

青霉素发酵过程的2DPCA建模及监测*

梁小凡^a, 熊伟丽^b, 杨薇薇^a, 徐保国^b

(江南大学 a. 物联网工程学院 自动化系; b. 轻工过程先进控制教育部重点实验室, 江苏 无锡 214122)

摘要: 针对多向主元分析(multi-way principal component analysis, MPCA)算法用于间歇过程实施监控时需要将三维数据转换为高阶的二维矩阵,从而易导致算法的计算量大,且会丢失一些有用信息的情况进行了研究,提出了一种新的间歇过程故障诊断方法——二维主成分分析法(2-dimensional principal component analysis, 2DPCA)。该方法首先利用各个批次的二维矩阵构造协方差矩阵,进而求得所有批次协方差矩阵的平均值进行建模,大大降低了计算复杂度,运算时间较MPCA缩短了19/20到3/4,且无须占用太多存储空间;同时,2DPCA计算协方差矩阵较MPCA更为准确,取协方差矩阵的平均值能够更加精确地反映不同类型的故障,在一定程度上增强了故障诊断的准确率。最后,通过将所提出的方法应用于青霉素发酵过程的监控中,验证了该算法的有效性和准确性。

关键词: 多向主元分析; 二维主元分析; 故障监测; 青霉素发酵

中图分类号: TP181

文献标志码: A

文章编号: 1001-3695(2014)04-1098-04

doi:10.3969/j.issn.1001-3695.2014.04.034

2DPCA modeling and monitoring in penicillin fermentation process

LIANG Xiao-fan^a, XIONG Wei-li^b, YANG Wei-wei^a, XU Bao-guo^b

(a. Dept. of Automation, School of IOT Engineering, b. Key Laboratory of Advanced Process Control for Light Industry Ministry of Education, Jiangnan University, Wuxi Jiangsu 214122, China)

Abstract: The three-dimensional data for batch process monitoring needed to transform as a vector in high-dimensional space using MPCA algorithm, resulting in the large amount of computation and the loss of some useful information. This paper proposed the new method for batch process fault diagnosis, which based on the two-dimensional principal component analysis (2DPCA). The method took advantage of the two-dimensional matrix in each batch to construct the covariance matrix, and obtained the average of the covariance matrix of all batches to model, which led to reduce the computational complexity, short the operation time from 19/20 to 3/4 compared with MPCA and took up the little storage space. At the same time, the 2DPCA approach was more accurate than MPCA when calculating the covariance matrix. 2DPCA was used to model with the covariance average of all batches, which accurately reflected the different faults and enhanced the accuracy of fault diagnosis. Finally, the proposed method was used in the penicillin fermentation process. The results demonstrate the validity and accuracy of the algorithm.

Key words: multi-way principal component analysis; two-dimensional principal component analysis; fault monitoring; penicillin fermentation

0 引言

青霉素是目前使用最多的广谱抗生素之一,其生产发酵是典型的间歇过程,具有变量繁多、过程变化剧烈等特征。青霉素发酵过程操作复杂,操作条件的细微变化、原料中混入杂质等异常状况都会影响最终产品的产量和质量。因此,对该生产过程实施在线监控及故障诊断具有重要的现实意义。

主元分析(PCA)^[1-3]是多元统计过程建模的核心技术,它只需要利用正常的生产数据建立模型,在处理高维数据时具有很大的优势,但是它只适用于连续生产过程。Nomikos等人^[4-6]提出了多向主元分析算法(MPCA),并将其成功应用于间歇过程,现有的间歇过程诊断方法多以此为基础。

MPCA主要应用于简单的间歇过程,为使其获得更广泛的应用,不少研究者提出了各种MPCA的改进算法,例如,针对间

歇过程多阶段的特征,Yu等人^[7]提出基于混合高斯模型划分阶段的方法;Zhao等人^[8,9]采用聚类算法将过程进行分段;Camacho等人^[10]提出自动分段的方法,以未被主元所表示的残差的平方和为指标,将间歇过程分为多段;但是MPCA、多阶段MPCA及其改进方法实质上都是一种线性方法,而青霉素发酵是一个复杂的非线性系统,如果将普通的MPCA方法应用于非线性问题,可能导致重要信息的丢失。基于此,有些学者提出了非线性的MPCA算法^[11-14],但这些算法比较复杂,并且难以处理多变量的间歇过程。

上述各种改进算法能在一定程度上改变传统MPCA对线性过程的依赖和过程不确定程度对算法带来的影响,但这些算法有一个共同特点,即在建模的过程中都需要将三维数据转换成二维数据。当处理的数据量大、维数高时,这给特征提取造成了困难,在进行主成分分析时将会耗费大量的时间,而且三

收稿日期: 2013-07-01; **修回日期:** 2013-08-02 **基金项目:** 国家自然科学基金资助项目(21206053,21276111);中国博士后基金资助项目(2012M511198);江苏高校优势学科建设工程项目(PAPD)

作者简介: 梁小凡(1987-),女,硕士研究生,主要研究方向为生产过程性能监视、故障诊断(qiantofan@126.com);熊伟丽,女,副教授,主要研究方向为工业过程模型化与控制;杨薇薇,女,硕士研究生,主要研究方向为传感器网络;徐保国,男,教授,博导,主要研究方向为过程控制与优化。

维转换成二维矩阵时会丢失掉很多有用信息,容易造成较高的漏报和误报率。因此,为了降低误诊断率,实现准确的实时监测,本文提出了一种新的故障检测方法——二维主成分分析算法,它在处理高维、高度耦合数据时具有独特的优势,越来越受到研究人员和现场工程师的青睐,并对其进行了相应的卓有成效的研究工作。在文献[15~19]中,研究者广泛地把2DPCA应用到人脸图像处理系统中,得到了很好的识别效果。针对该情况,本文结合2DPCA算法的优点,把其应用于间歇过程故障监测与诊断中。该算法首先对各个批次的二维数据矩阵进行标准化处理,然后求其协方差矩阵,最后取其协方差的平均值建立模型,并对新一批的数据进行监测。通过在青霉素发酵过程中的应用研究,结果表明该方法能简化特征值与特征向量的计算,保留多变量生产过程的重要信息,提高算法的运算速度,效果明显优于传统的MPCA算法。

1 MPCA

在间歇生产过程中,其过程数据通常可以表示为一个三维数据矩阵 $X(I \times J \times K)$ 。其中 I 、 J 和 K 分别表示操作批次、过程变量和反应时间序列。MPCA算法是将三维数据沿时间轴方向展开为一个二维矩阵的形式 $X(I \times JK)$ (图1)。为了利用该算法实现间歇过程建模,首先对建模数据进行预处理,然后将三维数据进行二维展开,这里按批次进行展开,得到 K 个时间片矩阵 $X_i(I \times J)$ 。接着按式(1)分别对 K 个时间片矩阵进行主元分析,得到 K 个负载矩阵 $P_k(J \times R)$,其中 $R(R < J)$ 是根据主成分累积贡献率所保留的主成分个数。

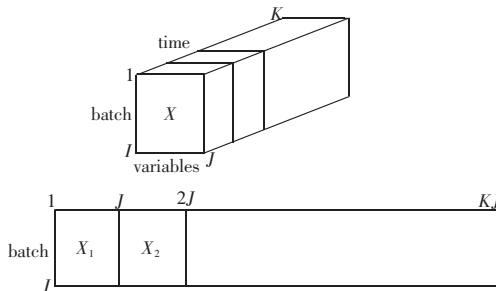


图1 MPCA算法示意图

基于二维矩阵展开的PCA建模原理为

$$\begin{aligned} X_k &= T_k P_k^T + E_k \\ \hat{X}_k &= T_k P_k^T \\ E_k &= X_k - \hat{X}_k \end{aligned} \quad (1)$$

模型建立后,通过计算平方预测误差(SPE)和Hotelling T^2 统计量来实现对间歇生产过程的实时监控。在采样时刻为 k 、置信水平为 α 的条件下,SPE统计量定义为

$$\begin{aligned} \text{SPE}_{k,\alpha} &= g_k \chi_{h_k,\alpha}^2 \\ g_k &= \frac{v_k}{2m_k} \\ h_k &= \frac{2m_k^2}{v_k} \end{aligned} \quad (2)$$

其中: m_k 、 v_k 分别是每个采样时刻SPE样本的均值和方差; $\chi_{h_k,\alpha}^2$ 是检验水平为 α 、自由度为 h_k 条件下的卡方分布临界值。

2 2DPCA

2.1 基本原理

2DPCA是在传统PCA基础上改进的一种数据降维方法。

与MPCA算法不同的是,2DPCA直接是以间歇过程各批次数据矩阵为分析对象,不需要将间歇过程的三维数据矩阵转换成二维向量,这样避免了高维向量的产生,使得计算复杂度大大降低,加快了特征提取的速度。研究表明,2DPCA的运算时间是MPCA方法的 $1/20 \sim 1/4$;同时,在批生产过程中,直接用各个批次矩阵来构造协方差矩阵,使得生成的协方差矩阵规模比MPCA的小且精确,提高了故障识别率。

在间歇生产过程中,2DPCA是以各批次数据(各批次数据组成一个二维数据矩阵)为工作对象,分别对每个二维矩阵进行标准化,然后求出相应的协方差,接着求得所有批次协方差矩阵的平均值,最后对协方差的平均值进行主元分析。

2.2 建模步骤

首先在成功历史批次的模型参考数据库中选取 I 个正常批次,每个批次的维数是 $k \times j$,则数据集可表示为

$$X = \{X_1, X_2, \dots, X_I\}$$

其中:每一个样本 X_i 为 $k \times j$ 的矩阵。

对每一个 X_i 分别进行标准化计算,形成均值为0、方差为1的数据矩阵,即

$$X'_i = [X_i - (1, 1, \dots, 1)^T M] \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \dots, \frac{1}{\sigma_j}\right) \quad (3)$$

式中: $M = [m_1, m_2, \dots, m_k]$; $M \in R^{1 \times k}$ 为变量的均值; $[\sigma_1, \sigma_2, \dots, \sigma_k]$ 为对应变量的标准差。

对经规范化处理后的 I 个矩阵分别求取协方差矩阵:

$$Y_i = \frac{(X'_i)^T \times (X'_i)}{k-1} \quad (4)$$

接着计算 I 个协方差矩阵的平均值:

$$\bar{Y}_i = \left(\sum_{i=1}^I Y_i \right) / I \quad (5)$$

最后对经过一系列处理得到的 $\bar{Y}_i$ 矩阵建立PCA过程的统计模型。

$$\bar{Y}_i = t_1 p_1^T + t_2 p_2^T + \dots + t_j p_j^T = TP^T + E \approx TP^T \quad (6)$$

其中: t_k 为主元变量; p_k 为负荷分量; $T = [t_1, t_2, \dots, t_m]$ 称为得分矩阵; $P = [p_1, p_2, \dots, p_n]$ 为负荷矩阵。

2.3 SPE统计量及控制限的确定

SPE统计量主要反映了某时刻测量数据对主元模型的偏离程度,过程变量的均值、方差或变量之间的相关关系矩阵发生变化都会引起SPE统计量超出控制限。故在多变量统计过程控制中,常用SPE统计量对过程进行监测。

基于2DPCA的SPE控制限与PCA算法是一样的。SPE统计量的控制限可按式(7)计算:

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

式中: $\theta_1 = \sum_{j=R+1}^n \lambda_j^1$, $\theta_2 = \sum_{j=R+1}^n \lambda_j^2$, $\theta_3 = \sum_{j=R+1}^n \lambda_j^3$, $h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$; λ_i 为协方差矩阵 $\bar{Y}_i$ 的特征值; C_α 是正态分布置信度为 α 的统计; R 是保留的主元个数; n 是模型的全部主元个数。

在对新批次进行在线监测时,要计算出新批次观测变量 x_{new} 第 k 时刻的SPE统计量 $\text{SPE}_{\text{new},k}$,如式(8)所示:

$$\text{SPE}_{\text{new},k} = e_{\text{new},k} e_{\text{new},k}^T = x_{\text{new},k}^T (I - PP^T) x_{\text{new},k} \quad (8)$$

其中: I 是 $n \times n$ 的单位矩阵, $e_{\text{new},k}$ 为新批次第 k 时刻采样值 $x_{\text{new},k}(J \times 1)$ 的偏差。

3 青霉素发酵过程仿真

青霉素作为一种抗生素,具有广泛的临床医用价值,其生产过程是一个典型的非线性、动态间歇生产过程,它是在合适的培养基、PH 值、温度、空气流量等发酵条件下生长和合成抗生素的代谢活动。其发酵流程如图 2 所示。本研究在 Pensim v 2.0 仿真平台中实现,通过该平台可以仿真不同初始条件下青霉素生产过程的各个变量。在实验中,设定过程发酵周期为 200 h,采样间隔为 1 h,选择 10 个主要过程变量(空气流量、搅拌功率、底物流加速度、溶解氧浓度、培养基容量、二氧化碳浓度、PH 值、发酵温度、发酵热、冷却水流量)来综合表征青霉素发酵菌体生长和产物合成状况,选取 30 个正常批次作为离线模型参考数据库,得到三维数据矩阵 $X(30 \times 10 \times 200)$ 。

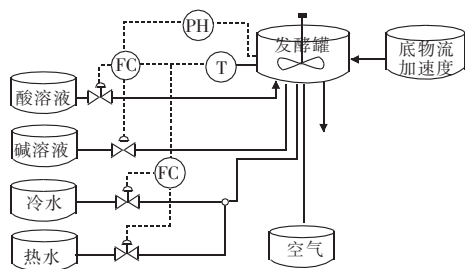


图2 青霉素发酵流程

2DPCA 与 MPCA 都是基于一个历史轨迹数据库进行建模的,该数据库要能够代表间歇生产过程一段时间内的正常操作过程。本文选取 30 个正常批次数据建立离线模型后,用该模型与未来间歇过程的时间轨迹数据相比较,由此可以监测和判断新批量生产过程是否发生异常。

在对青霉素发酵过程进行在线监测时,本文选取 30 个正常批次和 20 个故障批次作为校验数据进行仿真实验,每个批次含 10 个过程变量和 200 个采样点,并且假定每个批次是等长的。这里分别用普通的 MPCA 与本文提出的 2DPCA 两种算法对正常和故障批次进行在线监测。图 3(a)(b)是分别用两种算法对第 10 个正常批次的 SPE 监测图。从图 3 可以看出,两者都对该批次作出了正确监测。第 12 个批次设定的故障是在反应的第 30 个小时对底物流加速度引入 5% 的阶跃扰动,直到第 50 个小时结束。图 4(a)(b)分别是 MPCA 与 2DPCA 算法在线监控第 12 个故障批次的结果。通过比较可以看出, MPCA 算法把第 12 个故障批次监测为正常,而 2DPCA 的监测结果是故障,所以需要验证哪种算法对故障进行了准确判断。这里本文采用故障变量贡献图法对故障原因进行判定,实验结果如图 5 所示。从图 5 中可以看出,变量 3(底物流加速度)是引起故障的主要原因,这与事实相符,证明了 2DPCA 算法的有效性,并在一定程度上提高了模型诊断的准确性。

为了进一步说明 2DPCA 算法的有效性,本文对选取的 30 个正常批次和 20 个故障批次进行故障监测,综合得出两种方法对新批次数据的 SPE 统计量监测率,监测结果如表 1 所示。从表中显示的监测率可以明显看出,2DPCA 对正常批次的 SPE 监测率高达 90%,而传统的 MPCA 算法的监测率为 70%;2DPCA 对故障批次的 SPE 监测率高达 95%,而传统的 MPCA 算法的监测率只有 75%,可见,2DPCA 算法在进行青霉素发酵生产过程的在线监测时,监测的准确率远远高于传统的 MPCA

算法,大大降低了过程的误警率和漏报率,提高了过程的生产效率。

表 2 是在 MATLAB 2009b 平台下分别利用两种算法进行故障检测所需 CPU 时间对比。由表 2 可以看出,2DPCA 算法的故障识别时间要比 MPCA 短,原因在于 2DPCA 直接用协方差矩阵大小为 30×30 来提取过程数据主成分的,而 MPCA 方法需要将 30 个批次的三维数据矩阵 $30 \times 10 \times 200$ 转换成二维矩阵 30×2000 ,增加了特征提取时间,使得算法的计算复杂度变高。

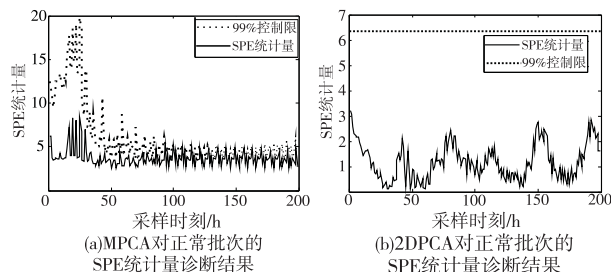


图3 MPCA和2DPCA对正常批次的诊断结果

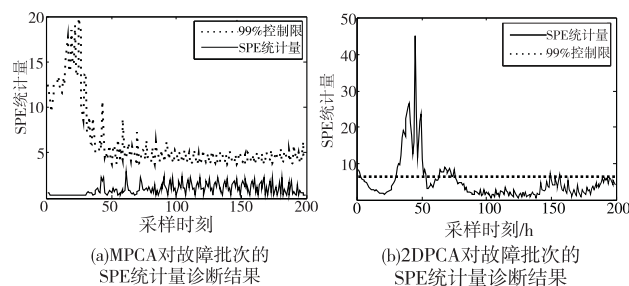


图4 MPCA和2DPCA对故障批次的诊断结果

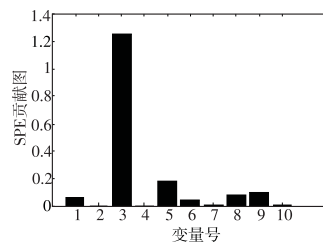


图5 变量贡献图

表1 识别准确率对比

仿真研究	2DPCA	MPCA
正常批次总数	30	30
故障批次总数	20	20
监测正常批次	27	21
监测故障批次	19	15
正常批次监测率/%	90	70
故障批次监测率/%	95	75

表2 故障诊断时间对比

算法	2DPCA	MPCA
协方差矩阵大小	30×30	$2\,000 \times 2\,000$
主元特征矩阵大小	30×4	$2\,000 \times 4$
故障识别时间	0.267 8	0.671 5

4 结束语

本文利用 2DPCA 算法对间歇过程进行监控。该方法不需要对间歇过程的三维数据进行展开,而是直接对每个批次的二维数据矩阵进行处理,能有效解决由生产过程中数据量繁琐、庞大,导致在进行故障监测时一部分重要信息的丢失,致使监测结果不够准确、产生漏报的现象等问题。通过对青霉素发酵生产过程的实际应用,结果表明 2DPCA 方法是有效的,能准确

地对观测数据进行监控。

参考文献:

- [1] JACKSON J E. A user's guide to principal components[M]. New York: Wiley, 1991.
- [2] KOURTI T, MACGREGOR J F. Process analysis, monitoring and diagnosis, using multivariate projection methods[J]. *Chemometrics and Intelligent Laboratory Systems*, 1995, 28(1): 3-21.
- [3] WANG X Z. Data mining and knowledge discovery for process monitoring and control [C]//Advances in Industrial Control. London: Springer, 1999: 1-251.
- [4] NOMIKOS P, MACGREGOR J F. Monitoring batch processes using multiway principal component analysis[J]. *American Institute of Chemical Engineers Journal*, 1994, 40(8): 1361-1375.
- [5] 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.
- [6] NOMIKOS P, MACGREGOR J F. Multivariate SPC charts for monitoring batch processes[J]. *Technometrics*, 1995, 37(1): 41-59.
- [7] YU Jie, QIN S J. Multi-way Gaussian mixture model based multiphase batch process monitoring[J]. *Industrial & Engineering Chemistry Research*, 2009, 48(18): 8585-8594.
- [8] ZHAO Chun-hui, WANG Fu-li, MAO Zhi-zhong, et al. Improved batch process monitoring and quality prediction based on multiphase statistical analysis[J]. *Industrial & Engineering Chemistry Research*, 2008, 47(3): 835-849.
- [9] ZHAO C, WANG F. Adaptive monitoring method for batch processes based on phase dissimilarity updating with limited modeling data[J]. *Industrial & Engineering Chemistry Research*, 2007, 46(14): 4943-4953.
- [10] CAMACHO J, PICO J. Multi-phase principal component analysis for batch processes modeling[J]. *Chemometrics and Intelligent Laboratory Systems*, 2006, 81(2): 127-136.
- [11] YU Jie. Nonlinear bioprocess monitoring using multiway kernel localized Fisher discriminate analysis[J]. *Industrial and Engineering Chemistry Research*, 2011, 50(6): 3390-3402.
- [12] TIAN Xue-min, ZHANG Xiao-ling, DENG Xiao-gang, et al. Multiway kernel independent component analysis based on feature samples for batch process monitoring[J]. *Neurocomputing*, 2009, 72(7~9): 1584-1596.
- [13] ZHANG Ying-wei, ZHOU Hong, QIN S J. Decentralized fault diagnosis of large-scale process using multi block kernel principal component analysis[J]. *Acta Automatica Sinica*, 2010, 36(4): 593-597.
- [14] ZHAO Chun-hui, GAO Fu-rong, WANG Fu-li. 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.
- [15] YANG Jian, ZHANG D, FRANGI A F, et al. Two-dimensional PCA: a new approach to appearance-based face representation and recognition[J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2004, 26(1): 131-137.
- [16] 韩柯, 朱秀昌. 基于二维PCA的人脸识别方法研究[J]. *杭州电子科技大学学报*, 2007, 27(1): 69-72.
- [17] 李艳芳, 费洪晓. 基于2DGabor小波与2DPCA的人脸识别方法[J]. *湖南科技学院学报*, 2008, 29(8): 72-74.
- [18] 袁宁, 吴小俊, 王士同, 等. 一种模块化2DPCA和CSLDA相结合的人脸验证算法[J]. *计算机研究与发展*, 2008, 30(3): 286-289.
- [19] 吴清江. 一种基于2D-DWT和2D-PCA的人脸识别方法[J]. *计算机应用*, 2006, 26(9): 59-78.

(上接第1097页)识别方法具有需求训练样本少、训练时间短、诊断速度快等优点,为小样本下的刀具磨损状态在线监测与剩余寿命预测提供了有效的解决方案,具有较高实用价值。

d) 刀具磨损量识别精度还取决于学习样本的代表性,随着学习样本数量的增加,预测精度将进一步提高。针对时序监测结果波动性大的问题,可以考虑对时序观察结果进行滤波。

参考文献:

- [1] Jr VALLEJO A G, NOLAZCO-FLORES J A, MORALES-MENENDEZ R, et al. Tool-wear monitoring based on continuous hidden Markov models[C]//Proc of the 10th Iberoamerican Congress Conference on Progress in Pattern Recognition, Image Analysis and Applications. Berlin: Springer, 2005: 880-890.
- [2] LI Wei-lin, FU Pan, CAO Wei-qing. Tool wear states recognition based on frequency-band energy analysis and fuzzy clustering[C]//Proc of the 3rd International Workshop on Advanced Computational Intelligence. 2010: 162-167.
- [3] TETI R, JEMIELNIAK K, O'DONNELL G, et al. Advanced monitoring of machining operations[J]. *CIRP Annals-Manufacturing Technology*, 2010, 59(2): 717-739.
- [4] ROTH J T, DIURDIANOVIC D, YANG Xiao-ping, et al. Quality and inspection of machining operations; tool condition monitoring[J]. *Journal of Manufacturing Science and Engineering*, 2010, 132(4): 1-16.
- [5] SNR D, DIMLA E. Sensor signals for tool-wear monitoring in metal cutting operations: a review of methods[J]. *International Journal of Machine Tools and Manufacture*, 2000, 40(8): 1073-1098.
- [6] 王国锋, 李启铭, 秦旭达, 等. 支持向量机在刀具磨损多状态监测中的应用[J]. *天津大学学报*, 2011, 44(1): 35-39.
- [7] 关山, 王龙山, 聂鹏. 基于EMD与LS-SVM的刀具磨损识别方法[J]. *北京航空航天大学学报*, 2011, 37(2): 144.
- [8] SHI Dong-feng, GINDY N N. Tool wear predictive model based on least squares support vector machines[J]. *Mechanical Systems and Signal Processing*, 2007, 21(4): 1799-1814.
- [9] WIDODO A, YANG B S. Support vector machine in machine condition monitoring and fault diagnosis[J]. *Mechanical Systems and Signal Processing*, 2007, 21(6): 2560-2574.
- [10] 李文元, 闫海华, 姚宏杰. 粒子群优化的最小二乘支持向量机在通信装备故障预测中的应用[J]. *微电子学与计算机*, 2013, 30(2): 99-102.
- [11] 姚全珠, 蔡婕. 基于PSO的LS-SVM特征选择与参数优化算法[J]. *计算机工程与应用*, 2010, 46(1): 134-136.
- [12] 宋宏耀, 宋宏兵, 崔秀政. 基于PSO优化最小二乘支持向量机的热工系统辨识[J]. *电力科学与工程*, 2009, 25(10): 43-46.
- [13] 李文莉, 李郁侠. 基于粒子群最小二乘支持向量机的水文预测[J]. *计算机应用*, 2012, 32(4): 1188-1190.
- [14] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. *Neural Processing Letters*, 1999, 9(3): 293-300.
- [15] KENNEDY J, EBERHART R. Particle swarm optimization[C]//Proc of International Conference on Neural Networks. 1995: 1942-1948.