15 卷 3 期
1999 年 7 月

生 物 工 程 学 报
Chinese Journal of Biotechnology

Vol.15 No.3
July 1999

多元统计分析方法在链霉素发酵中的应用

陈元青　陈　琦　王树青

(浙江大学工业控制技术国家重点实验室　杭州　310027)

摘　要　将多元统计方法中的最重要的方法之一主元分析法用于实际工业链霉素发酵数据的分析，该方法能有效地将原来较多的相关变量所包含的大部分信息用少数不相关的变量来反映，从而简化链霉素发酵中的控制变量。建立规则基模式识别系统，将链霉素发酵过程分三个时期。利用多元回归方法，将链霉素发酵分三段进行产物浓度预测，取得了较好效果。

关键词　主元分析，链霉素发酵，多元回归

分类号　Q81　　**文献标识码**　A　　**文章编号**　1000-3061(1999)03-0368-72

在抗生素发酵过程中，影响菌体生产和产物形成的因素很多，各参数之间相互关联，变化其中某一个参数，常会引起其它参数的变化。在实际工业中，积累了大量过程生产数据，大多数情况下是弃之不用的。这主要是因为这些数据关联程度高，很难从表面上判断过程特点和规律，从而指导实践生产。但是过程很多重要信息都隐含在这些大量数据中，因此用基于数据的模型监控发酵过程显得很有必要[1,2]。工业上链霉素发酵基本依靠生产经验，使生产中链霉素发酵水平时高时低。本文利用主元分析法[3]将链霉素发酵过程中大量生产数据矩阵简化，揭示影响链霉素发酵的主要过程变量。分析这些主要变量建立规则基模式识别系统，识别链霉素发酵过程三个阶段，再利用多元线性回归分析[4]预测产物浓度，为实际生产提供具体指导。

1　过程方法以及结果

1.1　主元分析法

设 $X=(x_1, x_2\cdots x_m)^T$ 是一个均值为 u 方差为 Σ 的随机向量，即 $X\sim(u, \Sigma)$。则主元变换为下面的变换：

$$X \rightarrow t = p^T(x-u) \tag{1}$$

其中 P 是正交的，且有

$$P^T\sum P = A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_m \end{pmatrix} \tag{2}$$

其中 $\lambda_1, \lambda_2\cdots\lambda_m$ 为协方差阵 Σ 的特征值，且有 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant 0$，$p_1, p_2\cdots p_m$ 为单位化正交特征向量，即 X 的主元是以 Σ 的单位化正交特征向量为系数的线性组合。这

收稿日期：1998-06-02，修回日期：1999-02-02。

样,X 的第 i 个主元就可定义为向量 t 的第 i 个元素,即

$$t_i = p_i^T(x - u) \tag{3}$$

其中 p_i 是 P 的第 i 列,称之为第 i 个主元的载荷向量,t_i 称之为分数向量。

X 的主元有很多重要性质[5]:

1. $E(t_i)=0$, 2. $\mathrm{var}(t_i)=\lambda_i$, 3. $\mathrm{cov}(t_i,t_j)=0(i\neq j)$, 4. $\mathrm{var}(t_1)\geqslant\mathrm{var}(t_2)\geqslant\cdots\geqslant\mathrm{var}(t_m)$, 5. $\sum_{i=1}^{m}\mathrm{var}(t_i)=tr\sum$, 6. $\prod_{i=1}^{m}\mathrm{var}(t_i)=|\sum|$, 7. X 的任一标准线性组合的方差都不会大于 λ_1, 8. 比值 $\sum_{i=1}^{k}\lambda_i/\sum_{i=1}^{m}\lambda_i$ 表示前 K 个主元对总体方差的贡献率, 9. 主元分量依赖于测量初始变量的尺度。

在实际应用中,往往根据性质 9,对变量组 $X=(x_1,x_2\cdots x_m)^T$ 进行标准化处理,即:

$$X^* = D_\sigma^{-1}(x - u) \tag{4}$$

其中 $D_\sigma=\mathrm{diag}(\sigma_1,\sigma_2\cdots\sigma_m)$, u 为 X 的均值向量,σ_i 为 x_i 的标准差。

假设一个工业过程对 m 个过程变量进行了 n 次测量,形成了 $m\times n$ 矩阵 X,标准化处理得到 X^*,有

$$\sum{}^* = 1/n\, X^* X^{*T} \tag{5}$$

特征分解得到　　$$\sum{}^* = P^T\Lambda P \tag{6}$$

其中 Λ 为特征值矩阵,这是一个以 $\lambda_1,\lambda_2\cdots\lambda_m$ 为对角元素的对角矩阵,其中 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant0$。P 为特征向量矩阵,$p_i$ 为 λ_i 对应的特征向量。

由于前 K 个主元($K<P$)反映了 $X=(x_1,x_2\cdots x_m)^T$ 中的大部分信息,我们就达到了用少数几个不相关变量表征多数相关变量所包含的信息的目的。

为了说明前 K 个主元概括原变量信息的大小,这里通常根据性质 8 来确定,当 K 个主元的方差贡献率达 80% 以上时,就可认为 X 中的大部分信息已被前 K 个主元所包含。

1.2　主元分析法用于链霉素发酵

链霉素发酵过程机理比较复杂,过程变量多。工业上链霉素发酵过程测量变量有: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。但是这些变量是相互关联的,它们各自对最后产物合成所作的贡献也是不一样的。如果对所有变量都进行控制,并预测最后产物浓度,是极其复杂和不切实际的。主元分析法能使一个维数很大的数据矩阵简化,揭示其主要结构。我们选用链霉素发酵过程变量为 8 个:碳源浓度(c)、氮源浓度(n)、pH、粘度(S)、罐温(T(℃))、发酵时间(t(h))、碳氮源浓度比(c/n)、空气流量(F)组成数据矩阵 $X(8\times n)$,用不同批次产物浓度有好有坏的数据进行主元分析,揭示影响最终产率的主要过程变量。主元分析法中的降维方法很多[6,7],此处是利用对载荷矩阵的分析:即从第一个载荷向量 p_1 中选取绝对值最大的系数所对应的变量,然后将变量从数据集中剔除,再继续用主元分析法分析剩余的数据集合,用同样的判据选择第二个变量。重复此过程,依据主元性质 8,选定前 4 个主元,它们对总体方差的贡献率达到 80% 以上,可认为此 4 个主元能充分表达考察过程。分析结果表明,影响链霉素发酵的主要过程变量有:发酵时间、碳氮源浓度比、碳源浓度和粘度。

$$\begin{bmatrix} c \\ n \\ \mathrm{pH} \\ S \\ T(℃) \\ t(\mathrm{h}) \\ c/n \\ F \end{bmatrix} \xrightarrow{\text{主元分析}} \begin{bmatrix} c \\ S \\ c/n \\ t(\mathrm{h}) \end{bmatrix}$$

1.3 回归分析用于产物浓度预测

前面用主元分析法得到的4个变量是影响发酵产物的主要因素，我们可以通过它们去预测发酵产物浓度，回归方程如式(7)所示，回归预测结果见图1，2。从图上可以看出，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。

$$\begin{cases} P = X \cdot B \\ X = [cStc/n] \\ B = [-9992\ 2570\ 114\ 172\ -73159]T \end{cases} \tag{7}$$

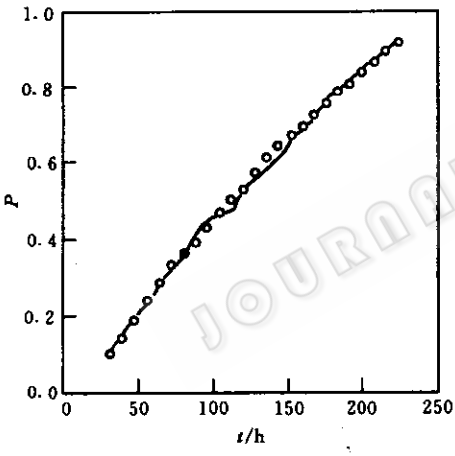

图1 多元回归拟合

Fig.1 Multivariate regression

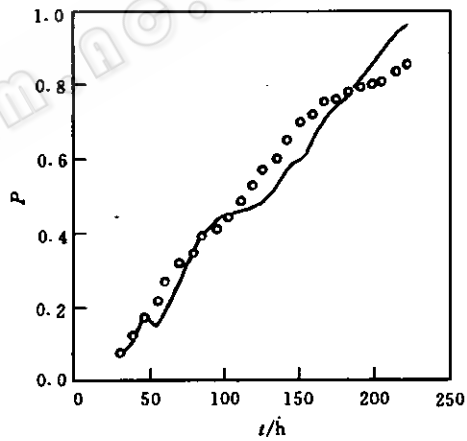

图2 多元回归检验

Fig.2 Multivariate regressive rectification

1.4 运用规则基进行发酵过程分段

分析链霉素发酵过程数据可知，氮源浓度在发酵过程中基本维持不变，因此，碳氮源浓度比的变化与碳源浓度变化趋势基本一致。如图3所示，碳源浓度、碳氮源浓度比均有明显的三段线性趋势。发酵前期($t<t1$)，碳源浓度、碳氮源浓度比均维持在较高水平，并且保持基本恒定。发酵中期($t1<t<t2$)，碳源浓度、碳氮源浓度比下降，维持在一适中水平上，并且基本保持恒定。发酵后期($t>t2$)，碳源浓度、碳氮源浓度比持续下降。在整个发酵期，粘度基本呈上升趋势。根据该过程主要特点，可以建立一规则基模式识别系统，以碳源浓度、碳氮源浓度比变化以及发酵时间作参考。

分析过程数据可知，链霉素发酵过程碳源浓度、碳氮源浓度比在不同时期，其值都有

一固定范围，在分段转折点时，其值有一大的跳跃。因此建立的模式识别系统如下：

if ($\Delta c > a, c > b1$) and ($\Delta(c/n)) > c, c/n > d1$) and ($t > e1$) then $t = t1$

if ($\Delta c > a, b2 < c < b1$) and ($\Delta(c/n)) > c$, $d2 < c/n < d1$) and ($t > e2$) then $t = t2$

经过大量数据分析，对链霉素发酵过程，上述阈值可取为：$[a, b1, b2, c, d1, d2, e1, e2] = [0.5, 3.5, 2, 10, 40, 30, 50, 150]$。

找出了 $t1$、$t2$，也即找出了三个不同阶段的时间分隔点。输入发酵过程数据，可将其分成三个不同发酵时期，即生长期($t < t1$)，生产期($t1 < t < t2$)，衰亡期($t > t2$)。

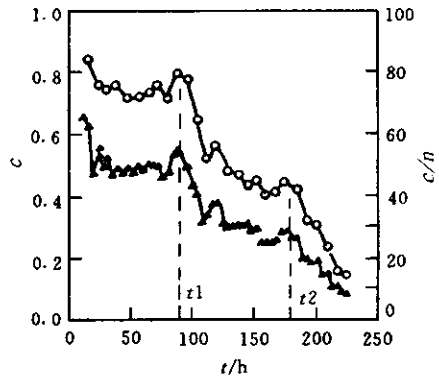

图 3　c 浓度和 c/n 浓度比随时间变化图

Fig. 3　Carbon concentration and carbon/nitrogen concentration rate varying with time

－ ○ － cabon conc, －▲－ c/n

1.5　混合模型用于产物浓度预测

混合模型处理过程如图 4 所示，当在线和离线变量测量值输入数据集后，利用已建立的规则基模式识别系统将链霉素发酵过程分成三段，每一段利用多元线性回归法进行产物浓度预测。这样，就将非线性发酵过程用简单的分段线性函数来表示，简化过程模型，便于过程监测和控制。回归方程如式(8)，预测结果见图 5，6。利用主元分析法得到 4 个变量，分段预测发酵各阶段的产物浓度误差较小。因此我们通过监控发酵各阶段的主要变量，利用混合模型预测发酵产物浓度，指导过程生产。

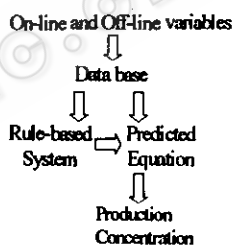

图 4　混合模型处理过程

Fig. 4　Mixed model proces

$$\begin{cases} P1 = X \cdot B1 \quad (t < t1) \\ P2 = X \cdot B2 \quad (t1 < t < t2) \\ P3 = X * B3 \quad (t > t2) \\ X = [cStc/n] \\ B1 = [610 - 778 - 164\ 178\ 20284]T \\ B2 = [7317\ 365 - 669\ 125 - 54864]T \\ B3 = [38000 - 530 - 400 - 10 - 280630]T \end{cases} \tag{8}$$

2　结　论

本文将主元分析法用于实际工业链霉素发酵过程分析，该方法能有效的减少过程控制变量，将影响链霉素发酵的众多变量的大部分信息抽提出来，用 4 个过程变量来表示，即碳源浓度、氮源浓度、发酵时间、碳氮源浓度比。为发酵过程建模和控制带来方便。利用已建立的规则基系统可将链霉素发酵过程分成三个阶段表示，即生长期、生产期和衰退期。链霉素发酵的非线性过程可近似用这三段时期线性模型表示，并利用多元回归分析

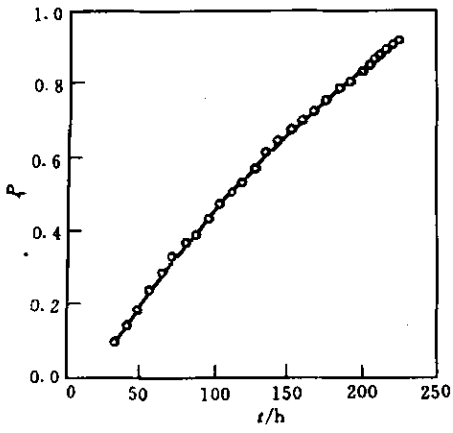

图5 分段多元回归拟合

Fig.5 Segmented multivariate regression

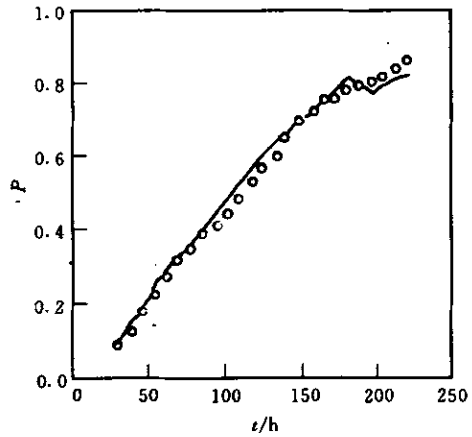

图6 分段多元回归检

Fig.6 Segmented multivariate regressive rectification

进行产物浓度预测,取得了较好的效果。

参 考 文 献

[1] Mark R. Warnes, Jarmila Glassey, Gary A. Montague *et al*. *Process Biochemistry*, 1996, **131**(2), 147～155.

[2] A. Sabten, G. L. M. Koot, L. C. Zullo. *Computers Chem. Engng*., 1997, 21, Suppl., sll23～sll29.

[3] 方开泰.《实用多元统计方法》,上海:华东师范大学出版社,1986.

[4] 周纪芗.《回归分析》,上海:华东师范大学出版社,1993.

[5] Michael H. Kaspar, W. Harmon Ray. *AIChE Journal*, 1992, **38**(10), 1593～1608.

[6] J. Kresta, J. F. MacGregor, T. E. Marlin. AIChE Meeting, 1989.

[7] Paul Nomikos, John F. MacGregor. *AIChE Journal*, 1994, **40**(8): 1361～1375.

Multivariate Statistical Analysis for Streptomycin Fermentation

Chen Yuanqing Chen Qi Wang Shuqing

(*National Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou* 310027)

Abstract Principal Component Analysis, a method of multivariate statistical analysis, is used for analyzing the datum of actual industrial streptomycin fermentation. This method is ideal for analysis of large multivariate datasets with highly correlated and noisy measurements by compressing it into a lower dimension space which contains most of the variance of the original matrix. A rule-based identification system is developed for dividing streptomycin fermentation process into three phases. The method built by multivariate regressive method can predict the production concentration of streptomycin fermentation according to three different phases. Good result is obtained.

Key words Principal component analysis, streptomycin fermentation, multivariate regressive method