正交递归选择法及其在转炉炉龄研究中的应用 *

朱尔一　杨芃原　邓志威　黄本立

(厦门大学化学系，厦门 361005)

摘　要　本文提出了一种新的变量筛选法 — 正交递归选择法，该法可以得到预报能力较强的模型，即 PRESS(预报残差平方和) 值较低的模型。转炉炉龄问题作为研究实例，并与逐步回归正向选择法及偏最小二乘回归法进行了比较。

关键词　变量筛选　PRESS 判据　逐步回归　PLS 回归　Gram-Schmidt 正交化

目前主成分回归 (PCR) 和偏最小二乘回归 (PLS) 在许多学科领域中有着广泛的应用，尤其 PLS 回归法最为流行。这两种方法在处理实际问题时能有效地避免复共线问题，并能得到预报能力较强的模型 [1] 。然而用这两种方法处理未经变量筛选的数据时，得到模型的预报能力往往不如逐步回归等变量筛选法求得的模型预报能力强 [2] 。

为得到预报能力较强的模型，变量筛选是十分重要的。目前通常可用的变量筛选法有逐步回归和所有可能回归两类。其中逐步回归中包括三种基本算法，正向选择法，逆向剔除法和逐步回归法 [3] 。在正向选择法中自变量被逐个引入模型，在逆向法中自变量从模型中被逐个删除，逐步回归法为前两种方法的组合，既有变量引入，又有变量删除。本研究工作表明这类方法在建立用于预报目的的模型时，特别当处理过拟合数据时往往不能找到最优模型，即以 PRESS(Prediction error of squares) 值作为变量筛选判据时，难以找到 PRESS 值为最低的模型。另外所有可能回归法处理问题时所需要的计算量较大，若自变量数为 n, 则所有可能的模型有 2^n 种，当 $n > 10$, 计算在普通微机上难以进行。

本文提出一种新的变量筛选法 — 正交递归选择法。该法不同于逐步回归，逐步回归中三种算法在寻找最优模型时，采用单方位搜索方式，即每次搜索在上一次选中变量上增加或删除一个变量。而本文提出的方法采用多方位搜索方式，即每次循环选出若干个候选变量，每个候选变量下次循环又产生各自候选变量。由于采用多方位搜索方式，使搜索范围增加，因此在实际应用中能较有效地找到最优模型。另外该法的计算量又大大小于所有可能回归法的计算量。正交递归选择法在计算中还用到了数据正交化分解，其目的一是便于变量筛选，二是可避免矩阵求逆运算以减少计算量。

另一方面本文采用 PRESS 值作为变量筛选判据，为了衡量对新样本点的预报能力，目前广泛使用 PRESS 判据 [4] 。本文通过处理转炉炉龄问题，将所提出的正交递归选择法与

本文于 1993 年 1 月 19 日收到初稿，1993 年 2 月 15 日收到修改稿。

* 国家博士后基金资助项目。

逐步回归中正向选择法及 PLS 回归法作一比较，结果表明正交递归选择法可得到 PRESS 值较低的模型。

一、原理与方法

1.1 PRESS 判据

变量筛选判据有许多种，但对衡量模型的预报能力，PRESS 判据较为灵敏。PRESS 判据值定义如下

$$\mathrm{PRESS}=\sum_{i=1}^{m}(y_i-\hat{y}_{-i})^2=\sum_{i=1}^{m}e_{-i}^2 \tag{1}$$

式中 m 为样本数，y_i 为第 i 个样本的目标变量值，$\hat{y}_{-i}$ 为第 i 个样本不参加回归时得到的模型对该点的预测值，e_{-i} 为预报残差。PRESS 值越低意味着模型的预报能力越强。在求 PRESS 值的实际计算中可用以下公式 [4]

$$\mathrm{PRESS}=\sum_{i=1}^{m}\left(\frac{e_i}{1-h_{ii}}\right)^2 \tag{2}$$

式中 e_i 为普通残差，$e_i=y_i-\hat{y}_i$，$\hat{y}_i$ 为所有样本参加回归得到模型的预测值，h_{ii} 定义为 Mahalanobis 距离。

1.2 Gram-Schmidt 正交化

在逐步回归法中对于各子模型所用的回归系统模型与多元线性回归用的模型是相同的。即

$$\mathbf{b}=(\mathbf{X^TX})^{-1}\mathbf{X^Ty} \tag{3}$$

$\mathbf{b}$ 的分量等于变量引入数。

在正交递归选择法中为避免矩阵 $\mathbf{X^TX}$ 求逆及便于变量筛选，在计算中对矩阵 $\mathbf{X}$ 采用了 Gram-Schmidt 正交化分解。Gram-schmidt 正交化过程为使一组不正交的矢量 $\mathbf{x}_1,\mathbf{x}_2,\cdots$ 经线性变换成一组两两正交矢量 $\mathbf{t}_1,\mathbf{t}_2,\cdots$ 的数学方法，其算式如下：

$$\mathbf{t}_1=\mathbf{x}_1\mathbf{t}_2=\mathbf{x}_2-P_{21}\mathbf{t}_1\vdots\mathbf{t}_i=\mathbf{x}_i-P_{i1}\mathbf{t}_1-\cdots-P_{ii-1}\mathbf{t}_{i-1} \tag{4}$$

其中为了使 $(\mathbf{t}_1^T\mathbf{t}_2)=0$，取 $P_{21}=(\mathbf{x}_2^T\mathbf{t}_1)/(\mathbf{t}_1^T\mathbf{t}_1)$，为使 $(\mathbf{t}_i^T\mathbf{t}_1)=\cdots=(\mathbf{t}_i^T\mathbf{t}_{i-1})=0$，取

$$P_{ij}=(\mathbf{x}_i^T\mathbf{t}_j)/(\mathbf{t}_j^T\mathbf{t}_j)\quad(j=1,2,\cdots,i-1;\quad i=1,2,\cdots,n) \tag{5}$$

经以上算式计算可完成矩阵 $\mathbf{X}$ 的正交化分解，即

$$\mathbf{X}=\mathbf{T}P^{\mathbf{T}} \tag{6}$$

其中 $P^{\mathbf{T}}$ 为上三角阵

$$P^{\mathbf{T}}=\begin{pmatrix}1 & P_{21} & \cdots & P_{n1}\\ 0 & 1 & \cdots & P_{n2}\\ & & \vdots & \\ 0 & 0 & \cdots & 1\end{pmatrix} \tag{7}$$

式 (6) 可写成

$$\mathbf{T} = \mathbf{X}(P^{\mathbf{T}})^{-1} = \mathbf{XR} \tag{8}$$

上式中 $\mathbf{R}$ 也为上三角阵。所求得模型的回归系数可表示为

$$\mathbf{b} = R(T^TT)^{-1}T^T\mathbf{y} \tag{9}$$

上式求得的回归系数与式 (3) 求得的相同，因为 $R(T^TT)^{-1}T^T = (X^TX)^{-1}X^T$ 。另外在求式 (2) 时， $h_{ii} = \mathbf{a}_i^T(X^TX)^{-1}\mathbf{a}_i, \mathbf{a}_i$ 为矩阵 X 的第 i 行矢量，可用 $h_{ii} = \mathbf{b}_i^T(T^TT)^{-1}\mathbf{b}_i$ 代替， $\mathbf{b}_i$ 为矩阵 T 的第 i 行矢量，注意到式 (6) 和 (8) 有

$$h_{ii} = \mathbf{a}_i^T(X^TX)^{-1}\mathbf{a}_i = \mathbf{b}_i(T^TT)^{-1}\mathbf{b}_i \tag{10}$$

因此在上式计算中 X^TX 求逆用对角阵 T^TT 求逆代替，使计算大大简化。

另一方面采用 Gram-Schmidt 正交化便于变量筛选。

1.3 有序 Gram-Schmidt 正交化

首先我们注意到 Gram-Schmidt 正交化分解有以下特性。若根据 (8) 式将矩阵 X 分解， R 为上三角阵，分解得到的第一个矢量或隐变量 $\mathbf{t}_1$，只与 $\mathbf{x}_1$ 有关，与其它变量无关，第二个矢量 $\mathbf{t}_2$ 只与矩阵 X 中前两个变量有关，第 i 个矢量 $\mathbf{t}_i$ 也只与 X 中前 i 个变量有关，与其它变量无关。因此矩阵 X 中变量不同的排列将得到不同的分解结果，若不限制 X 中变量的排序， Gram-Schmidt 正交分解不是唯一的。

若按以下顺序排列矩阵 X 中的变量，取 $\mathbf{t}_1$ 为 X 中与目标变量 $\mathbf{y}$ 的相关系数最大的变量，选 $\mathbf{t}_2$ 时，先按式 (4) 中 $\mathbf{t}_2 = \mathbf{x}_2 - P_{21}\mathbf{t}_1$ 计算，这里 $\mathbf{x}_2$ 在确定前为待选变量，而 $\mathbf{t}_2$ 表示 $\mathbf{x}_2$ 扣除掉与 $\mathbf{t}_1$ 重迭的部分，再求与 $\mathbf{y}$ 的相关系数，按其与 $\mathbf{y}$ 的相关系数为最大确定 $\mathbf{x}_2$，由此下去，对 $\mathbf{t}_1$ 先按式 (4) 计算，再求与 $\mathbf{y}$ 的相关系数，由相关系数为最大确定 $\mathbf{x}_i$，直到选出所需要的变量。在选出每一个变量步骤中都可按式 (9) 确定回归模型，按式 (10) 和 (2) 求得 PRESS 值。按以上步骤可完成正交化分解，可得到唯一的分解结果，这个结果中变量的排列顺序刚好与逐步回归正向选择法选出变量的顺序完全相同，以上分解过程称之为有序 Gram-Schmidt 正交化 [5] 。

在有序 Gram-Schmidt 正交化过程中按与目标变量 $\mathbf{y}$ 的相关系数大小排序，分解得到的隐变量 $\mathbf{t}, \mathbf{t}, \cdots$ 为两两正交的，同时又按与 $\mathbf{y}$ 的相关系数大小排列，与 $\mathbf{y}$ 相关系数最大的变量 $\mathbf{t}$ 排在第一个，其次排在第二，排在后面的变量与 $\mathbf{y}$ 的相关性较小，并认为这些变量含噪音成分多，一般这些变量会使模型的 PRESS 值升高，在变量筛选中可根据 PRESS 判据将这些变量删除。

然而，在处理实际问题时我们发现用有序 Gram-Schmidt 正交化法作变量筛选，有时并不能得到 PRESS 值为最低的模型。我们认为其原因是，变量与目标变量 $\mathbf{y}$ 的相关系数只能粗略地衡量变量含噪音的多少，当两变量与 $\mathbf{y}$ 的相关系数很接近时，就较难分辨哪个变量含噪音多或少。而我们希望正交分解得到的隐变量 $\mathbf{t}_1, \mathbf{t}_2, \cdots$ 按含噪音多少的顺序排列，排在前面的变量含噪音少，后面的变量含噪音多，以便于除去含噪音多的变量，为此目的我们提出了正交递归选择法。

1.4 正交递归选择法

用正交递归选择法处理问题时也先挑选与目标变量 $\mathbf{y}$ 相关系数大的变量，当与 $\mathbf{y}$ 相关系数最大的几个变量与 $\mathbf{y}$ 的相关系数大小差不多时，此时不仅是要用到相关系数最大的变量，而是这些相关系数大小差不多的变量也都要用到，这些变量都被归为候选变量，正交递归选择法基本计算过程如下：

首先挑选第一层变量，先计算 X 中所有变量与 $\mathbf{y}$ 的相关系数，选出若干个相关系数最大的变量为候选变量，可按下列条件

$$(\gamma_{\max} - \gamma_j)/\gamma_{\max} < 0.1 \tag{11}$$

式中 $\gamma_{\max}$ 为求出的最大相关系数，γ_j 为第 j 个变量与 $\mathbf{y}$ 的相关系数。

然后挑选第二层以下变量，当选第一层变量时，先按式 (4) 计算，即 $\mathbf{t}_l = \mathbf{x}_l - P_{l1}\mathbf{t}_1 - \cdots - P_{ll-1}\mathbf{t}_{l-1}$ 式中 $\mathbf{t}_1, \cdots, \mathbf{t}_{l-1}$ 为前面选出变量对应的隐变量 (有很多组)，$\mathbf{x}_l$ 为选剩下的变量。这样可求得每个 $\mathbf{x}_l$ 对应的 $\mathbf{t}_l$，并求出每个 $\mathbf{t}_l$ 与 $\mathbf{y}$ 的相关系数。再根据式 (11) 条件选出相关系数最大的若干个为候选变量。在挑选过程中，由于本层选出的候选变量是在上一层每一个候选变量上生成的，因此计算过程中可形成具有树形结构的许多变量组合，计算需要递归进行，最后可形成一棵树，其中树的每个分枝代表一中变量组合。

正交递归选择法计算过程可表示为一棵树的形成过程，先选出 $\mathbf{t}_1$ 的若干个候选变量，而每个候选变量又能生成下一层的候选变量，由此随着层数增加树随之而长大，而树的每一个分枝都可表示为一种变量组合，对于每种变量组合可根据式 (10) 和 (2) 求得一个 PERSS 值，通过比较可求出每一层 PRESS 值最低的变量组合。

正交递归选择法中需截去 PRESS 值较大的和变量组合相同的树分枝，以减少计算量。

在正交递归选择法中存在两种极端情况。一种是当每层计算中候选变量只取一个与 $\mathbf{y}$ 相关系数最大的变量时，其计算选出的变量顺序与逐步回归正向选择法所得到的变量顺序相同。另一种是当每层计算中候选变量取所有可能变量时，则计算与所有可能回归法相同。因此正交递归选择法可通过控制候选变量数目控制计算量，并希望用最小的计算量找到 PRESS 值为最低的模型。

二、转炉炉龄问题研究

某钢铁公司炼钢转炉的炉龄约为 1000 炉钢 / 炉，按 30 炉 / 天炼钢规模，大约一个月就需停炉一次，更换炉衬，每次更换炉衬需消耗大量资金。为减少消耗，厂方希望建立炉龄的预测模型，以便适当调节工艺参数以延长炉龄。厂方提供数据共 33 组 (样本)，每组数据中有 6 个自变量：$\mathbf{x}_1$ 喷补料量，$\mathbf{x}_2$ 吹炼时间，$\mathbf{x}_3$ 炼钢时间，$\mathbf{x}_4$ 钢水含 Mn 量，$\mathbf{x}_5$ 渣中含铁量，$\mathbf{x}_6$ 作业率，目标变量为炉龄 (炼钢炉次 / 炉)。由于自变量与目标变量间可能存在非线性关系，在建模时除了考虑各自变量线性因子外，还需考虑自变量的非线性因子，将自变量的平方项及二次交叉项全部参加建模，共有 27 个因子，为了便于计算将各非线性因子从新编号如下：

变量	非线性因子	变量	非线性因子	变量	非线性因子
x_7	x_1^2	x_{14}	$x_2\ x_3$	x_{21}	$x_3\ x_6$
x_8	$x_1\ x_2$	x_{15}	$x_2\ x_4$	x_{22}	x_4^2
x_9	$x_1\ x_3$	x_{16}	$x_2\ x_5$	x_{23}	$x_4\ x_5$
x_{10}	$x_1\ x_4$	x_{17}	$x_2\ x_6$	x_{24}	$x_4\ x_6$
x_{11}	$x_1\ x_5$	x_{18}	x_3^2	x_{25}	x_5^2
x_{12}	$x_1\ x_6$	x_{19}	$x_3\ x_4$	x_{26}	$x_5\ x_6$
x_{13}	x_2^2	x_{20}	$x_3\ x_5$	x_{27}	x_6^2

再用各种方法处理，对以上 27 个因子进行筛选以找出最优变量子集。

我们采用了 PLS 回归法，逐步回归正向选择法，及本文提出的正交递归选择法处理以上炉龄问题，其结果见表 1 。其中在 PLS 法中求出每引入一个正交分解产生的隐变量对应模型的 PRESS 值，结果见表 1 中的第一列。在逐步回归正向选择法中，每次引入一个自变量，并给出对应模型的 PRESS 值，结果见表 1 中的第二列，括号中的数为每次选入的自变量的编号。在正交递归选择法中，每个变量引入数对应一个 PRESS 值为最低的变量组合，结果见表 1 中的第三列，后面括号中的数为从所有候选变量组合中选出的 PRESS 值为最低的一组变量编号。

表 1 三种方法求得的 PRESS 值

变量数	PLS	逐步回归	正交递归
		正向选择	选择
0	15.80	15.80	15.80
1	9.65	11.94 (14)	11.94 (14)
2	9.97	10.85 (10)	10.33 (3,12)
3	9.72	11.05 (3)	9.97 (3,12,6)
4	9.98	11.37 (18)	8.60 (3,10,7,22)
5	11.17	12.19 (7)	6.86 (3,10,18,7,22)
6	11.94	5.51 (22)	5.51 (14,10,3,18,7,22)
7	10.43	6.16 (11)	4.99 (3,10,18,7,22,15,11)
8	7.91	5.44 (15)	4.83 (14,10,13,3,7,22,11,12)
9	7.20	5.61 (4)	4.82 (14,10,13,3,7,19,12,11,22)
10	7.20	6.53 (2)	4.54 (14,10,13,3,7,4,12,11,22,15)
11	7.64	7.62 (1)	4.63 (14,10,13,3,7,4,12,11,22,15,6)
12	6.31	11.11 (23)	5.05 (14,10,13,3,7,4,27,16,22,15,11,12)
13	13.05	10.56 (16)	5.02 (14,10,7,19,13,3,23,27,22,5,11,8,24)
14	12.80	9.08 (20)	5.43 (14,10,7,19,13,3,24,5,22,11,23,1,8,27)
15	11.25	9,91 (25)	6.20 (14,10,7,19,13,3,6,11,22,12,23,16,8,20,2)
16	9.02	11.73 (9)	6.38 (14,10,7,19,13,3,23,22,18,16,11,8,6,20,25,21)

从表 1 的 PRESS 值计算结果可知，三种方法中正交递归选择法求得的最小 PRESS 值为最低。最小 PRESS 值为 4.54, 对应的变量子集中的 10 个变量为 14, 10, 13, 3, 7, 4, 12, 11, 22, 15 。而用逐步回归正向选择法则难以找到 PRESS 值为最低的变量子集或模型，并且对于每个变量引入数，该法得到模型的 PRESS 值均高于正交递归选择法求得的 PRESS 值。另外用 PLS 回归法求得模型的 PRESS 值，均高于正交递归选择法求得的 PRESS 值，并且用正交递归选择法求得 PRESS 值为最低的变量子集后，再用 PLS 法处理，其结果见表 2, 从表 2 中的结果可知，PLS 回归法不能得到 PRESS 值更低的模型。本工作用正交递归选择法求得 PRESS 值最低的模型为

表 2　PLS 法求得的 PRESS 值

变量数	PRESS
0	15.80
1	9.92
2	10.81
3	9.94
4	11.25
5	10.57
6	7.77
7	8.23
8	8.75
9	5.61
10	4.54

$$y = 50144 + 126.5x_2x_3 + 140.5x_1x_4 - 139x_2^2 - 2390x_3 - 8757x_1^2 - 131x_4 + 11.4x_1x_6 - 124x_1x_5 - 0.457x_4^2 + 7.85x_2x_4$$

三、结　语

从模型的预报能力来看，在多元回归分析中变量筛选是很重要的，由于正交递归选择法可找到 PRESS 值较低的模型，计算量又不是很大，远远小于所有可能回归法的计算量，所得模型便于对各因子作灵敏性分析，因此认为该法是一种较实用的变量筛选方法。

附：转炉炉龄问题数据

No.	x_1	x_2	x_3	x_4	x_5	x_6	y
1	0.2922	18.5	41.4	58.0	18.0	83.3	1030
2	0.2672	18.4	41.0	51.0	18.0	91.7	1006
3	0.2685	17.7	38.6	52.0	17.3	78.9	1000
4	0.1835	18.9	41.8	18.0	12.8	47.2	702
5	0.2348	18.0	39.4	51.0	17.4	57.4	1087
6	0.1386	18.9	40.5	39.0	12.8	22.5	900
7	0.2083	18.3	39.8	64.0	17.1	52.6	708
8	0.4180	18.8	41.0	64.0	16.4	26.7	1223
9	0.1030	18.4	39.2	20.0	12.3	35.0	803
10	0.4893	19.3	41.4	49.0	19.1	31.3	715
11	0.2058	19.0	40.0	40.0	18.8	41.2	784
12	0.0925	17.9	38.7	50.0	14.3	66.7	535
13	0.1854	19.0	40.8	44.0	21.0	28.6	949
14	0.1963	18.1	37.2	46.0	15.3	63.0	1012
15	0.1008	18.2	37.0	46.0	16.8	33.9	716
16	0.2702	18.9	39.5	48.0	20.2	31.3	858
17	0.1465	19.1	38.6	45.0	17.8	28.1	826
18	0.1353	19.0	38.6	42.0	16.7	39.7	1015
19	0.2244	18.8	37.7	40.0	17.4	49.0	861
20	0.2155	20.2	40.2	52.0	16.8	41.7	1098
21	0.0316	20.9	41.2	48.0	17.4	52.6	580
22	0.0491	20.3	40.6	56.0	19.7	35.0	573
23	0.1487	19.4	39.5	42.0	18.3	33.3	832
24	0.2445	18.2	36.6	41.0	15.2	37.9	1076
25	0.2222	18.4	37.0	40.0	13.7	42.9	1376
26	0.1298	18.4	37.2	45.0	17.2	44.3	914
27	0.2300	18.4	37.1	47.0	22.9	21.6	861
28	0.2436	17.7	37.2	45.0	16.2	37.9	1105
29	0.2804	18.3	37.5	46.0	17.3	20.3	1013
30	0.1970	17.3	35.9	46.0	13.8	57.4	1249
31	0.1840	16.2	35.3	43.0	16.6	44.8	1039
32	0.1679	17.1	34.6	43.0	20.3	37.3	1502
33	0.1524	17.6	36.0	51.0	14.2	36.7	1128

参 考 文 献

[1] Haaland, D.M. and Thomas, E.V., *Anal. Chem.*, **60**, 1193(1988).
[2] Kowalski, K.G., *Chemometrics Intell. Lab. Syst.*, **5**, 129(1991).
[3] Weisberg, S., *Applied Linear Regression*, 2nd ed., Wiley, New York, 1985.
[4] Myers, R.H., *Classical and Modern Regression with Application*, Duxbury Press, Boston, Massachusetts, 1986.
[5] 朱尔一, "化学模式识别新方法研究", 博士论文, 中国科学院上海冶金研究所, 1991.

ORTHOGONALIZATION RECCURENCE SELECTION METHOD AND ITS APPLICATION IN THE RESEARCH TO AGE OF CONVERTER IN STEEL MAKING

Zhu Eryi Yang Pengyuan Deng Zhiwei Huang Benli
(*Department of Chemistry Xiamen University*, Xiamen 361005)

ABSTRACT A new model selection method named Orthogonalization Reccurence Selection (ORS) method is proposed. By applying this method the model with higher predictive ability can be obtained or the lower PRESS statistic values of the model can be achieved. The comparison is made between OGS method and forward selection method in the stepwise regression as well as PLS methods through the example which involves the age of converter in steel making.

KEYWORDS Model selection PRESS criteria Stepwise regression PLS regression Gram-Schmidt Orthogonalization