一种基于信噪比确定主元个数的方法①

项亚南, 潘 丰

(江南大学 轻工过程先进控制教育部重点实验室, 无锡 214122)

摘 要: 多段主元分析(MPCA)是针对间歇进行故障诊断一种行之有效的方法. 在 MPCA 中主元个数的确定是模型的关键, 关系到主元模型的可靠性、准确性、完备性. 传统的累积方差贡献率(CPV)方法确定主元个数主观性较大并且没有考虑故障因素. 为了提高检测性能, 有效的提取主元, 文中提出一种信噪比(SNR)与 MPCA 相结合选取间歇过程主元个数的方法, SNR 表明的是故障诊断的灵敏度和主元个数的影响关系, 在确保主元信息充分描述数据的基础上, 该方法考虑了故障的信息对主元个数的影响来选取主元. 将此方法应用于青霉素间歇发酵过程故障诊断中, 仿真结果表明 T^2 统计量和 SPE 统计量的响应曲线对故障更加敏感, 有效地提高了故障诊断的准确率.

关键词: 多段主元分析; 故障诊断; 累计方差贡献率; 信噪比; 青霉素间歇发酵过程

Method Based on the Fault Signal Noise Ratio to Determine the Number of Principal Component

XIANG Ya-Nan, PAN Feng

(Key Laboratory of Advanced Process Control for Light Industry, Jiangnan University, Wuxi 214122, China)

Abstract: Multi-way principal component analysis (MPCA) is an effective method for fault diagnosis in batch processes. In MPCA, the determination of principal component numbers(PCs) is the key to the model, which concerns the reliability, accuracy, completeness of PCA model. The traditional method, using CPV to determine PCs, is too subjective and does not consider the failure factors. In order to improve the detection performance, and effectively extract principal component, this paper proposes a method that is combing SNR with MPCA to select PCs in batch process, SNR indicates the relationship between the sensitivity of fault diagnosis and PCs. On the basis that the principal information fully describes the data ,and considering the influence of fault information on CPs, then it selects principal component. Applying this method to fault diagnosis in penicillin batch fermentation process, the simulation results show that the response curves of T^2 statistics and SPE statistics are more sensitive to fault, which effectively improves the accuracy rate of fault diagnosis.

Key words: multi-way principal component analysis; fault diagnosis; CPV; SNR; penicillin batch fermentation process

主元分析(PCA)是一种数据降维和特征提取的方法, 通过把高维的数据空间投影到低维的数据空间, 得到描述数据的主要的表达部分. 主元个数关系到主元模型描述原数据的信息量程度, 恰到好处的选择主元个数直接关系到故障监控的性能[1], 主元选取的太少, 无法表征数据特点, 主元选取的太多, 又会将测量的噪声包含进去, 不利于诊断.

间歇过程与连续过程不同, 间歇过程无稳定的工作点, 一般是由一个稳定的状态转化到另一个稳定状态, 有些或者根本就不存在稳态的工作点, 青霉素发酵过程就是一个典型的间歇过程[2]. PCA 处理的是二维数据矩阵, 而多批次发酵过程构成的是三维数据. 20 世纪 90 年代 Nomikos 和 MacGregor[3]首次提出了一种多段的主元分析(Mutil-way Principal Component Analysis; MPCA)

① 基金项目: 国家自然科学基金(61273131);江苏省产学研联合创新资金(BY2013015-39)
收稿时间:2014-06-04;收到修改稿时间:2014-06-30

模型来监控间歇过程. 文献[4,5]提出了一种基于 MPCA 采样时刻展开的方法, 把传统的 MPCA 进行分段, 用累计方差贡献率(cumulative percent variance;CPV)的方法来提取主元来做故障诊断, 但是没有考虑故障信息对主元个数的影响. 文献[6]将故障诊断率结合复相关系数来选取主元, 提高了诊断的准确率, 但是计算各个过程变量与主元之间的复相关系数比较复杂, 计算量比较大.

针对 CPV 方法没有充分考虑故障监控的性能及故障信息对主元个数的影响的不足. 本文提出一种基于信噪比(Signal Noise Ratio;SNR)的方法来确定间歇过程的主元个数, 这种方法考虑故障类型对主元个数的影响来提取主元个数. 基于青霉素发酵 Pensim2.0 软件仿真平台, 应用于青霉素间歇发酵过程故障诊断中, 分别用该方法和 CPV 方法选取主元, 并对比它们故障诊断的监控效果.

1 多段主元分析方法

一个典型的间歇过程, 过程数据描述为 $X=I\times J\times K$, 其中 I 表示批次, J 表示变量个数, K 表示采样时刻序列, 由于发酵过程高度的非线性、复杂的动态性能和时段过程特性, 如果只用一个主元MPCA来进行监控整个的批次, 会使得监控过程变得不准确. 多阶段主元分析的方法考虑对 X 按照一个轴向展开, 进行合理的分段[7], 把三维的数据空间分解为二维数据空间, 这样就使得过程变得更加细化.

对三维矩阵 X 按照采样时刻进行分解, 分解的思路如图1所示, $X=I\times J\times K$ 的PCA分解如下式: $X=I\times KJ=TP^T+E$, $P(KJ\times R)$ 是负载矩阵, $T(I\times R)$是得分矩阵, $E(I\times KJ)$是残差子空间, 反映的是过程噪声和干扰, R 是主元数目.

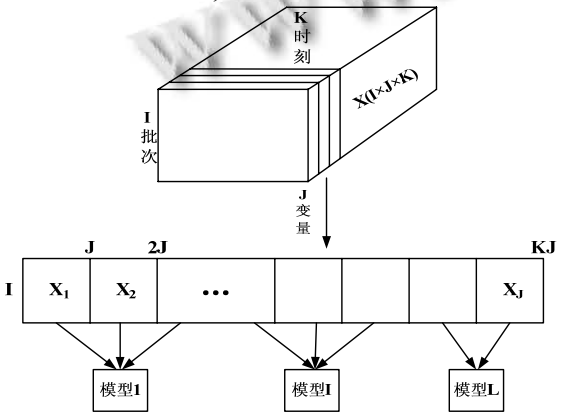


图 1 基于 MPCA 的模型采样时刻展开示意图

2 主元个数选择的方法

PCA 的作用是将数据降维, 以排除多重信息的叠加影响, 实现在主元 $i(i\geq\alpha)$ 中选取 α 个相关的变量来概括 i 个变量的绝大部分信息.

2.1 CPV 确定主元的个数

对 $X\in R^{n\times m}$ 标准化处理, 得到均值为0方差为1的标准化矩阵, $X_i=[x_{i1},x_{i2},\cdots,x_{in}]^T\ i=1,2,\cdots,m$, 其中 m 代表包含了 m 个传感器的测量样本, n 为每个传感器各有 n 个独立的采样数据. 标准化后的样本的协方差矩阵表示为:

$$\mathrm{cov}(x)=\frac{1}{n-1}X_i^TX_i \tag{1}$$

对 $\mathrm{cov}(x)$ 正交分解, 分解为 $\mathrm{cov}(x)=P\Lambda P^T$, $\Lambda=diag(\lambda_1,\lambda_2,\cdots,\lambda_m)$, Λ 为 $\mathrm{cov}(x)$ m 个特征值按照降序排列($\lambda_1>\lambda_2>\cdots>\lambda_m$)组成的对角矩阵 $P=[P_1,P_2,\cdots,\ P_m]$是 $\mathrm{cov}(x)$ 的特征值对应的特征向量组成的矩阵.

特征值 λ_i 的大小反应主元成分的解释程度, 较大的特征值描述了系统状态的主要变化特征, 即前 α 个主元解释的数据占整个数据的比例, 主元个数决定了负载矩阵 $P\in R^{\mathrm{m}\times\alpha}$ 的维数, 据此再得到得分矩阵 $T=XP$, 较小特征值对应部分描述了系统的随机噪声.

主元累计方差贡献率描述为:

$$CPV=\sum_{\mathrm{i}=1}^{\alpha}\lambda_{\mathrm{i}}\Big/\sum_{\mathrm{i}=1}^{\mathrm{m}}\lambda_{\mathrm{i}}\geq\beta \tag{2}$$

β 是设定值, 一般选择 85%及以上, 当 i 取 1, 2, 3…α 到满足(2)式时, 此时的选取的主元个数即为 α.

2.2 SNR 与 MPCA 结合确定主元的个数

假设故障信号记为 $A\xi_j$, 故障信号输出描述为 $x_{i,j}=\tilde{x}_{i,j}+A\xi_j$, $\tilde{x}_{i,j}$ 表示没有发生故障时标准化后的正常数据, A 表示故障的大小, ξ_j 表示的是故障方向. SPE 指标表征某个时刻数据在残差空间偏离主元模型的程度, 定义为:

$$SPE=\left\|x_{i,j}{}^T-\tilde{x}_{i,j}{}^T\right\|^2 \tag{3}$$

其中 $x_{i,j}$, $\tilde{x}_{i,j}$ 分别表示 i 时刻, 第 j 个变量的观测值和模型的预测值, 在发生故障时, SPE 统计量还可以描述为:

$$SPE=\left\|x_{i,j}{}^T-x_{i,j}{}^TP_\alpha P_\alpha{}^T\right\|^2$$

$$= \left\| \left(\tilde{x}_{i,j} + A\xi_j\right)^T - \left(\tilde{x}_{i,j} + A\xi_j\right)^T P_\alpha P_\alpha^T \right\|^2 \quad (4)$$

其中 $P(j \times \alpha)$, 令 $P_\alpha^* = E_{j,j} - P_\alpha P_\alpha^T$, 因为 $\tilde{x}$ 是标准化后的数据, 因此 $\tilde{x} = 0$, $A = 1$, 得到 $SPE = \left\| \xi^T P_\alpha^* \right\|^2$, 所以 SPE 只与主元的个数和故障的方向有关.

SPE 统计量下的信噪比的定义为:

$$SNR_{SPE} = SPE / SPE_\alpha = \frac{\left\| \xi^T P_\alpha^* \right\|^2}{SPE_\alpha} \quad (5)$$

多元数据模型的基础是依赖过程建立的一个正常工况的数据库[8], 其中 SPE_α 是反应数据正常的控制限.

统计量 $SPE_\alpha = \left(\nu / 2\mathrm{m}\right)\chi^2_{2\mathrm{m}^2/\nu,\ \theta}$ m 、ν 分别是每个采样时刻 SPE 样本的均值和方差, $\chi^2_{2\mathrm{m}^2/\nu,\ \theta}$ 是检验水平为 θ 、自由度为 $2\mathrm{m}^2/\nu,\ \theta$ 的卡方分布的临界值. P_α^* 与主元的个数有关的.

T^2 统计量是由主元空间建立的, 表示的含义是主元模型描述主元数据的程度, 它的表达式为:

$$T^2 = \sum_{i=1}^{K} \left(\frac{t_{i,j}}{\sigma_j} \right) = x_{i,j}{}^T P \lambda^{-1} P^T x_{i,j}{}^T \quad (6)$$

$t_{i,j}$ 是第 i 个样本在模型中 j 个隐含变量的主成分, σ_j 是 j 隐含变量主模型成分的标准差. T^2 统计量的信噪比(SNR_{T^2})的表达式为:

$$SNR_{T^2} = T^2 / T_\alpha^2 = \frac{\left\| \xi_j{}^T P \lambda^{-1} P \xi_j \right\|^2}{T_\alpha^2} \quad (7)$$

其中 T^2 的控制限为 $T_\alpha^2 = \frac{\alpha\left(n^2-1\right)}{n-\alpha} F_{\alpha,n-\alpha,\theta}$, n 是样本的个数, $F_{\alpha,n-\alpha,\theta}$ 是检验水平为 θ 、带有自由度为 α 和 $n-\alpha$, F 分布的临界值. 易知分子也是一个与主元个数相关的式子.

SPE 和 T^2 统计量的监控性能往往是不同的, 也就是说, 同一个主元个数很难同时满足两个监控指标的要求. 由上述的表达式可知, 定义信噪比时把 SPE 和 T^2 分开来定义, 使得主元个数的选择更加合理. 分子是故障信息, 分母是正常工况下的信息, 分子看作是故障信号, 分母看作是故障的噪声. 只需计算统计量的比值大小关系, 计算比较简便.

SNR 由主元个数和故障方向决定, 它的大小反映出了模型对故障的灵敏度. 在决定主元个数之前首先确定故障的方向, 如果故障发生在第 i 个变量, 则故障的方向定义为 1, 其他方向定义为 0, 即 $\xi = [0,\cdots,0,1,0,\cdots,0]$. 当信噪比的值取最大值时, 即对应为最佳的主元个数.

3　青霉素间歇发酵过程建模与监控

3.1 应用环境和过程的建模

本文结合应用青霉素发酵 Pensim2.0 软件仿真平台[9]模拟现场的发酵过程. 如图 2 所示是仿真平台应用器, 包括发酵罐、搅拌电机、温度控制、空气流量、pH 控制等必备组成部分. 先建立一个正常的数据库, 选取主元个数建立模型, 计算 T^2 和 SPE 两个统计量的控制限.

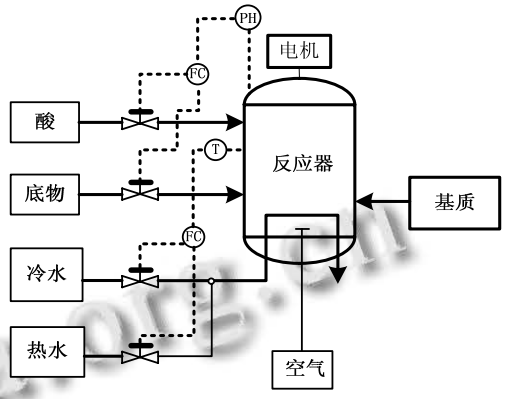


图 2　青霉素发酵流程图

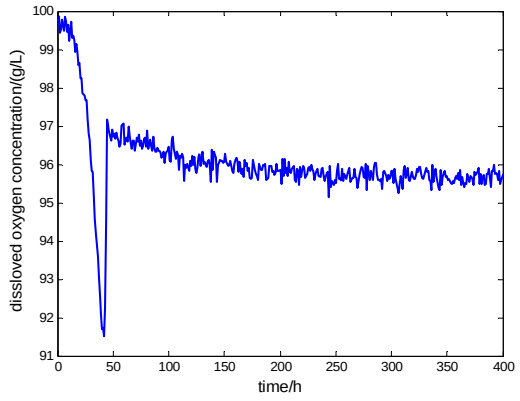
根据生产工艺和各个变量在生产过程中的影响程度, 从青霉素发酵过程中的 18 个变量中选取 10 个主要的过程变量来建模并监控. 选取的过程变量分别是空气流量、搅拌功率、底物流加速度、反应器温度, PH 值、O_2 浓度、反应体积、CO_2 浓度、反应产生的热量以及冷却水速, 各过程变量如表 1 所示.

每个批次历时 400h, 每小时采集一次数据(即采样 400 次/批), 大量的实验表明青霉素发酵可以分成四个反应过程, 青霉素的发酵是个好氧过程, pH 与底物浓度和菌体相关, 罐体的温度反映发酵过程中菌体的活性程度, 所以可以通过溶氧的浓度和 pH 值以及罐体的温度对批次进行分段.

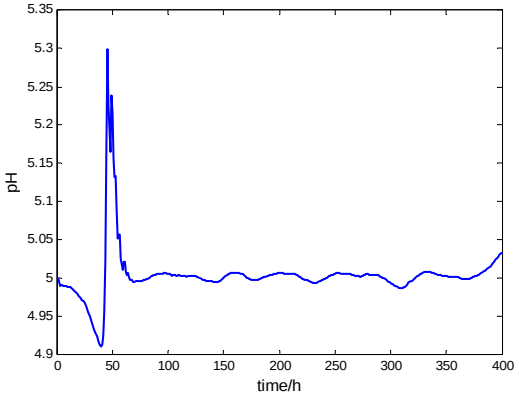
表 1 青霉素发酵中的过程变量

编号	过程变量	编号	过程变量
1	空气流量(L/h)	6	O_2 浓度(%)
2	搅拌功率(W)	7	反应体积(L)
3	底物流加速度(L/h)	8	CO2 浓度(m mol/L)
4	反应器温度(K)	9	反应产生的热量(Kcal)
5	pH 值	10	冷却水速(L/h)

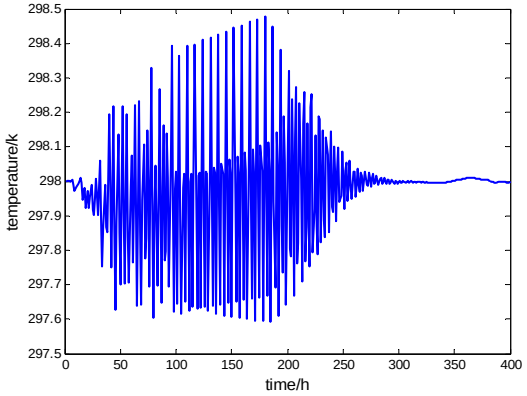
选取 30 组正常批次过程的数据, 建立的 MPCA 三维数据集合为 $X(30\times10\times400)$, 再将三维的模型按照采样时刻分解展开得到 400 个时间片矩阵 $X_1(30\times10)$, $X_2(30\times10)$, …, $X_{400}(30\times10)$. 分别用基于 CPV 和基于 SNR 的方法来提取主元个数, 比较两者的在线监控图, 对比各自的检测性能. 图 2 中的(a,b,c)是青霉素发酵反应过程中的 3 条曲线, 采用趋势分析的方法分段[10], 把发酵过程分成以下部分: 0~40h(a,b 可看出), 菌体的调整期; 40~100h(a,b,c), 菌体的指数增长期;100~230h(c), 青霉素的合成期; 230~400h 是衰亡期(c).



(a) 溶氧的过程曲线



(b) pH 值的过程曲线



(c) 温度的过程曲线

图 3 控制过程曲线图

3.2 主元的选取

Pensim2.0 仿真可设置三种类型的故障, 分别是底物流加速率、空气流量以及搅拌功率, 三个变量都可以设置为阶跃扰动或斜坡扰动, 而且故障的引入时间到终止的时间, 故障的幅值都是可以进行设置的.

指数的增长时期(100~230h)是菌体最活跃的时期, 文中假设在该时段发生底物流加阶跃故障. 分别用 CPV 方法和 SNR 方法确定这段时间的主元个数, 分别见公式(2), 和公式(5)、(7), 仿真结果分别如图 4、图 5 所示.

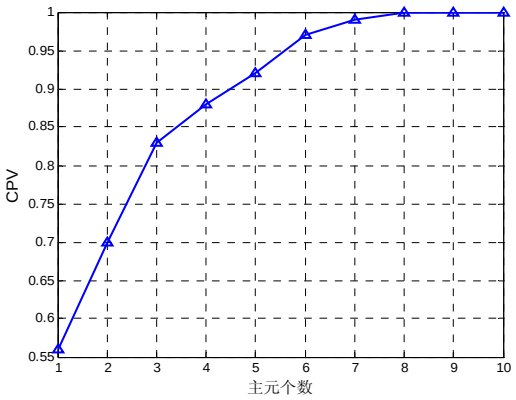
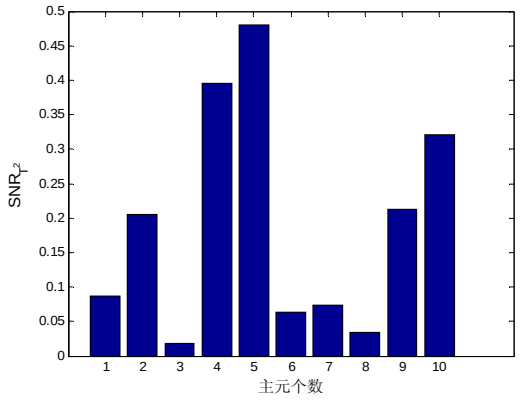


图 4 CPV 方法选取主元

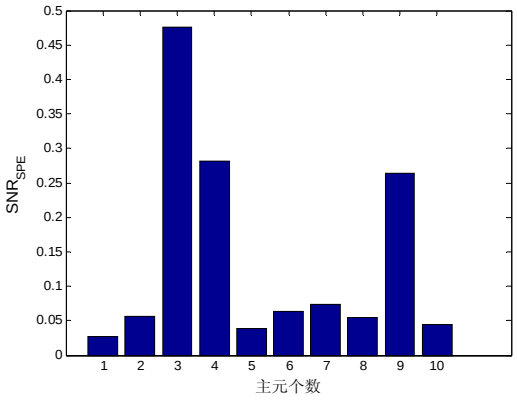
基于 CPV 确定主元个数如图 4 所示, 当主元个数取 4 个时 CPV 的值超过 85%, 所以, 基于 CPV 方法选取的主元个数为 4.

基于 SNR 方法选取主元个数如图 5 所示, 图 5(a)直方图表示统计量 T^2 不同的主元个数对应的信噪比大小, 图 5(b)直方图表示统计量 SPE 不同的主元个数对

应的信噪比大小. 直方图值越大表示灵敏度越高, 由图 5 知 T^2 统计量选取主元个数为 5 时, T^2 灵敏度最高, 则 T^2 统计量的主元个数选为 5, SPE 统计量中当主元的个数取 3 时, SPE 统计量的主元个数选为了 3, 此时 SPE 统计量的灵敏度最高.



(a) T^2 统计量信噪比直方图



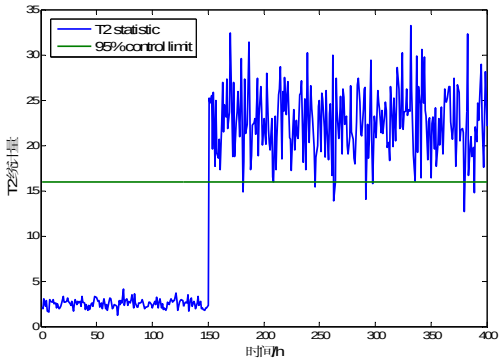
(b) SPE 统计量信噪比直方图

图 5 SNR 方法选取主元

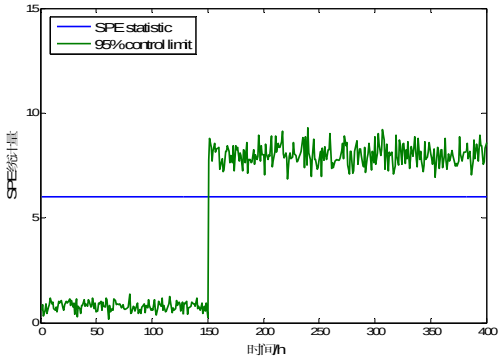
3.3 故障的监控

为了检验采用故障信噪比选取主元个数方法的检测性能, 在新批次的 150h 时加入一个底物流加速度为+5%的阶跃扰动故障直到发酵结束. 在 MPCA 监控中, 必须要对未完成的批次进行预估, 本文采用 Nomikos 和 MacGregor[7]提出的第二种方法, 即预估数据采用当前变化法来填补, 首先假设未完成的数据和参考数据的均值和偏差保持一致, 用当前的数据来预测未完成的数据.

采用 CPV 和 SNR 选取主元个数分别得到各自的 T^2 统计量和 SPE 统计量的监控图, 仿真结果如图 6、图 7 所示.

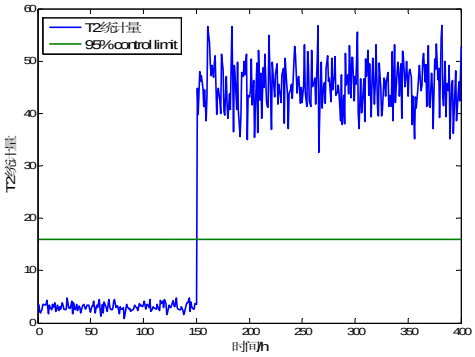


(a) T^2 统计曲线

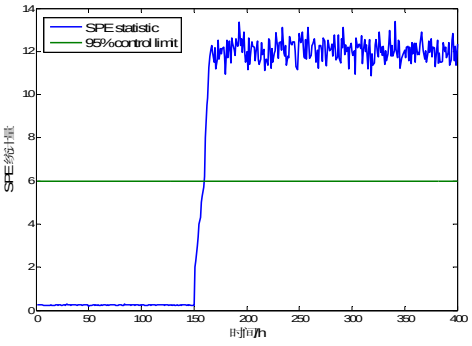


(b) SPE 统计曲线

图 6 基于 CPV 方法 T^2 和 SPE 的统计量监控图



(a) T^2 统计曲线



(b) SPE 统计曲线

图 7 基于 SNR 方法 T2 和 SPE 的统计量监控

从图 6 和图 7 可以看出 SNR 方法相对 CPV 方法的优越性, 图 6(a)中 T^2 统计量在延时 10h, 即在 160h 时刻左右检测到了故障, 但在 T^2 统计量以后的监控中出现了较多的虚报处. 图 7(a), 基于 SNR 的方法也在 160h 左右的时候检测出了故障, 并且在以后数据的监控没有出现误报, 故障诊断准确性有很大的提高. SPE 图 7(b)较图 6(b), 改进后的统计量的监控性能相比之前方法的监控性能有更高的灵敏度. 底物是菌体生长和产物合成所必须的物质, 它的变化造成的影响是间接的, 通过培养基的体积、氧气含量、CO_2 的浓度等表现出来, 所以故障发生是有延时的, 与实际情况相符.

4 结语

MPCA 的方法针对间歇反应机理, 通过合理的分段建模, 分段处理逼近非线性来提高监控性能. CPV 确定主元的方法因为没有考虑故障因素, 主元个数的选定没有针对性. SNR 确定主元的方法考虑了故障的方向, 并且分开定义 SPE 和 T^2 统计量的信噪比来选取主元个数, 同时兼顾 T^2 统计量、SPE 统计量的检测性能, 这种主元选取的方法更加符合过程特征, 使得主元的选择更加合理. 仿真结果表明基于 SNR 选取主元的方法比基于 CPV 选取主元的方法有效地降低了故障诊断的误报率, 具有更好的优越性. 应用在青霉素间歇发酵过程中具有很好的指导意义.

参考文献

1 张佳,孙巍,赵劲松,孙美红.多段 MPCA 法监测间歇过程的故障.计算机与应用化学,2010,27(3):298–302.

2 赵娟平,陈健,姜长洪.青霉素发酵过程建模研究.计算机仿真,2008,25(2):80–82.

3 Nomikos P, MacGregor JF. Monitoring batch process using mutil-way principle component analysis. American Institute of Chemical Engineers, 1994, 40(8): 1368–1375.

4 Choi SW, Morris J, Lee IB. Dynamic model-based batch process monitoring. Chemical Engineering Science, 2008, 63(3): 622–636.

5 肖应旺.改进的 MPCA 批过程在线监测方法.控制工程, 2011,18(2):299–303.

6 张新荣,熊伟丽,徐保国.基于 PCA 的发酵过程监控模型主元数的确定.计算机测量与控制,2009,17(6): 1120–1122, 1131.

7 王姝.基于数据的间歇过程故障诊断及预测方法研究[学位论文].哈尔滨:东北大学,2010.

8 Nomikos P, Mac Gregor JF. Muhivariate SPC charts for monitoring batch processes. Technometries, 1995, 37(1): 41–59.

9 刘毅,王海清.Pensim 仿真平台在青霉素发酵过程的应用研究.系统仿真报,2006,18(12):3524–3527.

10 刘敏华,萧德云.基于趋势分析和 SDG 模型的故障诊断.控制理论与应用,2006,23(2):306–310.