

文章编号 1000-5013(2005)01-0076-04

# 样本筛选与操作优化的可视化方法实现

华 丽 肖美添

(华侨大学材料科学与工程学院, 福建 泉州 362021)

**摘要** 可视化方法通过将多维空间数据映射到二维平面上,依据“物以聚类”属性对二维平面中各样本点进行分类、识别,找出离群样本并将其剔除掉.与此同时,产生目标函数等值线.沿着等值线值增大或减小的趋势方向寻优,就很容易地确定出最优点或最优化操作区域.以某卷烟厂生产烟丝为例进行分析和计算.结果表明,文中筛选样本数据方法中,离群点的剔除提高优化结果的准确度,最佳操作条件的确定可为工业生产或科学实验提供决策支持.

**关键词** 可视化方法, 样本筛选, 离群样本, 操作优化, 烟丝

**中图分类号** O 232

**文献标识码** A

工业生产和科学实验中,数据在不断堆积,这些数据有些是有用的,有些是无用.无用数据来源主要有以下 3 种. (1) 测量、记录不准. (2) 无代表性数据的混入. (3) 额外因素干扰. 显然,当采用这些不可靠或无代表性的数据总结规律,势必会导致严重失误. 因此,如何剔除无用数据,利用有用数据来挖掘信息,从中找出生产或科研工作中规律性的东西,是数据处理工作中的一个重要环节. 这一步处理得好坏,关系到后继工作的顺利进行. 目前数据处理通常采用的方法是模式识别法<sup>[1]</sup>. 这种方法是应用 PCA<sup>[2]</sup>或偏最小二乘法等投影方法,来实现多维空间样本数据向二维平面的映射. 然后,将此平面分为好、中、坏区,通过选择好区样本进行操作优化. 这种方法缺点是,PCA 主要是实现线性映射,非线性映射能力很差. 而偏最小二乘法则是在投影时保证样本点间距的平方和最小,在加和平方时可能会使数据空间信息失真,结果也会导致严重不准. 鉴于上述方法的不足,有必要研究一种新的数据处理方法.

## 1 结构模型

此方法原理是采用降维映射<sup>[3]</sup>技术,在保证数据的空间拓扑结构不变的原则下,将多维空间样本数据非线性映射到二维平面上. 然而,依据“物以聚类”属性对二维平面中样本数据进行分类、识别,找出离群样本并将其剔除掉. 与此同时,产生目标函数等值线,沿着等值线值增大或减小的方向操作,还可实现对目标函数的优化. 它是集数据筛选和目标优化于一体的方法,详细原理见文献[4].

多维空间向二维平面的非线性映射,是通过建立一种特殊设计的三层前馈 BP 网络模型来实现的,如图 1 所示. BP 网络的误差反向传播特性,使其具有自检测、自记忆及自容错等能力,特别适用于模式识别、复杂的非线性函数关系的拟合与映射等. 它是从大量实验数据中总结经验规律的有力手段. 文中对这种网络稍作了改进,主要体现在对网络的权值调节,采用的是二级调节形式. 二级调节作用是尽可能使误差随传播过程的延续而逐渐减小,直至最后达到精度要求,从而保证在映射过程中空间信息不会失真. (1) 设输入矢量为  $X$ ,  $Y$  为实验的目标函数,  $X_i = [x_{i1} \ x_{i2} \ x_{i3} \ \dots \ x_{im}]^T$ , 每个分量对应一种组成因素. 设有  $m$  种组成因子,当有  $n$  组数据时,  $i = n$ . 模型输出为  $Y_i$ , 并确定中值  $Y_0$  和相对误差限,输入矢量首先在一级权向量  $W$  作用下映射到二维平面  $z_1, z_2$  上. 其中  $Z_1 = W_1 X_i$ ,  $Z_2 = W_2 X_i$ ,  $W_1 = [w_{01} \ w_{11} \ w_{21} \ \dots \ w_{m1}]$ ,  $W_2 = [w_{02} \ w_{12} \ w_{22} \ \dots \ w_{m2}]$ . (2) 在权系数  $v$  作用下,使非线性扩展矢量

收稿日期 2004-05-09

作者简介 华 丽 (1974-), 女, 助教, 硕士, 主要从事化工系统工程的研究. E-mail: txuehua @163. com

F 进行 2 次拟合,当然也可以 3 次.其分量  $f$  基于 S 型函数,显然要比一次线性逼近效果好.但并不是拟

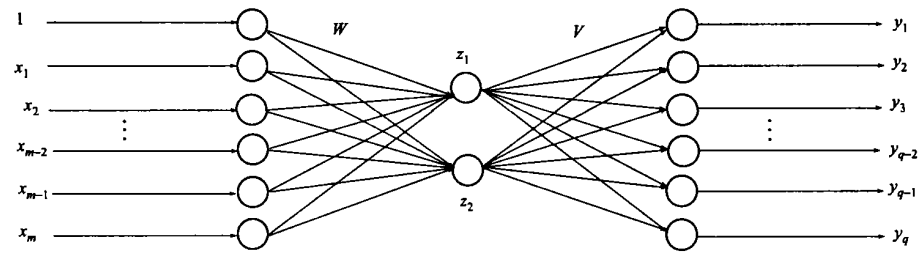


图 1 三层神经网络结构模型

合次数越高越好.因为次数越高,反而会使误差叠加增大.2 次或 3 次较为合适,最后到输出  $Y_t$ . 有

$$Y_t = [y_1 \quad y_2 \quad y_3 \quad \dots \quad y_q], \quad Y_t = VF, \quad Y_0 = Y[(y_{\max} + y_{\min})/2],$$
$$F = [1 \quad f(z_1) \quad f(z_2) \quad f(z_1 z_2) \quad f(z_1^t) \quad f(z_2^t)]^T, \quad t = 2, 3, \quad f(x) = 1 + e^{(-x)}$$
$$V = [V_1 \quad V_2 \quad V_3 \quad \dots \quad V_f]_{q \times f}, \quad V_k = [v_{k1} \quad v_{k2} \quad v_{k3} \quad \dots \quad v_{k(f+1)}], \quad k = 1, 2, 3, \dots, q.$$

(3) 网络训练开始时,需对各级权向量设初值.当无任何先验知识时,其分量可取某一实数范围内具有等概率分布的随机值.(4) 依次输入样本  $X_i$   $X$  及其对应实验目标函数值  $Y$ ,与结果  $Y_t$  进行比较,误差反向处理.有

$$E = \min \frac{1}{2} (Y - Y_t)^2.$$

(5) 采用一种优化算法来训练网络.(6) 判断所有的样本是否训练完一轮,若未完成一轮,则返回(1). (7) 判断每次输出误差中最大值是否超越误差限,若超越误差限,则返回(1);反之,则输出结果并产生目标函数等值线.此时,整个过程结束,用数据结果来查询所得的性能指标满足精度要求.

以下为离群样本的判别法则.(1) 依据“物以聚类”的属性,异常样本往往表现为“离群样本”.当某区域主要为 A 类样本,却有少量 B 类样本混入时,可认为少量 B 类样本为异常样本;反之亦然.(2) 当某一个或极少数样本远离群体且目的指标并不是极大或极小时,以中值  $Y_0$  为定界线分为两类,在映射平面上分别记为“A 或 \* ”和“B 或 o ”表示.规定当样本数据  $X_i = [x_{i1} \quad x_{i2} \quad \dots \quad x_{im}]^T$  的目标指标值  $Y_i$   $Y_0$  时,  $X_i$  “A 或 \* ”类;当  $Y_i < Y_0$  时,  $X_i$  “B 或 o ”类.然后,运用人工神经网络将样本数据降维映射到二维平面上,直接观察出“A 或 \* ”与“B 或 o ”的聚类情况,即可判别出离群样本而剔除掉离群点.

值得注意的是,由于映射到平面上的映射方向是随机的,每次映射在平面上的图像并不一定相同,即“A 或 \* ”和“B 或 o ”类的分布区域是有变化的,对可疑样本有必要进行两次映射.若连续两次或两次以上都表现为离群样本,则可将其剔除掉.

2 应用实例

某卷烟厂生产烟丝,为了提高烟丝产率,做了大量实验,实验数据见文献[5].碎丝率是一个重要经济指标,降低碎丝生产率,就可大大提高烟丝产率.其中  $t$  为目标函数,即碎丝生产率,其值愈低愈佳.结合工程背景以碎丝生产率小于 5.4 %者为优类样本.为了使生产的产品质量高,就必须尽可能地使碎丝率控制在低于 5.4 %以下.因此采用可视化方法来寻找离群样本,并将其剔除掉,进一步为优化操作作准备.影响目标函数碎丝率( $t$ )主要因素是温度( $x_1$ )、一次储存时间( $x_2$ )、一次储存后含水量( $x_3$ )、二次储存时间( $x_4$ )、二次储存后含水量( $x_5$ )、第 1 道工序(一次储存工序)的含水量( $x_6$ )、第 2 道工序(二次储存工序)的含水量( $x_7$ ).用可视化方法对这些数据进行处理,将这组数据以矩阵形式输入给计算机,在 MATLAB 操作平台下编写离群样本筛选和操作优化程序语言,得出相应的输出结果.

3 结果与讨论

对数据进行模拟计算,降维映射输出各个样本点,如图 2 所示.图 2(a), (b), (c), (d) 都有两类样本聚集区,即“ \* ”类聚集区和“o ”类聚集区.当目标函数值小于或等于 5.4 %时,为“ \* ”类聚集区;而当目

标函数值大于 5.4 % 时,为“o”类聚集区.从图中可看出,两类样本之间有混杂.说明原始实验中,由于某

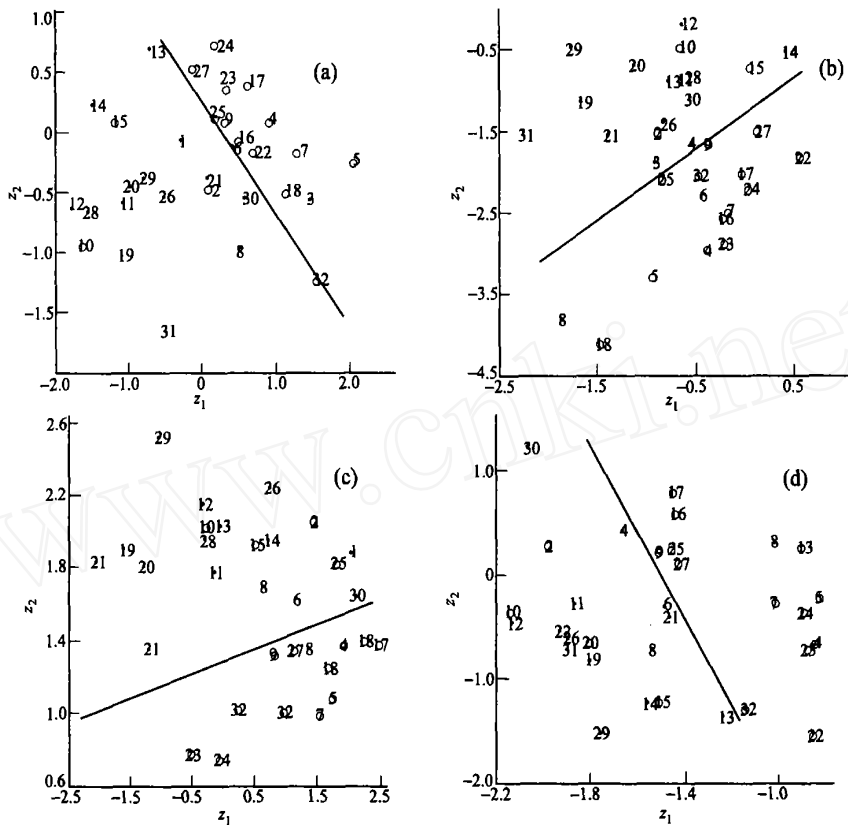


图 2 离群样本的识别

种原因导致有些数据测量不准,误差很大,用这些不准的数据来总结规律,势必会带来很大偏差.因此,应先根据离群样本判别法则,找出离群样本.其中图 2(a)中“\*”类样本聚集区中有 3 个离群样本点(2 #, 10 #, 15 #),而“o”类样本聚集区有 1 个离群样本点(8 #).图 2(b)中离群样本为 2 #, 10 #, 15 #, 8 #, 6 #;图 2(c)离群样本为 2 #, 10 #, 15 #, 25 #, 8 #;图 2(d)离群样本为 2 #, 10 #, 15 #, 8 #.综合 2,可知,2 #, 10 #, 15 #, 8 # 4 个点为离群样本点.为了使优化结果更准确、可靠,应当将离群样本剔除掉.图 3 为其剔除离群样本后映射结果.图中“\*”类样本全部聚集在右上方,两类样本之间没有任何混杂,样本的分类效果很好.说明原始点 2 #, 10 #, 15 #, 8 # 就是离群样本点.所以实验数据经过可视化方法处理以后,能有效地去除离群样本点,为后继优化工作作准备.图 4 为剔除离群样本后函数拟合曲线.从图可看出,计算值与真实值非常逼近.因此,离群样本的剔除能大大地提高优化结果的准确度,使模拟中的计算值与真实值之间误差达到最小.目标函数等值线如图 5 所示,样本点作了 1, 2, 3, ..., 28 标号,黑色实线为目标等值线.为了寻找优化方向,以第 7 #, 26 # 样本为基准进行预测,分别取步长为 1.2 或 1.3,得到图中的两个星号点.其预测结果,如表 1 所示.表中“\*”表示外推方向,为外推步长.

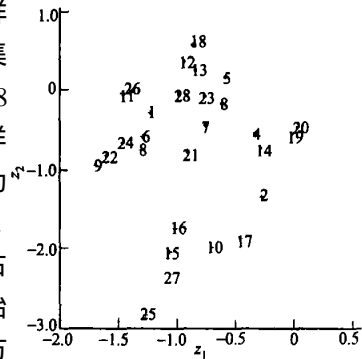


图 3 剔除离群样本后的映射图

表 1 中,第 1 个星号点的碎丝率  $t = 3.960\%$  比  $5.400\%$  小,说明其优化方向在等值线图中就是靠近下方稍微偏左的方向.越向下,其目标等值线值越小,生产的碎丝产率越低,烟丝产率越高,产品质量

表 1 外推预测结果

参照点	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$y$	$z_1$	$z_2$
7 26 1.2	18	1	29.64	5.50	29.42	32	12.28	3.960 7	- 0.177 6	- 1.867 7
7 26 1.3	17	1	29.76	5.25	29.33	32	12.27	3.777 8	- 0.184 3	- 1.935 4

量越好(如图中另一个星号点碎丝率  $t$  值为 3.777 8 %). 所以, 实际工作中应沿着等值线(值为 4.45)下方操作, 都符合工艺生产碎丝率低于 5.4 % 的要求. 这为卷烟厂提供了很好地决策支持.

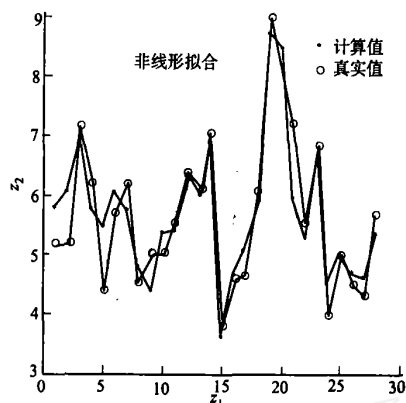


图 4 函数拟合图

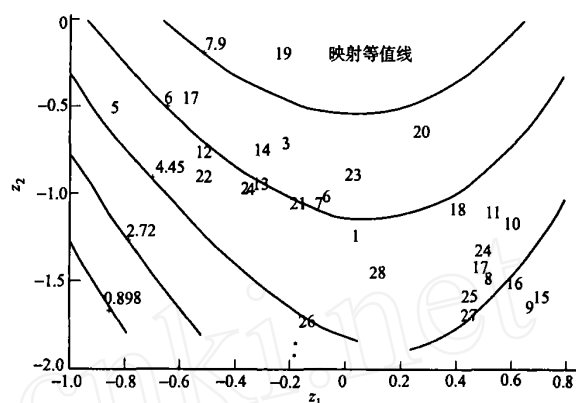


图 5 目标函数等值线分布图

## 4 结束语

由此可见, 可视化技术为筛选样本数据提供了一种新的途径. 特别是当各变量(因子)与目标之间关系不是十分明显, 要找出两者的定性关系, 需做大量实验, 产生大量数据时, 可视化方法处理大量数据是十分有效的. 而且离群样本去除之后, 能大大地提高优化结果的准确度, 这种集样本筛选与操作优化于一体的方法研究是一种新的发展方向.

## 参 考 文 献

- 1 陈念贻, 钦 佩, 陈瑞亮等. 模式识别方法在化学化工中的应用[M]. 北京: 科学出版社, 2000. 9 ~ 366
- 2 Chem J, Bondoni J A, Romagnol J A. Robust PCA and normal region if multivariate statistical process monitoring[J]. AIChE J., 1996, 42:3 563 ~ 3 566
- 3 鄢烈祥, 华 丽. 工业过程操作优化可视化方法——降维分析法[J]. 武汉理工大学学报(自然科学版), 2002, 24(7): 79 ~ 82
- 4 华 丽. 可视化优化方法及其应用的研究[D]: [学位论文]. 武汉: 湖北工学院, 2003. 5 ~ 27
- 5 陈念贻. 模式识别优化方法及其应用[M]. 北京: 中国石油化工出版社, 1997. 308 ~ 309

## Realizing Sieving Samples and Optimizing Operation by Visualized Method

Hua Li Xiao Meitian

(College of Material Science and Engineering, Huaqiao University, 362021, Quanzhou, China)

**Abstract** For realizing sieving samples and optimizing operation by visualized method, the authors map the data of multidimensional distribution onto two-dimensional plane; classify and recognize various sample points on two-dimensional plane, find the outliers and reject them. Mean while, the authors get the contour line of objective function; and search along trend and direction of increase or decrease of contour line value by which the optimum point or optimized operating zone can easily be determined. The production of cut tobacco in a cigarette factory is taken as example. As shown by reckoning, the rejection of outliers in sieving data of samples improves the accuracy of optimized results, the determination of optimal operating condition provides industrial production and scientific experiment with support of decision-making.

**Key words** visualized method, sieving samples, outliers, optimized operation, cut tobacco