广义关联分析——兼论灰色关联的本质

王旭升，葛龙进
(武汉中国地质大学研究生院，湖北 武汉　430074)

摘要：　在全等关联概念的基础上，提出了广义关联和映射关联的概念和算法，将统计线性相关和灰色关联统一在映射关联中，证明灰色关联是一种保比关联，提出了影响的定量概念．最后用实例证明不同的关联方式，对数据序列的关联排序结果也不同．

关键词：　关联；相关；映射；灰色关联

Generalized Relation Analysis
——Other to Discuss the Essence of Gray Relation

WANG Xu-sheng，GE Long-jin
(Graduate School，China University of Geosciences，Wuhan 430074)

Abstract：　Based on the definition of congruent relation，the algorithm and concept of generalized relation (reflective relation) were put forward，and statistic linear correlation and gray relation were unified in the reflectional relation. The gray relation was proved to be a kind of proportional relation and a quantitative concept for 'reaction' was also built in the paper. At last，an example showed that the taxis of data series' relation is different with different relation mode.

Keywords：　relation；correlation；reflection；gray relation

关联分析或相关分析，是数据处理和系统分析中最普遍的方法之一，尤其在灰色系统理论中具有关键地位，应用非常广泛．但是，由于实际应用中的一些问题，近年来人们逐渐认识到灰色关联分析的缺陷，怀疑它不具有规范性[1]，研究者提出了许多对灰色关联度计算的改进方法[2]，如面积关联度、斜率关联度、相对变率关联度、绝对关联度等[3,4]，这种局面不能不说有点混乱，但预示了灰色关联的本质问题所在．本文将从一个完全不同的角度审视这个问题，得到一些结论，希望能和大家讨论．

目前，统计理论有相关的概念和相关系数的算法，与灰色系统理论中的关联概念和灰关联度算法，两者并不相容．关联，从广义上讲，是指集合之间的联系，数学上通常指不同数据对象之间相互影响、相互依赖的关系．我们认为，任何两个对象之间都是有联系的，只不过联系的方式和速度因条件而异，希望从一般的意义上建立关联的量化概念和算法，这就是广义关联．在广义关联体系下，统计理论的线性相关和灰色系统理论的灰色关联都是某种特定的关联．

广义关联以全等关联为基础．

1　全等关联

全等关联是指两个向量(序列)接近完全相等的程度，其几何意义是相互重合的程度．对任意两个向量(或序列，以后不再指明) $X=(x_1,x_2,\cdots,x_n)$、$Y=(y_1,y_2,\cdots,y_n)$，全等关联用全等关联度 $R=(X,Y)$ 来度

收稿日期：1998-02-24

量, 元素之间的全等关联用关联系数 $\gamma(x_i, y_i)$ 刻画, 这沿用了灰色系统理论的一些思想 .

全等关联取的是距离的反义, 但并不等效, 不能直接用距离来说明关联 . 我们有如下定义:

定义 1　x, y, z, u 为实数, 存在函数 $\gamma \in (0, 1)$, 满足:

1) $\gamma(x, x) = 1$;

2) $\gamma(x, y) = \gamma(y, x)$;

3) 若 $\gamma(x, u) \geq \gamma(y, u)$, 且 $\gamma(y, u) \geq \gamma(z, u)$, 则 $\gamma(x, u) \geq \gamma(z, u)$.

称 $\gamma(x, y)$ 为 x, y 的关联系数 .

定义 2　X, Y, Z, U 均为 n 维向量, 存在单值函数 $R_= \in (0, 1)$, 满足:

1) $R_=(X, X) = 1$;

2) $R_=(X, Y) = R_=(Y, X)$;

3) 若 $R_=(X, U) \geq R_=(Y, U)$, 且 $R_=(Y, U) \geq R_=(Z, U)$, 则 $R_=(X, U) \geq R_=(Z, U)$.

称 $R_=(X, Y)$ 为 X, Y 的全等关联度 .

构造关联系数的算法有很多, 其中, 把它表示为距离的函数是最基本的方法 . 而全等关联度的一般算法是, 首先对向量作变换:

$$(X, Y) \to \triangle(X, Y) = (\triangle_1, \triangle_2, \cdots, \triangle_i, \cdots, \triangle_n),\quad \triangle_i = |x_i - y_i|;$$

并存在 n 维零向量 $O(0, 0, \cdots, 0)$, 不妨将 $\triangle(X, Y)$ 也简写为 $\triangle$, 令 $R_=(X, Y) = R_=(\triangle, O)$.

计算 $\triangle(X, Y)$ 与 O 的全等关联度有如下两种方法:

1) 加权平均距离法, 取

$$\beta = \frac{1}{n}\left(\sum_{i=1}^{n} \omega_i \triangle_i^p\right)^{(1/p)} \tag{1}$$

$$R_=(X, Y) = \gamma(\beta, 0)$$

其中 β 为 $\triangle$ 与 O 的加权平均距离, ω 为权系数, $0 \leq \omega \leq 1$. 把关联系数表示为:

$$\gamma(\beta, 0) = 1 - a\beta^b, \qquad a, b \text{ 为特定系数} . \tag{2}$$

在统计学中经常使用均方差:

$$\beta = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \triangle_i^2} \tag{3}$$

2) 加权平均关联系数法, 计算 $\gamma_i = \gamma(\triangle_i, 0)$, $i = 1, 2, \cdots, n$ 如邓氏关联系数[5]:

$$\gamma = (\lambda\triangle_{\max} + \triangle_{\min}) / (\triangle_i + \lambda\triangle_{\max}). \tag{4}$$

取 γ_i 的加权平均值为 X, Y 的全等关联度, 即

$$R_=(X, Y) = \frac{1}{n}\sum_{i=1}^{n} \omega_i \gamma_i, \quad \omega \text{ 为权系数} .$$

2　映射关联

广义关联实质就是以全等关联为基础的映射关联, 即向量之间能够通过某种映射变换接近完全相等的程度 . 映射可以用函数表示, 也可以用一系列串联的变换表示 .

定义 3　对 n 维向量 X, Y, 设有映射 f, 使

$$F = f(X, Y),$$

$F = (f_1, f_2, \cdots, f_n)$ 为一 n 维向量 . 则

$$R_f(X, Y) = R_=(F, O)$$

为 X, Y 之间关于 f 的映射关联度 . 其中 O 是 n 维零向量 .

映射关联度不一定满足对称性 $R_f(X, Y) = R_f(Y, X)$.

由全等关联和映射变换建构的这种广义关联概念和算法体系, 具有很大的灵活性 . 在此体系下, 关联度是一个不明确的量, 某映射关联度才有实在的意义, 这提高了关联分析的明晰程度, 回答了“怎样关联”

的问题.

下面是一些典型的映射关联:

1) 平行关联, 表示两个向量对应分量之差值保持一致的趋势, 在几何上表示平行的趋势.

平行关联度定义为 $R_{\#}$:

$$R_{\#}(X, Y) = R_{=}(X - A, Y - B) \tag{5}$$

其映射方式为:

$$\#: F = (X - A) - (Y - B),$$

其中, $A = (a, a, \cdots, a)$, $B = (b, b, \cdots, b)$ 均为 n 维向量, a, b 是特定的平移系数. 平行关联度反映的是 X 能够通过平移变换 $X + (B - A)$ 与 Y 相接近的程度.

2) 保比关联, 表示两个向量对应分量之比值保持一致的趋势, 几何意义是两个向量夹角为零的趋势.

保比关联度定义为 R_{α}

$$R_{\alpha}(X, Y) = R_{=}(aX, bY). \tag{6}$$

其映射方式为:

$$\alpha F = aX - bY,$$

a, b 为特定比例系数. 保比关联度反映的是向量 X 能够通过比例变换 aX/b 与 Y 相接近, 或者向量 Y 能够通过比例变换 bY/a 与 X 接近的程度.

3) 线性关联, 表示两个向量线性关联(统计学称线性相关)的程度.

线性关联度记为 $R_{\angle}$:

$$R_{\angle}(X, Y) = R_{=}(aX - C, bY - D),$$

其中, $C = (c, c, \cdots, c)$, $D = (d, d, \cdots, d)$ 均为 n 维向量, a, b, c, d 是特定系数.

其映射方式为:

$$\angle: F = (aX - bY) - (C - D);$$

或者

$$\angle: F = a(X - C') - b(Y - D').$$

其中 n 维向量 $C = (c', c', \cdots, c')$, $D = (d', d', \cdots, d')$, c', d' 也是特定系数.

如果取映射为:

$$h: \begin{cases} f_i = (x_i - \bar{x})/r_1 - (y_i - \bar{y})/r_2, & \alpha > 0; \\ f_i = (x_i - \bar{x})/r_1 + (y_i - \bar{y})/r_2, & \alpha < 0; \end{cases} \tag{8}$$

即统计学中的标准化变换. 其中 α 是用于反映正负相关的向量内积:

$$\alpha = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

比例系数 r_1, r_2 为标准协方差:

$$r_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad r_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

采用均方差(3)式来计算全等关联度, 有:

$$R_{\angle}(X, Y) = R_{=}(X/r_1, Y/r_2) = \gamma(\beta, 0),$$

$$\beta = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\triangle_i^2} = \sqrt{2 - \frac{2}{r_1 r_2}\left|\sum_{i=1}^{n} x_i y_i\right|} = \sqrt{2(1 - |r|)},$$

其中, r 为统计理论的线性相关系数:

$$r = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

同时, 由(2)式令 $\gamma(\beta, 0) = 1 - a\beta^b = 1 - \frac{1}{2}\beta^2$,

即

$$R_{\angle}(X, Y) = |r| \tag{9}$$

在此把(8)式 h 映射下的映射关联度称为统计线性关联度, 记为 $R_{h\angle}$, 由(9)式可知, 线性相关系数的绝对值就是统计线性关联度. 去掉(8)式中的数据平移项, 线性相关系数将退化成统计学的相似系数, 其绝对值即统计保比关联度, 记为 $R_{h\theta}$. 仅用平均值进行平移变换, 可以得到统计平行关联度, 记为 $Rh_{\#}$. 显然, 平行关联和保比关联是线性关联的特例, 是比线性关联约束更强的关联.

统计关联度具有对称性和规范性(关联度不随相应的线性变换而变化):

$$R_h(X,Y)=R_h(Y,X),$$
$$R_{h_{\#}}(X,Y)=R_{h_{\#}}(X+B,Y),$$
$$R_{h\theta}(X,Y)=R_{h\theta}(aX,Y),$$
$$R_{h\angle}(X,Y)=R_{h\angle}(aX+B,Y).$$

B 是以 b 为分量的常向量, a, b 为常数且 $a\neq 0$. 具有这些性质的关联度不妨称为标准关联度.

3　灰色关联是一种保比关联

定理　初值化处理的灰色关联是对应映射为 $g: F=X/x_1-Y/y_1$ 的保比关联.

以 Y 为参考序列, X 为比较序列, 灰色关联的初值化变换是

$$p: X\to X/x_1,\ Y\to Y/y_1$$

令 $X_1=X/x_1$, $Y_1=Y/y_1$, 则 x_i 与 y_i 的关联系数为

$$\gamma_i(p)=(\omega+\lambda m)/(\triangle_i+\lambda m),\qquad i=1,2,\cdots,n,$$

其中 $\omega=\min\{|x_i(k)-y_i(k)|\}$, $m=\max\{|x_i(k)-y_i(k)|\}$, $k=1,2,\cdots,n$, λ 为分辨系数, $\triangle_i=|x_1(i)-y_1(i)|$. 于是灰关联度为:

$$R_g(X,Y)=R_=(X/x_1,Y/y_1)=\frac{1}{n}\sum_{i=1}^{n}\gamma_i,$$

即以 x_1, y_1 为比例系数的保比关联度.

不难证明多个比较序列时的情况, 由此可见, 灰色关联又是保比关联的一个特例, 它反映的是参考序列 Y 与比较序列 X 对应元素比值保持 $y(1)/x(1)$ 的趋势, 它不是线性关联度而是保比关联度. $R_g(X,Y)$ 和统计保比关联度一样具有对称性和规范性:

$$R_g(X,Y)=R_g(Y,X),$$
$$R_g(X,Y)=R_g(aX,Y),$$
$$R_g(X,Y)\neq R_g(Ax+B,Y),\qquad a\neq 0,\ B\neq O,$$

这说明灰色关联度也属于标准关联度, 之所以有人认为它不满足规范性是因为误作线性关联度.

4　关联分析

在广义关联下, 关联分析有了更加广阔的空间, 例如: 1) 研究不同的关联度算法对关联度计算结果的影响; 寻找等效关联度; 2) 比较相同映射下多个向量(序列)的关联度, 给关联排序; 3) 对确定的向量(序列), 比较不同映射下的关联度, 给映射排序, 在系统建模中的选择使关联度达到最大的映射. 实际上, 无论是一般统计学, 还是灰色系统理论, 数据处理中的无量纲化和标准化过程中, 都包含映射变换, 因此关联分析或相关分析不能不和这种映射有关.

极大映射关联度可以作为系统建模的目标函数, 为模型的参数辨识服务. 可以证明, 由(8)式决定的映射变换, 是使线性关联度达到最大的一种映射, 根据这一映射能够确定 a, b, c, d 等变换系数的最佳值. 这也是最小二乘法得到的结果. 非线性回归常常被转换为线性, 得到的并不是严格最优解, 如用指数 $y=ae^{bx}$ 来拟合 X, Y 序列, 采用 $\{\ln(y_i)\}$ 与 $\{x_i\}$ 的线性回归得到的参数 a, b, 并不是使 $\{a\exp(bx_i)\}$ 与 $\{x_i\}$ 的全等关联度达到最大的参数, 因为变量代换影响了求参精度[6].

在此提出一个新概念:

影响 设向量 X, Y 之间对映射 $f(X, Y)$ 的关联度为 R, 并且由向量方程 $f(X, Y) = O$ 存在一阶可导函数 $y(x)$ 和 $x(y)$, 则

$$L_f(x/y) = R \cdot (\mathrm{d}y/\mathrm{d}x),\ L_f(y/x) = R \cdot (\mathrm{d}x/\mathrm{d}y)$$

分别称为关于 f 映射 X 对 Y 的影响 $L_f(x/y)$, 以及 Y 对 X 的影响 $L_f(y/x)$. 这对研究数据序列之间的相互关联和相互作用是有用的.

5 实例

表 1 中列出参考序列和比较序列的一批数据, 其关系如图 1, 表 2 是关联度计算结果.

表 1

序数	参考序列	比较序列		
N	Y	X_1	X_2	X_3
1	424.4	25.6	380.8	167.1
2	433.6	32.3	572.8	175.8
3	519.6	120.1	622.4	293.3
4	642.2	243.7	653.7	457.8
5	731.6	330.2	659.4	574.0
6	831.1	432.5	700.9	709.5
7	881.5	480.7	722.4	774.5
8	1072.0	671.8	862.5	1029.1

表 2

序数	与参考序列 Y 的关联度			关联度排序
N	X_1	X_2	X_3	
灰色关联度	0.73	0.99	0.94	$X_2 > X_3 > X_1$
平行关联度	0.99	0.62	0.70	$X_1 > X_3 > X_2$
统计保比关联度	0.95	0.99	0.98	$X_2 > X_3 > X_1$
线性相关系数	1.00	0.90	0.99	$X_1 > X_3 > X_2$

1) 灰色关联度

根据初值化序列计算灰关联系数, 分辨系数 λ 取 0.9, 灰关联度为灰关联系数的平均值. 计算得

$$R_g(X_1, Y) = 0.73,\quad R_g(X_2, Y) = 0.99,\quad R_g(X_3, Y) = 0.94,$$

即 X_2 与 Y 的灰关联度最大.

2) 平行关联度

采用(4)式计算关联系数:

$$\gamma_i(k) = (0.313 + 184.342\lambda)/[\triangle i(k) + 184.342\lambda],\quad i = 1, 2, 3;\ k = 1, 2, \cdots, 8,$$

其中 $\triangle_i(k) = \left|[x_i(k) - \overline{x}] - [y(k) - \overline{y}]\right|$, 184.342 和 0.313 分别为 $\triangle i(k)$ 的最大最小值. 由平均关联系数得平行关联度:

$$R_=(X_1, Y) = 0.99,\ R_=(X_2, Y) = 0.62,\ R_=(X_3, Y) = 0.70,$$

显然, X_1 与 Y 具有很好的平行关联, X_2 则相反.

3) 统计保比关联度

由统计学相似系数, 得 $R_{h\theta}(X_1, Y) = 0.95$, $R_{h\theta}(X_2, Y) = 0.99$, $R_{h\theta}(X_3, Y) = 0.98$, 其排序结果与灰关联度相同, 但是分辨效果要差.

4) 线性关联

计算统计线性相关系数:

$$r(X_1, Y) \approx 1.00,\ r(X_2, Y) = 0.90,\ r(X_3, Y) \approx 0.99$$

即 X_1, X_2, X_3 与 Y 之间都有很好的线性关联. 但相对而言 X_2 的关联最差.

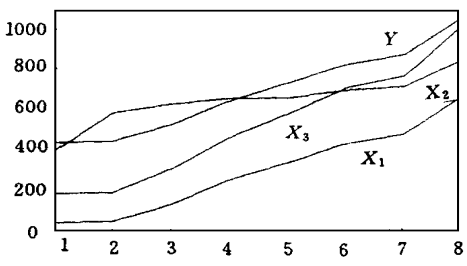

图 1 参考序列与比较序列的关系

参考文献:
[1] 肖新平. 关于灰色关联度量化模型的理论研究和评论. 系统工程理论与实践, 1997, 17(8): 76～81.
[2] 李明凉. 灰色关联度新判别准则及其计算公式. 系统工程, 1998, 16(1): 68～70.
[3] 张岐山. 灰关联熵分析方法. 系统工程理论与实践, 1996, 16(8): 7～11.
[4] 赵艳林等. 灰色关联分析的一种新的理论模型. 系统工程与电子技术, 1998, 20(10): 36～39.
[5] 邓聚龙. 灰色预测与决策. 武汉: 华中理工大学出版社, 1992.
[6] 李鸿仪. 线性回归中变量代换对回归精度的影响及清除. 数学的实践与认识, 1994, 24(3): 44～49.

(上接第 41 页)

第七步　一致逼近算法过程, 输出逼近系数如下:

- 8.183476*e* + 01	- 2.83241*e* + 01	- 4.457695*e* + 01	- 0.000000*e* + 00
- 9.009301*e* + 01	- 2.880612*e* + 01	- 1.273961*e* + 00	2.457784*e* + 01
- 6.750706*e* + 01	- 2.229463*e* + 01	- 3.603187*e* + 00	1.694623*e* + 01
- 3.670105*e* + 01	6.018483*e* + 00	- 2.759708*e* + 01	8.274547*e* + 00
- 4.174231*e* + 02	- 4.721543*e* + 02	- 6.164308*e* + 02	1.488026*e* + 02
- 2.928755*e* + 02	3.340412*e* + 01	- 6.559219*e* + 01	3.375131*e* + 02
- 2.175894*e* + 03	5.767623*e* + 02	- 3.546701*e* + 02	7.958941*e* + 02
- 4.234000*e* + 01	- 1.777559*e* + 01	- 1.400290*e* + 01	1.202591*e* + 01

第八步　最后评价结果

5.600000e + 01　5.200000e + 01　4.000000e + 01　2.200000e + 01
3.600000e + 02　1.200000e + 02　1.098000e + 03　2.600000e + 01

第九步　最优排序

{广东, 江苏, 浙江, 河北, 山西, 内蒙, 甘肃, 辽宁}

第十步　意义解释: 由最终排序可以看出, 广东自改革开放以来就是我国的经济窗口, 国家对其投资大, 政策优惠, 所以广东的经济发展在 8 个省中居于首位. 江浙一带工业发达, 乡镇企业也很活跃, 经济水平也在前列. 而我国中部的河北、山西发展居中游. 辽宁向来是我国的重工业基地, 但近几年来, 重工业不景气, 所以辽宁发展有些滞后.

参考文献:
[1] 帕失利迪斯(美). 结构模式识别. 上海: 上海科学技术文献出版社.
[2] 汪应洛. 系统工程理论方法与应用. 北京: 高等教育出版社, 1992
[3] 杨维权等. 多元统计分析. 北京: 高等教育出版社, 1990.
[4] 赵希男. 主成分分析法评价功能浅析. 系统工程, 1995, 13(2): 24～27.
[5] 朱孔来. 评价指标的非线性无量纲模糊处理方法. 系统工程, 1996, 14(6): 58～62.
[6] 姜旭平等. PCA 方法及其在多准则评估模型中的应用. 系统工程理论与实践, 1997, 17(4): 110～115.
[7] 严鸿和等. 专家评分机理与最优综合评价模型. 系统工程理论与实践, 1989, 9(2): 19～23.