云南大学学报 (自然科学版), 2004, 26 (4): 292~296 CN 53-1045/N ISSN 0258-7971
Journal of Yunnan University

数据矩阵条件指数的影响评价*

张文专[1,2], 石 磊[2]
(1. 贵州财经学院 信息学院, 贵州 贵阳 550003; 2. 云南大学 统计系, 云南 昆明 650091)

摘要: 考察了删除单个数据点对条件指数的影响, 导出了度量数据点对条件指数影响大小的 1 种诊断统计量——样本影响函数的近似计算公式, 一个实例被用来说明该方法的可行性.

关键词: 单点剔除; 影响评价; 数据矩阵; 条件指数
中图分类号: O 212.4 文献标识码: A 文章编号: 0258-7971(2004)04-0292-05

统计诊断是使用统计方法解决实际问题全过程的一个不可缺少的环节. 自诞生以来, 它一直受到许多统计工作者的青睐. 这方面的参考文献特别多, 诸如: 石磊和王学仁[1]研究了多元分析中的局部影响分析; 黄梅等[2]研究了双向分类混合交互效应模型中均值滑动模型的异常值检验; 唐年胜等[3]研究了多元线性回归中的异常值检验; 唐年胜等[4]研究了多元加权约束线性回归的影响分析; Dai Lin 等[5]研究了具有一致协方差的增长曲线模型的局部影响分析; 杨丽等[6]研究了随机约束线性回归模型参数估计的影响分析等.

线性回归模型中回归系数的 LS 估计具有许多良好的性质, 但是设计阵 X 呈病态, 即 X 的列向量近似地线性相关时, $X'X$ 接近奇异, 从而使得估计的精度降低, 所以在对线性回归模型中回归系数进行估计之前, 分析设计阵的数据结构是必要的. Belsley[7] 在《条件诊断》一书中利用条件指数 (Condition Indexes) 考察了数据矩阵的列中有复共线性关系的组数以及共线性的程度. 但这种分析对强影响点(Influence Point) 或异常值(Outlier) 非常敏感. 如果数据中存在强影响点或异常值, 那么应用这种分析方法得到的结论可能会产生误导, 所以有必要寻找考察条件指数中强影响点或异常值的诊断方法. 迄今为止, 还未见对条件指数进行影响评价的有关报道. 本文研究删除单个数据点对条件指数的影响, 导出了度量数据点对条件指数影响大小的样本影响函数的近似计算公式.

本文第 1 节介绍条件指数的定义及复共线性诊断的思想; 第 2 节讨论删除个别数据点对条件指数产生的影响; 第 3 节用一个实例说明该方法的应用.

1 条件指数

考虑数据矩阵

$$X = \begin{pmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & & \vdots \\ x_{1n} & \cdots & x_{pn} \end{pmatrix} = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}, \qquad (1)$$

其中 $x'_i = (x_{1i}, \cdots, x_{pi})$, $i = 1, \cdots, n$ 表示第 i 个数据点. 记 $B = X'X = \sum_{i=1}^{n} x_i x'_i$, 则 B 为 $p \times p$ 阶的半正定矩阵; 如果 B 的 p 个不同的特征值按从大到小的顺序排列分别为: $\lambda_1, \cdots, \lambda_p$, 它们对应的单位特征向量分别为 $v_1, \cdots, v_p$, 则定义统计量

$$\eta_k = \left(\frac{\lambda_1}{\lambda_k}\right)^{1/2}, \quad (k = 1, \cdots, p) \qquad (2)$$

为矩阵 X 的 p 个条件指数. 显然, $\eta_1 = 1$, 且 $\eta_1 < \cdots < \eta_p$.

文献[7] 指出: 大条件指数的个数就等于数据矩阵 X 的列中近似共线性关系的组数. 当条件指数的值在 5 ~ 10 这个范围时, X 的列中有弱共线

* 收稿日期: 2004-01-22
基金项目: 国家自然科学基金资助项目(10261009); 贵州省教育厅自然科学研究基金资助项目(2002231).
作者简介: 张文专(1966-), 男, 湖南人, 讲师, 博士, 主要从事应用数理统计方面的研究.
石 磊(1965-), 男, 云南人, 教授, 博士生导师, 主要从事数理统计及应用方面的研究.

性关系; 条件指数值在30以上时, X 的列中有强共线性关系.

下一节将推导度量数据点对各条件指数影响大小的统计量 —— 样本影响函数的近似计算公式.

2　条件指数的影响评价

若由(1)式表示的数据矩阵被删除第 i 个数据点以后记为 $X(i)$, 则相应地记 $B(i) = X'(i)X(i)$, $B(i)$ 的 p 个特征值为 $\lambda_{(i)1}, \cdots, \lambda_{(i)p}$, 与之对应的单位特征向量为 $v_{(i)j}$, $X(i)$ 的 p 个条件指数为

$$\eta_{(i)k} = \left(\frac{\lambda_{(i)1}}{\lambda_{(i)k}}\right)^{1/2}, \quad (k = 1, \cdots, p). \tag{3}$$

为得到条件指数的样本影响函数, 我们首先给出如下的代数学引理(见文献[8]).

引理 1　设 A 为实对称矩阵, λ 为 A 的单重特征根, v 为对应的特征向量, A 的 ε 扰动可表示为 ε 的幂级数

$$A(\varepsilon) = A + \varepsilon A^{(1)} + \frac{\varepsilon^2}{2}A^{(2)} + o(\varepsilon^3), \tag{4}$$

其中 $A^{(1)}, A^{(2)}$ 为实对称矩阵, 则 $A(\varepsilon)$ 的特征根 $\lambda(\varepsilon)$ 可表示为 ε 的幂级数

$$\lambda(\varepsilon) = \lambda + \varepsilon\lambda^{(1)} + \frac{\varepsilon^2}{2}\lambda^{(2)} + o(\varepsilon^3), \tag{5}$$

其中 $\lambda^{(1)} = v'A^{(1)}v$, $\lambda^{(2)} = v'[A^{(2)} - 2A^{(1)}(A - \lambda I)^+ A^{(1)}]v$, I 为单位矩阵, $(A - \lambda I)^+$ 为 $(A - \lambda I)$ 的加号逆.

由(1)式知

$$B = X'X = \sum_{i=1}^{n} x_i x'_i,$$

从而

$$B(i) = X'(i)X(i) = \sum_{j \neq i} x_j x'_j = B - x_i x'_i = B - \frac{1}{n-1}[(n-1)x_i x'_i], \tag{6}$$

将(6)式跟(4)式相对照, 我们知道: 在数据矩阵 X 中删除第 i 个数据点 x_i 后, 矩阵 B 受到 $\varepsilon = -(n-1)^{-1}$ 的扰动, 这时 $B^{(1)} = (n-1)x_i x'_i$, $B^{(2)} = 0$.

引理 2　数据矩阵 X 中删除第 $i(i = 1, \cdots, n)$ 个数据点 x_i 以后, 矩阵 B 的特征值 $\lambda_1, \cdots, \lambda_p$ 的扰动展开式可表示为

$$\lambda_{(i)j} = \lambda_j + \varepsilon\lambda_j^{(1)} + \frac{\varepsilon^2}{2}\lambda_j^{(2)} + o(\varepsilon^3), \tag{7}$$

其中 $\varepsilon = -(n-1)^{-1}$, $\lambda_j^{(1)} = (n-1)a_{ij}^2$, $\lambda_j^{(2)} = -2(n-1)^2 a_{ij}^2 \sum_{l \neq j} a_{il}^2 (\lambda_l - \lambda_j)^{-1}$, $a_{il}(l = 1, \cdots, p)$ 为 x_i 在特征坐标下的投影, 即

$$x_i = \sum_{k=1}^{p} a_{il}v_l \text{ 或 } a_{il} = x'_i v_l.$$

证明　因为 $x_i = \sum_{k=1}^{p} a_{ik}v_k$, 所以

$$x_i x'_i = \sum_{k,l} a_{ik}a_{il}v_k v'_l. \tag{8}$$

由引理 1, (8) 及 $B^{(1)} = (n-1)x_i x'_i$ 得

$$\lambda_j^{(1)} = v'_j B^{(1)} v_j = v'_j[(n-1)\sum_{k,l} a_{ik}a_{il}v_k v'_l]v_j,$$

注意到 $v'_k v_j = \begin{cases} 1 & (k = j) \\ 0 & (k \neq j) \end{cases}$,

于是 $\lambda_j^{(1)} = (n-1)a_{ij}^2$.

由实对称矩阵的谱分解有

$$B = \sum_{k=1}^{p} \lambda_k v_k v'_k,$$

$$I = \sum_{k=1}^{p} v_k v'_k,$$

这里的 I 是单位矩阵, 从而

$$B - \lambda_j I = \sum_{k \neq j} (\lambda_k - \lambda_j) v_k v'_k,$$

所以

$$(B - \lambda_j I)^+ = \sum_{k \neq j} (\lambda_k - \lambda_j)^{-1} v_k v'_k, \tag{9}$$

由 $B^{(1)}$ 和 $B^{(2)}$ 的定义, 引理 1 及(8), (9)式有

$$\begin{aligned}\lambda_j^{(2)} &= v'_j[B^{(2)} - 2B^{(1)}(B - \lambda_j I)^+ B^{(1)}]v_j = \\ &v'_j\{-2[(n-1)x_i x'_i](B - \lambda_j I)^+ \cdot \\ &[(n-1)x_i x'_i]\}v_j = \\ &v'_j\{-2\Big[(n-1)\sum_{k,l} a_{ik}a_{il}v_k v'_l\Big] \cdot \\ &\Big(\sum_{k \neq j} (\lambda_k - \lambda_j)^{-1} v_k v'_k\Big) \cdot \\ &\Big[(n-1)\sum_{k,l} a_{ik}a_{il}v_k v'_l\Big]\}v_j = \\ &-2(n-1)^2 a_{ij}^2 \sum_{k \neq j} a_{ik}^2 (\lambda_k - \lambda_j)^{-1}.\end{aligned}$$

证毕

引理 3　数据矩阵 X 中删除第 $i(i = 1, \cdots, n)$ 个数据点 x_i 以后, X 的条件指数 $\eta_1, \cdots, \eta_p$ 的扰动展开式可表示为

$$\eta_{(i)k} = \eta_k + \varepsilon\eta_k^{(1)} + \frac{\varepsilon^2}{2}\eta_k^{(2)} + o(\varepsilon^3), \tag{10}$$

其中 $\varepsilon = -(n-1)^{-1}$,

$$\eta_k^{(1)} = \frac{1}{2}(n-1)\eta_k(\lambda_1^{-1}a_{i1}^2 - \lambda_k^{-1}a_{ik}^2), \tag{11}$$

$$\eta_k^{(2)} = (n-1)^2\eta_k \cdot$$

$$[-\lambda_1^{-1}a_{i1}^2\sum_{l\neq 1}a_{il}^2(\lambda_l-\lambda_1)^{-1}+\lambda_k^{-1}a_{ik}^2\sum_{l\neq k}a_{il}^2(\lambda_l-\lambda_k)^{-1}+\frac{1}{4}(3\lambda_k^{-1}a_{ik}^2+\lambda_1^{-1}a_{i1}^2)\cdot(\lambda_k^{-1}a_{ik}^2-\lambda_1^{-1}a_{i1}^2)]. \quad (12)$$

证明 由(3)和(7)可知

$$\eta_{(i)k}=\lambda_{(i)1}^{1/2}\lambda_{(i)k}^{-1/2}=[\lambda_1+\varepsilon\lambda_1^{(1)}+\frac{\varepsilon^2}{2}\lambda_1^{(2)}+o(\varepsilon^3)]^{1/2}\cdot[\lambda_k+\varepsilon\lambda_k^{(1)}+\frac{\varepsilon^2}{2}\lambda_k^{(2)}+o(\varepsilon^3)]^{-1/2}=\lambda_1^{1/2}\lambda_k^{1/2}[1+\varepsilon\lambda_1^{-1}\lambda_1^{(1)}+\frac{\varepsilon^2}{2}\lambda_1^{-1}\lambda_1^{(2)}+o(\varepsilon^3)]^{1/2}[1+\varepsilon\lambda_k^{-1}\lambda_k^{(1)}+\frac{\varepsilon^2}{2}\lambda_k^{-1}\lambda_k^{(2)}+o(\varepsilon^3)]^{-1/2}=\eta_k[1+\varepsilon\lambda_1^{-1}\lambda_1^{(1)}+\frac{\varepsilon^2}{2}\lambda_1^{-1}\lambda_1^{(2)}+o(\varepsilon^3)]^{1/2}[1+\varepsilon\lambda_k^{-1}\lambda_k^{(1)}+\frac{\varepsilon^2}{2}\lambda_k^{-1}\lambda_k^{(2)}+o(\varepsilon^3)]^{-1/2}. \quad (13)$$

利用公式

$$(1+x)^\alpha=1+\alpha x+\frac{\alpha(\alpha-1)}{2!}x^2+o(x^3),$$

可得

$$[1+\varepsilon\lambda_1^{-1}\lambda_1^{(1)}+\frac{\varepsilon^2}{2}\lambda_1^{-1}\lambda_1^{(2)}+o(\varepsilon^3)]^{1/2}=1+\varepsilon(2\lambda_1)^{-1}\lambda_1^{(1)}+\frac{\varepsilon^2}{2}\{(2\lambda_1)^{-1}\lambda_1^{(2)}-(2\lambda_1)^{-2}[\lambda^{(1)}]^2\}+o(\varepsilon^3), \quad (14)$$

$$[1+\varepsilon\lambda_k^{-1}\lambda_k^{(1)}+\frac{\varepsilon^2}{2}\lambda_k^{-1}\lambda_k^{(2)}+o(\varepsilon^3)]^{-1/2}=1+\varepsilon(-2\lambda_k)^{-1}\lambda_k^{(1)}+\frac{\varepsilon^2}{2}\{(-2\lambda_k)^{-1}\lambda_k^{(2)}+\frac{3}{4}[\lambda_k^{-1}\lambda_k^{(1)}]^2\}+o(\varepsilon^3). \quad (15)$$

将(14),(15)代入(13),整理即可得

$$\eta_{(i)k}=\eta_k+\varepsilon\cdot\frac{1}{2}\eta_k[\lambda_1^{-1}\lambda_1^{(1)}-\lambda_k^{-1}\lambda_k^{(1)}]+\frac{\varepsilon^2}{2}\cdot\frac{1}{2}\eta_k\cdot\{\lambda_1^{-1}\lambda_1^{(2)}-\lambda_k^{-1}\lambda_k^{(2)}+\frac{1}{2}[3\lambda_k^{-1}\lambda_k^{(1)}+\lambda_1^{-1}\lambda_1^{(1)}]\cdot[\lambda_k^{-1}\lambda_k^{(1)}-\lambda_1^{-1}\lambda_1^{(1)}]\}+o(\varepsilon^3). \quad (16)$$

根据引理 2,我们有

$$\begin{cases}\lambda_1^{(1)}=(n-1)a_{i1}^2,\\ \lambda_1^{(2)}=-2(n-1)^2a_{i1}^2\sum_{l\neq 1}a_{il}^2(\lambda_l-\lambda_1)^{-1},\end{cases} \quad (17)$$

$$\begin{cases}\lambda_k^{(1)}=(n-1)a_{ik}^2,\\ \lambda_k^{(2)}=-2(n-1)^2a_{ik}^2\sum_{l\neq k}a_{il}^2(\lambda_l-\lambda_k)^{-1}.\end{cases} \quad (18)$$

将(17)式和(18)式代入(16)式,即可得到(10)式. 证毕

由文献[8]中关于样本影响函数的定义及引理 3 即可得到以下的定理 1.

定理 1 数据矩阵的条件指数 $\eta_1,\cdots,\eta_n$ 相对于第 $i(i=1,\cdots,n)$ 个数据点的样本影响函数(Sample Influence Functin)为

$$F_i(\eta_k)=\eta_k^{(1)}-\frac{1}{2}\frac{1}{n-1}\eta_k^{(2)}+o\left(\frac{1}{(n-1)^2}\right), \quad (19)$$

其中 $\eta_k^{(1)}$, $\eta_k^{(2)}$ 的定义分别与(11),(12)相同.

$F_i(\eta_k)$ 刻画了第 i 个数据点对第 k 个条件指数的影响. 因为它是一维统计量,所以绝对值较大的 $F_i(\eta_k)$ 对应的数据点 x_i 对 η_k 有较大的影响. 用上述 $F_i(\eta_k)$ 展开式中的前 2 项来近似代替 $F_i(\eta_k)$,我们通过比较 $F_i(\eta_k)(i=1,\cdots,p)$ 的近似绝对值大小就可以探测出 η_k 的强影响点.

由(11),(12)及(19)可得 $F_i(\eta_k)$ 的近似计算公式

$$F_i(\eta_k)\approx\eta_k^{(1)}-\frac{1}{2}(n-1)^{-1}\eta_k^{(2)}=\frac{1}{2}(n-1)\eta_k\{\lambda_1^{-1}a_{i1}^2\sum_{l\neq 1}a_{il}^2(\lambda_l-\lambda_1)^{-1}-\lambda_k^{-1}a_{ik}^2\sum_{l\neq k}a_{il}^2(\lambda_l-\lambda_k)^{-1}+(\lambda_1^{-1}a_{i1}^2-\lambda_k^{-1}a_{ik}^2)\cdot[1+(4\lambda_1)^{-1}a_{i1}^2+3(4\lambda_k)^{-1}a_{ik}^2]\}. \quad (20)$$

3 实例分析

本节用上述方法对一个实际数据矩阵的条件指数进行影响评价. 所选数据(见表 1)是 Shi[11]分析过的地质数据,为了寻找富含金的地点,有关人员调查了 6 个与金的蕴藏量相关的因素: Cu, As, Pb, Ni, W, Sb, 其中 As 的单位是 10^{-8},其余的单位都是 10^{-6},通过不断更换地点进行观测,得到了 44 组观测值.

搜集这组数据的当初目的是想研究金的蕴藏量与这 6 个变量之间的关系, 所以, 在表 1 中我们还列出了实际含金指标. 但我们所考虑的数据矩阵 X 是由前 6 个指标的 44 组观测值组成的, 它的 6 个条件指数如表 2 所示. 由第 6 个条件指数 η_6= 31.324 2 知, 这个数据矩阵存在严重的共线性关系; 结合其它的条件指数可知, 这个数据矩阵共有 4 组近似共线性关系.

由(2) 式中条件指数的定义可知, η_1 不受任何数据点的影响. 根据(20) 式分别计算各数据点对 η_2, …, η_6 的影响之值, 再以这些值的绝对值为纵坐标, 以数据点的序号为横坐标描点作图如图 1. 从图 1 中的(a) 可直观地判断: 原数据的第 27 号、15 号点是最大的条件指数 η_6 的强影响点; 从(b) 和(d) 可直观地判断: 原数据的第 27 号、43 号点是条件指数 η_3, η_5 的强影响点; 从(c) 可直观地判断: 原数据的第 22 号、27 号、39 号点是条件指数 η_4 的强影响点; 从(e) 可直观地判断: 原数据的第 22 号、27 号、39 号、42 号点是条件指数 η_2 的强影响点. 第 27 号点是 η_2, …, η_6 的公共强影响点, 删除 27 号点以后的数据矩阵 X 的条件指数如表 3 所示.

表 **1**　**7** 个变量的地质数据

Tab. 1 Geology data for 7 variables

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Cu	70	70	70	100	70	60	50	70	70	80	70	100	75	70	100	100	70	60	75	100	200	100
As	150	100	100	100	150	100	100	500	100	150	150	100	100	100	300	100	100	100	150	100	350	1 500
Pb	60	75	40	45	45	60	45	70	40	75	75	60	60	50	40	45	40	40	60	45	80	70
Sb	70	20	45	50	70	30	40	30	30	35	35	60	100	60	200	25	50	50	100	60	30	250
Ni	700	60	100	100	180	70	260	500	50	300	300	70	70	500	500	120	60	50	130	300	1 300	500
W	10	5	5	10	20	8	7	15	5	7	7	10	10	7	10	5	10	7	15	7	25	1 000
Au	165	13	6	9	9	9	7	240	18	14	14	5	16	7	450	4	11	5	10	9	31	1 000

No.	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
Cu	100	50	100	100	150	100	150	60	40	300	100	150	100	250	200	50	100	75	70	130	160	130
As	200	100	250	200	200	100	200	100	100	300	200	250	100	100	150	100	2 000	100	100	1 500	400	200
Pb	50	130	50	60	140	50	130	200	25	100	70	100	70	130	400	15	60	70	60	60	700	130
Sb	50	50	30	80	50	30	70	25	130	50	50	40	20	40	200	20	50	30	30	30	60	50
Ni	200	150	2 500	120	7 200	140	300	120	2 000	2 500	800	120	230	600	200	2 500	500	400	120	350	130	350
W	30	7	10	30	20	10	15	10	5	25	20	30	10	25	30	5	40	10	20	500	10	30
Au	63	5	386	40	6	10	11	4	3	14	125	21	15	11	136	2	1 000	506	10	9	13	271

表 **2**　实例分析中数据矩阵 X 的条件指数

Tab. 2 The conditional indexes of data matrix X in the real example

η_1	η_2	η_3	η_4	η_5	η_6
1.000 0	2.848 0	9.107 4	11.637 9	21.341 4	31.324 2

表 **3**　实例分析中删除 **27** 号点以后的数据矩阵 $X(27)$ 的条件指数

Tab. 3 The conditional indexes of data matrix $X(27)$ after deleting the No. 27th data point

η_1	η_2	η_3	η_4	η_5	η_6
1.000 0	1.868 6	5.687 3	7.151 9	14.335 6	19.269 8

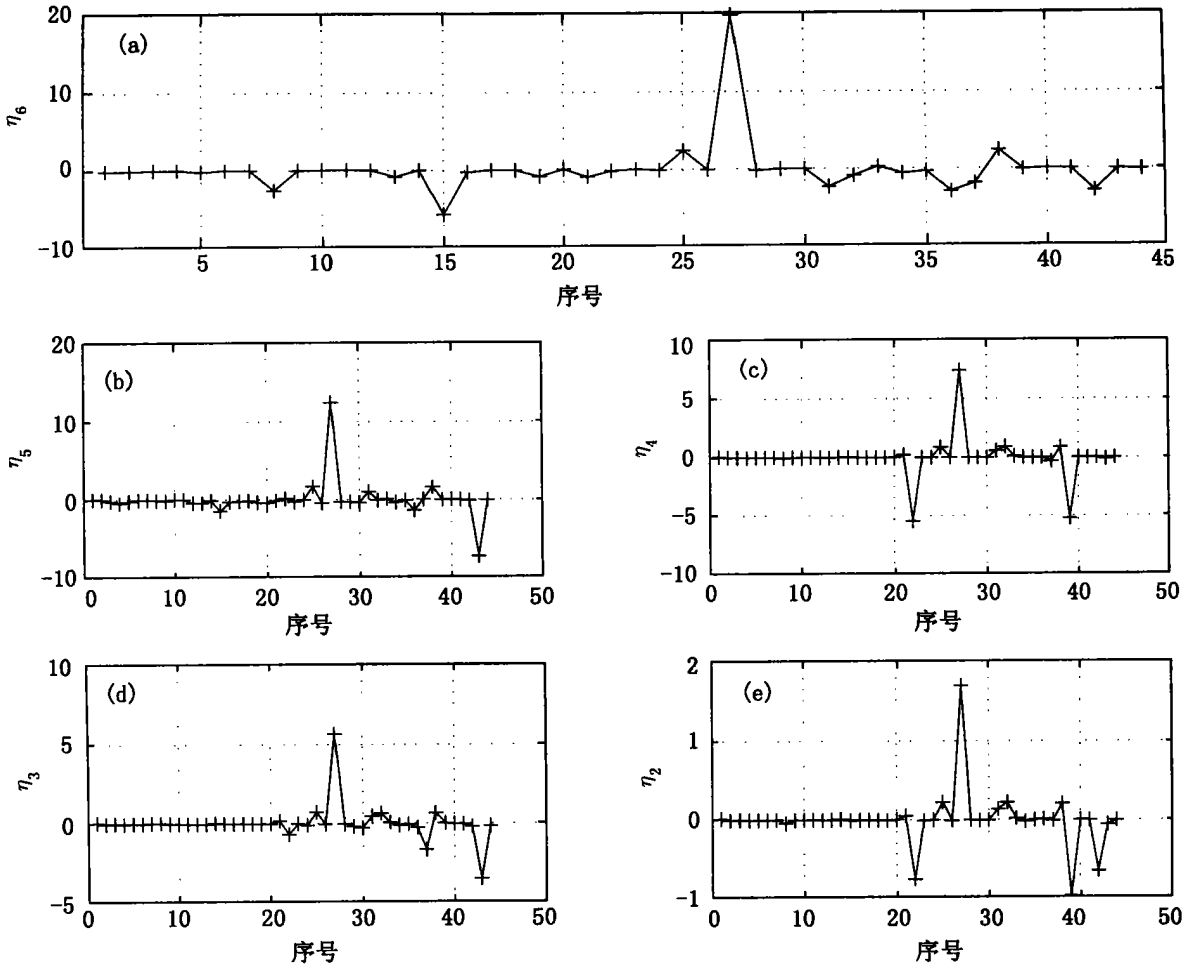

图 1 数据点对各条件指数的影响值

Fig. 1 The magnitude of influence of data points on each condition index

将表2与表3进行对比可知, 27号点被删除以后 $\eta_2, \ldots, \eta_6$ 均发生明显的变化, 说明27号点确实是 $\eta_2, \ldots, \eta_6$ 的强影响点, 同时也说明了上述方法的可行性.

参考文献:

[1] SHI L, WANG X R. Assessment of local influence in multivariate analysis[J]. Acta Mathematica Scientia, 1996, 16(3): 257—270.

[2] 黄 梅, 杨春芸, 何利平, 等. 双向分类混合交互效应模型中均值滑动模型的异常值检验[J]. 云南大学学报(自然科学版), 1999, 21(6): 465—468.

[3] 唐年胜, 王 娅, 杨 勇. 多元约束线性回归中异常值检验[J]. 云南大学学报(自然科学版), 2000, 22(3): 161—164.

[4] 唐年胜, 王 娅, 罗贤奎. 多元加权约束线性回归的影响分析[J]. 云南大学学报(自然科学版), 2000, 22(2): 92—94.

[5] DAI Lin, BAI Peng, CHEN Jiarr bao, et al. Local influrence of growth cure model with uniform covariancestructure[J]. 云南大学学报(自然科学版), 2000, 22(2): 87—91.

[6] 杨 丽, 陈炳生, 宋邵云. 随机约束线性回归模型参数估计的影响分析[J]. 云南大学学报(自然科学版), 2000, 22(5): 325—329.

[7] BELSLEY D A, KUH E, WELSCH R E. Regression Diagnostics[M]. New York: John Wiley, 1980.

[8] 韦博成, 鲁国斌, 史建清. 统计诊断引论[M]. 南京: 东南大学出版社, 1991.

(下转第 300 页)

[2] ESTER M, KRIEGEL H P, SANDER J. Algorithns and applications for spatial data mining[EB/OL]. http://www.dbs.informabik.nni-muenchen.de, 2003-10-15.

[3] ESTER M, KRIEGEL H P, XU X. Knowledge discovery in large spatial databases: focusing techniques for efficient class identification. Proc 4th Int Symp on Large Spatial Databases [EB/OL]. http://www.dbs.informabik.nni-muendon.de, 2003-10-15.

[4] LANGLEY P. An analysis of bayesian classifiers[A]. Proc of the National Conf on Artiticial Intelligence (AAAI'92) [C]. Menlo Park, CA: AAAI Press, 2002. 223—228.

[5] ELKAN C. Boosting and naïve Bayesian learning: [Technical Report No. CS97-557] [EB/OL]. http://www.cs.ucsd.edu, 2003-10-15.

[6] HAN J, KAMBER M. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2001.

[7] ESTER M, GUDLACH S, KRIEGEL H P, et al. Database primitives for spatial data mining[EB/OL]. http://www.dbs.informabik.nni-muenchen.de, 2003-10-15.

[8] 谭 旭, 王丽珍, 卓 明. 侧外挖掘研究[J]. 云南大学学报(自然科学版), 2001, 23(5A): 61—64.

[9] 谭 旭, 王丽珍, 卓 明. 利用决策树发掘分类规则的算法研究[J]. 云南大学学报(自然科学版), 2000, 22(6): 415—419.

One spatial classification arithmetic based on naïve Bayesian classifier

ZHAO Qin-yi, WANG Li-zhen, ZHOU Li-hua

(Department of Computer Science and Technology, Yunnan University, Kunming 650091, China)

Abstract: Spatial classification is one main case of spatial data mining. Finding effective spatial classification algorithm is an important problem of spatial classification. It is presented a new spatial classification algorithm based on neighbor graph and naïve Bayesian classifier. The algorithm takes the attributes of classifying object and its neighbor objects into account when classifying spatial objects. The algorithm have low computing cost and high classifying accuracy.

Key words: spatial classification; spatial neighbor relation; naïve Bayesian classifier

* *

(上接第 296 页)

Assessment of influence in condition indexes of data matrix

ZHANG Wen-zhuan[1,2], SHI Lei[2]

(1. Institute of Infomation, Guizhou Finance and Economics College, Guiyang 550001, China;
2. Department of Statistics, Yunnan University, Kunming 650091, China)

Abstract: The influence of data points on the condition indexes of data matrix is studied via deleting cases one at a time. A approximate formula of sample influence function has obtained. An example is used to illustrate the method.

Key words: case deletion; condition indexes; data matrix; influence assessment

MSC(2000): 62J99; 62J20; 62J05