经验交流

我国"硒与健康"数据库数据质量评价

杜　敏　　白迺彬
（中国科学院生态环境研究中心）

数据库数据质量评价是数据库设计与实现的必要基础工作之一. 鉴于数据源往往是兼容并蓄，数据质量如何应向用户交待，以期在应用时做为重要参考. 不久前，我们实现了我国"硒与健康"数据库[1]. 首先应用常见的统计方法进行了单因素数据质量评价[2]. 继而参照不久前 Mcgien 的工作[3]，尝试应用模式识别方法建立了多维数据的评价模式，对"硒与健康"数据库中所储多因素数据进行了质量评价.

一、数据库数据现状

数据库以我国公开刊物和文集为数据源. 为了使数据有可比性，我们对文献数据进行了抽提、整理、单位统一变换和格式规范化处理. 经目标分析，确定实体集，选定域. 域中相互关系项以字段形式传达了如下硒样品信息：样品名、硒含量数值、单位、产地、健康状况、分析方法及文献索引号等诸项. 当前，已储近五千个记录，遍及 28 个省市、约 300 个县的广大地区，主要反映的是克山病、大骨节病、硒中毒及非病区样品含硒量数据，也有某些反映克汀病、氟中毒症、某类癌症、心血管病及长寿地区样品含硒量数据，括及大气、岩石、土壤、水、粮食、药物及人体样品等四百余种之多.

库内数据集合粗略存在如下规律：数据数值大小与不同健康属性区域分布有关；同类地区与样品种类有关；同类地区、同样样品数值大小受样品分析方法影响不大. 可在同类地区、同种样品内进行单因素数据质量评价. 再者，我们看到，克山病、大骨节病和硒中毒等往往不仅与单类样品(如土壤、饮用水、粮食，副食品等)的含硒量有关，而是主要与人群所处环境中整体硒水平有关. 有时，单因素质量评价中看到某项数据离群，但置于多因素质量评价中考察，往往归于正常. 有幸的是，数据库目前所储克山病、大骨节病及非病区之环境样品的各主要数据，大多备齐，可以在单因素评价的基础上进行多因素数据质量评价.

二、单因素数据质量评价

1.原理与方法

所谓单因素数据质量评价，系指对可比数值集合的均衡性、变异性进行表征，求出频率分布，指明总体质量水平及可能的扬弃数据项，最终对质量进行控制.

平均数 $\bar{X}$ 代表样本数为 n 的变量 X_i 的平均水平

$$\bar{X} = \frac{\sum X_i}{n} \quad (i = 1, 2, \cdots, n) \qquad (1)$$

极差 R 是数值集合变量 X_i 的最大值 X_{max} 和最小值 X_{min} 之差，表示其最大变异程度，

$$R = X_{max} - X_{min} \qquad (2)$$

标准差 s 概括了所有变量 X_i 之离散程度，它对极数反映灵敏，但受其影响却较小，是一种较为精确和稳定的变异指标，

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} \quad (i = 1, 2, \cdots, n) \qquad (3)$$

均数与标准差密切相关，但内含相反；前者反映集中趋势，后者反映离散程度. 故从标准差和均数结合出发，用 $\bar{X} \pm s$ 代表 X_i 之精密度；用 $\bar{X} \pm 1.96s$ 来估计占 95% 的正常值范围. 如果某项数据超出 $\bar{X} \pm 2s$ 范围则提醒应该检查较正，讨论离群原因；若超出 $\bar{X} \pm 3s$ 范围则存在非偶然因素，可考虑将此离群值舍弃，检查失误原因，如是否可能录入数据库时操作有误，或其它.

对于单位不同的样品、或不同类样品之间，受均值或量纲影响难用标准差比较其变异特性. 可用变异系数 cv 比较各类样品的离异程度，

$$cv = \frac{s}{\bar{X}} \times 100\% \qquad (4)$$

变异系数 cv 实质上是相对标准差.

2.计算结果

由数据库检索而出的我国全部克山病、大骨节病和非病区饮用水、土壤、玉米、小麦、大米和人发硒

表 1 数据库数据质量指标

	样品	饮用水 (ppb)	土壤 (ppm)	玉米 (ppm)	小麦 (ppm)	大米 (ppm)	人发 (ppm)
克山病区	样本数 n	6	36	53	33	33	342
	均值 $\bar{X}$	0.750	0.122	0.008	0.016	0.018	0.117
	标准差 s	1.460	0.072	0.007	0.017	0.013	0.082
	最大值 X_{max}	3.700	0.275	0.041	0.078	0.070	1.000
	最小值 X_{min}	0.001	0.005	0.002	0.004	0.006	0.021
	极差 R	3.699	0.270	0.039	0.074	0.064	0.979
	$\bar{X}\pm 2s$ 范围%	100	97	94	94	94	95
	$\bar{X}\pm 3s$ 范围%	100	100	98	97	97	99
	变异系数 cv%	1.95	0.59	0.88	1.06	0.72	0.70
大骨节病区	样本数 n	54	15	47	45	7	187
	均值 $\bar{X}$	0.127	0.078	0.009	0.018	0.014	0.108
	标准差 s	0.158	0.035	0.010	0.020	0.005	0.093
	最大值 X_{max}	1.100	0.157	0.071	0.121	0.020	0.730
	最小值 X_{min}	0.004	0.004	0.001	0.002	0.005	0.010
	极差 R	1.096	0.153	0.070	0.119	0.015	0.714
	$\bar{X}\pm 2s$ 范围%	98	93	98	98	100	95
	$\bar{X}\pm 3s$ 范围%	100	100	98	100	100	97
	变异系数 cv%	1.24	0.45	1.11	1.11	0.36	0.86
非病区	样本数 n	109	122	79	70	53	376
	均值 $\bar{X}$	1.404	0.234	0.036	0.056	0.064	0.305
	标准差 s	7.984	0.145	0.043	0.071	0.116	0.262
	最大值 X_{max}	83.000	0.660	0.250	0.488	0.709	1.863
	最小值 X_{min}	0.001	0.005	0.003	0.004	0.006	0.027
	极差 R	82.999	0.655	0.247	0.484	0.703	1.836
	$\bar{X}\pm 2s$ 范围%	99	96	96	94	96	97
	$\bar{X}\pm 3s$ 范围%	100	100	98	99	96	98
	变异系数 cv%	5.69	0.62	1.19	1.27	1.81	0.86

含量数据 1667 个均值、标准差、极差、离散范围及变异系数列于表 1.

表 1 可见：数据总体 95% 处于 $\bar{X}\pm 2s$ 范围内，超出 $\bar{X}\pm 3s$ 的数据仅占总数的 1%. 说明数据库所收集的数据有一定质量，利于引用. 从变异系数 cv 指标来看，若做不同类地区、不同种样品离散程度相对比较，饮用水的数据质量相对较差. 可能有以下三种原因，饮用水样品来源不同：或井水、或雨水、或沟水、或河水、或泉水，含硒量有所波动；取样后，若置于容器内时间较久，易于吸附或挥发，重复性差；因含量较其它样品低三个数量级，含量致微，亦应允许有较大的离散程度. 有趣的是，土壤样品数据质量最好，这或许说明土壤品种虽多，但含硒量在同类地区大致稳定；也许是易于分析操作.

三、多因素数据质量评价

1.原理与方法

应用模式识别进行多维数据质量评价的原理是将每个因变量看成由多个变量组成的多维空间中的一个点，不同类属性点群将在多维空间内自然聚集. 借助超平面可以在多维空间内将不同类的点集分开；或者将多维空间点群非线性映照到可见的二维空间，用曲线类间分开. 那么，偏离理应归属类的点就是离群点. 可据其化学含义决定取舍或做出合理解释，这些“离群”或“叛类”的点往往会衍生出新概念或新规律. 具体做法如下：

两类数据超平面分离法 (PLAN)[4].

多维空间两类数据点群间有超平面界面方程

$$g_i(X)=W_{i0}+\sum_{k=1}^{m}W_{ik}X_k \quad (i=1,2). \qquad (5)$$

对于 M 维空间任一点 Y，若

$$g_1(Y)-g_2(Y)=(W_1-W_2)\cdot Y=W\cdot Y=0, \qquad (6)$$

则恰恰落在界面方程上. 若 $W \cdot Y > 0$, Y 则应属于类1; 若 $W \cdot Y < 0$, Y 则应属于类2. 关键是求解权重向量 W. 首先排出大量训练点集让计算机学习, 再令两类训练点重心连线的中垂面法矢量为 W 之起始值, 应用错分点反馈修正法进行迭代修正, 直至达到最大正确分类率为止. 建立分类面方程后, 可代入未知归类点的增广模式向量 Y, 计算机将自动预报其归属类.

非线性映照法 (NLM)[4,5].

在尽可能保持样本点间距离不变的条件下, 把 M 维空间中的样本点映照到可显示的二维空间, 由它们在二维空间中的分布来实现对样本分类和识别. 通过未知点在图上类的位置进行归属预报.

对于多维样本矩阵 $X_{ik}(i = 1, 2, \cdots, n; k = 1, 2, \cdots, m)$, 定义降维产生误差之误差函数为:

$$E = \frac{1}{\sum_{i<j}^{n} d_{ij}^{*}} \sum_{i<j}^{n} \frac{(d_{ij}^{*} - d_{ij})^2}{d_{ij}^{*}} \quad (i, j = 1, 2, \cdots, n), \qquad (7)$$

式中, d_{ij}^{*} 为 M 维空间 i 与 j 点间距离; d_{ij} 为映照

表 2　我国各省重要克山病、大骨节病区及非病区饮用水、土壤、粮食及人发硒水平

地　区	类别*	土　壤	饮用水	粮　1	粮　2	人　发	预报类别 (PLAN)	预报类别 (NLM)
1	0	0.293	1.115	0.034	0.037	0.482	0	0
2	0	0.260	0.685	0.029	0.061	0.286	0	0
3	0	0.248	0.100	0.135	0.065	0.280	0	0
4	0	0.293	0.400	0.104	0.179	0.238	0	0
5	0	0.135	0.146	0.042	0.059	0.367	0	0
6	0	0.372	0.086	0.013	0.018	0.434	0	0
7	0	0.089	0.314	0.018	0.038	0.220	1	1/0
8	0	0.199	0.015	0.013	0.030	0.339	0	0
9	0	0.132	0.100	0.010	0.030	0.396	0	0
10	0	0.316	0.117	0.033	0.082	0.295	0	0
11	0	0.213	0.431	0.015	0.047	0.270	0	0
12	0	0.208	0.837	0.030	0.045	0.232	0	0
13	0	0.250	0.100	0.008	0.011	0.256	0	0
14	0	0.293	1.115	0.045	0.037	0.482	0	0
15	0	0.248	0.100	0.050	0.065	0.280	0	0
16	0	0.316	0.117	0.041	0.082	0.295	0	0
17	1	0.006	0.035	0.005	0.010	0.060	1	1
18	1	0.006	0.035	0.005	0.018	0.057	1	1
19	1	0.089	0.115	0.005	0.013	0.091	1	1
20	1	0.053	0.115	0.010	0.021	0.150	1	1
21	1	0.170	0.100	0.004	0.005	0.156	1	1
22	1	0.081	0.064	0.004	0.007	0.087	1	1
23	1	0.075	0.102	0.013	0.038	0.097	1	1
24	1	0.122	3.700	0.031	0.078	0.068	1/0	0
25	1	0.084	0.100	0.010	0.011	0.115	1	1
26	1	0.084	0.078	0.028	0.077	0.223	1	1
27	1	0.048	0.160	0.010	0.013	0.127	1	1
28	1	0.048	0.160	0.008	0.013	0.132	1	1
29	1	0.095	0.277	0.014	0.020	0.094	1	1
30	1	0.085	0.277	0.005	0.007	0.063	1	1
31	1	0.147	0.022	0.107	0.018	0.860	0	1/0
32	1	0.037	0.100	0.018	0.065	0.090	1	1
33	1	0.084	0.100	0.011	0.011	0.115	1	1
34	1	0.084	0.078	0.011	0.077	0.223	1	1

* 0 表示非病区; 1 表示病区.

后二维空间内 i 与 j 点间距离，即

$$d_{ij}^{*}=\sqrt{\sum_{k=1}^{m}(X_{ik}-X_{jk})^{2}}\quad(i,j=1,2,\cdots,n) \tag{8}$$

和

$$d_{ij}=\sqrt{(Y_{i1}-Y_{j1})^{2}+(Y_{i2}-Y_{j2})^{2}}\quad(i,j=1,2,\cdots,n). \tag{9}$$

首先，选用主成份分析的前两个主成份组成初始二维空间，然后按最优化使误差函数 E 达到极小，用牛顿迭代法求解其中的 $2N$ 个未知数 Y_{i1} 和 Y_{i2}.

两个程序均用 Fortran 语言编写，在 VAX-11/780 计算机上实现运算. 对于 5×34 的样本矩阵，运算一次仅费机时几秒.

2.计算结果

首先，依省名、病名和样品名进行样品硒含量数值之协同检索，建立起我国各省重要克山病区、大骨节病区和非病区的饮用水、土壤、粮食及人发样品含硒量数值集合，含样本数约 1700 个. 继而，逐省、逐病区、逐样品类进行单因素统计平均处理，以避免取样随机性对数据类水平之影响，且使问题简单化和清晰化. 所得到的我国各省市重要克山病、大骨节病区及非病区的主要环境样品饮用水、土壤、两种粮食及人发硒含量均值列于表 2.

饮用水含硒量单位为 ppb；其余为 ppm.

应用两种方法计算表 2 数据结果如下：

两类分类方法处理均达到良好分类结果.PLAN 法求得最大正确分类率界面方程为：

$$W\cdot Y=0.068+0.160X_1+0.010X_2+0.016X_3+0.028X_4+0.127X_5=0 \tag{10}$$

按其分类正确分类率可达 0.941. NLM 法得到非线性映照图为图 1. 正确分类率亦达 0.941. 说明两种模式识别方法均可将为多因素(环境)所制约的因变量(健康状况)正确分类. 在环境与健康关系水平上看，所论数据库数据质量是良好的、可信的.

另外，两类方法错分结果相近，由错分点能标明感兴趣的离群点. PLAN 法指出错分点是 7 点（非病区）和 31 点(病区). 但从迭代结果分析，由于 24 点在分类界面两侧变动，正确分类率在 0.941 和 0.912 之间摆动，24 点亦可错分之可能. NLM 法

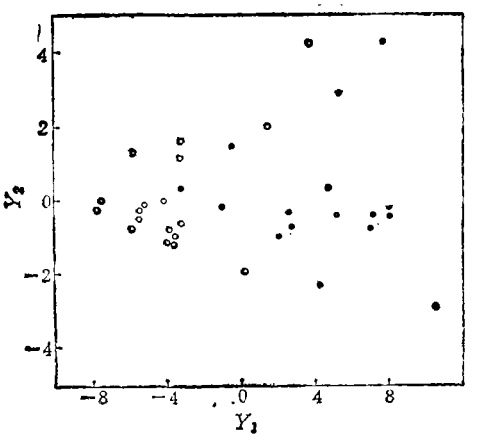

图 1 非线性映照图（△非病区；×病区）

则指出 24 点(病区)明显错分. 7 点(非病区) 和 31 点(病区)亦可错分 (表 2). 若综合两种方法，对于 7 点，NLM 法为非线性分类，有可能将其与病区分开，归类于非病区；而 PLAN 法为线性分类，实际情况往往局限于不能满足线性条件. 最后，24、31 两点，它们居于较高含硒量（图 1 右侧代表硒水平高区）的区域，理应成为非病区，现在实际上是落在病区，的确是离群点. 经查对，它们分别是内蒙与新疆. 这或者与该区样本少有关，或许有未揭示的原因待进一步研究.

四、小 结

无论是单因素，还是多因素数据质量评价，表明我国“硒与健康”数据库数据质量较好、可信，有助于数据库支持“硒与健康”研究与用户应用. 模式识别方法是数据库多因素数据质量评价有效技术之一.

参 考 文 献

[1] 白遵彬等，中国地方病学杂志，7(3)，144(1988).
[2] 郑宋，中国环境监测，(1)，40(1984).
[3] Breen, J. J. et al., *Environmental Applications of Chemometrics*, PP.16—33, ACS, Washington D. C., 1985.
[4] Kowalski B. R. et al., *J. Am. Chem. Soc.*, **94** (16), 5632(1972).
[5] 白遵彬，环境化学，6(1)，78(1987).

（收稿日期：1988 年 3 月 1 日）