文章编号:1671-4598(2019)08-0049-05 DOI:10.16526/j.cnki.11—4762/tp.2019.08.011 中图分类号:TP274 文献标识码:A

基于混合准则的软测量建模辅助变量选择方法

郭 明,陈伟锋
(浙江工业大学 信息工程学院,杭州 310013)

摘要:软传感器在工业中被广泛应用于预测与产品质量密切相关的关键过程变量,这些变量很难在线测量;要建立一个高精度的软传感器,选择合适的辅助变量是至关重要的;针对这个问题,通过耦合训练集的BIC准则以及验证集的MSE准则得到一个混合整数非线性规划问题,并将该MINLP问题分成内外两层结构,外层采用遗传算法对二元整数变量进行寻优,内层在整数变量固定之后退化成了较易于求解的非线性规划问题;在此基础上经过进一步分析提出了基于混合准则的变量选择方法,然后将所得辅助变量子集代入BP神经网络进行软测量建模;最后,通过4组案例对所提出方法进行验证;结果表明,所提出方法建立的软测量模型具有较好的预测性能。

关键词:关键过程变量,辅助变量选择,混合整数非线性规划,BIC准则

Mixed Criterion Based Secondary Variables Selection for Soft Sensor

Guo Ming, Chen Weifeng
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310013, China)

Abstract: Soft sensors are widely used in industry to predict key process variables that are closely related to product quality, and these variables are difficult to measure online. To build a high-precision soft sensor, it is important to choose the appropriate auxiliary variables. Aiming at this problem, this paper obtains a mixed integer nonlinear programming problem by coupling the BIC criterion of the training set and the MSE criterion of the verification set, and divides the mixed integer nonlinear programming problem into two layers, the inner and outer layers, and the outer layer uses the Genetic Algorithm (GA). The integer variable is optimized, and the inner layer degenerates into an easier to solve nonlinear programming problem (NLP) after the integer variable is fixed. Based on this analysis, a variable selection method based on hybrid criteria is proposed. Then the subset of secondary variables obtained is substituted into BP neural network for soft sensor modeling. Finally, the proposed method is validated by four actual cases. The results show that the soft-measurement model established by the proposed method has better prediction performance.

Keywords: key process variables; secondary variable selection; mixed-integer nonlinear programming (MINLP); Bayesian information criterion

0 引言

近年来,在现代生产过程中,对产品质量的要求越来越高,必须对与产品质量密切相关的关键变量进行实时检测。但是,在线分析仪表价格昂贵、维护保养复杂;而通过离线实验室分析结果存在滞后大等原因,将导致控制质量的性能下降,难以满足生产要求。为了解决这个问题,以推断控制为基础的软测量建模方法及其应用技术取得了广泛的关注[1-3]。

软测量建模的基本思想就是选择一组与主导变量相关的且易测量的辅助变量,并构造关于辅助变量和主导变量的数学模型,实现对主导变量的在线估计[4-5],其中最为关键的问题之一就是如何选取合适的辅助变量。目前,国内外对辅助变量选择进行了大量的研究。其中,基于统计技术的变量选择方法被较多的采用。2006年,Emet等人[6]提出了一种直接优化AIC准则,将变量选择描述成一个混合整数非线性(MINLP,Mixed Integer Nonlinear Programming)优化问题,该方法可以找到具有较优建模效果的辅助变量子集,但是由于目标函数为非线性且非凸,当候选辅助变量过多时,会导致求解时间过长,甚至难以找到最优解;2017年,Jian等人[7]在MINLP优化问题的基础上,提出了一种基于BIC准则的嵌套式MIQP的变量选择方法,该方法大大缩短了求解时间,但是该方法的求解结果容易陷入局部最优。

除此之外,建立具有出众预测性能的软测量模型仍然是一件困难的工作。一方面,现代工业通常存在很强的非线性,导致主成分回归[8-9],偏最小二乘[10-11]等线性软测量模型的预测精度下降[12];另一方面,现代生产过程中,通常存在多个重要且难以测得的主导变量。因此建立有非线性解释能力的多输出软测量模型极为重要,而神经网络凭借网络拓扑结构和非线性计算能力,广泛应用于软测量建模、模式识别、预测等领域[13-14],2018年,Qiu等人[15]提出了一种基于深层神经网络的多输出软测量建模方法,其

收稿日期:2019-02-19; **修回日期**:2019-02-29。
基金项目:国家自然科学基金项目(51404211)。
作者简介:郭 明(1994-),男,浙江金华人,硕士研究生,主要从事工业数据分析方向的研究。
陈伟锋(1984-),男,浙江宁波人,博士,副教授,主要从事复杂系统建模与优化方向的研究。

核心在于通过 VIP 方法进行辅助变量选择，然后将所获得辅助变量子集代入深度神经网络进行多输出软测量模型建立，该方法所建立污水处理模型具有较优的预测性能，但是通过 VIP 方法选择辅助变量需要选取一个合适的 VIP 阈值，阈值过小，使得选取辅助变量过多，会导致模型过拟合；而阈值过大，使得选取辅助变量太少，从而导致模型欠拟合。

本文在嵌套式 MIQP 的基础上进一步简化，将 MINLP 问题分成内外两层结构，外层采用启发式算法（本文采用遗传算法（GA，Genetic Algorithm）对二元整数变量进行寻优，内层在整数变量固定之后退化成了最小二乘求解（LS，Least Square），进一步分析提出了基于 GA 和 LS 的变量选择方法（GA－LS），实验结果表明，该方法能够较好地避免局部最优的情况方法，而且当候选辅助变量过多时，该方法能够以更快的速度获得更优的辅助变量子集。但是，实验结果表明该方法存在精度不够的问题，即使用 BIC 准则虽然能够较好的估计预测误差，但是在某些数据集中与真实预测误差仍存在较大差距。在后续研究中，为了更好的估计预测误差，本文通过耦合训练集的 BIC 准则以及验证集的 MSE 准则用于更精确的估计预测误差，并且仍将其描述为 MINLP 优化问题，并进一步分析提出了基于混合准则的变量选择方法（GA－NLP），该方法能够获得更优的辅助变量子集。从而建立预测性能更好的模型。

综上所述，本文在基于 BIC 准则的 MINLP 优化问题的基础上，提出了 GA－LS 和 GA－NLP 两种辅助变量选择方法。并且将所得到的辅助变量子集通过 BP 神经网络建立软测量模型，实验结果表明：通过 GA－LS 方法能够以较快的速度获得能够具有较优预测性能模型的辅助变量子集；而通过 GA－NLP 虽然求解时间较长，但是所获得的辅助变量子集能够建立预测性能更优的模型。

1 MINLP 以及 MIQP 原理

1.1 MLR 模型及评价准则

多变量统计分析方法，如主成分回归[12-13]，多元线性回归，偏最小二乘[14-15]等，是最常用的软测量模型。其中，MLR 模型基于其简便的分析表达式的特点[14]，被广泛用于辅助变量选择。MLR 模型表示如下：

$$\begin{aligned} &\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}\\ &E(\boldsymbol{\varepsilon})=0\\ &Cov(\varepsilon)=\sigma^2 I \end{aligned} \tag{1}$$

其中：

$$\boldsymbol{Y}=\begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_n \end{pmatrix},\boldsymbol{\beta}=\begin{pmatrix} b_0\\ b_1\\ \vdots\\ b_n \end{pmatrix},\boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n \end{pmatrix},\boldsymbol{X}=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1n}\\ 1 & x_{21} & \cdots & x_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{m} \end{pmatrix}$$

对于软测量模型，其主要任务是预测未知数据。建立模型的质量应根据其泛化性能进行评估。因此，在评估预测模型时，需要关注的应该是测试数据的预测误差，而不是训练数据的误差[7]。但是精确地测量测试数据的预测误差是不可能的，只能通过其它方法对测试数据的预测误差进行估计。其中一种方法就是计算模型的复杂性，然后将其添加到模型训练误差中。而对于线性模型，模型中变量的数量可以表征模型的复杂性。因此，本文选取上述 BIC 准则作为软测量评价准则，其定义如下：

$$BIC=-2\ln L+p\ln n \tag{2}$$

其中：L 为似然函数，由于本文使用 MLR 模型用于变量选择，似然函数 L 定义如下[7]：

$$\ln L=-\frac{n}{2}\ln\left(\frac{1}{n}\left\|y-X\beta\right\|^2\right)+C \tag{3}$$

1.2 MINLP 及 MIQP 方法

辅助变量选择旨在选择出主导变量密切相关的辅助变量子集。Emet 等人[6]为了实现这个目的，引入一组 0－1 决策变量 z_j，$j=1,2,\cdots,m$ 用于选择辅助变量，若第 j 个变量被选中，则 $z_j=1$，否则 $z_j=0$。然后，通过引入大 M 约束可以实现变量选择的目的：

$$-Mz_j\leqslant b_j\leqslant Mz_j\quad(j=1,2,\cdots,m) \tag{4}$$

其中：M 为一个足够大的正数，$-M$ 和 M 分别为回归系数向量 b_j 的上下界。

由于 BIC 是一个估计真实预测误差的有效指标，故将 BIC 准则作为模型的目标函数，最小化 BIC/AIC 准则，可以将变量选择问题表示为如下 MINLP 问题：

$$\begin{aligned} &\min J=GIC\\ &s.t.\quad \varepsilon_i=y_i-\left(b_0+\sum_{j=1}^{m}b_jx_{ij}\right)\quad(i=1,2,\cdots,n)\\ &-Mz_j\leqslant b_j\leqslant Mz_j\quad(j=1,2,\cdots,m)\\ &z_j\in\{0,1\}\quad(j=1,2,\cdots,m) \end{aligned} \tag{5}$$

值得注意的是，由于 MINLP 优化问题中的目标函数是一个非线性且非凸的函数，当候选变量数量过大时（$m>40$），将难以找到最优解。2009 年，Hastie 等人的研究[1]表明随着模型复杂度的增加，测试误差会先降低；但当复杂度高于某一临界值时，测试数据的预测效果却越来越差。Jian 等人基于这个原理在 MINLP 优化问题基础上，进一步简化，提出了一种嵌套式 MIQP 的变量选择方法，表示如下：

$$\begin{aligned} &\min J=GIC\\ &s.t.\quad \min J=\sum_{i=1}^{n}\varepsilon(k)_i^2\\ &s.t.\quad \varepsilon_i=y_i-\left(b_0+\sum_{j=1}^{m}b_jx_{ij}\right)\quad(i=1,2,\cdots,n)\\ &\sum_{j=1}^{m}z_j=k\quad z_j\in\{0,1\}\quad(j=1,2,\cdots,m)\\ &-Mz_j\leqslant b_j\leqslant Mz_j\quad(j=1,2,\cdots,m) \end{aligned} \tag{6}$$

该优化问题通过外层目标函数，参数化所选变量个数 k，并在内层中，持续求解一个 MIQP 问题，直至外层目标函数结果变差为止。

2　GA－LS 及 GA－NLP 方法

2.1　GA－LS

本文将 MINLP 问题分成内外两层结构，外层采用启发式算法（本文采用遗传算法（Genetic Algorithm，GA））对二元整数变量进行寻优，内层在整数变量固定之后退化成了较易于求解的非线性规划问题（Nonlinear Programming，NLP）。在此基础上经过进一步分析提出了基于 GA 和最小二乘（Least Squares，LS）的变量选择方法（GA－LS）。

首先，通过固定每一次进行建模的辅助变量子集时，原 MINLP 优化问题进一步简化为 NLP 问题，而该 NLP 问题的本质就是最小二乘求解；然后，通过搜索算法找到具有最优预测性能（GIC）的辅助变量子集，而 GA[17] 具有直接对结构对象进行操作的特点，正适合用来搜索最优辅助变量子集。GA－LS 的计算步骤总结如下：

1）数据预处理，对数据集进行归一化处理，并将数据集按照 7：3 的比例分为训练集和测试集，训练集用于辅助变量选择，测试集用于验证所选子集效果；

2）随机生成种群，即等概率 0、1 编码的标准化矩阵，矩阵中行向量代表候选变量个数 m，列向量代表遗传算法种群大小 N。并指定遗传算法最大迭代次数 500。

3）对于一组给定的有 m 个候选辅助变量的数据集，通过遗传算法种群个体固定了一个有 p 个辅助变量的子集时，原 MINLP 优化问题进一步简化为一个 NLP 问题：

$$\min n\log\left(\frac{\sum_{i=1}^{n}\varepsilon^2}{n}\right)+p*\ln n$$

$$s.t.\quad \varepsilon=y_i-(b_0+\sum_{j=1}^{p}b_jx_{ij})\quad (i=1,2,\cdots,n)\qquad (7)$$

4）其中 p 已知，故式（7）中的 $p*\ln n$ 是一个常数。故该 NLP 问题实质为均方误差最小化问题：

$$\min\sum_{i=1}^{n}\varepsilon^2$$

$$s.t.\quad \varepsilon=y_i-(b_0+\sum_{j=1}^{p}b_jx_{ij})\quad (i=1,2,\cdots,n)\qquad (8)$$

即简化为最小二乘法求解，其求解结果如下：

$$\beta=(X^TX)^{-1}X^TY\qquad (9)$$

当目标数据集为多输出数据集时，即主导变量为 $H=(Y_1,Y_2,\cdots,Y_h)$，则此时的求解结果为：

$$\beta=(X^TX)^{-1}X^TY\qquad (10)$$

5）建立子集模型后，通过式（4）计算个体的适应度值，表达如下：

$$fval=\sum_{j=1}^{h}BIC(j)\qquad (11)$$

用于评价该子集模型的预测性能。

6）计算出种群中各个个体的适应度后，保留适应度最优个体，共 R 个。

7）对其余个体进行交叉和变异操作，其中选交叉算子为 0.85，变异算子为 0.02。

8）一轮遗传迭代结束后，求出最佳个体，并与上一轮求得的最佳个体比较，较优个体留下。转到第 1）步，开始新一轮的迭代。

9）达到 GA 设定迭代次数，则迭代结束。

2.2　GA－NLP

上述 GA－LS 方法中的广义信息标准（GIC）虽然能够较好的估计预测误差，但是不够精确。于是本文通过耦合训练集的 BIC 准则和验证集的 MSE 准则用于更精确的估计预测误差。进一步提出了 GA－NLP 方法，该方法在 GA－LS 方法基础上对步骤 1、2、4、5 进行改进，改进如下：

1）数据预处理，对数据集进行归一化处理，并将数据集按照 5：2：3 的比例分为训练集、验证集和测试集，训练集、验证集用于辅助变量选择，测试集用于验证所选子集效果；

2）通过耦合训练集的 BIC 准则和验证集的 MSE 准则用于更精确的估计预测误差，仍表达为 MINLP 优化问题，其表达如下：

$$\min J=n_1\ln\left(\frac{1}{n_1}\sum_{i=1}^{n_1}{\varepsilon_1}^2\right)+p\ln n_1+n_2\ln\left(\frac{1}{n_2}\sum_{i=1}^{n_2}{\varepsilon_2}^2\right)$$

$$s.t.\quad {\varepsilon_1}^i=y_i-\left(b_0+\sum_{j=1}^{m}b_jx_{ij}\right)\ (i=1,2,\cdots,n_1)$$

$${\varepsilon_2}^i=y_i-\left(b_0+\sum_{j=1}^{m}b_jx_{ij}\right)\ (i=1,2,\cdots,n_2)$$

$$-Mz_j\leqslant b_j\leqslant Mz_j\quad (j=1,2,\cdots,m)$$

$$z_j\in\{0,1\}\quad (j=1,2,\cdots,m)\qquad (12)$$

式中，n_1，n_2 分别为训练集和验证集的过程数据长度，ε_1，ε_2 分别为训练集和验证集的模型预测误差。

4）对于一组给定的有 m 个候选辅助变量的数据集，通过遗传算法种群个体固定了一个有 p 个辅助变量的子集时，原 MINLP 优化问题进一步简化为一个 NLP 问题：

$$\min J=n_1\ln\left(\frac{1}{n_1}\sum_{i=1}^{n_1}{\varepsilon_1}^2\right)+p\ln n_1+n_2\ln\left(\frac{1}{n_2}\sum_{i=1}^{n_2}{\varepsilon_2}^2\right)$$

$$s.t.\quad {\varepsilon_1}^i(j)=y_i-\left(b_0+\sum_{j=1}^{m}b_jx_{ij}\right)\ (i=1,2,\cdots,n_1)\qquad (13)$$

当目标数据集为多输出数据集时，即主导变量为 $H=(Y_1,Y_2,\cdots,Y_h)$，则需要多次求解 NLP 问题。

5）通过求解 NLP 问题建立子集模型，通过式（14）计算个体适应度值；

$$fval=\sum_{i=1}^{h}J(i)\qquad (14)$$

式中，$J(i)$，$i=1,\cdots,h$ 是 H 中每个主导变量 Y，$i=1,\cdots,h$ 对应的 NLP 求解结果。

3　结果与讨论

3.1　实验数据

本文从 UCI 数据库中选取了 3 组数据集以及 1 组废水

处理数据集[18]（WWTP）进行了仿真实验。其中，数据集 WWTP 有四个输出变量可以被预测（生物需氧量、化学需氧量、悬浮固体和沉积物）。

对于 CCPP[19]数据集，本文在原始数据集的基础上生成了二阶多项式特征，对于数据集 Crime[20]和 WWTP 数据集，原始数据集中包含缺失值的变量被剔除。在辅助变量选择前，对所有实验数据进行标准化处理，即它们的列均值（每一个过程变量的均值）都为 0，方差都为 1。

3.2 单输出测试

本文使用 CCPP 及 Crime 两个数据集作为单输出测试用例。为了评估 GA－LS 和 GA－NLP 的性能，本文对该算法进行了实验仿真及分析，并与 MINLP－MLR、MIQP－MLR 两种方法进行比较，其中 MINLP－MLR 使用 BARON 求解器进行求解；MIQP－MLR 使用 CPLEX 求解器进行求解。求解的结果通过 BP 模型进行建模，并且采用测试集的均方根误差 $RMSEP$ 和测试集的模型决定系数 R^2P 两个指标对模型的性能进行评价。两个指标定义如下：

$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15}$$

$$R^2P = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{16}$$

式中，y_i 为训练集中第 i 个样本的实际值，$\hat{y}_i$ 为训练集的模型预测值，$\bar{y}$ 为测试集中实际值的均值。三种方法比较结果如表 2 所示。

在表 2 中，显示了 4 种方法的预测效果。其中，p 表示最终辅助变量子集的变量个数；$RMSEP$ 和 R^2P 为预测模型评价指标，其中，$RMSEP$ 的值越小越好，R^2P 的值越接近 1 越好。CPU（s）代表该方法进行变量选择所使用的时间。所有比较方法的最佳 $fval$/ R^2P/$RMSEP$ 值和最小时间成本用粗体字表示。

由表 2 可得，本文所提出两种方法所得预测精度优于 MINLP 以及 MIQP 两种方法。其中又以 GA－NLP 方法所得预测精度最高。

综上所述，通过 GA－LS 方法能够以较快的速度获得能够具有较优预测性能模型的辅助变量子集；而通过 GA－NLP 虽然求解时间较长，但是所获得的辅助变量子集能够建立预测性能更优的模型。

3.3 多输出测试

本文使用 WWTP 数据集作为多输出测试用例。为了评估 GA－LS 以及 GA－NLP 方法的性能，本文对该算法进行了实验仿真及分析，并与 VIP 方法进行比较。三种方法比较结果如表 3 所示。

在表 3 中，显示了 3 种方法的预测效果。其中，p 表示最终辅助变量子集的变量个数；$RMSEP$ 和 R^2P 为预测模型评价指标，其中，$RMSEP$ 的值越小越好，R^2P 的值越接近 1 越好。所有比较方法的最佳 R^2P/$RMSEP$ 值用粗体字表示。

由表 3 可得，本文所提出两种方法所得预测精度优于 VIP 方法。其中又以 GA－NLP 方法所得预测精度最高。

预测输出曲线如图 1～4 所示。

表 1 UCI 数据集

数据集	数据集简写	n	m	原始数据集
Combined Cycle Power Plant	CCPP	9568	14	Combined Cycle Power Plant
Crime	Crime	1994	99	Communities and Crime
Biological demand of oxygen	RD－DBO－G	379	34	Water Treatment Plant
Cemical demand of oxygen	RD－DQO－G	379	34	Water Treatment Plant
Supended solids	RD－SS－G	379	34	Water Treatment Plant
Sdiments	RD－SED－G	379	34	Water Treatment Plant

表 2 单输出数据集预测结果

Case	m	n	Model	k	f	$RMSEP$	$R2P$	CPU(s)
CCPP	9568	14	MINLP(BIC)	13	－18395.2	0.204	0.945	469.4
			MIQP(BIC)	4	－18073.1	0.208	0.943	**86.8**
			GA－LS(BIC)	8	**－18408.5**	0.251	0.947	350.7
			GA－NLP	8	/	**0.199**	**0.948**	1457.7
Crime	1994	99	MINLP(BIC)	14	－1047.5	0.549	0.662	18000.8
			MIQP(BIC)	9	－1148.2	0.575	0.630	3508.5
			GA－LS(BIC)	14	－1161.9	0.554	0.656	499.2
			⟧GA－NLP	14	/	0.537	0.678	2336.8

表 3　多输出数据集预测结果

Case	m	n	Model	k	RMSEP	R2P
RD－DBO－G	379	34	VIP	8	0.432	0.565
			GA－LS(BIC)	15	0.039	0.996
			GA－NLP(BIC)	16	**0.034**	**0.998**
RD－DQO－G	379	34	VIP	8	0.097	0.984
			GA－LS(BIC)	15	0.113	0.976
			GA－NLP(BIC)	16	**0.041**	**0.999**
RD－SS－G	379	34	VIP	8	0.323	0.804
			GA－LS(BIC)	15	0.064	0.992
			GA－NLP(BIC)	16	**0.031**	**0.998**
RD－SED－G	379	34	VIP	8	0.150	0.837
			GA－LS(BIC)	15	0.072	0.963
			GA－NLP(BIC)	16	**0.062**	**0.970**

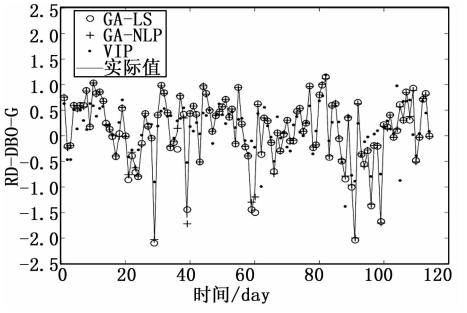

图 1　RD－DBO－G 的预测输出与实际输出

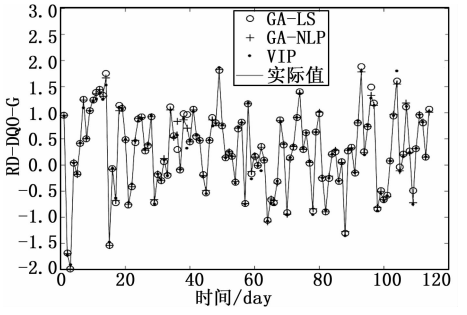

图 2　RD－DQO－G 的预测输出与实际输出

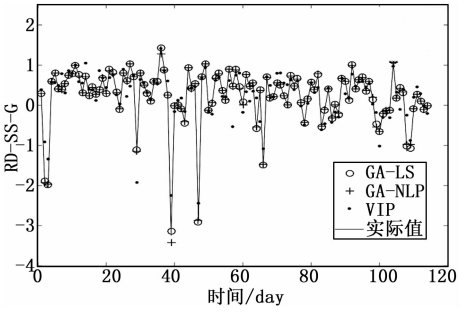

图 3　RD－SS－G 的预测输出与实际输出

用过上面 4 幅预测输出与实际输出对比图可以发现，本文提出的两种方法所得预测输出明显优于 VIP 方法所得预测输出；而所提出的耦合准则（BIC＋MSE）方法所得结果也优于单一准则（BIC）方法所得结果。

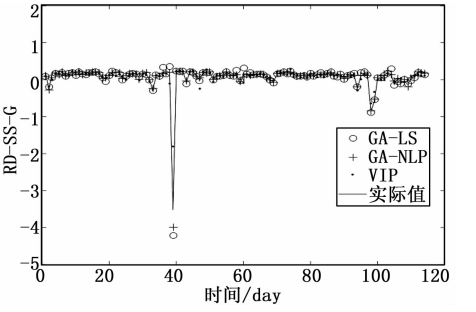

图 4　RD－SS－G 的预测输出与实际输出

4　总结

辅助变量选择对于构建软传感器非常重要。为了选择最佳的辅助变量子集，提出了一种遗传算法结合 MINLP 问题的辅助变量选择方法（GA－LS），并在 GA－LS 的基础上，通过耦合训练集的 BIC 准则以及验证集的 MSE 准则提出了一种更精确的辅助变量选择方法（GA－NLP），并将所得辅助变量子集通过 BP 神将网络建立软测量模型。与其他方法相比，本文所提出的方法能够很好保证所选变量的质量。通过 4 组数据集的实验结果表明，该方法可以得到具有良好泛化能力的模型。本文还介绍了该方法在污水处理厂案例上的应用，结果表明，所提出的变量选择方法能够好的与关键变量相关性高且变量数尽可能少地辅助变量子集，从而建立预测性能良好的模型。

参考文献：

[1] Joseph B, Brosilow C B. Inferential control of processes: PartI Steady state analysis and design [J]. AICHE Journal, 1978, 24 (3): 485－491.

[2] Brosilow C, Tong M. Inferential control of processes: PartII The structure and dynamics of inferential control system [J]. AICHE Journal, 1978, 24 (3): 492－500.

(下转第 58 页)