的集成偏最小二乘改进算法 EMWPLS_PSO。PSO 是处理非线性连续

收稿日期:2017－11－02　　网络出版日期:2017－12－11
基金项目:国家自然科学基金项目(61203177);浙江省自然科学基金项目(LY17F030024)
作者简介:马仕强(1993－),男,浙江金华人,硕士研究生,主要从事工业软测量建模方面的研究。
通信作者:任　佳,E-mail:jren@zstu.edu.cn

优化问题、组合优化问题和混合整数非线性优化问题的有效优化工具，拥有算法简洁、易于实现、参数少且不需要梯度信息等优势[13]。PSO 的适应度函数选择是参数寻优的关键，本文对此进行了研究。EMWPLS_PSO 算法充分考虑时序数据的自身特性，利用移动窗口对数据集进行划分，辨识出各状态突变时刻。权值计算过程中结合 TIM(Just-in-time)思想，得到的权值更加合理。冗余模型检查和 PSO 参数优化使最后的集成模型结构最优化，降低运算量，提高计算效率。

1 EMWPLS 算法实现

EMWPLS 算法的建模步骤为：首先采用移动窗口法对数据集进行移动分割，根据分割结果对各子数据集单独建模；然后引入模型冗余检测技术对建立的子模型进行剪枝；最后根据贝叶斯理论进行模型的集成。

本文使用以下三个指标进行模型性能评价[14]，以衡量预测值与真实值的契合度：误差平方根(The root mean square error，*RMSE*)、相对误差平方根(the relative RMSE，*RE*)和最大绝对误差(The maximum absolute error，*MAE*)，具体公式可以表示为：

$$RMSE=\sqrt{\sum_{t=1}^{N_t}(y_{\mathrm{pre},t}-y_{\mathrm{true},t})^2/N_t} \tag{1}$$

$$RE/\%=\sqrt{\sum_{t=1}^{N_t}\left(\frac{y_{\mathrm{pre},t}-y_{\mathrm{true},t}}{y_{\mathrm{true},t}}\right)^2/N_t}\times 100 \tag{2}$$

$$MAE=\max\{|y_{\mathrm{pre},t}-y_{\mathrm{true},t}|,t=1,2,\cdots,N_t\} \tag{3}$$

其中：$y_{\mathrm{pre},t}$和 $y_{\mathrm{true},t}$分别代表第 t 组测试集的模型预测值和真实输出值；N_t 表示样本数。

1.1 基于移动窗口法建立局部模型

建立局部模型首先确定宽度为 W 的原始窗口，原始窗口中数据集记为 $W_{\mathrm{ini}}=\{X_{\mathrm{ini}},Y_{\mathrm{ini}}\}$，其中 $X_{\mathrm{ini}}\in\mathbf{R}^{W*m}$，$Y_{\mathrm{ini}}\in\mathbf{R}^{W*1}$分别表示输入变量和输出变量。利用 PLS 对 W_{ini}建模，得到模型 f_{ini}。然后将窗口下移一步，新得到的数据集记为 $W_{\mathrm{sft}}=\{X_{\mathrm{sft}},Y_{\mathrm{sft}}\}$。通过判定条件 1 判断模型 f_{ini}是否适用于 W_{sft}，若适用，则窗口继续下移直至条件不再满足为止。最后得到第一个子数据集$\{X_1,Y_1\}$，利用 PLS 对其建模可得到第一个子模型 f_1。

判决条件 1 可以表述为：

$$E_{\mathrm{ini}}=f_{\mathrm{ini}}(X_{\mathrm{ini}})-Y_{\mathrm{ini}} \tag{4}$$

$$E_{\mathrm{sft}}=f_{\mathrm{ini}}(X_{\mathrm{sft}})-Y_{\mathrm{sft}} \tag{5}$$

$$T=\sqrt{W}(\overline{E_{\mathrm{sft}}}-\overline{E_{\mathrm{ini}}})/\sigma_{\mathrm{sft}} \tag{6}$$

$$\chi^2=(W-1)\sigma_{\mathrm{sft}}^2/\sigma_{\mathrm{sft}}^2 \tag{7}$$

其中：$\overline{E_{\mathrm{ini}}}$，$\sigma_{\mathrm{ini}}^2$分别表示 E_{ini}的均值和方差，$\overline{E_{\mathrm{sft}}}$，$\sigma_{\mathrm{sft}}^2$分别表示 E_{sft}的均值和方差。如果$\overline{E_{\mathrm{ini}}}$与$\overline{E_{\mathrm{sft}}}$近似相等，且 σ_{ini}^2与 σ_{sft}^2近似相等，则认为 E_{ini}和 E_{sft}是否近似相等，此时模型 f_{ini}同样适用于 W_{sft}。为判定 E_{ini}和 E_{sft}是否近似相等，本文利用 t-分布和 χ^2-分布进行判定，其中 $T\sim t(W-1)$，$\chi^2\sim\chi^2(W-1)$。设置一个显著性水平 α，即 $P(|T|<\lambda_t)=1-\alpha$，$P(\chi^2<\lambda_{\chi^2})=1-\alpha$。当满足条件 $|T|<\lambda_t$ 且 $\chi^2<\lambda_{\chi^2}$时，认为模型 f_{ini}适用于 W_{sft}，否则窗口停止下移。

1.2 冗余模型删除

依照 1.1 小节介绍的方法继续建立 f_2，f_3，…，至遍历所有数据。为了解决子模型建立过程中的模型冗余问题，本文引入一个删除冗余模型的步骤：当子模型个数大于 2 个时，通过判决条件 2 判定模型间是否存在冗余，若冗余则用新模型取代旧模型，并将旧模型删除。

判决条件 2 可以表述为：

将当前数据集分别代入新模型和旧模型，并计算其误差、误差的均值与方差，并利用 t-分布和 χ^2-分布判定预测误差是否近似，用公式表示为：

$$E_{\mathrm{new}}=f_{\mathrm{new}}(X_{\mathrm{new}})-Y_{\mathrm{new}} \tag{8}$$

$$E_l=f_l(X_{\mathrm{new}})-Y_{\mathrm{new}} \tag{9}$$

$$T_l=\sqrt{N_{\mathrm{new}}}(\overline{E_l}-\overline{E_{\mathrm{new}}})/\sigma_l \tag{10}$$

$$\chi_l^2=(N_{\mathrm{new}}-1)\sigma_l^2/\sigma_{\mathrm{new}}^2 \tag{11}$$

其中：f_l，f_{new}分别表示第 l 个旧模型和新模型，X_{new}，Y_{new}分别表示用于建立新模型的输入与输出，$\overline{E_l}$，σ_l^2 分别表示 E_l 的均值和方差，N_{new}表示用于建立新模型的样本数。

对 $T_l\sim t(N_{\mathrm{new}}-1)$，$\chi_l^2\sim\chi^2(N_{\mathrm{new}}-1)$设置一个显著性水平 α，得到两个阈值 λ_t^{new}，$\lambda_{\chi^2}^{\mathrm{new}}$。当满足条件 $|T|<\lambda_t^{\mathrm{new}}$ 且 $\chi_l^2\sim\chi_{\chi^2}^{\mathrm{new}}$时，认为 E_{new}、E_l 近似，新模型与第 l 个旧模型之间存在冗余，则用新模型取代第 l 个旧模型。

子模型的建立与冗余检验交替进行，直到遍历全部数据集，最终可得到 L 个子模型。整个算法的流程如图 1 所示。

1.3 集成 PLS 算法实现

得到 L 个子模型后利用集成学习法对新样本进行预测，具体步骤为：a) 计算新样本在每个子模型中的估计值；b) 赋予每个子模型权值；c) 加权得到最终的集成学习预测值。这一部分算法的关键是如何定义步骤 b)中的权值。本文使用 Shao 等[15]提

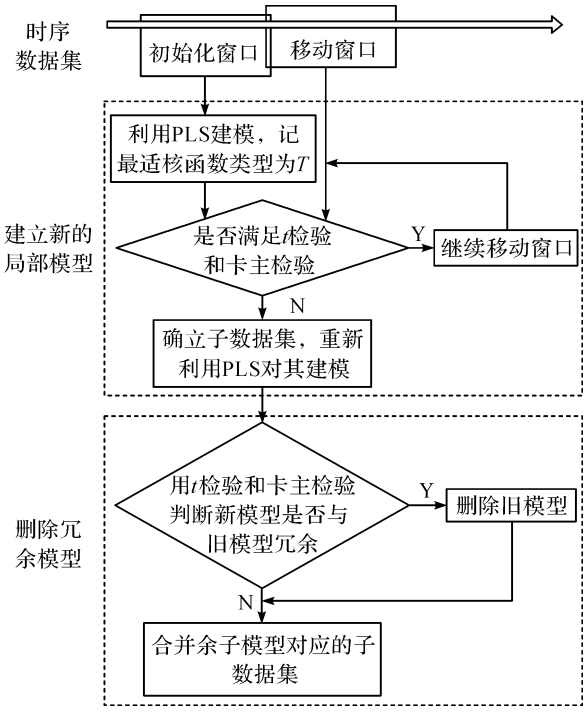


图 1　局部子模型构建算法流程

出的方法定义权值。对输入 x_q,定义指标 $J^{(l)}$ 评价每个子模型的预测能力,$J^{(l)}$ 越大表示第 l 个模型的预测误差越大,可用公式表示为:

$$J^{(l)}=\gamma e_0^{(l)}+\frac{(1-\gamma)\sum_{k=1}^{K}s_k e_k^{(l)}}{\sum_{k=1}^{K}s_k}=\gamma J_1^{(l)}+(1-\gamma)J_2^{(l)} \tag{12}$$

其中:$l=1,2,3,\cdots,L$;K 表示从历史数据中选取与 x_q 最近的点的个数,K 个邻近点可表示为 $\{x_k,y_k\}$,$k=1,2,3,\cdots,K$;$e_0^{(l)}$,$e_k^{(l)}$ 分别表示为点 (x_0,y_0)、(x_k,y_k) 代入第 l 个子模型得到的误差平方,用公式可以表示为:

$$e_0^{(l)}=(f_l(x_0)-y_0)^2 \tag{13}$$

$$e_k^{(l)}=(f_l(x_k)-y_k)^2 \tag{14}$$

其中:(x_0,y_0) 表示历史数据中最新的一组数,γ 是连接系数,且 $0<\gamma<1$,γ 的大小和 x_0 与 x_q 之间的距离有关,可表示为:

$$\gamma=\exp(-\rho d(x_q\cdot x_0)) \tag{15}$$

其中:ρ 为可调参数。

类似地,s_k 和 x_q 与 x_k 之间的距离有关,计算公式为:

$$s_k=\exp(-d(x_q\cdot x_k)/\sigma_d) \tag{16}$$

其中:$d(\cdot)$ 表示两点间的欧氏距离,σ_d 表示 x_q 与历史数据的距离的方差。

由式(12)可以看出,$J^{(l)}$ 的大小与 $J_1^{(l)}$ 和 $J_2^{(l)}$ 有关。$J_1^{(l)}$ 表示 (x_0,y_0) 代入各个子模型后得到的预测误差,可理解为从时域角度的误差分析,$J_2^{(l)}$ 表示 (x_k,y_k) 在各子模型中的预测误差的加权量,作为从空间上的误差分析。$J^{(l)}$ 同时考虑了时域和空间域的误差变化,得到的误差变化将更加全面合理。分析式(15),当 x_0 与 x_q 之间距离较大时(可视为出现数据突变的情况),γ 的值较小,此时 $J_2^{(l)}$ 的比重将增加。反之,当时序数据前后时刻变化较小,则 $J_1^{(l)}$ 的大小对 $J^{(l)}$ 的影响较大。从时序数据自身特征考虑,$J_1^{(l)}$ 应当在 $J^{(l)}$ 中占较大比重,Ni 等[16]对此做过研究,结果表明 γ 的值在[0.875,0.975]区间时预测效果最好。

由于 $J^{(l)}$ 代表模型的预测误差,$J^{(l)}$ 越大则表示分配给第 l 个模型的权值应当越小,本文用 $g^{(l)}$ 表示:

$$g^{(l)}=\exp(-\psi J^{(l)}) \tag{17}$$

其中:ψ 为可调参数。

最后对所有子模型集成:

$$Y_{pre}=\sum_{l=1}^{L}P(f_l\mid x_q)f_l(x_q) \tag{18}$$

其中:$f_l(x_q)$ 是 x_q 在第 l 个子模型中的预测值,$P(f_l|x_q)$ 是由贝叶斯推理得到的后验概率。

$$P(f_l\mid x_q)=\frac{P(f_l)P(x_q\mid f_l)}{\sum_{l=1}^{L}P(f_l)P(x_q\mid f_l)} \tag{19}$$

其中:$P(f_l)$ 和 $P(x_q|f_l)$ 分别代表先验概率和第 l 个模型能准确预测 x_q 的可能性。

$$P(f_l)=N_l/\sum_{l=1}^{L}N_l \tag{20}$$

$$P(f_l\mid x_q)=g^{(l)} \tag{21}$$

其中:N_l 表示建立第 l 个模型所用到的样本数量。

综合式(18)—(21),可以得到

$$Y_{pre}=\frac{\sum_{l=1}^{L}N_l g^{(l)}f(x_q)}{\sum_{l=1}^{L}N_l g^{(l)}} \tag{22}$$

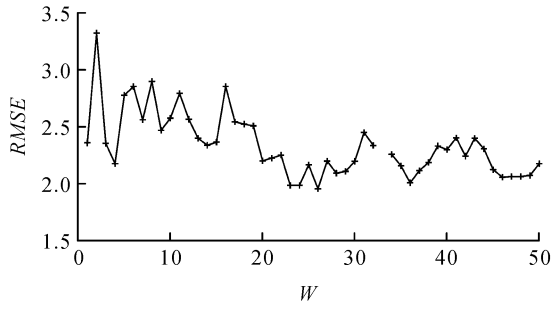
2　基于 PSO 的参数优化

2.1　模型参数对预测效果的影响

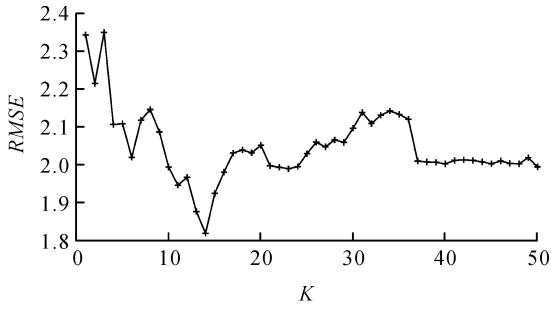
在利用 EMWPLS 算法建模的过程中会涉及 4 个关键的可调参数:移动窗口初始大小 W、邻近点数量 K 以及模型集成时的参数 ρ 和 ψ。移动窗口初始大小 W 与最终构建的子模型数量密切相关,若 W 较大,则可能导致不同状态的数据被归为一类,从而会影响子模型预测效果;反之若 W 太小,子模型数量过多,则会增加模型的复杂度、降低运行效率。类似地,如果 K 太小,可能导致模型过拟合,增大模型的预测误差;反之若 K 太大,则会影响相连时刻的数据对各子模型预测能力的判断。在对子模

型进行集成时，ρ 和 ψ 大小也同样关键。由式(15)可知，如果 ρ 的值较大，则相应的 γ 很小(ρ 无穷大时，γ 趋向于 0)，导致式(12)中 J_1 在 J 中所占的比例较小。同理，式(7)中的 ψ 不宜过大，因为当 ψ 无穷大时，g 趋向于 0，而在式(22)中可以看出 $g(l)$不能都为零(分母不能为零)。

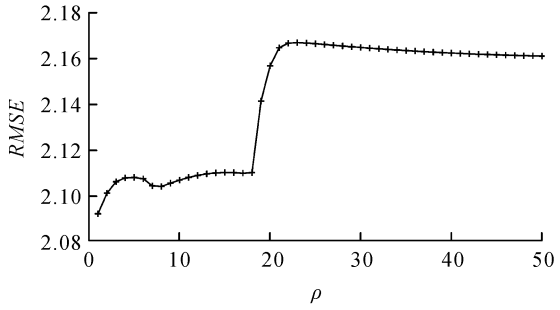
本文以标准数据集中的 abalone 数据集[17]为例，分析四个参数对模型预测精度的影响。图 2—图 4 分别给出了四个参数与模型预测误差指标 *RMSE*、*RE* 以及 *MAE* 的对应关系曲线，从图中可以观察得到：四个参数对预测效果均有较大影响，且它们之间呈现出较复杂的非线性关系。对于不同数据集，其影响关系也不同。因此本文提出一种基于 PSO 进行参数自动寻优的解决思路。



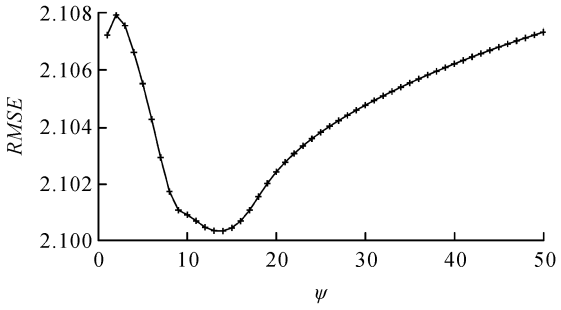
(a) *W*与*RMSE*的关系曲线



(b) *K*与*RMSE*的关系曲线

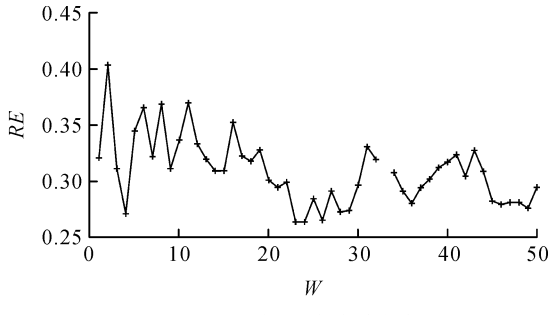


(c) ρ与*RMSE*的关系曲线

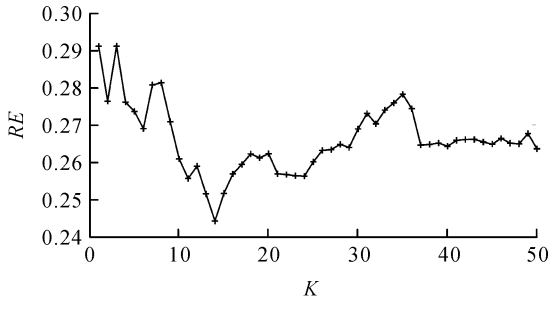


(d) ψ与*RMSE*的关系曲线

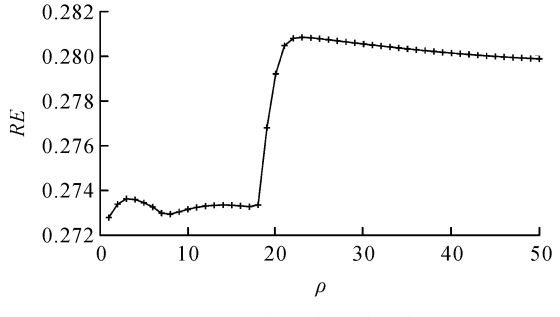
图 2 W，K，ρ，ψ 四个参数与 *RMSE* 的关系曲线



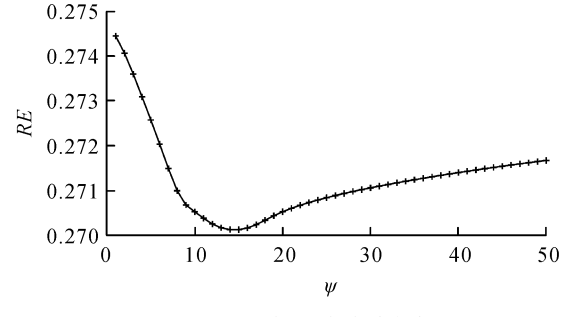
(a) *W*与*RE*的关系曲线



(b) *K*与*RE*的关系曲线

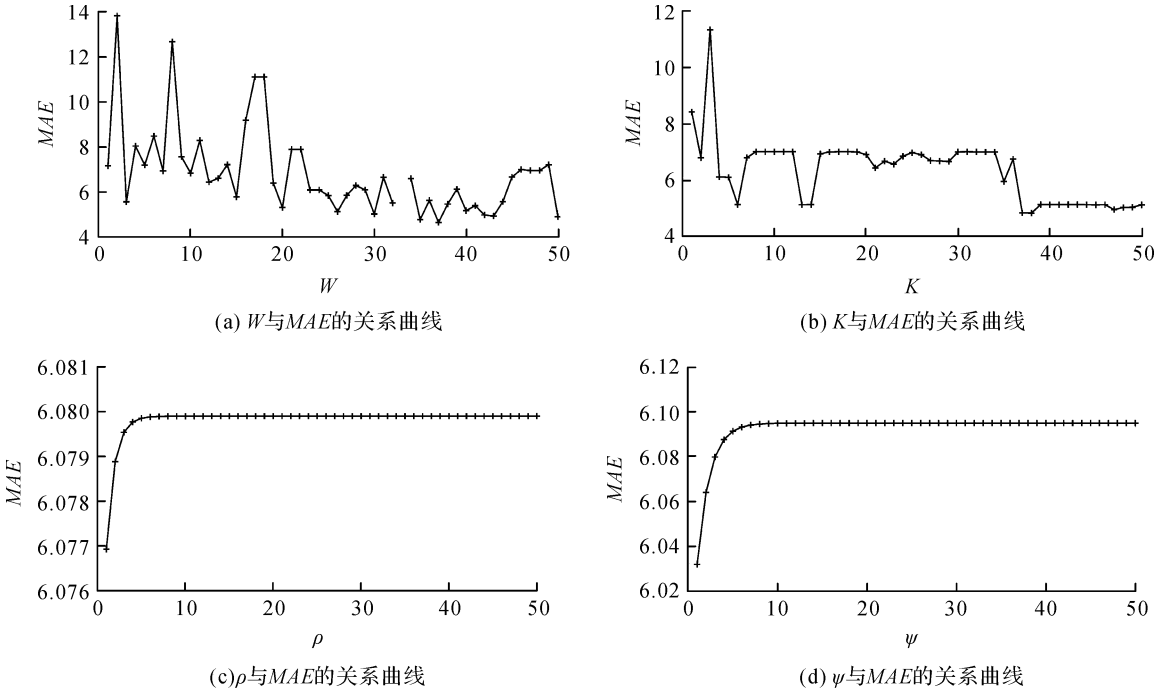


(c) ρ与*RE*的关系曲线



(d) ψ与*RE*的关系曲线

图 3 W，K，ρ，ψ 四个参数与 *RE* 的关系曲线



(a) W与MAE的关系曲线　(b) K与MAE的关系曲线

(c)ρ与MAE的关系曲线　(d) ψ与MAE的关系曲线

图 4　W,K,ρ,ψ四个参数与 MAE 的关系曲线

2.2　粒子群参数寻优

2.2.1　适应度函数选择

在使用粒子群算法时适应度函数的选择是一个关键的步骤,本文选用 $RMSE$、RE、MAE 来评价模型的性能,所以将其作为备选适应度函数,由于 MAE 受单个样本影响较大,故不考虑其作为适应度函数。分析式(1)—(2)可以得到结论:对于同一个模型而言,预测结果的 $RMSE$、RE 的变化趋势并非一直保持一致,它受误差统计分析时$|(Y_{\mathrm{pre},t}-Y_{\mathrm{true},t})/Y_{\mathrm{true},t}|>1$ 的样本数在所有样本中所占比例影响,当其比例高时 RE 的值可能会随着 $RMSE$ 的减小反而增大。图 5 显示了 abalone 数据集在 16 组不同参数下得到的对应 EMWPLS 模型的 $RMSE$、RE、MAE 指标曲线,从中观察发现 $RMSE$ 和 RE 的变化趋势并不一致(为对比二者趋势,图中 $RMSE$ 的数值已放大 10 倍,RE 数值放大 5 倍)。图 5 中曲线对应的具体数值详见表 1。在这 16 组数中第 10 组的 $RMSE$ 最小,第 14 组的 RE 最小。这表明分别用 $RMSE$ 和 RE 作为适应度值进行参数寻优会得到不同的结论。

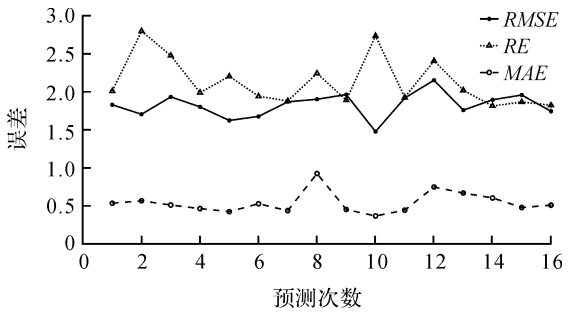


图 5　误差变化与 $RMSE$、RE、MAE 之间的关系图

表 1　不同参数下 abalone 数据集的 $RMSE$、RE、MAE 数值

编号	$RMSE$	RE	MAE
1	0.1828	0.4016	0.5350
2	0.1703	0.5591	0.5667
3	0.1932	0.4948	0.5115
4	0.1801	0.3974	0.4660
5	0.1622	0.4402	0.4256
6	0.1675	0.3881	0.5286
7	0.1867	0.3758	0.4388
8	0.1901	0.4480	0.9253
9	0.1963	0.3777	0.4539
10	0.1476	0.5465	0.3690
11	0.1918	0.3852	0.4454
12	0.2150	0.4807	0.7487
13	0.1758	0.4037	0.6692
14	0.1894	0.3625	0.6047
15	0.1957	0.3728	0.4798
16	0.1744	0.3644	0.5099

可见适应度函数的选择会直接影响参数的寻优结果。选取几组不同参数对 abalone 数据集进行 EMWPLS_PSO 建模,并依照式(1)—(3)计算出各组的 $RMSE$、RE、MAE,选出 $RMSE$ 相同 RE 不同、$RMSE$ 不同 RE 相同、$RMSE$ 和 RE 都不相同的两两对应的几组数据,记录于表 2,并绘制图 6—图 8 的曲线。通过比较可发现若单独选其中一个值作为适应度值 $RMSE$ 比 RE 更合适。当 $RMSE$ 值相同时,RE 值越小则模型效果越好。根据上述结果笔者认为将 $RMSE$ 和 RE 相结合效果更佳。本文提出

$Z=p*RMSE+(1-p)*RE$ 作为适应度函数，其中 p 为连接系数，且在 Z 中 $RMSE$ 应该占更大的比重。

表 2 不同参数下模型的 *RMSE*、*RE*、*MAE* 值对比

编号	*RMSE*	*RE*	*MAE*
1	0.1482	0.5103	0.3476
2	0.1475	0.3033	0.4492
3	0.1631	0.4412	0.4262
4	0.2220	0.4411	0.7592
5	0.1479	0.5101	0.3478
6	0.1944	0.3878	0.5003

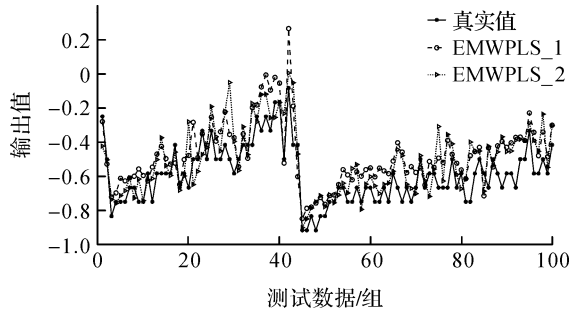


图 6 *RMSE* 相同 *RE* 不同时实际输出与预测输出

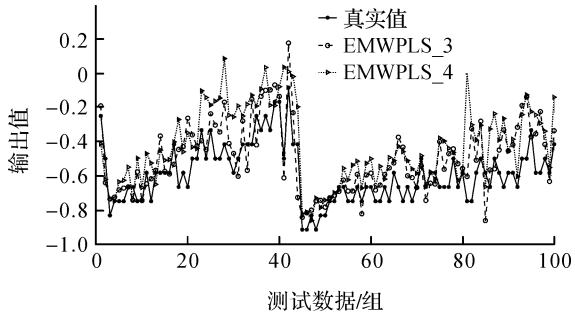


图 7 *RMSE* 不同 *RE* 相同时实际输出与预测输出

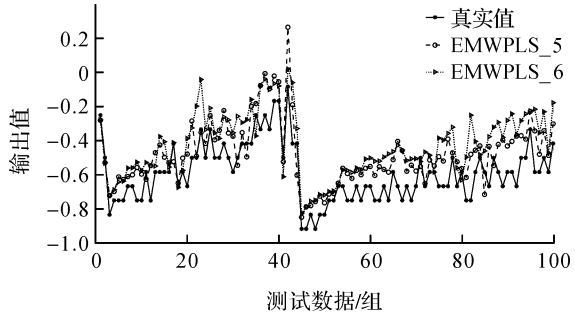


图 8 *RMSE*、*RE* 均不同时实际输出与预测输出

2.2.2 寻优过程

本文选择粒子群算法的适应度函数为：$Z=p*RMSE+(1-p)*RE$，优化目标为 W，K，ρ，ψ 四个关键参数。基于粒子群算法的参数寻优过程如下：

a) 参数随机初始化（种群粒子数设为 20，迭代次数设为 50 次，各参数设置合理上下限，初始位置与速度在参数上下限范围内随机给定）；

b) 计算适应度值，更新粒子的历史最优和全局最优位置；

c) 根据迭代公式更新粒子的位置和速度，如果超出边界值，则赋予其边界值；

d) 判断是否达到最大迭代次数和全局最优位置满足最小界限，若不满足则返回步骤 b)，反之结束寻优。

3 工业数据集测试

本文使用工业数据集 Debutanizer_data[18] 验证模型效果，该数据集共 700 组样本，7 个输入 1 个输出，将前 500 组样本作为训练集，后 100 组样本作为验证集，最后 100 组样本作为测试集。表 3、表 4 记录了其中 6 次粒子群寻优得到的参数及其误差分析，选取结果最好的第一组参数作为 EMWPLS_PSO 的最终参数，并与常规 PLS 以及 BP 算法进行对比。

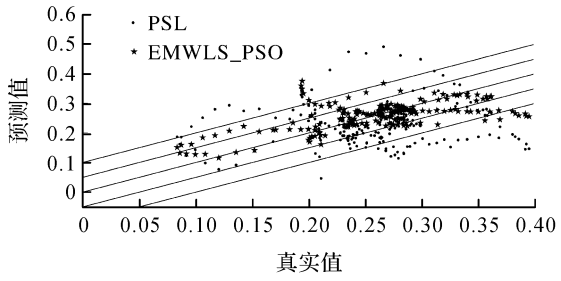
表 3 Debutanizer_data 数据集重复 6 次寻优参数记录表

编号	参数				适应度值
	W	*K*	*P*	*Q*	
1	14	64	0.0221	7049	0.0528
2	35	50	0.0523	6677	0.0839
3	20	29	0.0140	2480	0.0719
4	49	67	0.2174	6409	0.0792
5	42	53	0.2070	548	0.0891
6	13	69	0.4404	3559	0.0695

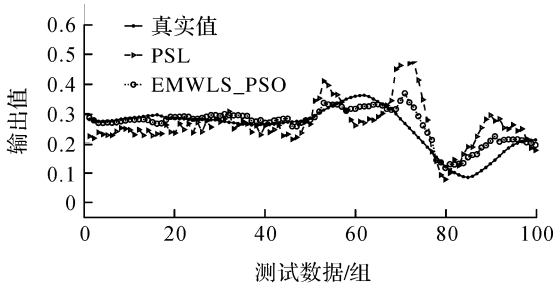
表 4 Debutanizer_data 数据集重复 6 次寻优误差分析表

编号	训练集			验证集			测试集		
	RMSE	*RE*	*MAE*	*RMSE*	*RE*	*MAE*	*RMSE*	*RE*	*MAE*
1	0.11	0.41	0.43	0.03	0.23	0.10	0.05	0.24	0.18
2	0.14	0.58	0.35	0.06	0.29	0.14	0.08	0.34	0.21
3	0.08	0.24	0.38	0.05	0.24	0.10	0.12	0.45	0.29
4	0.09	0.26	0.36	0.06	0.28	0.14	0.06	0.27	0.20
5	0.11	0.53	0.41	0.06	0.33	0.14	0.12	0.45	0.30
6	0.08	0.27	0.39	0.05	0.27	0.12	0.06	0.28	0.15
PLS	0.08	0.32	0.33	0.08	0.44	0.24	0.09	0.41	0.25

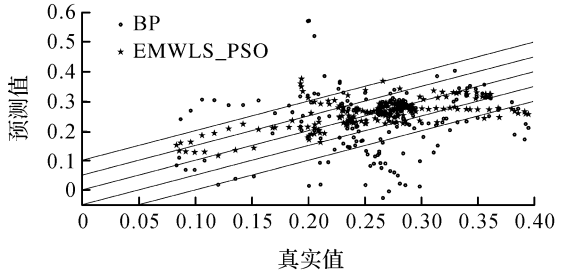
图 9 给出了 EMWPLS_PSO 算法与常规 PLS、BP 算法预测结果的对比曲线。图 9 中点越接近中间的 $Y=X$ 直线说明模型的预测值与真实值越吻合,模型预测效果越好。将真实值记为 Y_{real},预测值记为 Y_{pre}。记 $diff_{0.1}=P(|Y_{real}-Y_{pre}|\leqslant 0.1)$,常规的 PLS 得到的 $diff_{0.1}$ 为 0.75,EMWPLS_PSO 对应的 $diff_{0.1}$ 为 0.93。记 $diff_{0.05}=P(|Y_{real}-Y_{pre}|\leqslant 0.05)$,常规的 PLS 得到的 $diff_{0.05}$ 为 0.45,EMWPLS_PSO 对应的 $diff_{0.05}$ 为 0.785。可以得到结论:集成移动窗口技术对常规 PLS 的预测精度有很大的提高。图 9(b)是 PLS 模型、EMWPLS_PSO 模型的预测值与真实值的对比图,通过对比可发现相对于 PLS,EMWPLS_PSO 的预测趋势更准确,与真实值更吻合。图 9(c)—(d)是同作为非线性算法的 BP 与 EMWPLS_PSO 的预测结果对比图。其中 BP 的 $diff_{0.05}$ 为 0.47,$diff_{0.1}$ 为 0.68。由对比曲线可以看出,EMWPLS_PSO 的预测效果也优于 BP。表 5 记录了 PLS、EMWPLS_PSO 和 BP 算法的 *RMSE*、*RE*、*MAE* 值,从中可以得出结论,较之常规 PLS 和 BP 算法,EMWPLS_PSO 的预测误差最小。



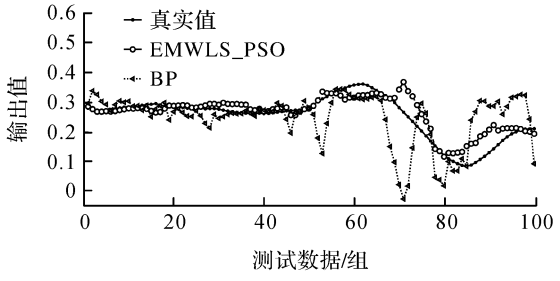
(a) 常规PLS、EMWPLS_PSO预测值-真实值散点图



(b) 常规PLS、EMWPLS_PSO预测对比曲线



(c) BP、EMWPLS_PSO预测值-真实值散点图



(d) BP与EMWPLS_PSO预测对比曲线

图 9　EMWPLS_PSO 与常规 PLS、BP 算法的预测结果对比

表 5　常规 PLS、EMWPLS_PSO、BP 的 *RMSE*、*RE*、*MAE* 对比结果

EMWPLS_PSO			PLS			BP		
RMSE	*RE*	*MAE*	*RMSE*	*RE*	*MAE*	*RMSE*	*RE*	*MAE*
0.05	0.24	0.18	0.09	0.41	0.25	0.08	0.54	0.25

本文将 EMWPLS_PSO 与改进前的 PLS 算法作纵向比较,同时横向比较了非线性的 BP 算法。EMWPLS_PSO 算法在 PLS 基础上性能有极大的改善,较好地克服了 PLS 对非线性数据拟合能力差的问题。同时,相比于纯数据驱动的传统神经网络建模方法,在小样本建模方面 EMWPLS_PSO 拥有更高的预测精度。

4　结　论

本文将移动窗口技术与集成学习的思想相结合,提出一种 EMWPLS_PSO 软测量算法。在利用移动窗口法建立局部模型时,增添了局部模型的冗余检查及删除的步骤,更好地提高了模型的效率和性能。此外,该模型同时应用时域和空间域上的历史数据对动态数据变化趋势进行预测,在避免过度拟合的同时进一步提高了模型的准确性。最后,为了保证模型在处理不同数据时拥有最佳预测精度,本文采用粒子群算法对参数进行自动寻优。通过以上技术的结合,很好地改善了 PLS 对线性相关性较差的时序数据的建模效果。

参考文献:

[1] Kaneko H, Funatsu K. Database monitoring index for adaptive soft sensors and the application to industrial process[J]. AIChE Journal, 2014, 60(1): 160-169.

[2] Souza F A A, Araujo R, Mendes J. Review of soft sensor methods for regression applications[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 152: 69-79.

[3] Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry [J]. Computers & Chemical Engineering, 2008, 33(4): 795-814.

[4] Kadlec P, Grbic R, Gabrys B. Review of adaptation mechanisms for data-driven soft sensors[J]. Computers & Chemical Engineering, 2011, 35(1): 1-24.

[5] 张宏伟,李鹏飞,景军锋,等.基于即时学习的软测量建模实时性改进[J].西安工程大学学报,2014,28(6):750-754.

[6] 石怀涛,刘建昌,张羽,等.基于相对变换PLS的故障检测方法[J].仪器仪表学报,2012,33(4):816-822.

[7] Vijayakumar S, Aaron D S, Schaal S. Incremental online learning in high dimensions [J]. Neural Computation, 2005, 17(12): 2602-2634.

[8] Kaneko H, Funatsu K. Ensemble locally weighted partial least squares as a just-in-time modeling method [J]. AIChE Journal, 2016, 62(3): 717-725.

[9] Kaneko H, Funatsu K. Applicability domain based on ensemble learning in classification and regression analyses [J]. Journal of Chemical Information and Modeling, 2014, 54(9): 2469-2482.

[10] 田慧欣,李坤,孟博.一种用于软测量建模的增量学习集成算法[J].控制与决策,2015,30(8):1523-1526.

[11] Wang X, Kruger U, Irwin G W. Process monitoring approach using fast moving window PCA[J]. Industrial & Engineering Chemistry Research, 2005, 44(15): 5691-5702.

[12] Kennedy J. Encyclopedia of machine learning [M]. Springer, 2010: 760-766.

[13] 蒋晓岫,任佳,顾敏明.多维度惯性权重衰减混沌化粒子群算法及应用[J].仪器仪表学报,2015,36(6):1333-1341.

[14] Willmott C J, Matsuura K. Advantages of the mean absolute error(MAE) over the root mean square error (RMSE) in assessing average model performance[J]. Climate Research, 2005, 30(1): 79-82.

[15] Shao W M, Tian X M, Wang P. Local partial least squares based online soft sensing method for multi-output processes with adaptive process states division [J]. Chinese Journal of Chemical Engineering, 2014, 22(7): 828-836.

[16] Ni W D, Tan S K, Ng W J, et al. Localized, adaptive recursive partial least squares regression for dynamic system modeling[J]. Industrial & Engineering Chemistry Research, 2012, 55(23): 8025-8039.

[17] Tan P J, Dowe D L. Mml inference of decision graphs with multi-way joins and dynamic attributes [J]. Australasian Joint Conference on Artificial Intelligence, 2003: 269-281.

[18] Fortuna L, Graziani S, Rizzo A, et al. Soft sensors for monitoring and control of industrial processes [M]. London: Springer-Verlag, 2007: 229-231.

Improved ensemble partial least-squares algorithm based on moving-windows and particle swarm optimization

MA Shiqiang, REN Jia, ZHAO Mengen

(Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In order to overcome the poor fitting effect of traditional partial least squares(PLS) algorithm pn nonlinear data, an improved algorithm based on moving window technique and particle swarm optimization (PSO) EMWPLS_PSO was proposed. The moving window was used to determine the mutation time of the time series data so as to divide the original data set. Besides, the redundant checking steps were added to simplify the model structure, and the PSO was introduced to optimize the key parameters and improve the model performance. The algorithm in the paper was proven by testing an industrial data set, Debutanizer_data. The testing result shows that the algorithm is more accurate and stable in processing time series and nonlinear data. It also proves the soft measurement modeling algorithm based on EMWPLS_PSO has practicability and operability in the industry field.

Key words: soft measurement; partial least squares; locally weighting; moving-window; particle swarm optimization

(责任编辑:康 锋)