文章编号: 1001-0920(2005)05-0571-04

改进PCA 在发酵过程监测与故障诊断中的应用

肖应旺, 徐保国
(江南大学 通信与控制工程学院, 江苏 无锡 214036)

摘 要: 提出一种改进的主元分析(PCA)方法. 利用主元相关变量残差统计量代替平方预测误差Q 统计量, 并采用累积方差贡献率及复相关系数确定PCA 模型的主元数. 将改进的主元分析法应用于粘菌素发酵过程监测和故障诊断中, 仿真结果表明改进的PCA 方法避免了Q 统计量的保守性, 并保证了主元子空间中的信息存量. 与一种基于特征子空间的系统性能监控方法相比较, 改进的PCA 方法具有更强的有效性.

关键词: 主元分析; 统计过程监测; 发酵; 故障诊断
中图分类号: TP273 **文献标识码**: A

Application of improved PCA to fermentation process monitoring and fault diagnosis

XIAO Ying-wang, XU Bao-guo
(School of Communication and Control Engineering, Southernyangtze University, Wuxi 214036, China. Correspondent: XIAO Ying-wang, E-mail: ymy19701030@163.com)

Abstract: An improved principal component analysis (PCA) is presented which uses principal-component-related variable residual (PVR) statistic to replace Q-statistic. The principal components in PCA is decided by virtue of cumulative percent variance and multi-correlation coefficients. The improved PCA is applied to mycetozoan fermentation process monitoring and fault diagnosis. The simulation result shows that the improved PCA can avoid the conservation of Q-statistical test and ensure enough information in principal component subspace. Compared with a system performance monitoring based on characteristic subspace, the improved PCA is more effective.

Key words: PCA; statistical process monitoring; fermentation; fault diagnosis

1 引 言

发酵过程比较复杂, 其过程变量多且相关性强, 难以建立精确的机理模型来实现正确的故障监测. 另一方面, 大多发酵过程是不可逆的, 如果监控不准确, 很难使以后的发酵过程达到预期的目标. 因此对发酵过程进行准确监测和故障诊断十分重要.

近年来, 主元分析法(PCA)在化工过程的数据处理、故障诊断中得到了广泛应用[1,2]. 它是将多个相关变量转化为少数几个相互独立变量的一种有效分析方法[3]. 它不依赖于过程机理, 只需通过过程数据的信息进行统计建模. 本文利用主元相关变量残差(PVR)统计量代替通常的平方预测误差Q 统计量, 并用累积方差贡献率及复相关系数确定PCA 模型的主元数. 将该主元分析法应用于粘菌素发酵过程的监测和故障诊断中的仿真结果表明, 改进的PCA 能够提供更详细的过程变化信息, 提高了对故障原因的识别能力. 同时利用它确定的主元数保证了主元子空间中的信息存量.

2 PCA 过程统计监测模型

首先, 取一段正常工况生产下的过程数据, 并对其进行预处理后构成数据矩阵$X_{n\times m}$, n 为采样次数, m 为测量变量数. 由于同一变量采用的量纲不同, 在进行主元分析前, 需要对$X_{n\times m}$ 作标准化处理, 处理

收稿日期: 2004-06-21; **修回日期**: 2004-09-20.
基金项目: 国家科技攻关计划课题的子课题项目(2001BA204B01-03).
作者简介: 肖应旺(1970—), 男, 湖南常德人, 博士生, 从事过程监测与故障诊断、智能控制技术的研究; 徐保国(1950-), 男, 江苏淮阴人, 教授, 博士生导师, 从事过程控制、智能仪表等研究.

后的矩阵记作$X'_{n\times m}$. 建立 PCA 的过程统计模型

$$X'_{n\times m} = \hat{X}'_{n\times m} + E_{n\times m} = \hat{T}_{n\times k}\hat{P}^{\mathrm{T}}_{m\times k} + \tilde{T}_{n\times(m-k)}\tilde{P}^{\mathrm{T}}_{m\times(m-k)}. \quad (1)$$

式中: $\hat{X}'_{n\times m}$ 为 $X'_{n\times m}$ 的估计, $E_{n\times m}$ 为残差矩阵, $\hat{T}_{n\times k}$ 和 $\hat{P}_{m\times k}$ 分别为主元得分和负荷矩阵, $\tilde{T}_{n\times(m-k)}$ 和 $\tilde{P}_{m\times(m-k)}$ 分别为残差得分和负荷矩阵, $k(k<m)$ 为主元数

通过对矩阵 $X'_{n\times m}$ 的协方差矩阵 $R = X'^{\mathrm{T}}_{n\times m}X'_{n\times m}/(n-1)$ 进行特征向量分析, 得到负荷向量矩阵 $P_{m\times m}$, 则

$$R = P_{m\times m}\Sigma P^{\mathrm{T}}_{m\times m},\ T_{n\times m} = X'_{n\times m}P_{m\times m}. \quad (2)$$

式中: 对角矩阵 $\Sigma = \mathrm{diag}(\lambda_i), \lambda_i (i = 1, 2, \cdots, k)$ 为 $X'_{n\times m}$ 在新坐标系 $P_{m\times m}$ 相应方向上的方差

这样, k 维的主元空间代替了原来的 m 维过程数据空间, 并且消除了过程变量之间的相关性 通过对主元得分和残差变化进行分析, 即可在低维的主元空间中实现对多元统计过程的监测 具体过程是建立关于主元得分和残差两个统计量, 即 T^2 和 Q 统计量 其中对于某时刻新的测量样本 $X_{1\times m}$, T^2 统计量定义为

$$T^2 = \hat{t}_{1\times k}\Sigma^{-1}\hat{t}^{\mathrm{T}}_{1\times k} = X_{1\times m}\hat{P}_{m\times k}\Sigma^{-1}\hat{P}^{\mathrm{T}}_{m\times k}X^{\mathrm{T}}_{1\times m} \leq \frac{k(n-1)}{n-k}F_{k,n-k,\alpha} \quad (3)$$

对于某时刻新的测量样本 $X_{1\times m}$, Q 统计量定义为

$$Q = E_{1\times m}E^{\mathrm{T}}_{1\times m} = X_{1\times m}(I - \hat{P}_{m\times k}\hat{P}^{\mathrm{T}}_{m\times k})X^{\mathrm{T}}_{1\times m} \leq Q_\alpha \quad (4)$$

式中: $\hat{t}_{1\times k}$ 为 $X_{1\times m}$ 的主元值; α 为检验水平, $F_{k,n-k,\alpha}$ 为对应于检验水平是 α, 自由度是 $k, n-k$ 条件下 F 分布的临界值; $E_{1\times m}$ 为 $X_{1\times m}$ 的残差; Q_α 为 Q 统计量的 α 置信限[3].

在信息抽取过程中, 合理确定主元个数非常重要 在实际应用中, 可采用交叉检验法或累积方差贡献率(CPV) 方法确定主元个数[3]. 本文采用CPV 结合复相关系数确定 PCA 模型的主元数

将表示样本变量与主元的相关程度定义为复相关系数 $\rho(X'_i, T)$ (或 r_i)[4], 即

$$r_i = \rho(X'_i, T) = \left(\sum_{j=1}^{k}\lambda_j\hat{p}^2_{i,j}\right)^{1/2},\quad i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, k. \quad (5)$$

式中: X'_i 为 $X'_{n\times m}$ 的第 i 个列向量, T 为 k 个主元向量组成的主元矩阵, λ_j 为 $X'_{n\times m}$ 的协方差阵的特征值, $\hat{p}_{i,j}$ 为 $\hat{P}_{m\times k}$ 中的元素 主元包含每一变量的信息程度可用 $\rho(X'_i, T)$ 衡量, 同时此参数的变化情况反映了主元信息的变化程度 因此利用复相关系数可评价主元模型的近似程度和信息量[5].

3 主元相关变量残差统计量

T^2 和 Q 检验只能监测过程是否发生了变化, 不能直接提供引起变化的原因, 难以对故障进行识别, 变量贡献图[6] 实际给出的也只是定性的信息 文献[7] 提出了一种基于特征子空间的系统性能监控方法, 获得了比传统 PCA 更好的监控性能 但它仍不能有效判断过程变化究竟是由工况改变, 还是由过程故障引起的, 即不能确定过程中是否出现了故障 而本文的改进 PCA 方法能够准确地判断过程变化的原因 仿真结果表明, 将与主元显著相关的过程变量的残差构成一个新的统计量以取代 Q 统计量进行过程故障的检测与诊断是有效的

通过对新的测量数据进行 T^2 和 Q 检验, 可判断过程是否发生了变化 该检验将出现 4 种结果: 当 Q 统计量发生大的变化时, 说明 PCA 统计模型代表的正常工况下的变量之间关系被破坏, 有过程故障(或传感器故障) 发生; 当 T^2 统计量发生大的变化而 Q 没有明显变化时, 说明各变量之间的关系仍得到(近似) 满足, 但过程发生了某种变化, 既可能是由工况改变引起的, 也可能有过程故障发生, 只是未显著改变变量之间的关系 为此, 本文采用一个新的统计量来取代 Q 统计量进行过程故障的检测和诊断

设过程中有 m 个测量变量, 其中 s 个与主元显著相关, 称其所构成的残差为主元相关过程变量残差(PVR) 统计量; 其余 $m-s$ 个被测变量构成的残差称为一般过程变量残差(CVR) 统计量 类似 Q 统计量, 这两个统计量分别为

$$\mathrm{PVR} = X_s(I - P_sP^{\mathrm{T}}_s)X^{\mathrm{T}}_s, \quad (6)$$

$$\mathrm{CVR} = X_{m-s}(I - P_{m-s}P^{\mathrm{T}}_{m-s})X^{\mathrm{T}}_{m-s}. \quad (7)$$

式中 s 和 $m-s$ 分别为过程数据矩阵和负荷矩阵中对应于主元相关过程变量和一般过程变量的取值 实际上, Q 统计量值恰好分成了 PVR 和 CVR 两部分, 故PVR 和CVR 统计量控制限不采用 Q_α 计算, 而利用

$$\mathrm{PVR}_\alpha = \omega_{\mathrm{PVR}}Q_\alpha,\ \mathrm{CVR}_\alpha = \omega_{\mathrm{CVR}}Q_\alpha \quad (8)$$

计算 式中

$$\omega_{\mathrm{PVR}} + \omega_{\mathrm{CVR}} = 1,\quad \omega_{\mathrm{PVR}} = 1 - \sum_{i\in\mathrm{PV}}\gamma_i\Big/\sum_{i=1}^{m}\gamma_i \quad (9)$$

其中: γ_i 利用式(5) 计算, PV 表示与主元相关的过程变量

选取 PV 变量一种较好的方法是利用式(5) 计算主元与各过程变量之间的复相关系数[4], 根据复相关系数的大小选取与主元显著相关的过程变量

至此, T^2 检验、PVR 和CVR 检验共同构成了用于过程检测和故障诊断的改进的PCA. 其中 T^2 统计

和 PVR 统计反映的是与主元显著相关的变量信息, CVR 统计则主要反映与主元无明显相关的变量信息. 可见, 将Q 统计量反映的信息划分为更细致的两部分, 使得改进的 PCA 能充分刻画过程的变化, 增强了 PCA 的故障诊断能力.

4　粘菌素发酵过程的故障检测与诊断

工业上粘菌素发酵过程的测量变量有: 发酵周期、相对校价、总糖浓度、还原糖浓度、氨基氮浓度、菌量、溶解氧、罐温、pH. 现在对过程进行监测和故障诊断, 具体步骤如下:

1) 对于正常工况下某批次的历史数据, 因为数据的采集和传送会受到各种随机扰动和噪声的影响, 数据有可能存在失真现象, 所以对数据需要进行量化还原、大误差剔除等预处理. 预处理后的数据构成数据矩阵 $X_{100\times 9}$, 即上述的 9 个测量变量, 共 100 组采样值; 然后对数据矩阵 $X_{100\times 9}$ 作标准化处理, 得 $X'_{100\times 9}$.

2) 求 $X'_{100\times 9}$ 的协方差矩阵 $X'^{\mathrm{T}}_{100\times 9}X'_{100\times 9}/(100-1)$, 并对其进行特征向量分析, 得特征值 λ_i ($i=1,2,\cdots,9$) 和与之对应的标准特征向量矩阵 $P_{9\times 9}$.

3) 利用 CPV 和复相关系数(式(5)) 确定主元个数, 计算结果如表 1 所示. 表 1 中, 当 CPV = 87.78% > 85% 时, 主元数为 3, 表明 3 个主元足以满足 PCA 中的信息量. 为进一步确定取 3 个主元数的有效性, 再考察复相关系数, 在表 2 中, 当主元数 ≥ 3 时, 复相关系数的平均值超过了 0.9, 说明主元阵已拥有 90% 的信息. 尽管可取更多的主元数, 而且 CPV 和复相关系数显然也将更大, 但数据的简化程度不是所要求的最优主元数.

4) 在确定主元数为 3 后, 得到 λ_i ($i=1,2,3$) 相对应的主元负荷向量矩阵 $\hat{P}_{9\times 3}$.

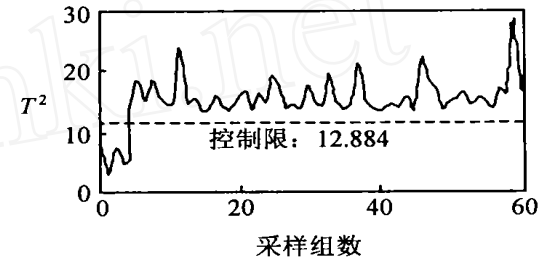
5) 根据式(3) 和式(4) 不等式的右边, 计算出 T^2 统计量和 Q 的控制限分别为 12.884, 11.406, 此时 $\alpha=99\%$.

6) 在本例中取复相关系数 $\gamma_i>0.85$ 的过程变量作为主元显著相关的过程变量, 即 PV 变量. 由表 1 得 PV 变量为: 相对效价、还原糖浓度、菌量、pH, 其余为一般变量, 即 CV 变量. 根据式(8) 和(9) 得 PVR 检验控制限为 4.801, CVR 为 6.605. 它们之和正好等于上面的 Q 控制限, 说明计算结果正确.

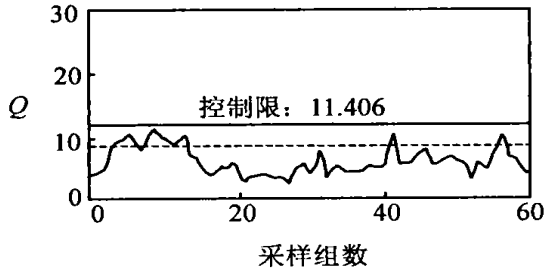
表 1　累积方差贡献率和复相关系数表

主元数	λ_i	$\sum_{i=1}^{k}\lambda_i$	$\sum_{i=1}^{k}\lambda_i\Big/\sum_{i=1}^{i=m}\lambda_i$/%
1	4.287 6	4.287 6	42.87
2	3.120 4	7.408 0	74.08
3	1.370 1	8.778 1	87.78
4	0.591 2	9.369 3	93.69
⋮			
8	0.000 1	10.000 0	100.00

7) 首先改变过程工况, 将补料速率适当增大. 取该工况下的 60 组采样数据, 根据文献[7] 方法以及式(3) 和式(4) 不等式左边, 计算出每组采样数据的 Q 和 T^2 统计量, 得到图 1 所示的基于特征子空间的系统性能监控方法的 Q 和 T^2 检验结果.



(a)　T^2 统计检验



(b)　Q 统计检验

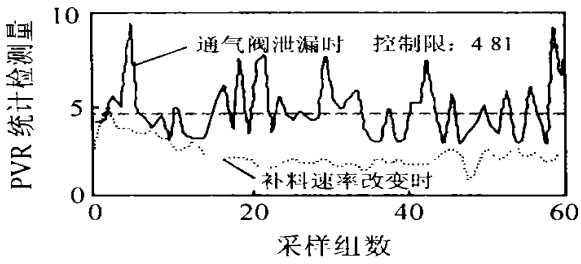
图 1　补料速率改变时文献[7] 的Q 和 T^2 检验结果

可见从 T^2 图检测出过程发生了变化, 而 Q 图没有变化, 这与实际情况相符. 但仅根据图 1 难以判断过程的变化是由工况改变引起的, 还是由过程故障引起的. 因为如果故障发生在与主元关系不紧密的过程变量(即 CV 变量) 上, 也可能出现图 1 的检测结果. 故现在用 PVR 和 CVR 统计代替 Q 统计, 判断过程的变化究竟是由工况改变引起的, 还是过程故障引起的.

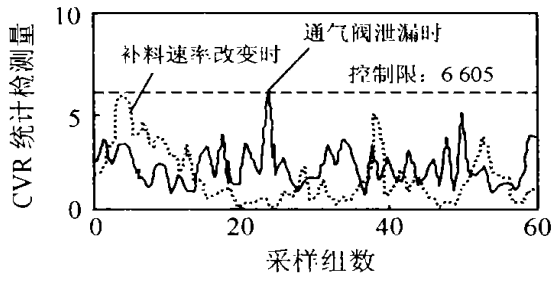
表 2　复相关系数值

	发酵周期	相对效价	总糖浓度	还原糖浓度	氨基氮浓度	菌量	溶解氧	罐温	pH
$\gamma_i(k=3)$	0.826	0.974	0.848	0.887	0.772	0.939	0.843	0.758	0.956

8) 根据式(6) 和式(7) 计算出每组采样数据的PVR 和CVR 统计量, 其检测结果如图 2 所示



(a) 主元相关过程变量残差统计检测



(b) 一般过程变量残差统计检测

图 2 补料速率改变、通气阀泄漏时改进 PCA 的检测结果

PVR 图检测结果进一步确定了PV 变量的残差未发生变化, 而CVR 图的检测结果则表明CV 变量也未发生变化, 所以可确定是正常工况变化引起的T^2 检测图发生变化, 排除了是过程故障的可能性

9) 现在进一步检测改进 PCA 的故障识别能力. 假设在某段时间通气空压机的阀门发生了泄漏, 这对 PV 中的变量相对效价、菌量和溶氧有直接影响. 文献[7] 方法的T^2 图检测出过程发生了变化, 但Q 图仍没有变化(检测结果图略). 出现这种情况的原因是Q_α控制限包括所有的误差信息, 即等于PVR 和 CVR 控制限之和, 具有较大保守性. 当 PV 变量发生的变化较小, CV 变量的变化又不明显时, 二者之和有可能不会超过Q_α, 这样Q_α将掩盖 PV 变量发生的变化

从图 2 的通气阀泄漏时改进 PCA 的检测结果可见, PV 变量确实发生了变化. 由此可认为PCA 模型描述的过程变量已发生了变化, 过程中存在故障.

5 结 论

本文针对发酵过程难以建立精确机理模型的特点, 利用 PCA 建模, 结合模型对粘菌素发酵过程进行监测和故障诊断, 采用新的统计量代替Q 统计量, 避免了其保守性, 并采用了累积方差贡献率结合复相关系数确定主元数. 仿真表明, 改进的PCA 对过程故障具有准确的识别能力.

参考文献(References)

[1] Nomikos P, MacGregor J F. Monitoring batch process using multi-way principle component analysis[J]. *J of American Institute Chemical Engineer*, 1994, 40 (8): 1361-1369.

[2] Kourti T, Lee J, MacGregor J F. Analysis, monitoring and fault diagnosis of batch process using multi-block and multi-way PLS[J]. *J of Process Control*, 1995, 5 (4): 277-283.

[3] 张杰, 阳宪惠. *多变量统计过程控制*[M]. 北京: 化学工业出版社, 2000: 44-76.

[4] 孙文爽, 陈兰祥. *多元统计分析*[M]. 北京: 高等教育出版社, 1994: 35-82.

[5] 李元, 谢植, 王纲. 基于故障重构的PCA 模型主元数的确定[J]. *东北大学学报(自然科学版)*, 2004, 25(1): 20-23.
(Li Y, Xie Z, Wang G. Determination of principal components in PCA model on basis of fault reconstruction [J]. *J of Northeastern University (Natural Science)*, 2004, 25(1): 20-23.)

[6] MacGregor J F, Jaeckle C, Kiparissides C. et al. Process monitoring and diagnosis by multiblock PLS methods [J]. *J of American Institute Chemical Engineer*, 1994, 40(5): 826-838.

[7] 郭明, 王树青. 基于特征值空间的系统性能监控与工况识别[J]. *化工学报*, 2004, 55(1): 151-154.
(Guo M, Wang S Q. System performance monitoring and region identification based on characteristic subspace[J]. *J of Chemical Industry and Engineering*, 2004, 55(1): 151-154.)

(上接第570 页)

[11] Liu J S, Chen R. Sequential monte carlo methods for dynamic systems [J]. *J of the American Statistical Association*, 1998, 93(443): 1032-1044.

[12] Doucet A, Godsill S J, Andrieu C. On sequential simulation-based methods for Bayesian filtering [J]. *Statistics and Computing*, 2000, 10(3): 197-208.

[13] Andrieu C, de Freitas N, Doucet A, et al. An introduction to MCMC for machine learning [J]. *Machine Learning*, 2003, 5(1/2): 5-43.