文章编号: 1002-1175(2009)02-0209-06

基于改进的 OFA-MPCA 的监控方法*

边福强[1-] 高 翔[1] 苑明哲[2]

(1 沈阳化工学院信息工程学院, 沈阳 110142; 2 中国科学院沈阳自动化研究所, 沈阳 110015)

(2008 年 6 月 30 日收稿; 2008 年 7 月 16 日收修改稿)

Bian FQ, Gao X, Yuan MZ. Monitoring based on improved OFA-MPCA. ***Journal of the Graduate School of the Chinese Academy of Sciences*****, 2009, 26(2): 209~214**

摘 要 多向主元分析(MPCA)是利用多变量统计方法从纷杂的海量数据信息中提取出能够准确表征数据信息的几个主元, 并通过投影法来降低数据的维数, 主要应用于间歇生产过程中. 在实际的间歇生产过程中, 由于各种原因导致各批次异步造成它们运行时间的不一致, 而无法直接建立有效的统计模型, 正交函数近似(OFA)是一种基于正交基的投影变换技术, 通过对原始数据进行 OFA 处理后, 可以用投影系数来描述原始数据所具有的特征, 并且可以达到轨迹同步化和压缩数据量的目的. 对 OFA 法进行了部分改进, 并结合 MPCA 法对典型的间歇过程——青霉素发酵过程进行了仿真研究. 结果表明, 改进的 OFA 计算速度有了极大的提高, 且改进的 OFA-MPCA 法能完好地对各批次进行同步、建模并得出准确的监视结果.

关键词 正交函数近似, 多向主元分析, 间歇过程, Pensim

中图分类号 TP277

1 引言

间歇生产过程是一个动态变化过程, 它的过程反应时间限定在固定的时间段内, 一般不同的间歇生产过程差别很大. MPCA[1] 法是由 Nomikos & MacGregor 首次开发的针对间歇过程进行监控的一种方法, 该法不需要了解间歇生产过程的物理机理, 也不用去归纳系统的规则, 甚至不要求精确的数学模型, 是三维数据条件下主元分析法的扩展. 然而, 在实际的批量生产过程中用于建模的各历史批次长度是不一致的, 也就无法直接使用 MPCA 进行建模和监控, 这要求我们必须想方设法对建模数据进行同步化处理.

多元统计模型的同步化方法基本上有: Shah, *et. al.* 提出的基于 PCA 的方法[2], 这种方法只有在 PCA 轨迹都具有全局线性这个假设下成立; Lakshminarayanan, *et. al.* 的扩展每一条轨迹与最长的轨迹的持续时间相匹配, 短的轨迹用该轨迹最后采集到的数据补齐[3], 不切实际; Kourti, *et. al.* 提出指示变量法[4], 但很难找到这样一个满足要求的单变量; Kassidas, *et. al.* 建立了 DTW 理论[5], 但其算法实现起来相当复杂; Junghui & Jialin 的 OFA 理论[6,7], 基本思想是把每一条变量轨迹用正交函数集及相应的投影系数进行近似表示, 并把投影系数作为这条变量轨迹的测量值, 从而达到同步化的目的. 其原理易于理解, 算法较简单, 程序执行速度快, 较为实用.

* 国家高技术研究发展计划(863 计划)(2006AA04Z185); 辽宁省教育厅项目(2005320)资助

- E-mail: fudy81@126.com

本文采用了 OFA 法, 并做了部分改进. 通过改进, OFA 法可以更快速而有效地处理运行时间长、采样样本大、变量方差波动大的间歇过程, 而且可以确保实时性以适用于某些具有实时性要求系统过程监控中. 仿真结果显示, 改进后的方法与改进前相比, 可行性有了明显提高, 可以快速地处理数据并保证处理结果的有效性, 从而更好地为产品的质量和企业的经济效益服务.

在 Junghui & Jialin 的文章中使用了 2 种模型进行验证[6]: 分别是某一催化化学反应和半导体制造工艺中的等离子体刻蚀过程进行了仿真研究, 这 2 种模型的运行时间较短, 相应的采样样本比较少, 属于快速反应过程; 2 种模型各自变量的相对幅值, 各自批次之间的相对波动都比较小. 我们研究的青霉素发酵过程的运行时间可达到几十甚至上百个小时, 相应的采样样本比较大, 变量多且变量间存在一定的内在联系, 各批次的运行时间也不同. 在这种模型的情况下研究该算法是具有一定的现实意义的. 在仿真时我们也对 OFA 法和传统的截取法进行了对比, 验证了应用 OFA 法进行同步化处理的优越性. 文中对传统截取法不能有效报告故障批次的原因做了一定的分析.

2 MPCA 和改进 OFA 理论简介

主元分析法主要用于处理连续过程, 是通过投影把高维的数据信息变成低维, 同时保留原始数据的所有特征. 而 MPCA 法主要处理间歇过程. 一般批处理过程(时间上同步) 中所采集的数据可以用一个三维矩阵 $\underline{\boldsymbol{X}}(I \times J \times K)$ 来表示. 其中 J 代表每个批处理过程中的测量变量数, K 代表对每一个测量变量采集的测量值的个数, I 代表批处理过程的批次数.

MPCA 对三维数据矩阵的一般处理方法是: 进行垂直切片处理$(I \times J)$, 即按照时间顺序组织成时间块, 每一个时间块代表一个批处理过程, 然后把每一个垂直切片$(I \times J)$ 沿右侧并排依次放置, 第一片对应于第一时刻的采样值$(k = 1)$, 第 K 片对应于第 K 时刻的采样值$(k = K)$. 三维数据矩阵展开的结果如式 (1) 所示:

$$\boldsymbol{X} = \begin{bmatrix} \overbrace{x_1, x_2, \cdots, x_J}^{i=1;\,k=1} & \cdots & \overbrace{x_1, x_2, \cdots, x_J}^{i=1;\,k=K} \\ \overbrace{x_1, x_2, \cdots, x_J}^{i=2;\,k=1} & \cdots & \overbrace{x_1, x_2, \cdots, x_J}^{i=2;\,k=K} \\ \vdots & & \vdots \\ \overbrace{x_1, x_2, \cdots, x_J}^{i=I;\,k=1} & \cdots & \overbrace{x_1, x_2, \cdots, x_J}^{i=I;\,k=K} \end{bmatrix}. \tag{1}$$

按照式 (1) 法处理后产生的二维矩阵 $\boldsymbol{X}$ 的维数是$(I \times JK)$, 与 PCA 的统计和算法相似, MPCA 也要把得到的二维矩阵 $\boldsymbol{X}$ 分解成得分向量$\boldsymbol{t}$ 与负荷矩阵$\boldsymbol{P}$ 的乘积加上误差矩阵$\boldsymbol{E}$ 的形式.

$$\underline{\boldsymbol{X}} = \sum_{k=1}^{r} \boldsymbol{t}_k \otimes \boldsymbol{P}_k + \underline{\boldsymbol{E}} \quad \text{or} \quad \boldsymbol{X} = \sum_{k=1}^{r} \boldsymbol{t}_k \boldsymbol{P}_k^{\mathrm{T}} + \boldsymbol{E}, \tag{2}$$

其中, $\boldsymbol{t}$ 代表批次之间的关系, $\boldsymbol{P}$ 代表所有的变量及它们的时间变量, r 是保留在 PCA 模型中的主元个数, $\otimes$ 代表 Kronecker 积. 经过这种主元分解处理后数据会极大地减少.

OFA 法是通过把原始数据在正交基函数上进行投影来提取数据的特征并达到同步化的目的. 该法把每一批次中的每一变量的轨迹线用一个函数 $F(t)$ 来表示. 函数 $F(t)$ 又可以用一个连续函数的正交函数集$\{\phi_n\}$来近似:

$$F(t) \approx F_n(C, t) = \sum_{n=0}^{N-1} \alpha_n \phi_n(t), \tag{3}$$

其中的系数, $C = \{\alpha_n\}$, $\alpha_n = \int F(t)\, \phi_n(t)\, \mathrm{d}t$ 是函数$F(t)$在每一个基函数上的投影. 因此, 绝大部分的原始过程变量 $F(t)$ 的批次工作特征被提取为隐含的关键参数 $C = \{\alpha_n\}$, 这个参数包含了每个批次中的每个变量的必须的工作条件.

在计算规格正交函数集$\{\Phi_n\}$时, 可以采用勒让德多项式、Hermite 多项式、Laguerre 多项式等计算一个规格正交序列或者采用 Gram- Schmidt 法直接计算一个规格正交矩阵. 然而, 直接计算规格正交矩阵很难确定其维数, 所以本文延续了 Junghui & Jialin 的方法[6], 并采用递推算法(可以加快运算速度) 求取 n 阶勒让德多项式构成 n 个规格正交序列. 另外, 因为实际的变量轨迹值都是由采样值构成的, 即都是离散化的, 所以相应地也需要对确定的每一个规格正交序列进行离散化处理, 即

$$\Phi_n(t) = [\Phi_n(t_1) \quad \Phi_n(t_2) \quad \cdots \quad \Phi_n(t_K)], \tag{4}$$

其中, K 为每个变量的采样样本数.

当采样样本很大时, 相应的对每一个规格正交序列进行离散化的计算量就会很大, 巨大的计算量很难确保监控的实时性.

本文提出了一种新的规格正交函数集的离散化处理方法并基于这样一个事实: 在计算第一个勒让德多项式时, 即勒让德多项式中的参数 n 等于 0 时, 计算出来的勒让德多项式值是一个常数, 由其计算出来的离散化的规格正交函数集中的元素都是相同的. 那么可不可以对其他的由勒让德多项式构成正交序列进行固定点离散化, 从而使每一个序列离散化后的元素各自相同呢? 本文采用了下面的处理方法.

参数 t 都选择了区间$[-1, 1]$上的固定随机数, 它只会随着 n 的增加, 即在计算下一个规格正交函数集时发生变化, 这样做既满足了勒让德多项式的正交性又有效地降低计算量, 特别是在一个有着成百上千采样样本的化工间歇过程中有效地降低了计算量. 这种处理方法可能会使正交基失去时序特性, 从而影响到投影系数 α_n 以及监控效果, 但仿真结果却验证了其并没有影响到投影系数和监控效果, 因为经过这种处理后, 每个行向量依然是正交的.

计算完每一批次的每一变量的隐含关键参数后, 重构系数矩阵 Θ 如下:

$$\Theta = \begin{bmatrix} \boldsymbol{C}_{1,1} & \boldsymbol{C}_{1,2} & \cdots & \boldsymbol{C}_{1,J} \\ \boldsymbol{C}_{2,1} & \boldsymbol{C}_{2,2} & \cdots & \boldsymbol{C}_{2,J} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{C}_{I,1} & \boldsymbol{C}_{I,2} & \cdots & \boldsymbol{C}_{I,J} \end{bmatrix}, \tag{5}$$

其中, $\boldsymbol{C}_{i,j} = [\alpha_{i,0} \quad \alpha_{i,1} \quad \cdots \quad \alpha_{i,N_{j-1}}]$ 代表 i 批次的测量变量j 的近似函数的系数向量——隐含的关键参数, N_j 是对变量j 进行映射满足条件时必须的项的个数. 系数矩阵 Θ 的维数$I \times N$, 其中 $N = \sum_{j=1}^{J} N_j$. 此方法就是把第 i 批次变量j 的测量值$\boldsymbol{X}_{i,j}$分解成

$$\boldsymbol{X}_{i,j} = \boldsymbol{C}_{i,j} \Phi_j, \tag{6}$$

其中,

$$\Phi_j = \begin{bmatrix} \phi_0(t_1) & \phi_0(t_1) & \cdots & \phi_0(t_1) \\ \phi_1(t_2) & \phi_1(t_2) & \cdots & \phi_1(t_2) \\ \vdots & \vdots & & \vdots \\ \phi_{N_{j-1}}(t_{N_{j-1}}) & \phi_{N_{j-1}}(t_{N_{j-1}}) & \cdots & \phi_{N_{j-1}}(t_{N_{j-1}}) \end{bmatrix}, \tag{7}$$

$t_1, t_2, \cdots, t_{N_{j-1}}$是区间$[-1, 1]$上的固定随机数.

本文提出的改进的 OFA 法对于处理具有较多采样样本的化工间歇过程, 或某些需要实时统计监控的过程是非常有利的. 在相同的仿真条件下对比改进的 OFA 法与原 OFA 法, 处理同样的一个具有 10 变量 400 采样样本的过程数据, 改进的 OFA 法仅需 0.25s, 而原 OFA 法则需要 11.016s 左右, 处理速度提高了近 50 倍. 应用改进的 OFA 法处理具有上千个采样样本的过程数据会更具有优势, 足以满足某些化工实时统计监控的需求. 完成所有批次的同步化处理后, 就可以应用 PCA 技术来提取组合系数中的相关特征, 并对新批次的情况进行评估了.

3 仿真实例

为了验证改进的OFA 法的正确性, 本文通过测试10 个批次的 SPE 检验及SPE 贡献图 2 个方面与原 OFA 法进行了仿真对比. 并与传统多元轨迹同步化方法进行对比得出OFA 法的优越性. 所谓传统的多元轨迹同步化方法是对各批次的各个变量轨迹或者进行截去处理使之与最短的轨迹相同, 或者补轨迹的最后值使之与最长的轨迹相同, 或者截取补充到指定的长度, 本文采用了截去法, 使所有批次的轨迹与最短批次的轨迹相同.

本次仿真试验的数据来源于青霉素发酵模拟软件 PenSim2. 0, 共采集了 60 个正常运行的批次, 各批次的仿真时间介于 350~ 450h 并近似呈正态分布, 采样时间是 1h, 并从PenSim2. 0 软件所模拟的 18 个过程变量的数据中选取出与产品浓度相关的 10 个变量的数据用于建模. 并采集了 6 个故障批次对本文采用的方法进行检验. 选用的 10 个变量如下: 通风速率, 鼓风机功率, 葡萄糖反馈温度, 溶解氧浓度饱和度, 培养容积, 二氧化碳浓度, pH 值, 温度, 产生热量, 冷水速率. 应用传统法时所有批次的仿真时间统一取 350h, 以满足 MPCA 方法建立监控模型的要求.

测试数据信息: 1~ 6 批次是故障数据, 7~ 10 批次是正常数据. 故障批次 1、2 为变量 2 故障, 阶跃幅值 10%, 斜坡幅值 1%, 故障时间从 45h 到 200h, 属于早、中期故障; 3、4 批次为变量 3 故障, 阶跃幅值 65%、斜坡幅值 0. 1%, 故障时间从 300h 到 400h, 属晚期故障; 5、6 批次为变量 3 故障, 阶跃幅值 70%、斜坡幅值 0. 1%, 故障时间从 45h 到 150h, 属早期故障.

在仿真过程中要用到以下 3 个公式[8, 9].

平方预测误差(SPE) 也称 Q 统计量, 其计算公式:

$$Q(i) = [\Theta(i) - \boldsymbol{t}_r(i)\boldsymbol{P}_r^{\mathrm{T}}][\Theta(i) - \boldsymbol{t}_r(i)\boldsymbol{P}_r^{\mathrm{T}}]^{\mathrm{T}}, \tag{8}$$

其中, r 为保留的主元数, i 是批次.

Q 统计控制限的计算公式:

$$Q_\alpha = \theta_1\left[\frac{c_\alpha\sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0(h_0 - 1)}{\theta_1^2}\right]^{1/h_0}, \tag{9}$$

其中, $\theta_i = \sum_{j=r+1}^{n}\lambda_j^i\ (i = 1, 2, 3)$, $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, λ_j 为 Θ 的协方差矩阵的特征值, c_α 是标准正态分布在置信水平 α 下的临界值, r 为保留的主元数, n 为全部主元数.

SPE 变量贡献图的计算公式:

对于第 i 个批次, 第 j 个变量的正交系数 $\boldsymbol{C}_{i,j}$, 其 SPE 的贡献率为:

$$\mathrm{Con}\_\boldsymbol{C}_{i,j} = \sum_{n=0}^{N_{j-1}}(\alpha_{i,n} - \hat{\alpha}_{i,n})^2, \tag{10}$$

这样, 在某一批次 i 中, 将各变量的贡献率画在一起就得到贡献图.

采用青霉素发酵过程的仿真数据, 验证基于改进的 OFA-MPCA 的监控方法的仿真结果如图 1 和图 2 所示.

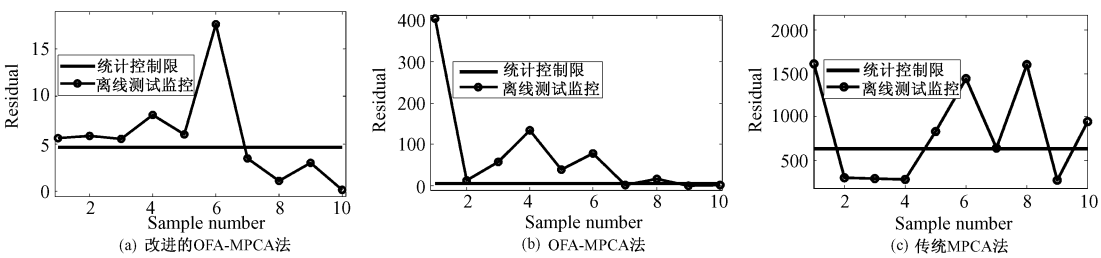

图 1 10 个测试批次 SPE 检验对比

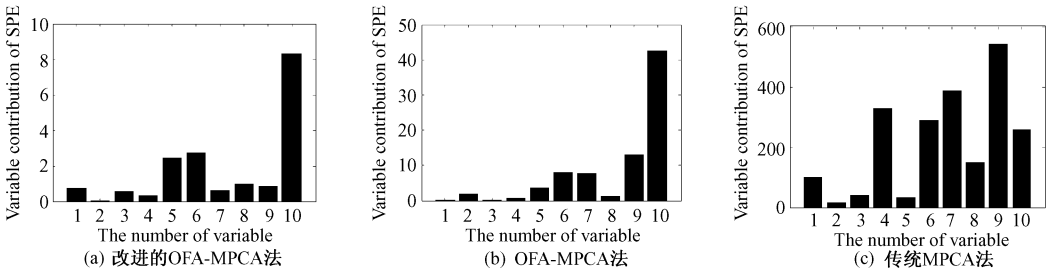

图 2　测试批次 6 的 SPE 变量贡献图对比

从图 1 中(a)、(b)2 个子 SPE 图里可以看出, 对于故障批次 1~ 6, 改进的 OFA 法与原 OFA 都检测出来了, 对于正常批次 7~ 10, 原 OFA 对批次 8 误报, (c) 图误报漏报都有; 从图 2 中的(a)、(b)2 个子 SPE 贡献图可以看出, 2 种方法在判断出现故障的变量时结果基本上一致, 由此得出改进的 OFA 方法的正确性, (c) 图没有很好地指出哪个变量出现了故障. 传统的同步化方法由于截去或补加了一部分采样数据, 导致运行批次的特征信息的丧失或改变, 在故障检测时存在很大的漏报误报情况, 由此可以看出 OFA 法的优越性.

4　总结

本文将正交函数近似法进行了改进, 并与多向主元分析方法相结合, 对青霉素发酵过程进行了监控, 验证了改进 OFA 法的可行性、快速性及优越性. 对于运行时间长、采样样本大、变量方差波动大的系统进行同步化处理或进行实时统计监控时, 本文提出的方法具有无可比拟的优越性. 本文提出的方法, 也可以应用在其他的多元统计方法中, 如 MPLS、MICA 等对质量数据进行监控或者对过程数据进行监控都具有研究意义.

参考文献

[1] Nomikos P, MacGregor JF. Monitoring batch process using multi-way principle component analysis. *Journal of AICHE*, 1994, 40(8): 1361~1369

[2] Shah SL, Randy M, Takada H, *et al*. Modelling and control of a tubular reactor: A PCA-based approach. In: Proceedings of the 5th IFAC Symposium on Dynamics and Control of Process Systems. Corfu, Greece, 1998. 17~22

[3] Lakshminarayanan S, Guidi RD, Shah SL, *et al*. Monitoring batch processes using multivariate statistical tools: extensions and practical issues. IFAC Triennial World Cong. San Francisco, 1996. 241~246

[4] Kourti T, MacGregor JF. Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, 1996, 28(4): 409~427

[5] Kassidas A, MacGregor, *et al*. Synchronization of batch trajectories using dynamic time warping. *Journal of AIChE*, 1998, 44(4): 864~875

[6] Chen JH, Liu JL. Post analysis on different operating time Processes using orthonormal function approximation and multiway principle component analysis. *Journal of Process Control*, 2000, 10(5): 411~418

[7] Chen JH, Liu JL. Multivariate calibration models based on functional space and partial least square for batch processes. IFAC (ChemFas-4) Conference. Chejudo, Korea, 2001. 161~166

[8] Jiang TH, *et al*. Fault detection and diagnosis in industrial systems. China Machine Press, 2003

[9] Zhang J, Yang XH. Multivariable statistal process control. Chemical Industrial Publication House, 2000

Monitoring based on improved OFA-MPCA

BIAN Fu-Qiang[1] GAO Xiang[1] YUAN Ming-Zhe[2]
(1 *Information Engineering School, Shenyang Institute of Chemical Technology, Shenyang* 110142, *China*;
2 *Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang* 110015, *China*)

Abstract Multiway Principal Component Analysis (MPCA) is a multivariable statistical approach, which can extract several principal components from the numerous of data to express the data information well, and is mainly used in batch process. In practice, for many reasons, the runtime of each batch is different from others so that the effective statistical model can not be built directly. Orthonormal Function Approximation (OFA) is a technique of project transformation based on orthonormal base, after OFA we can use the projection coefficient to express the characteristics of the original data and synchronize the trajectories of each historical batch and reduce the dimension. This paper presents some improvement on the OFA and combined the MPCA to model and monitor the typical batch process——Penicillin fermentation process. The simulation results show that the improved OFA can deal with data more quickly and the improved OFA-MPCA is able to synchronize the trajectories of all the batches, and monitor the batches perfectly.

Key words orthonormal function approximation, MPCA, batch process, pensim