

文章编号 :1009 - 038X( 2000 )05 - 0446 - 04

# 基于主元分析和模糊模型的链霉素发酵过程建模

张泉灵,金晓明,王树青,荣冈,陈元青

(浙江大学工业控制技术国家重点实验室,浙江杭州 310027)

**摘要:**基于主元分析和模糊模型,提出了一种简单而有效的链霉素发酵过程产物浓度的预报方法.该方法采用主元分析压缩关联程度高且含有测量噪声的实际工业生产数据,筛选出影响产物浓度的主要过程变量,构造了模糊分段线性模型的产物浓度估计器.与线性多元回归模型相比,模糊模型更适合作为间歇发酵过程的状态估计器.

**关键词:**链霉素发酵;主元分析;模糊模型;监控

中图分类号:R978.12

文献标识码:A

## Streptomycin Fermentation Process Modeling with Principal Component Analysis and Fuzzy Model

ZHANG Quan-lin, JIN Xiao-ming, WANG Shu-qing, RONG Gong, CHEN Yuan-qing  
(State Key Lab of Industrial Control Technology, Zhejiang University, Hangzhou 310027)

**Abstract:** Analysis, modeling and control for fed-batch fermentation process still remain a challenging issue. Based on Principal Component Analysis (PCA) and fuzzy model, a simple and efficient approach to monitor the fed-batch streptomycin fermentation is presented. The data obtained from industrial streptomycin fermentation process were preliminary analyzed with PCA so that the large multivariate data with highly correlated and noisy measurements can be compressed into a lower dimension space which contains most of the variance of the original matrix. Moreover, fuzzy model was used to construct a product (antibiotic) concentration estimator of the streptomycin fermentation process in that prior knowledge and expertise are important in fed-batch fermentation processes. The results of fuzzy model comparing with linear multivariate regression model indicate that the potential of fuzzy model as state estimator of all such industrial fed-batch processes.

**Key words:** streptomycin fermentation; principal component analysis; fuzzy model; monitoring

工业发酵过程的监控对确保该过程的安全生产和获得优质产品十分重要.然而这种生化过程具有高度的非线性和时变性、内在机理非常复杂而且一些重要的过程变量又不能在线测量,因此难以建立精确的机理模型来实现正确的监控.另一方面,

生化过程是有生命的过程,而且通常是不可逆的.如果一开始监控不佳,就很难使以后的发酵过程达到预期的目标.为保证发酵过程的一致性和可重复性,人们通常利用以往的经验进行生产并有着重视间歇操作过程严格程序化<sup>[1]</sup>.

收稿日期:1999-12-23,修订日期:2000-06-09.

作者简介:张泉灵(1973-),男,浙江温岭人,工学博士,讲师.

万方数据

近年来,工业发酵过程监控的研究主要集中在如何开发和使用带状态估计器的机理模型和基于知识的智能模型.这些模型可以为生化参数如:菌体浓度、基质浓度、产物浓度等的测量或一次变量的在线估计提供基于二次可测量变量的软仪表.建模的主要方法有:二次变量与一次变量之间的线性近似方法<sup>[2]</sup>、基于神经网络的方法<sup>[3]</sup>和模糊辅助神经网络方法<sup>[4]</sup>等.工业发酵过程中的丰富过程数据蕴含着过程的很多重要信息.然而,变量数目过多往往会增加问题的复杂性.因此,人们自然希望选取少量包含丰富信息的变量.主元分析(PCA)是一种将多个变量转化为少数变量的多元统计方法.它可将原来的所有变量综合成尽可能少的几个综合变量,这些互不相关的综合变量可以尽可能多地包含原来多个变量所反映的信息<sup>[5,6]</sup>.

模糊模型既可以表达由专家行为特征或专家经验中获取的启发式知识,也可以在专家知识无法用语言表达时,采用无导师的规则聚类算法从经验数据中获取启发式知识<sup>[7]</sup>.在用于工业发酵过程建模时,提高模糊模型表现能力的最直接方法就是在发酵过程的不同阶段分别采用分段线性模型来描述过程.

作者在综合主元分析方法和模糊模型的基础上,构造了链霉素发酵过程产物浓度状态估计器.主元分析主要用于重构降维的数据集,而模糊模型则基于过程动态特性信息对整个发酵过程进行逻辑划分,再用分段线性模型分别加以建模.这种新的建模方法为链霉素发酵过程的监控提供了一条有效途径.

## 1 过程描述

链霉素是一种氨基糖苷类抗生素,在医药应用中量大面广.链霉素发酵是利用特定的微生物(生产菌),在一定条件下使之生长繁殖,并在代谢过程中产生抗生素,然后用适当的化学手段将抗生素从发酵液中提取出来.链霉素发酵过程机理十分复杂,可能存在的酶反应超过几百个,影响因素繁多.链霉素发酵过程有以下特点(1)链霉素是次级代谢产物,菌体生产与产物形成不平行(2)理论产量难以用物料平衡来推算(3)生产稳定性差.总之,链霉素发酵过程的建模和控制有相当大的难度.链霉素工业发酵过程基本上依赖于人工经验进行操作,其结果是链霉素发酵水平时高时低.如何按照生物生长规律,提供合适的外部环境,实现最优控制是提高发酵单位的关键<sup>[8]</sup>.

对链霉素发酵而言,碳源、氮源是构成链霉素菌体和链霉素构成部分的重要来源.在发酵过程中需要不断补充这些营养物质,以保证基质处于适宜的浓度,从而满足菌体生长的需要.通过对链霉素工业发酵过程的在线测量和定期化验分析数据如:发酵时间、pH 值、温度、空气流量、碳浓度、氮浓度、粘度以及碳氮浓度比进行初步分析,我们发现链霉素发酵是一个可分为 3 个阶段的非线性过程.这 3 个阶段可以借助菌体生长和产物生成来加以刻划.图 1 给出了链霉素发酵过程中高产率和正常产率两种情况下产物浓度( $c_p$ )与碳源浓度( $c_c$ )、氮源浓度( $c_n$ )的变化情况.

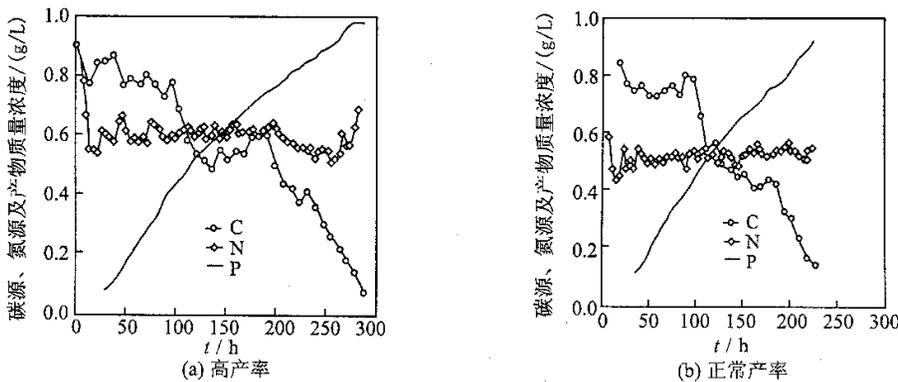


图 1 不同发酵批次菌体生长和产物变化情况

Fig.1 Relationship between cell growth and product formation in different fermentation batch

如图 1 可知,在链霉素发酵过程的前期,应控制发酵液中碳源浓度在一较高水平,使菌体快速生长.80 h 后,发酵液中碳源浓度控制在一适中水平上,缩短菌体生长期,使菌体生长速率降低,启动抗

生素合成酶系生成,使发酵转入生产期.碳源含量控制主要通过补糖量来控制,在发酵前期,80 h 内出现一补糖高峰,在 100~200 h 内补糖量略为前期一半,后期不补糖.发酵液中氮源量基本保持稳定,

补氮量适中,环境条件 pH 在生产期保持在 6.7 ~ 6.9 范围内,前期和后期略为升高,可通过补氨水来控制。

在此基础上,作者利用链霉素工业生产中成功发酵批次的数据来描述其运行情况并建立链霉素发酵过程产物浓度的经验模型。

## 2 产物浓度模型

在链霉素发酵过程中,产物浓度是一个关键指标,但它很难在线检测,因此有必要建立产物浓度的估计模型。工业链霉素发酵过程中的测量变量有:发酵时间、pH、碳源浓度、氮源浓度、效价、粘度、罐温、罐压、空气流量等,这些变量对链霉素的菌体生长和产物合成都有一定的影响,但是它们是相互关联的,而且有些变量对最后产物合成的贡献很小。剔除非关键变量不仅可以改善模型的预测能力,同时也有利于模型工业应用时的维护。作者采用主元分析法来分析和压缩成功批次的历史数据集,并利用模糊模型来构造产物浓度估计器。

### 2.1 主元分析

作为一种多元统计方法,主元分析可以将一组变量的原有特征进行组合,抽取所谓的主元作为新的特征,并在信息论的意义上保证熵达到极小值。因此,它在大量减少特征变量数的同时,仍能保留基本的信息。其简单的原理如下:

设  $X = (X_1, X_2, \dots, X_p)^T$  是一个  $p$  维随机向量,代表原变量组,且有二阶矩存在,记  $\mu = \epsilon(X)$ ,  $\Sigma = D(X)$ 。考虑其线性变换:

$$t_i = L_i^T X = l_{i2}x_2 + \dots + l_{ip}x_p, \quad i = 1, 2, \dots, p \quad (1)$$

且方差  $\text{var}(t_i) = L_i^T \Sigma L_i$ , 协方差  $\text{cov}(t_i, t_j) = L_i^T \Sigma L_j, \quad i, j = 1, \dots, p, \quad i \neq j$ ,  $\text{var}(t_1)$  越大,表明  $t_1$  中包含的信息越多。若上述变换满足:

$$\text{cov}(t_i, t_j) = 0, \quad \text{var}(t_1) > \text{var}(t_2) > \dots > \text{var}(t_p) \quad (2)$$

则  $t_1, t_2, \dots, t_p$  分别称为  $X$  的第一主元,第二主元, ..., 第  $p$  主元。此时,方差  $\text{var}(t_i)$  即为协方差阵  $\Sigma$  的特征值  $\lambda_i$ , 且  $L_1, L_2, \dots, L_p$  为对应的单位化正交特征向量。

这样,  $X$  的主元是以  $\Sigma$  的单位化正交特征向量为系数的线性组合,即:

$$X = TL^T \quad (2)$$

$X$  的主元有很多重要性质,其中有:

- ①  $X$  的任一标准线性组合的方差都不会大于  $\lambda_i$ ;
- ② 比值  $\frac{\lambda_k}{\sum_{i=1}^p \lambda_i}$  表示前  $k$  个主元对总体方差

的贡献率;③ 主元随原始变量标准化的不同而有变化;

性质③对于实际应用有重要意义,需要依据一定的标准化原则对变量组  $X = (x_1, x_2, \dots, x_p)^T$  进行标准化处理,即:

$$X^* = D_\sigma^{-1}(X - \mu) \quad (3)$$

其中  $D_\sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p)$ ,  $\sigma_i$  为  $x_i$  的标准差。

在实际应用时,协方差阵一般是未知的,可以取  $p$  个过程变量的  $n$  次测量值,形成  $n \times p$  的矩阵  $X$ , 标准化处理得到  $X^*$ , 即:

$$\hat{\Sigma}^* = n^{-1} X^{*T} X^* \quad (4)$$

特征分解得到:

$$\hat{\Sigma}^* = L^T \Lambda L \quad (5)$$

由于前  $k$  个 ( $k < p$ ) 主元反映了  $X = (x_1, x_2, \dots, x_p)^T$  中的大部分信息,这样就可以用少数几个不相关变量表征原有变量所包含的信息。

作者选用链霉素发酵过程变量为 8 个:发酵时间  $t$ 、pH、碳源浓度( $c_C$ )、氮源浓度( $c_N$ )、粘度( $S$ )、罐温( $T$ )、空气流量( $F$ )、碳氮浓度比( $c_C:c_N$ )组成数据矩阵  $X(8 \times n)$ ,用不同批次产物浓度的数据进行主元分析,图 2 是变量载荷图。

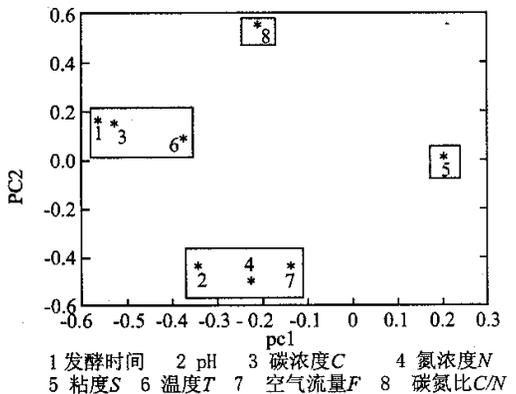


图 2 变量载荷图

Fig.2 The loading plot of variables

从图 2 中可看出,变量分成 4 个聚类,同一聚类中,一般选出第一载荷向量中系数较大的一个,也可根据工艺要求具体选出其中一主要变量。主元分析的结论是:

发酵时间、碳氮浓度比、氮浓度和粘度是影响链霉素发酵的主要过程变量,因此产物浓度模型可表述为:

$$P = f(t, S, c_C, c_C:c_N) \quad (6)$$

### 2.2 线性回归模型

根据主元分析结果,采用 4 个输入变量来建立

链霉素发酵过程的产物浓度线性回归模型,回归方程如下:

$$\begin{cases} P = X * B \\ X = [1NStc_C : c_N] \\ B = [-9992 \ 2570 \ 114 \ 172 \ -73159]T \end{cases} \quad (7)$$

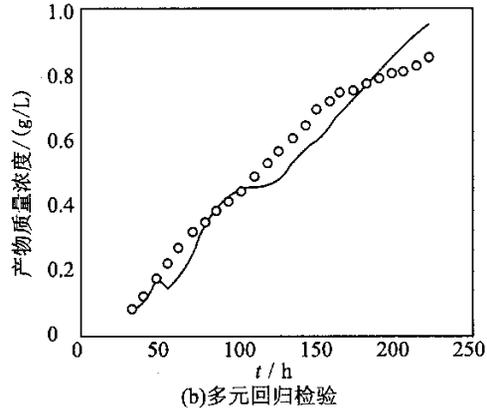
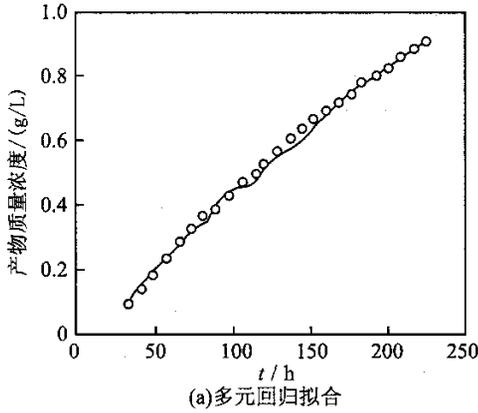


图 3 线性回归模型

Fig.3 The model of linear regression

### 2.3 模糊模型

分析链霉素发酵过程数据可知,氮源浓度在发酵过程中通过补料基本维持不变,因此,碳氮浓度比的变化与碳源浓度变化趋势基本一致.碳源浓度、碳氮浓度比过程状态变量均有明显的三段线性趋势.根据该过程这一特点,可以建立一模糊规则基发酵阶段识别系统,以碳浓度、碳氮浓度比变化以及发酵时间作参考.

将模糊模式识别和多元线性回归模型的优点相结合,可以得到一个模糊分段线性模型的产物浓度状态估计器.该模糊模型是由多个规则基线性模型组成的非线性模型,其前提部分是模糊的,而结论部分则是线性的.隶属函数决定了对相应规则的

线性回归模型共使用 35 个数据集,其中 25 个用于建模,剩余 10 个用于模型检验.回归预测结果见图 3.从图中可以看出,由于发酵过程的非线性,用简单的线性方程去拟合,误差比较大.

置信程度.

模糊 IF-THEN 规则具有如下的形式:

$R_j$  :IF  $c_C$  is  $A_i$  and  $c_C/c_N$  is  $B_i$  THEN

$$P = b_{0j} + b_{1j}N + b_{2j}S + b_{3j}t + b_{4j}c_C/c_N \quad j = 1 \dots A \quad (8)$$

此处,  $c_C$  和  $c_C/c_N$  是标准化输入变量,  $P$  是输出变量而  $A_i$  和  $B_i$  ( $i = 1, 2$ ) 是相应的输入模糊集.此外,发酵时间与  $C_C$  和  $C_C/C_N$  的变化率可作为发酵阶段模糊识别的约束变量.

模糊模型开发时使用数据集的方式与线性回归模型相类似,其结果见图 4.对比图 4 和图 3 可知,采用模糊模型作为发酵过程状态估计器是准确而有效的.

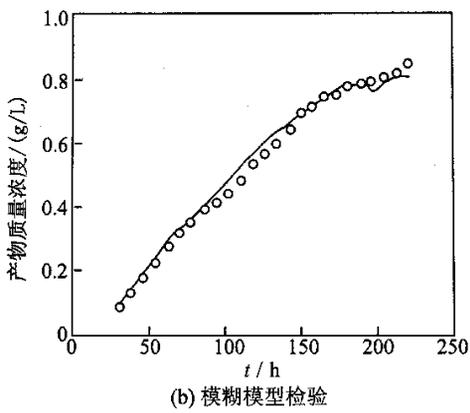
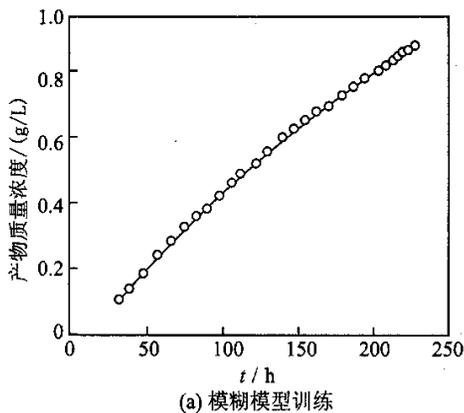


图 4 模糊模型

Fig.4 The fuzzy model

### 3 结 论

从链霉素发酵过程实际批报数据出发,分析链霉素发酵过程的主要特点,利用主元分析方法对众多变量进行降维处理,找出影响过程的主要变量。

在此基础上将链霉素发酵过程分成 3 段,即发酵前期、发酵中期和后期,并利用模糊分段线性模型进行产物浓度预测。由于本方法只是利用了工业实际数据,原理简单且方便实用,对机理复杂的生化过程分析、建模和监控十分有效。

### 参考文献

- [1] NOMIKOS P, MACGREGOR J F. Monitoring of Batch Processes Using Multi-Way Principal Component Analysis[ J ]. *AICHE J*, 1994, 40 :1361 ~ 1375.
- [2] THAM M T, MONTAGUE G A, MORRIS A J. Soft-sensors for process estimation and inferential control[ J ]. *J. Process Control*, 1991, 1 : 13 ~ 14.
- [3] MASSIMO C D, MONTAGUE G A, WILLIS M J. Towards improved penicillin fermentation via artificial neural networks[ J ]. *Computers Chem Engng*, 1992, 16( 4 ), 283 ~ 291.
- [4] SIMUTIS R, HAVLIK I, LUBBERT A. Fuzzy-aided Neural Network for Real-time State Estimation and Process Prediction in the Alcohol Formation Step of Production-Scale Beer Brewing[ J ]. *Journal of Biotechnology*, 1993, 27 :203 ~ 215.
- [5] SABTEN A, KOOT G, ZULLO L C. Statistical data analysis of a chemical plant[ J ]. *Computers Chem Engng*, 1991, 21 :1123 ~ 1129.
- [6] TAKAGI H, SUGENO M. Fuzzy identification of systems and its applications to modeling and control[ J ]. *IEEE Trans*, 1985, 15 :115 ~ 132.
- [7] MARDIA K V, KENT J T, BIBBY J M. Multivariate analysis[ M ]. London : Academic Press, 1982.
- [8] 陈元青.生化过程优化分析、建模和监控的研究[D].杭州:浙江大学,1999.