

基于二维主元分析的间歇过程故障诊断

孔晓光*, 郭金玉, 林爱军

(沈阳化工大学 信息工程学院, 沈阳 110142)

(* 通信作者电子邮箱 cxz119@sohu.com)

摘要:传统的多向主元分析(MPCA)已广泛应用于监视多变量间歇过程。在MPCA算法中,三维的间歇过程数据需要转换为高维的二维向量,导致计算量和存储空间大,同时不可避免地丢失一些重要信息。因此,提出一种新的基于二维主元分析(2DPCA)的故障诊断方法。由于每个批次的间歇过程数据是一个二维向量(矩阵),应用以各个批次矩阵为分析对象的2DPCA算法,避免矢量化,存储空间和存储需求小;另外,2DPCA采用各个批次的协方差的平均值来进行建模,能够更加准确地反映出不同类型的故障,在一定程度上增强了故障诊断的准确性。半导体工业实例的监视结果说明,2DPCA方法优于MPCA。

关键词:间歇过程;故障诊断;主元分析;多向主元分析;二维主元分析

中图分类号: TP181; TP277.3 **文献标志码:** A

Fault diagnosis for batch processes based on two-dimensional principal component analysis

KONG Xiaoguang*, GUO Jinyu, LIN Aijun

(College of Information Engineering, Shenyang University of Chemical Technology, Shenyang Liaoning 110142, China)

Abstract: Multiway Principal Component Analysis (MPCA) has been widely used to monitor multivariate batch process. In MPCA method, the batch data are transformed as a vector in high-dimensional space, resulting in large computation, storage space and loss of important information inevitably. A new batch process fault diagnosis method based on the two-Dimensional Principal Component Analysis (2DPCA) was presented. Essentially, every batch data was presented as a second order vector, or a matrix. In this case, 2DPCA could be used to deal with the two-dimensional batch data matrix directly instead of performing vectorizing procedure with low memory and storage requirements. In addition, 2DPCA was used to model with the covariance average of all the batches, which accurately reflected the different faults and enhanced the accuracy of fault diagnosis to a certain extent. The monitoring results of an industrial example show that the 2DPCA method outperforms the conventional MPCA.

Key words: batch process; fault diagnosis; Principal Component Analysis (PCA); Multiway Principal Component Analysis (MPCA); two-Dimensional Principal Component Analysis (2DPCA)

0 引言

间歇过程具有变量繁多、变量之间相互关联、过程变化剧烈且存在有限操作周期等特点。多元统计分析是间歇过程进行监控和故障诊断的一种有效方法。Nomikos等^[1]于1995年曾经提出用多向主元分析法(Multiway Principal Component Analysis, MPCA)解决间歇过程的离线和在线监控的问题,这种方法随后得到了很大发展。

现有的间歇过程诊断方法多基于MPCA^[2-3]以及多向偏最小二乘法(Multiway Partial Least Squares, MPLS)^[4]。普通的MPCA针对整个批次建立模型应用于故障监测,但是一方面,由于实际间歇生产过程可以根据操作或反应现象的不同而分成多个阶段,每个阶段的数据结构不同,这时用单一的模型不能表现原始数据的信息。针对这个缺点,Undey等^[5]提出了将间歇过程按照不同操作分成多阶段的MPCA方法,在每个阶段分别建立MPCA模型。另一方面,MPCA是一种线性方法,MPCA提出的主元是原始变量的线性组合,但实际间歇过程的变量往往是非线性的,如果将普通的MPCA方法应用于非线性问题,可能导致一些重要信息的缺失。针对

MPCA这个缺点,有研究者提出了非线性的MPCA方法^[6-12],但这种方法比较复杂,并且难以处理有大量变量的过程,因此,Camacho等^[13]提出了multi-phase PCA方法,以未被主元所表示的残差的平方和为基准将间歇过程分成多个阶段,每个阶段的非线性关系都能用局部线性建模方法来表示,但此方法计算复杂,且方法中用到的诸多参数的选择缺乏统一的标准。

上面的方法在建模过程中都需要将三维数据转换为二维数据,当数据量大、维数高时,难免会有少量有用信息的丢失。因此,本文提出一种新的故障检测方法,利用二维主元分析(2-Dimensional Principal Component Analysis, 2DPCA)算法^[14]对三维数据进行标准化,再取其协方差的平均值建立模型,并对新数据进行检测。经过仿真实验,证实了2DPCA方法明显优于传统的MPCA方法。

1 二维主元分析算法

2DPCA算法是一种数据特征提取方法,与PCA算法的不同之处在于,2DPCA算法是先以各个批次矩阵为分析对象建立模型,而PCA算法是以各个批次经过转换成的二维矩阵为

收稿日期:2012-08-14;修回日期:2012-10-26。 基金项目:国家自然科学基金资助项目(61174119)。

作者简介:孔晓光(1974-),女,辽宁沈阳人,讲师,博士,主要研究方向:生产过程性能监视、故障诊断;郭金玉(1975-),女,山东高唐人,副教授,博士,主要研究方向:故障诊断、生物特征识别;林爱军(1968-),女,辽宁沈阳人,助教,主要研究方向:检测技术、故障诊断。

分析对象。

1.1 2DPCA 算法的基本原理

批过程体现为重复性的生产过程,其数据集合比连续生产过程数据集合多一维“批量”元素,具有序贯特性,因此可用三维数据矩阵 $X(N, J, K)$ 代表批过程数据集合,其中: N 为批次数, J 为变量数, K 为采样点数。

与 MPCA 算法不同, 2DPCA 方法是分别对过程的各个批次进行标准化, 求出其协方差, 然后求所有批次协方差的平均值, 最后对协方差的平均值进行主元分析, 具体建模步骤如下:

1) 选取 N 个成功历史批次数据, 每一批次数据组成矩阵 $X_i(K, J)$ 。

2) 对每一个 X_i 矩阵进行标准化的计算。

$$\bar{X}_i = [X_i - (1, 1, \dots, 1)^T M] \text{diag}\left(\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_K}\right) \quad (1)$$

其中: $M = [m_1, m_2, \dots, m_K]$ 为变量的均值, $s = [s_1, s_2, \dots, s_K]$ 为变量的标准差。

3) 将 N 个标准化后的矩阵分别求取协方差, 然后求取平均协方差矩阵。

$$\bar{A}_i = \bar{X}_i^T * \bar{X}_i / (K - 1) \quad (2)$$

$$\bar{A} = \left(\sum_{i=1}^N \bar{A}_i \right) / N \quad (3)$$

4) 运用 PCA 方法对 \bar{A} 进行主元分析。

矩阵 \bar{A} 可以分解为 J 个向量的外积之和, 即

$$\bar{A} = t_1 p_1^T + t_2 p_2^T + \dots + t_J p_J^T \quad (4)$$

在式(4)中, $t_i \in \mathbf{R}^K$ 被称为得分向量, $p_i \in \mathbf{R}^J$ 称为负荷向量。 X 的得分向量也叫作 X 的主元。式(4)也可写为下列矩阵形式:

$$\bar{A} = T P^T \quad (5)$$

其中: $T = [t_1, t_2, \dots, t_m]$ 称为得分矩阵, $P = [p_1, p_2, \dots, p_n]$ 称为负荷矩阵。

1.2 基于 2DPCA 监视的统计量及其控制限

对于一个新的批次的被监视的观测变量 x_{new} 来说, Q 统计量可以定义为:

$$Q_i = e_i e_i^T = x_{\text{new}}^T (I - P_a P_a^T) x_{\text{new}} \quad (6)$$

其中: e_i 是 E 的第 i 行, I 是 $n \times n$ 的单位矩阵。 Q 统计量代表了数据中没有被主元模型所解释的变化。当 Q 统计量过大时, 说明过程中出现了不正常情况, 从过程正常运行数据所建立的模型已经不再适用。控制限的计算建立在一定的假设基础上, 当检验水平为 α 时, Q 控制限^[15] 可以按下式计算:

$$Q_\alpha = \theta_1 \left[\frac{C_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}} \quad (7)$$

其中:

$$\theta_i = \sum_{j=a+1}^n \lambda_j^i; i = 1, 2, 3 \quad (8)$$

$$h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2} \quad (9)$$

C_α 是正态分布在检验水平为 α 下的临界值。

2 仿真实验

MPCA 与 2DPCA 建模方法是基于一个能够代表间歇生产过程一段周期内正常操作的历史轨迹数据库。统计学模型建立后, 用该模型与未来间歇过程的时间轨迹数据相比较,

由此检测和判断新批量生产过程是否正常。

本文针对半导体工业实例——A1 堆腐蚀过程^[16] 运用 MPCA 和 2DPCA 方法进行了大量仿真实验。数据有 108 个正常硅片和 21 个故障硅片构成。由于两个批次过程数据的大量丢失, 本文应用 107 个正常硅片和 20 个故障硅片进行仿真研究。每个批次含 17 个过程变量。每个批次是不等长的, 持续时间在 95 ~ 112 s 变化。为了消除传感器中初始的波动影响, 去除开始的 5 个样本, 保留 85 个样本以适应最短的批次。随机抽取 96 个正常批次为建模数据, 其余 11 个正常批次为校验数据。以 96 个正常批次数据分别用两种方法进行建模, 并对 11 个校验批次数据和 20 个故障批次进行故障检测, 平方预测误差 (Squared Prediction Error, SPE) 检测结果如表 1 所示。两种方法对正常批次的检测结果如图 1 所示, 对故障批次的检测结果如图 2 所示, 图中虚线对应 SPE 95% 的控制限。

表 1 SPE 检测结果

算法	总校验 批次数	检测正常 批次数	正常批次 检测比例/%	总故障 批次数	故障 批次数	故障批次 检测比例/%
MPCA	11	7	64	20	16	80
2DPCA	11	8	73	20	20	100

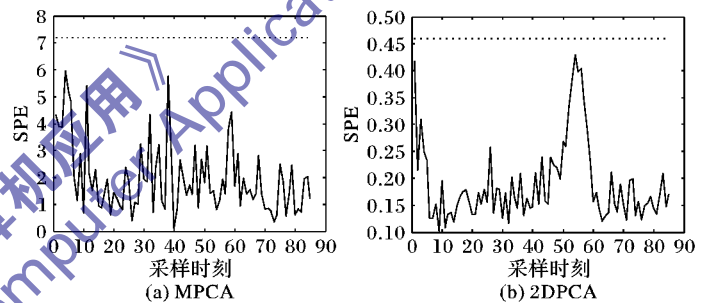


图 1 两种方法对正常第 68 批次的 SPE 检测图

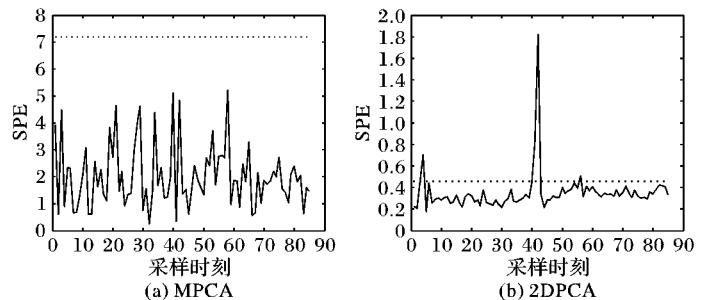


图 2 两种方法对故障第 9 批次的 SPE 检测图

从表 1 (数据监测故障表) 中显示出的检测数据比例可以明显看出, 2DPCA 对正常批次的 SPE 检测率为 73%, 而传统的 MPCA 的检测率为 64%; 2DPCA 对故障批次的 SPE 故障检测率为 100%, 而传统的 MPCA 的检测率为 80%。证明运用 2DPCA 方法进行检测的准确率远远高于传统的 MPCA 方法, 大大提高了检测故障率, 减少了故障的误报率和漏报率。

由图 1 和图 2 可以看出, 2DPCA 和 MPCA 方法对正常第 68 批次的检测结果是相同的, 而对故障第 9 批次的检测结果是不同的。MPCA 方法把故障第 9 批次检测为正常批次, 而 2DPCA 的检测结果是故障批次, 与事实相符, 验证了 2DPCA 算法的有效性。

对故障第 9 批次, 2DPCA 方法在 41 时刻的故障变量贡献图如图 3 所示, 从中可以看出变量 6 是引起故障的主要原因。

表 2 是在 Matlab 7.0 平台下利用 MPCA 和 2DPCA 方法进行故障检测所需的 CPU 时间 (CPU 为 2.00 GHz, RAM 为

512 MB)。由表2可知,与MPCA方法相比,2DPCA方法特征向量维数降低了,减少了存储空间;2DPCA方法故障检测时间低于MPCA方法。由于MPCA方法需要将96个批次的三维维数数据矩阵 $96 \times 85 \times 17$ 转换成二维矩阵 96×1445 ,再进行特征提取,增加了时间;而2DPCA直接计算二维矩阵 85×17 的协方差矩阵,再进行特征提取,减少了计算量,节省了时间。

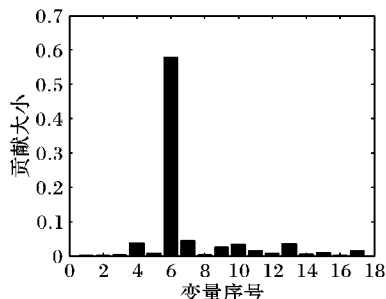


图3 2DPCA方法的变量贡献图

表2 MPCA和2DPCA故障检测对比

方法	协方差矩阵大小	特征向量大小	故障检测时间/s
MPCA	1445×1445	1445×3	0.024
2DPCA	17×17	17×3	0.010

3 结语

传统的MPCA检测故障方法可以有效地对批次生产过程中存在的故障进行检测和诊断,但是由于在实际生产过程中数据量的繁琐、庞大,导致在进行故障检测时一部分重要信息的丢失,致使检测结果不够准确,产生漏报的现象;而利用2DPCA方法可以有效解决这样的问题,使故障检测结果更加准确;同时,2DPCA方法减少了计算量和存储空间,降低了故障检测时间。

参考文献:

- [1] NOMIKOS P J F, MacGREGOR J F. Multivariate SPC charts for monitoring batch process [J]. *Technometrics*, 1995, 37(1): 41 - 59.
- [2] NOMIKOS P, MacGREGOR J F. Monitoring of batch processes using multi-way principal component analysis [J]. *AIChE*, 1994, 40(8): 1361 - 1375.
- [3] NOMIKOS P. Detection and diagnosis of abnormal batch operation based on multi-way principal component analysis [J]. *ISA Transactions*, 1996, 35(3): 259 - 266.
- [4] NOMIKOS P, MacGREGOR J F. Multi-way partial least squares in monitoring batch processes [J]. *Chemometrics and Intelligent Laboratory Systems*, 1995, 30(1): 97 - 108.
- [5] UNDEY C, CINAR A. Statistical monitoring of multistage, multi-phase batch processes [J]. *IEEE Control Systems Magazine*, 2006, 22(5): 40 - 52.
- [6] DONG D, MCAVOY T J. Nonlinear principal component analysis — based on principal curves and neural networks [J]. *Computers and Chemical Engineering*, 1996, 20(1): 65 - 78.
- [7] ZHANG Y W, QIN S J. Fault detection of nonlinear processes using multiway kernel independent component analysis [J]. *Industrial and Engineering Chemistry Research*, 2007, 46(23): 7780 - 7787.
- [8] YU J. Nonlinear bioprocess monitoring using multiway kernel localized Fisher discriminant analysis [J]. *Industrial and Engineering Chemistry Research*, 2011, 50(6): 3390 - 3402.
- [9] TIAN X M, ZHANG X L, DENG X G, *et al.* Multiway kernel independent component analysis based on feature samples for batch process monitoring [J]. *Neurocomputing*, 2009, 72(7/8/9): 1584 - 1596.
- [10] GE Z Q, SONG Z H. Batch process monitoring based on multilevel ICA-PCA [J]. *Journal of Zhejiang University: Science A: Applied Physics and Engineering*, 2008, 9(8): 1061 - 1069.
- [11] ZHANG Y W, QIN S J. Decentralized fault diagnosis of large-scale process using multiblock kernel principal component analysis [J]. *Acta Automatica Sinica*, 2010, 36(4): 593 - 597.
- [12] ZHAO C Y, GAO F R, WANG F L. Nonlinear batch process monitoring using phase-based Kernel-Independent Component Analysis-Principal Component Analysis (KICA-PCA) [J]. *Industrial and Engineering Chemistry Research*, 2009, 48(20): 9163 - 9174.
- [13] CAMACHO J, PICÓ J. Multi-phase principal component analysis for batch processes modeling [J]. *Chemometrics and Intelligent Laboratory Systems*, 2006, 81(2): 127 - 136.
- [14] YANG J, ZHANG D, FRANGI A F, *et al.* Two-dimensional PCA: a new approach to appearance-based face representation and recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(1): 131 - 137.
- [15] 张晓庭, 方开泰. 多元统计分析[M]. 北京: 科学出版社, 1999.
- [16] HE Q P, WANG J. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes [J]. *IEEE Transactions on Semiconductor Manufacturing*, 2007, 20(4): 345 - 354.
- [7] HASPEL S, de CAUSMAECKER P, STØLEVIK M, *et al.* First International nurse rostering competition 2010 [C]// *Proceedings of the 8th International Conference for the Practice and Theory of Automated Timetabling*. Belfast, Northern Ireland: Queen's University Belfast, 2010: 498 - 501.
- [8] 陆蕊君, 鲍学英, 王惠芬. 护士排班方法的改进与实践[J]. *当代护士: 综合版*, 2007(2): 23 - 23.
- [9] 王爱莲, 刘筱英. 儿科病房一日护理密集时段的调查及对护士排班的启示[J]. *中国护理管理*, 2007, 7(12): 51 - 52.
- [10] 刘晓荣, 陈国良, 顾仁萍, 等. 护士排班决策支持系统模型的研究[J]. *解放军护理杂志*, 2006, 23(2): 7 - 9, 12.
- [11] 沈吟东, 陈名晖, 邓婕. 利用矩阵向量化变换求解护士排班问题 [C]// *中国控制与决策会议论文集*. [S. l.]: ISTP, 2008: 1019 - 1022.
- [12] MLADENOVIC N, HANSEN P. Variable neighborhood search [J]. *Computers and Operations Research*, 1997, 24(11): 1097 - 1100.
- [13] 董红宇, 黄敏, 王兴伟, 等. 变邻域搜索算法综述[J]. *控制工程*, 2009, 16(S2): 1 - 5.
- [14] BURKE E K, de CAUSMAECKER P, PETROVIC S, *et al.* Variable neighborhood search for nurse rostering problems [M]// *Metaheuristics: Computer Decision-Making*. Norwell: Kluwer Academic Publishers, 2004: 153 - 172.
- [15] BURKE E K, CURTOIS T, POST G, *et al.* A hybrid heuristic ordering and variable neighbourhood search for the nurse rostering problem [J]. *European Journal of Operational Research*, 2008, 188(2): 330 - 341.
- [16] AICKELIN U, DOWSLAND K A. Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem [J]. *Journal of Scheduling*, 2000, 3(3): 139 - 153.

(上接第341页)