模式识别法在化工调优中的应用

程兆年* 汤锋潮 罗学才 张未名 陈念贻

(中国科学院上海冶金研究所，上海)

摘 要

模式识别调优的基本出发点是，以工艺参数为特征变量构筑模式空间，按调优目标划分样本的类别。采用模式特征抽提方法压缩工艺参数，找出影响目标的主要因素。分别对两类调优问题，即最优指标问题和最优方向问题，提出了寻找最优工况的具体处理方法。对多目标调优也作了简单的讨论。

一、引 言

模式识别一词具有广泛的含义。本文涉及的模式识别方法，是从空间区域划分和属性类别判断角度出发，处理多元数据的一种非函数方法。七十年代初，Isenhour和Kowalski开创性地将模式识别方法引入化学领域，处理谱分析数据获得成功[1,2]。此后，模式识别在化学领域逐渐得到广泛应用，并已成为化学计量学的重要研究内容之一[3-5]。

设有属于若干类别的N个样本，不同类别的样本具有不同的属性，设每个样本均有k个特征变量描述，称k个特征变量构成的k维空间为模式空间。本文涉及的模式识别基本原理是：如果样本的属性和描述样本的特征变量之间存在着对应关系，则不同类别的样本就处于模式空间中的不同区域。

近年来，笔者和合作者曾利用模式识别方法预报并发现了一批新的金属间化合物[7,8]。本文在以前工作的基础上，讨论将模式识别应用于工业调优。模式识别作为一种有效的信息分析手段，尤其有希望在产品质量控制和最佳工艺条件选择诸方面获实际效益。

二、模式识别在工业调优中的应用

本文涉及的工业调优是指优化工艺参数使生产过程达到预期的指标或希望达到的指标。

1978年，Albano等人发表题为“模式识别的四个水平”的文章[9]。第一水平是将样本分类成为若干确定类别之一，找出分类判据，判断未知样本属哪一类。第二水平是在第一水平基础上构筑各类样本在模式空间的区域边界，以确认样本属性或否。第三水平是第二水平加上一种能力，使样本的可测特征变量能与其外部（指模式空间之外）的一个连续性质相关，因此可由特征变量预报样本的外部性质。第四水平与第三水平的差别是：外部性质不再是一个，而是多个；多个外部性质构成了模式空间之外的性质空间，进一步建立起性质空间与模式空间的对应关系。为明显区别于性质空间，也常将模式空间称为特征空间。

1988-12-19收到初稿，1989-06-22收到修改稿。

* 通讯联系人。

生产指标可视为第三水平的外部性质。模式识别应用于工业调优的基本出发点是以生产工艺参数为特征变量构筑模式空间。设生产工况由 k 个工艺参数表示，组成 k 维模式空间。一种工况即一个样本，对应于 k 维模式空间中的一个点，称为样本点。以调优目标作为划分样本类别属性的依据。将指标分为若干档次，每一档次对应一种类别。

统计回归调优建立函数关系

$$U=U(x_1, x_2, \cdots, x_k) \qquad (1)$$

式中U为目标，$x_1, x_2, \cdots, x_k$为工艺参数。式（1）中出现的系数通过回归方法得到。统计回归调优已广泛应用于化工生产，并为实践证明是提高化工生产技术水平的一种有效手段。但在正常生产条件下，目标值可能仅在某一水平（正常生产要求达到的指标）附近波动。有时在所得样本集中，由统计回归得到的均方差S_R并不比目标值本身的均方差（即所谓零个特征变量下的均方差）S_T小很多，甚至很接近，此时统计回归调优不易奏效。其原因是目标变量U和特征变量$x_1, x_2, \cdots, x_k$都有一定的误差，误差的概率密度分布可能不满足正态分布。而本文讨论的模式识别调优采用交互模式识别方法，通过人机交互，结合人眼对映射结果的观察判断进行，从而避免了对误差概率密度分布知识的要求。

模式识别方法应用于工业调优大致有两个步骤：(1)特征抽提找出影响目标的主要变量；(2)寻找最优工况。后一步骤又可分为最优指标和最优方向两类问题。

三、应用中需解决的问题

1．特征抽提

为全面描述一生产过程，首先需大量引入描述生产工况的工艺参数，建立描述样本尽可能多的特征变量，然后为在生产中容易控制，需在众多的工艺参数中找出影响目标的主要因素，亦即抽提出尽可能少的特征变量而得到良好的分类结果。

特征抽提很大程度上影响识别和优化的水平。统计模式识别中特征抽提的一些常规方法，如逐步判别分析[10]等，均可在调优中用来提取主要特征参量。

一个简单的特征抽提方法在化工调优中很有实用性，称之为DDL(Distance from Discriminant Line)方法。要点如下：将生产工况分为优和非优两种类别，分别记为1类和2类。将全部样本通过Sammon建议的最优判别平面(ODP, Optimal Discriminant Plane)方法[11]线性映射到二维平面，在该平面上优与非优至少需有分类趋势。在此平面上按目测划一直线，称之为判别线，使两类样本大体处于判别线两侧，(图1)。通过ODP逆运算可得该判别线方程，并得到各样本点离判别线的距离D为

$$D=a_0+\sum_{i=1}^{k}a_ix_i \qquad (2)$$

式中对应于 i 特征变量的系数a_i和两类样本平均值差 $(\bar{x}_i^{(1)}-\bar{x}_i^{(2)})$ 的乘积绝对值$|a_i(\bar{x}_i^{(1)}-\bar{x}_i^{(2)})|$记为$P_i$。$P_i$的相对大小，反映 i 特征变量在D表达式中的作用，亦即反映了对分类的贡献。保留m个对应于较大P_i的特征变量，其余较小的舍去，则 k 维特征空间压缩为m维。取m维特征变量重新作ODP映射进行验证，以仍得到良好的分类效果为准。

DDL方法的基本考虑与类间散度特征选择法[12]相近，不同之处是DDL方法中的判别线按人机交互方式给出。

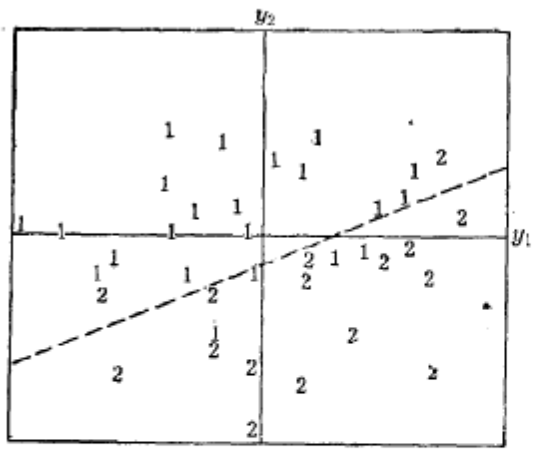

图 1　判别线一例（图中虚线为判别线，由人眼观察判断给出）
1—优类；2—非优类

2．最优指标

二维映射平面上的一个映射点，可对应高维特征空间中的无限多样本点(生产工况)。如何从二维映射回复到高维特征空间，从而得到最优工况，是模式识别调优需要解决的一个重要问题。从低维到高维，必须引入合理的条件才能得到确定解。这一条件可从实际生产过程产生的优类样本（生产结果较好者）有关信息中获取。

生产中常希望控制某一指标在某一特定值附近，如聚合过程往往要求产品粘度稳定在一个很小的范围。最优指标问题要求找出使指标稳定在最优值附近的工况（一组特征变量的取值）。有时这组特征变量中存在一部分不可控变量，则要求在不可控变量变化时，回答可控变量如何随不可控变量作相应改变。

控制质量指标是典型的最优指标问题。Kowalski早就指出，用模式识别解决材料生产问题是一个很有得益的潜在领域[12]，并且早就探讨了模式识别应用于质量控制的可能性[13]。实现模式识别第三水平的SIMCA方法，也可用来处理最优指标问题。

笔者及合作者曾作了应用模式识别控制顺丁橡胶门尼（ML）值的研究[14]，使用超平面识别模型方法建立了控制工艺参数的数学模型，并实现了优化控制。以下从另一角度（不同于文献[14]，更为清晰合理，也为了和下文衔接）简述最优指标问题超平面模型方法要点。(1)先将全部样本分为高于和低于最优指标两类进行特征抽提，得m个特征变量。(2)然后从全部样本中取出指标落在最优值附近（一个允许范围内）的一部分样本组成优类样本集。对优类样本集进行主成分分析(PCA, Principal Component Analysis, 例如,参见文献[10])，得优类样本集的一系列主坐标轴，以本征值大小为序分别用，$y_1, y_2, \cdots, y_m$表示。优类样本集在y_1方向伸展最宽，y_m方向最窄。故在y_1-y_m平面上的映射点分布于一窄长区域。如果以优类中心（平均值）为坐标原点，则分布在y_1轴附近。(3)最后，作为建模前的验证核实，余下的样本再分为指标过高和过低两类，将这两类非优样本也映射到优类样本产生的y_1-y_m平面。由于特征抽提时已引用指标信息，故过高类和过低类应大致分布于y_1轴两侧。如过高类和过低类没有分处y_1轴两侧，则需重新考虑特征抽提，直至出现两侧分布。

图2给出实现两侧分布一例：顺丁橡胶聚合的映射结果[14]，优类样本聚集在m维模式空间的一个超平面附近。该超平面可按PCA本征矢通过原始变量$x_1, x_2, \cdots, x_m$表示

$$y_m(x_1, x_2, \cdots, x_m)=0 \qquad (3)$$

如表示为超平面附近一厚度d的薄层

$$-d \leqslant y_m(x_1, x_2, \cdots, x_m) \leqslant d \qquad (4)$$

则就给出了优类样本的区域边界。模式空间中的不同区域已与外部性质——指标相关，属Albano的第三水平，一薄层对应了指标的一小区间。

为使生产达到预定的指标值，需建立各特征变量$x_1, x_2, \cdots, x_m$间满足的数学关系。在对应于最小本征值的y_m轴上，反映样本投影值的式(3)和式(4)给出了这一关系。因此式(3)和

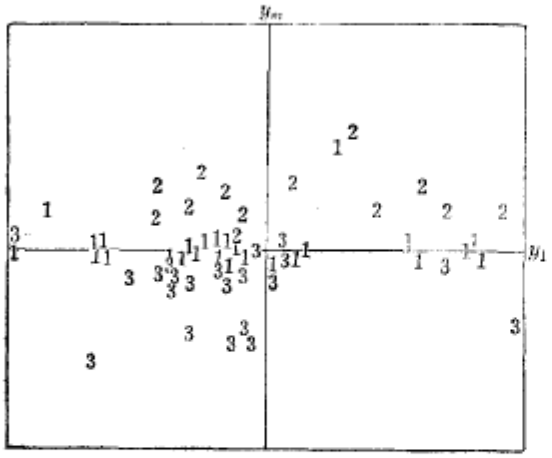

图 2　顺丁橡胶聚合的映射结果（取自文献[14]）
1—优级品；2—ML高；3—ML低

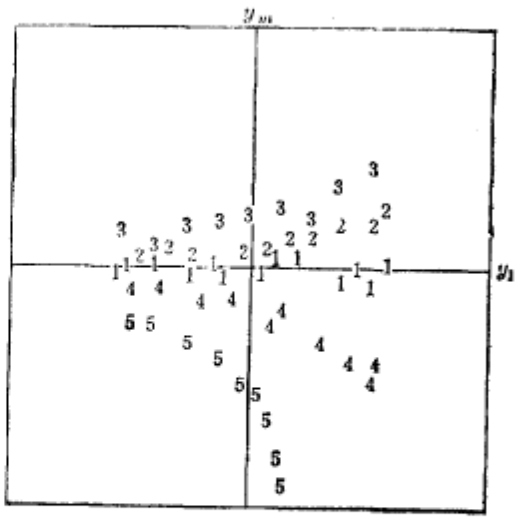

图 3　y_1-y_m平面上等指标分布例
1—最优；2,3—偏高；4,5—偏低

式(4)即控制生产指标的数学模型。至此，超平面方法解决了回复到工艺参数空间建模问题。

应该指出，优类样本点在y_m轴上的投影值与指标有确定的对应关系，但非优样本投影值一般与指标没有很好的对应关系。设想一个类别较多的样本集（将指标划为较多的档次，每一档次对应一个类别），在干扰甚小时，则可能出现图3所示的等指标分布。此时统计回归会给出很好的结果，统计回归建立的模型会更全面、更有效。但是为建立数学模型，就要求有较多的非优样本，有时会遇到困难。实际生产总希望生产优级品，生产非优品经济上会受损失。模式识别调优通过对优类样本集PCA建模有可能避免这一困难，并且已实际使用获显著经济效益[14]。

3．最优方向

实际生产过程常要求某一指标尽可能高或低，如要求生产成本越低越好。要回答的问题是：工况(一组工艺参数)应如何改变才能靠近所希望的目标。这就要求预报一个优化工况，也称为优化点。

预报模式空间中优化点的坐标$x_1, x_2, \cdots, x_m$，最简单的方法是将全部样本按指标划分为优与非优两类，取优类样本集中心（平均值$\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m$）作为预报值。进一步的改善，可在优类中心和非优中心连线方向前跨一小步作预报优化点。生产实施后得新样本点加入原样本集再作预报加以修正，如此循环进行。

一个稍复杂的方法称之为FOP(Forecast for Optimal Point)方法。为便于看出该方法的要点，以三维情形作为例子加以说明。

设样本由三特征变量x_1、x_2、x_3描述。如图4所示，先将全部样本按指标值分为优与非优两类，用Sammon方法映射到最优判别平面z_1-z_2

$$z_1=z_1(x_1, x_2, x_3) \qquad z_2=z_2(x_1, x_2, x_3)$$

在z_1-z_2平面上可见优与非优样本映射点的分布，图中上方是实际样本集。同时，优类中心(图中⊙)和非优类中心（图中⊗）也已映射到z_1-z_2平面。通过人机交互，目测估计预报优化点在z_1-z_2平面上的位置，如图中●所示，对应的坐标记为$z_1=\xi$，$z_2=\eta$。

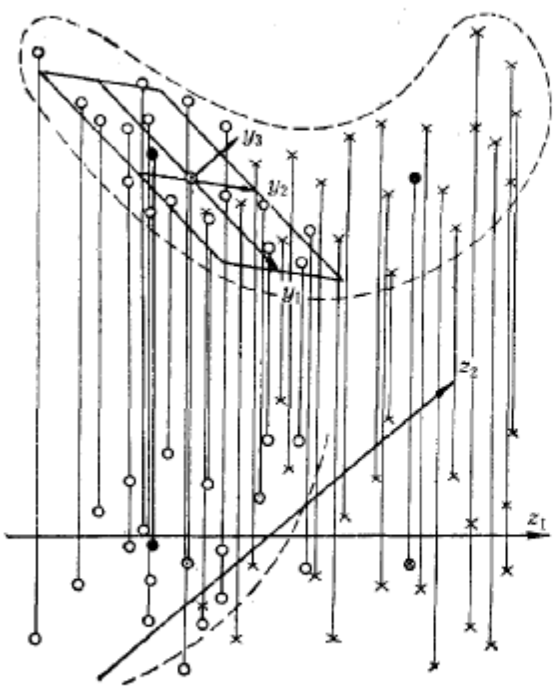

图 4　FOP方法示意图

○-优类样本; ×　非优样本; ⊙　优类中心;
⊗　非优中心; ●　预报优化点

过优化点垂直于z_1-z_2平面的直线上所有点，映射结果均在优化点。但实际生产并不经历这一无限长直线上的大多数点，即并非垂线上所有的点所对应的x_1、x_2、x_3在生产中都有实际意义。实际上，大量的点对应的工艺参数x_1、x_2、x_3在生产中达不到或不存在。仅仅是穿过样本集或处于样本集附近的一部分点才有实际意义，用作预报比较可靠。为了合理地找出一个实际工况的优化预报点，也必须引入优类样本的信息。方法如下：(1)对优类样本作主成分分析，以优类样本中心为原点得优类主轴y_1、y_2、y_3(以本征值大小为序)。如图所示，优类样本分布在y_1-y_2平面附近。(2)取垂线与y_1-y_2平面的交点为预报优化点，则预报点对应的x_1、x_2、x_3必定可以很好地与优类样本实际工况相配。

计算预报优化点对应的x_1、x_2、x_3值的方程为

$$\begin{cases} z_1(x_1, x_2, x_3)=\xi \\ z_2(x_1, x_2, x_3)=\eta \\ y_3(x_1, x_2, x_3)=0 \end{cases} \tag{5}$$

式中$y_3(x_1, x_2, x_3)=0$即y_1-y_2平面满足的方程。

同理，推广到m维。一般m维情形求解预报优化点的方程为

$$\begin{cases} z_1(x_1, x_2, \cdots, x_m)=\xi \\ z_2(x_1, x_2, \cdots, x_m)=\eta \\ y_3(x_1, x_2, \cdots, x_m)=0 \\ y_4(x_1, x_2, \cdots, x_m)=0 \\ \quad\cdots\quad\quad\cdots \\ y_m(x_1, x_2, \cdots, x_m)=0 \end{cases} \tag{6}$$

如x_1, x_2, …, x_m中有不可控变量，FOP方法亦能处理。设x_1为不可控变量，测得$x_1=v$，可在式(6)中去掉一个方程，一般地，去掉$y_3(x_1, x_2, \cdots, x_m)=0$，代之以$x_1=v$，解得的预报优化点在$z_1$-$z_2$平面上的映射位置保持不变。

FOP方法在生产中可循环使用。每次跨一小步，逐步使生产实现最优。

4. 多目标

多目标问题接近模式识别的第四水平。如同用模式识别方法建立模式空间和性质空间的对应关系，在调优中可建立模式空间和目标空间的对应关系。

但调优问题可不必涉及全部模式空间和全部目标空间之间建立对应关系。就实用来说，只需找出模式空间中一部分满足多最优目标的区域。取多项目标均优的样本组成优类样本

集，即可找出最优指标问题中的优指标方程和最优方向问题中的优化预报方程。于是，单目标问题中的方法可简单地推广到多目标问题。

四、应 用 实 例

在乳液法聚氯乙烯聚合反应过程中如何控制温度是工艺参数优化的主要目标。

由于聚合是间歇的，取一釜为一个工况，即一个样本。将整个聚合过程中每次加料之间的平均温度取为特征变量，共24个平均温度，构成24个特征。

特征抽取采用DDL方法。用以寻找影响生产的主要因素。数据处理表明，诱导期温度对反应进行有重要影响。诱导期共有六个温度参数$T_1, T_2, \cdots, T_6$，分别表示六个时间段(每时间段加入一批单体）的平均温度。衡量诱导成功水平的指标为聚合反应时间T_c，因为缩短T_c不仅意味着提高生产效率，而且意味着不出现明显爆聚，聚合反应稳定，产品质量稳定。

以六个温度参数为新的特征，以聚合反应时间为目标，多元线性回归给出

$$T_c = 50.21 + 0.3508T_1 - 0.0759T_2 - 1.2772T_3 + 0.1349T_4 + 0.0520T_5 + 0.1612T_6$$

$S_R = 1.526$，但$S_T = 1.698$，很接近。故此时回归难以奏效。

使用FOP方法，将所有样本按目标分类后映射到最优判别平面，在最优判别平面上目测一优化预报点，求解式(6)后在生产中实施。图5表示优化预报点给出的温度控制曲线，为比较，也列入了优类样本平均值给出的温度控制曲线。

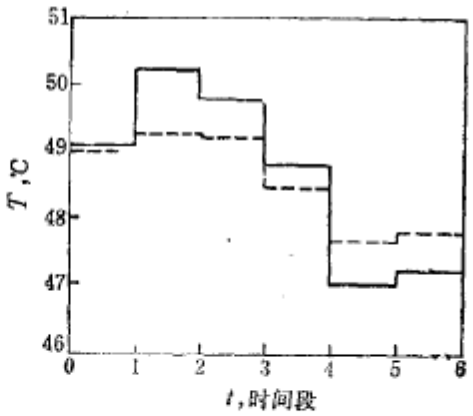

图 5　诱导期温度控制曲线

—— 优化预报；--- 优类样本平均

生产实施后产生新的样本，与原先样本一起作映射，再目测新的优化预报点，求解后在生产中再实施。

通过一个阶段对调优方案的实施，在不影响正常生产的前提下，有效地缩短了反应时间，减小爆聚出现率，对稳定生产、提高产质量起到了很好的作用。调优前(1985.12－1986.1)与调优后同期（1986.12—1987.1）相比，聚合时间平均缩短1.1小时，爆聚出现率显著下降，由17%下降为3.5%，质量稳定，产量提高。

五、小　　结

模式识别方法基于同类模式分布于高维模式空间邻近区域的原理，通过确认优类模式占据的空间区域找出优化决策。主要用于稳定生产波动。模式识别调优基于现有生产经验，但循环进行也可使工艺参数超出原有经验范围，使生产达到新的水平。

模式识别调优的一大特点是利用人机交互方式，充分利用人眼对平面分布的直觉判断能力，借以避免对误差概率密度分布知识的要求。但模式识别本着重于“识别”，得到映射结果就完成“识别”。而映射不是调优的目的，调优要求给出优化决策。为此，研究了寻找最优工况，亦即映射结果复原为生产工艺参数问题，提出并讨论了在复原中引入优类样本信息

的具体实现方法。

模式识别调优用作给出优化决策，在生产过程中起离线开环指导的作用。由模式识别调优给出的结果，如工艺参数之间的数学关系，也可应用于闭环反馈，作为在线优化控制的一个环节。

模式识别调优已累见实效，且很有前景。可视为一种调优方法，在S_R接近S_T时作为统计回归调优的补充。

参 考 文 献

〔1〕 Isenhour, T. L. and Jurs, P. C., *Anal. Chem.*, 43, 20A (1971).
〔2〕 Kowalski, B. R., *Anal. Chem.*, 47, 1152A (1975).
〔3〕 Kowalski, B. R., *Anal. Chem.*, 52, 112R (1980).
〔4〕 Frank, I. E. and Kowalski, B. R., *Anal. Chem.*, 54, 232R (1982).
〔5〕 Delaney, M. F., *Anal. Chem.*, 56, 261R (1984).
〔6〕 Ramos, L. S. Beebe, K. R., Carey, W. P. et al., *Anal. Chem.*, 58, 294R (1986).
〔7〕 陈念贻、谢雷鸣、施天生等，中国科学，7，836(1981)。
〔8〕 江乃雄、李庆芝、陈念贻等，中国科学，8，987(1982)。
〔9〕 Albano, C., Dunn, W. J., Edlund, U., et al., *Anal. Chem. Acta*, 103, 429(1978).
〔10〕 Enslein, K., Ralston, A. and Wilf, H. S., Statistical Methods for Digital Computer, Vol. 3 of Mathematical Methods for Digital Computer, John Wiley and Sons, Inc., 1977.
〔11〕 Sammon, J. W., *IEEE Trans. Comput.*, C-19, 826(1976).
〔12〕 钱学双，自动化学报，12，10(1986)。
〔13〕 Kowalski, B. R., *Chem. Tech.*, 4, 300(1974).
〔14〕 张未名、陈念贻、李再绵等，自动化学报，15，1(1989)。

Pattern Recognition in Process Optimization

Cheng Zhaonian, Tang Fengchao, Lou Xuecai, Chang Weiming and Chen Nianyi

(Shanghai Institute of Metallurgy, Academia Sinica, Shanghai)

Abstract

Pattern recognition method has been applied to solving the problems of the optimization of chemical processes. The basic starting-point of optimization by pattern recognition is that the technological parameters are used as feature variable to construct the pattern space and all samples are divided into a number of classes according to the production target to be optimized. In order to find out the key factors influencing the target, the method of feature extraction is adopted to reduce the dimensionality of the pattern space of technological parameters. Some specific methods are suggested to search for optimal production conditions for two kinds of optimization problems, i. e., the problem of optimal index and the problem of optimal direction, respectively. The problems of multiple targets are also discussed briefly. The methods of process optimization by pattern recognition have been applied to some factories and considerable technical-economic benefits have been achieved.