

文章编号: 1001-0920(2007)06-0707-04

时滞多变量系统 PCA 优化建模

贾明兴, 王福利, 何大阔, 牛大鹏

(东北大学 信息科学与工程学院, 沈阳 110004)

摘 要: 主元分析(PCA)在工业生产过程的产品质量控制与故障诊断等方面已得到广泛应用,然而当过程的变量间存在着未知时滞性时,必须确定数据间的对应关系,否则 PCA 模型将会不准.基于此,提出了 PCA 优化建模方法.该方法以过程变量间的时滞常数为优化变量,在分析 PCA 模型特点基础上,确定主成分个数和 SPE 统计量为综合目标函数,并建立模型约束条件,采用遗传算法求解.最后给出了仿真实例,证明了所提出方法的有效性.

关键词: 主元分析; 优化模型; 遗传算法; 时滞

中图分类号: TP277

文献标识码: A

PCA optimal modeling for time delay multivariable system

JIA Ming-xing, WANG Fu-li, HE Da-kuo, NIU Da-peng

(College of Information Science and Engineering, Northeastern University, Shenyang 110004, China. Correspondent:

JIA Ming-xing, E-mail: jiamingxing@ise.neu.edu.cn)

Abstract: As an effective method to extract correlations among variables, principal component analysis (PCA) is widely applied to multivariate statistical process monitoring, fault diagnosis and quality control. For process modeling with unknown time delay, time delay parameters are determined by using the conventional PCA method. Otherwise, the model will be so inaccurate that influences the monitoring capability. Therefore the PCA optimal modeling method is proposed, in which process variable time delay parameter is used as optimal variable. Based on the analysis characteristic of PCA model, the integration target function of PCs and SPE statistics is designed, and the restriction condition is constructed. The genetic algorithm is adopted for obtaining the solution of optimal model. A simulation result shows that effectiveness of the proposed method.

Key words: Principal component analysis; Optimal model; Genetic algorithm; Time delay

1 引言

主元分析(PCA)技术作为一种数据统计分析技术,已经应用于过程的建模、监测和故障诊断^[1,2].利用 PCA 模型能够抽取原始数据空间的主要变化信息,在保证数据信息丢失最少的情况下,大大降低原始数据空间的维数,消除变量间的非线性关联,降低噪声影响^[3].

基于 PCA 的过程监测和故障诊断是对过程的正常历史数据进行 PCA,得到主元子空间 P ,利用在线得到的过程数据以及 P 计算出当前数据的估计,进而得到平方预测误差(SPE 统计量),将它与阈值比较,以判断是否发生故障.该方法的一个基本假设前提是过程的变量在关系上是对齐的,然而许多工业生产过程的变量间存在着时滞性,如污水处理过程、啤酒发酵过程、烟气脱硫过程等^[4-6].常规

PCA 方法在处理这样的过程时,认为时滞常数是已知的,而实际上往往有个范围,但不精确,甚至不知其大小,故造成数据间的错位,建模精度不高,影响过程监测的能力.

为此,本文提出了 PCA 优化建模方法.该方法以过程变量间的时滞常数为优化变量,以 r 和 SPE 为优化目标,确立综合目标函数,建立优化数学模型.通过求解该优化问题辨识出时滞常数,进而获得过程的时滞 PCA 模型.

2 优化模型建立

对于过程变量时滞常数未知或不能准确确定的情况,如何建立优化模型来获得时滞常数的解,是优化 PCA 监测模型建立的关键问题,主要包括 3 部分:优化变量的选择、优化目标的建立、约束条件的建立.

收稿日期: 2006-04-30; 修回日期: 2006-07-31.

基金项目: 国家自然科学基金项目(60374003).

作者简介: 贾明兴(1972—),男,辽宁凌源人,副教授,博士,从事复杂工业过程的建模与故障诊断等研究;王福利(1957—),男,辽宁辽阳人,教授,博士生导师,从事复杂工业过程建模与优化、智能控制等研究.

2.1 优化变量的选择

假设已知过程变量正常情况下的数据矩阵为 $X_{n \times m}$, m 为变量个数, n 为采样点数, 数据具有相同的采样周期, 数据点在时间上是对齐的, X_n^i 为第 i ($i = 1, 2, \dots, m$) 个变量的 n 个采样点集. 若变量间由于工艺的缘故存在着不同滞后, 则变量间的对应关系可由图1表示. 图中 $t_1, t_2, \dots, t_n$ 表示采样时刻, $l_2, l_3, \dots, l_m$ 是以第1个变量为标准, 在数据上具有对应相关关系情况下, 第2, 3到 m 个变量的滞后时间. 可见, $l_2, l_3, \dots, l_m$ 有正有负, 正表示超前, 负表示滞后, 统称滞后.

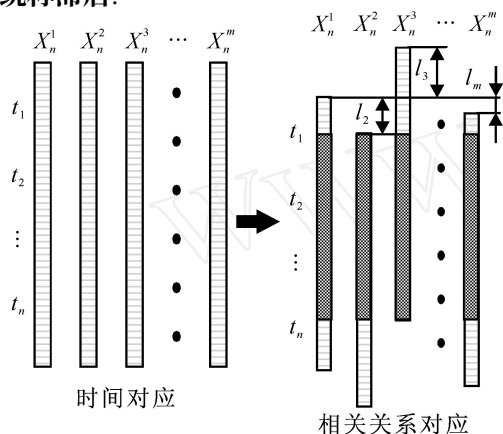


图1 变量间的对应关系

图1中, 时间对应关系表示的是原始建模数据矩阵 $X_{n \times m}$, 相关关系对应中阴影部分表示的是时滞处理后的有效建模数据矩阵, 记为 $X_{n' \times m}$, n' 为阴影部分数据个数, 有

$$n' = n - l_{\max}^+ - l_{\min}^- \quad (1)$$

式中: $l_{\max}^+$ 表示集合 $\{l_i, i = 2, 3, \dots, m\}$ 中正数的最大值, 若无正数, 则取值为0; $l_{\min}^-$ 表示集合 $\{l_i, i = 2, 3, \dots, m\}$ 中负数的最大值, 若无负数, 则取值为0; n' 由 n 和 $l_2, l_3, \dots, l_m$ 几个未知量决定, 因此选择 $l_2, l_3, \dots, l_m$ 为优化变量, 优化变量的个数为 $m - 1$.

2.2 优化目标建立

一个过程, 无论采用何种建模方法, 其真实模型是唯一的, 也是最优的. 就多元统计 PCA 建模方法而言, 针对图1所示的过程数据, 在不清楚过程变量滞后时间的情况下, 可以以数据阵 $X_{n \times m}$ 直接建模, 也可以以任意滞后时间 $l_2, l_3, \dots, l_m$ 所形成的数据矩阵 $X_{n' \times m}$ 建模.

2.2.1 SPE 最大值

对于一个模型, 评判其好坏的一个最普遍适用的指标是模型估计误差. 多元统计 PCA 建模中, 该项指标采用平方预测误差 SPE, 以原始数据矩阵 X 为例, 有

$$X =$$

$$t^1(p^1)^T + t^2(p^2)^T + \dots + t^r(p^r)^T + \dots + t^m(p^m)^T = TP^T + E \quad (2)$$

式中: p^i 为 $(m \times 1)$ 维负载 (loading) 向量, 亦是主成分的投影方向; t^i 是 $(n \times 1)$ 维得分 (score) 向量, 也称为主成分向量, $t^i = Xp^i, i = 1, 2, \dots, m$; r 为选取主成分个数, $r < m$; T 为得分矩阵, $T = [t^1, t^2, \dots, t^r]$; P 为负载矩阵, $P = [p^1, p^2, \dots, p^r]$; $E = X - TP^T$ 为在 r 个主成分下的残差. 于是

$$\text{spe}_i = e_i e_i^T \quad (3)$$

式中: spe_i 表示 x_i 的平方预测误差, 是 SPE 的第 i 个元素; x_i 为 X 的第 i 个元素; e_i 为 x_i 在 r 个主成分下的残差,

$$e_i = x_i - \hat{x}_i \quad (4)$$

$\hat{x}_i$ 为 x_i 在 r 维投影空间 P 下的重构值,

$$\hat{x}_i = x_i P P^T \quad (5)$$

这里, 选择 SPE 的最大值作为优化目标参数之一, 记为 $\text{spe}_{\max}$. $\text{spe}_{\max}$ 越小, 模型越接近真实过程.

2.2.2 主成分个数

主成分个数 r 代表了 PCA 模型投影空间的大小, 反映的是变量间相关的数量. 一般情况下, 根据累计贡献率来决定主成分个数. 令

$$\lambda^i = \frac{1}{n} \|t^i\|^2, i = 1, 2, \dots, m \quad (6)$$

式中 λ^i 为第 i 个主成分的方差, 有关系 $\lambda^1 > \lambda^2 > \dots > \lambda^m$, 贡献率为它与总方差的比值

$$\eta = \lambda^i / \sum_{j=1}^m \lambda^j \quad (7)$$

求 η ($i = 1, 2, \dots, m$) 的累积和称为累积贡献率, 当 r 个主成分的累积贡献率超过一定的指标后 (一般 80 % 足够), 就可以认为主成分个数为 r , 可以综合原数据足够多的信息.

选取不同的 $l_2, l_3, \dots, l_m$, 用累积贡献率法确定主成分个数 r , r 可能不同. 显然, 对于充足的数据, n' 足够大, r 值越小, 模型越接近真实过程, 因此选择 r 作为优化目标参数之一.

2.2.3 综合指标

由上述分析可见, 时滞系统 PCA 建模是一个多目标优化问题, 这增加了解题的难度. 就 $\text{spe}_{\max}$ 与主成分个数 r 两个目标而言, 其间具有如下关系:

- 1) $\text{spe}_{\max}$ 小 r 未必小, 同样, r 小 $\text{spe}_{\max}$ 未必小;
- 2) 在 r 相同情况下, $\text{spe}_{\max}$ 越小越好;
- 3) 若 r 小, 累积贡献率超出预定指标, 即使 $\text{spe}_{\max}$ 大也认为好.

为此, r 在目标中应占有较大的权重, 起主导作用, $\text{spe}_{\max}$ 占有较小权重, 起辅助作用. 进一步分析, r 为 $1 \sim m$ 之间的自然数, $\text{spe}_{\max}$ 为大于 0 的实数, 其

值可大可小.若将二者直接相加,则 spe_{max} 作用将大于 r ,与上述分析不符,为此可将 spe_{max} 转化为 $0 \sim 1$ 之间的数,则 spe_{max} 在目标中的影响将会小于 r .基于此设计综合指标为

$$r - e^{-spe_{max}}. \tag{8}$$

2.3 约束条件

该优化问题的变量就是过程变量的时滞常数,其范围往往是可以预知的,因此有约束

$$l_{i\min} \leq l_i \leq l_{i\max}, i = 2, 3, \cdots, m, \tag{9}$$

式中 $l_{i\max}$ 和 $l_{i\min}$ 为第 i 个时滞常数的上下限.

2.4 优化模型

时滞系统 PCA 优化模型为

$$\begin{aligned} \min(r - e^{-spe_{max}}), \\ l_{i\min} \leq l_i \leq l_{i\max}, i = 2, 3, \cdots, m. \end{aligned} \tag{10}$$

3 优化算法

关于优化问题的解法很多,有单纯形法、共轭梯度法、惩罚函数法、遗传算法、蚁群算法等,每种算法都有其自身的特点和适用对象.其中遗传算法是模拟达尔文的自然选择学说和自然界的生物进化过程的一种计算模型,它采用简单的编码技术来表示各种复杂的结构,并通过对一组编码表示进行简单的遗传操作和优胜劣汰的自然选择来指导学习和确定搜索的方向.这是一种新型的全局优化搜索算法,因为它直接对结构对象进行操作,不存在求导和函数连续性的限定,鲁棒性强、随机性、全局性以及适于并行处理,已广泛应用于神经网络、计算机科学、优化调度、运输问题、组合优化、机器学习、信号处理、自适应控制和人工生命等领域,并且遗传算法在实际应用中也取得了巨大成功^[7].

时滞系统 PCA 优化模型变量为过程变量间的滞后步数,是离散值,因此目标函数是不连续的,也就不存在导数.另外,从优化目标值的建立可以知道,目标值是变量的复杂的非线性隐式函数,难以解析表达,有可能存在局部极小.基于此特点,本文选择遗传算法求解.

Step1: 确定种群个数 n_g ,将可行解群体在约束条件下初始化,每一个可行解用一个向量 x_i 编码,称为一条染色体,这里采用二进制编码,编码长度由变量范围确定,向量的分量代表基因,它对应可行解的某一决策变量;

Step2: 计算群体中每条染色体 $l_{xi} (i = 1, 2, \cdots, n_g)$ 所对应的目标函数值,并以此计算适应值 $F_i = e^{-spe_{max}} - r$,按 F_i 的大小来评价该可行解的好坏;

Step3: 以优胜劣汰的机制,将适应值差的染色体淘汰掉,对幸存的染色体根据其适应值的好坏,按概率随机选择,进行繁殖,形成新的群体;

Step4: 通过杂交和变异的操作,产生子代,杂交是随机选择两条染色体(双亲),将某一点或多点的基因互换而产生两个新个体,变异是基因中的某一点或多点发生突变;

Step5: 对子代群体重复 Step2~Step4的操作,进行新一轮遗传进化过程,直到迭代收敛(适应值趋稳定)即可找到最优解或准最优解.

4 仿真研究

考虑如下时滞多变量系统:

$$\begin{aligned} x_1 &= [[0 : 0.1 : 20] \ 40 \times \text{ones}(1, 50) \rightarrow \\ &\quad \leftarrow 10 \times [20 \ :- \ 0.1 : 0]] + \text{rand}(1, 452), \\ x_2(k) &= 3x_1(k - 20) + (0.5 - \text{rand}), \\ x_3(k) &= \\ &\quad x_1(k - 30) + x_2(k - 10) + (0.5 - \text{rand}). \end{aligned}$$

式中: k 为采样时刻, rand 为 $0 \sim 1$ 间的随机数.

若以变量 x_1 为标准,则变量 x_2 超前 20 个采样时刻, x_3 超前 30 个采样时刻.以前 300 点进行常规 PCA 建模,得主成分个数为 2, SPE 如图 2 所示.

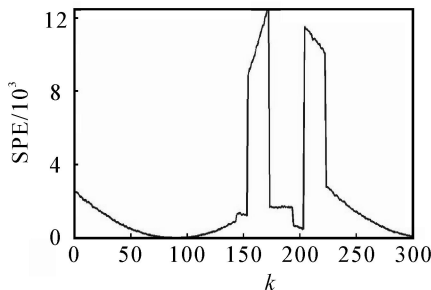


图 2 常规 PCA 建模 SPE

可见,主成分个数不能正确反应变量间的关系, SPE 变化起伏较大,这主要是未考虑变量间的滞后造成的.

采用本文提出的优化 PCA 建模方法,设定 l_2 变化范围为 $-40 \leq l_2 \leq 40$, l_3 变化范围为 $-60 \leq l_3 \leq 60$,遗传算法初始种群个数为 20,编码为二进制.优化最大迭代次数为 50,进化过程中, SPE 最大值如图 3 所示,主成分个数和适应度、种群平均适应度值如图 4 所示.得最优目标值为 -3.0×10^{-6} ,最优变

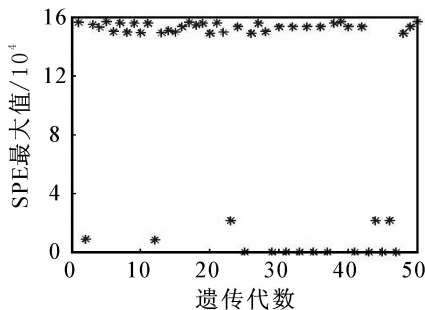


图 3 SPE 最大值

量值为 $l_2 = -20$, $l_3 = -30$, 此时主成分个数为 1, SPE 如图 5 所示. 可见所提出的方法是有效的.

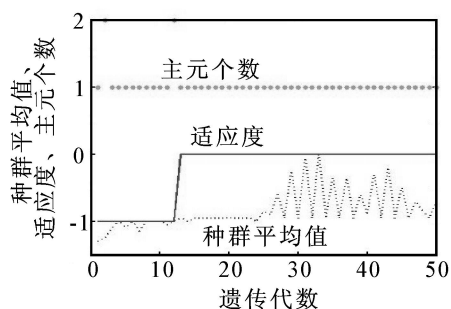


图 4 适应度、种群平均值和主成分个数

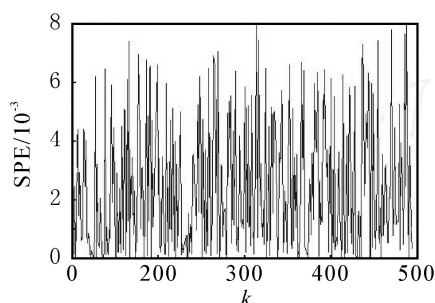


图 5 优化 PCA 建模 SPE

5 结 语

本文针对时滞过程 PCA 建模问题进行了研究. 在分析过程特性和 PCA 建模特性的基础上, 设计了 PCA 优化建模模型. 此模型以时滞常数为寻优变量, 以 $r - e^{-\text{spe}_{\max}}$ 为综合优化目标, 该目标充分体现了主成分个数和 SPE 指标的不同作用. 对于该优化问题, 采用遗传算法求解, 避免了局部最优和连续求导问题. 最后给出了实例, 以常规方法和优化方法分别进行仿真, 结果表明常规的方法对于时滞过程建模无能为力, 优化方法可以获取过程的时滞常数, 建立准确的模型.

参考文献 (References)

[1] Evan L R, Leo H. Fault detection in industrial

processes using canonical variate analysis and dynamic principal component analysis [J]. Chemometrics and Intelligent Laboratory Systems, 2000, 51(8): 81-93.

[2] Seongkyu Yoon, Mac Gregor J F. Statistical and causal model based approaches to fault detection and isolation [J]. AIChE J, 2000, 46(9): 1813-1824.

[3] Kourtis T, Mac Gregor J F. Multivariate SPC methods for process and product monitoring [J]. J. of Quality Technology, 1996, 28(4): 409-428.

[4] 薛福珍, 庞国仲, 林盛荣. 啤酒发酵过程的建模仿真与控制[J]. 中国科学技术大学学报, 2001, 31(4): 502-508.

(Xue F Z, Pang G Z, Lin S R. Modeling simulation and control for beer fermentation process [J]. J. of China University of Science and Technology, 2001, 31(4): 502-508.)

[5] 刘翔, 王文海, 熊斌, 等. 典型大时变时滞工业过程的鲁棒数字 PID 调节控制[J]. 自动化学报, 2002, 28(5): 784-788.

(Liu X, Wang W H, Xiong B, et al. Robust digital PID adjusting control for a general industrial process with large time-varying time delay [J]. Acta Automatica Sinica, 2002, 28(5): 784-788.)

[6] 杜树新. 污水生化处理过程建模与控制[J]. 控制理论与应用, 2002, 19(5): 660-666.

(Du S X. Modeling and control of biological wastewater treatment processes [J]. Control Theory and Applications, 2002, 19(5): 660-666.)

[7] 林旭梅, 梅涛. 一种基于自适应遗传算法的 CMAC 的学习率优化方法[J]. 系统仿真学报, 2005, 17(12): 3081-3084.

(Lin X M, Mei T. A kind of optimization based on adaptive GA for cerebellar model [J]. Articulation Acta Simulata Systematica Sinica, 2005, 17(12): 3081-3084.)

稿 约 声 明

本刊已加入《中国学术期刊网》(光盘版)、《万方数据——数字化期刊群》、重庆维普咨询公司《中文科技期刊数据库》和(中国·台湾)华艺数位艺术股份有限公司《中文电子期刊服务》. 其作者文章的稿酬本刊一次性付清. 凡不同意入编的稿件, 请作者在投稿时声明.