一种 FUZZY 相关分析方法

徐承伟

(昆明工学院自控系)

摘要 线性的相关分析方法只能处理线性相关与否的问题，它无法用来了解是否存在非线性的相关．本文中我们提出一种近似的相关分析方法——Fuzzy 相关分析法，它主要可用于建立 Fuzzy 模型时对 Fuzzy 变量间的相关程度加以度量．此法的原理建立在 Fuzzy 规则模型的基础之上，数学上并不严格．但大量的算例表明，这种 Fuzzy 相关分析方法有相当的实用价值．

关键词：相关分析，Fuzzy，Fuzzy 模型

1 引言

相关分析，乍一看是个很成熟的领域，几乎每一本线性回归的教科书中都包含了相关分析的经典内容．其大略如下：给定观测数据$\{y(k),\ x(k),\ k=1,\ \cdots,\ N\}$，则 x 和 y 间的相关系数为

$$r=l_{12}/(l_{11}l_{22})^{1/2}\in[-1,\ 1] \tag{1}$$

其中

$$l_{11}=\sum_k[x(k)-a]^2/N,\quad a=\sum_k x(k)/N$$

$$l_{22}=\sum_k[y(k)-b]^2/N,\quad b=\sum_k y(k)/N$$

$$l_{12}=\sum_k[x(k)-a][y(k)-b]/N$$

相关系数 r 表明了 x 与 y 间的线性相关程度．$|r|$愈接近 1，线性相关愈强，反之愈弱．$r=\pm1$ 是为完全线性相关，而 $r=0$ 为完全线性无关．$r>0$ 表明是正相关，即 $x\uparrow\rightarrow y\uparrow$；$r<0$ 是负相关，即 $x\uparrow\rightarrow y\downarrow$．式(1)的导出建立在假定的模型 $y=cx+e$ 上，其中 e 是个正态噪声．

当 y 同时受到多个 $x_1,\ \cdots,\ x_m$ 影响时，就存在 m 个相关系数 $r_1,\ \cdots,\ r_m$．这些系数不能直接用(1)式计算，而要用所谓“偏相关系数”的概念(见[1]，pp.429–434)来计算．在计算某一个 x，如 x_2，与 y 间的偏相关系数 r_2 时，需要考虑消除 $x_1,\ x_3,\ \cdots,\ x_m$ 对 y 的影响．这种偏相关系数的计算公式，仍然建立在 $y=f(x_1,\ \cdots,\ x_m)$为线性模型的假定之上．所以，可称之为“线性偏相关系数”．

如果 y 与 x 之间有非线性的函数关系，则用计算线性相关系数的公式得到的结果往往不说明问题．例如，可以找到一个非线性函数 $y=f(x)$，用它产生出样本数据来计算的话，线性相关系数却为 0.

Fuzzy 模型是用来描述系统行为的一类模型，它是一种本质非线性的模型．利用观测

收稿日期：1990—04—25，昆明工学院青年科学基金资助课题．

样本数据建立 Fuzzy 模型的工作是 Fuzzy 系统理论的内容之一. 建模时, 设模型输出为 y, 有时需要在一组候选的输入 x_1, …, x_m 中确定哪些 x 对 y 有显著影响, 而哪些 x 又是次要的, 可以忽略不计的(例如, 动态 Fuzzy 模型的结构辨识问题). 这一问题实际上可以用求取 y 和各 x 间的相关系数的方法来解决. 但 y 和各 x 间的关系是 Fuzzy 的, 因而是非线性的, 线性的一套相关系数公式多半不适用. 所以, 需要一套能用于 Fuzzy 变量间相关分析的方法.

本文提出一种 Fuzzy 相关分析方法, 用以根据观测样本数据求 Fuzzy 变量 y 和诸 x 间的(偏)相关系数. 此法本身是与呈推理规则形式的 Fuzzy 模型的建立密切相联的, 它的计算结果(Fuzzy 相关系数)只具有近似的意义. 不过对于 Fuzzy 模型来说, 要求严格的精确性是既无可能, 也无意义的.

2 Fuzzy 相关分析方法

Fuzzy 相关分析方法试图解决如下的问题: 给定观测样本$\{y(k), x_1(k), \cdots, x_m(k); k=1, \cdots, N\}$, 确定 y 与 x_1, …, x_m 间的 Fuzzy 偏相关系数 r_1, …, r_m. 这些系数 r_1, …, r_m 应归一化(如取值于[−1, 1]或[0, 1]), 并能指明 y 与 x_1, …, x_m 间的联系强度. 尤其是 r_1, …, r_m 的计算应不依赖于关于 y 与 x_1, …, x_m 间函数类型的假定, 即具备通用性.

既然 Fuzzy 相关分析意在为建立 Fuzzy 模型提供帮助, 我们先从建立 Fuzzy 模型讨论起. 为方便计, 先考虑 $m=1$ 的情形. 呈规则形式的 Fuzzy 模型可写作

$$\text{IF} \quad x=A(i) \quad \text{THEN} \quad y=B(i), \ i=1, \cdots, n \tag{2}$$

这里 $A(i)$, $B(i)$为 Fuzzy 常量, 共有 n 条规则. 从相关分析的角度看, 不仿把 $A(i)$和 $B(i)$简化为论域 X 和 Y 上的小区间. 设 X 上的小区间为 $A_1=(a_1, b_1)$ (即当 $a_1<x<b_1$ 时 $x=A_1$, 余类推), $A_2=(a_2, b_2)$, …, $A_p=(a_p, b_p)$; Y 上的小区间为 $B_1=(c_1, d_1)$, …, $B_p=(c_p, d_p)$. 这里假定了每一论域上的小区间数为 p. 从建模的角度看, 诸 A_i 应满足如下条件:

$$(1) \quad A_1 \cup A_2 \cup \cdots \cup A_p = X \quad (完备性) \tag{3a}$$

$$(2) \quad A_i \cap A_j = \phi(空集), \ i \neq j. \quad (语义单一性) \tag{3b}$$

对 B_i 有类似要求. 至此, 可以把规则集(1)表为一个 $p\times p$ 矩阵, 它的一般元素 $R(i, j)$取值或 0 或 1. 当 $R(i, j)=1$, 表明存在一条规则"IF $x=A_j$ THEN $y=B_i$"; 而 $R(i, j)=0$ 就说明无此规则. 规则的总数显然等于 R 中所有为 1 的元素个数. 由于 R 和规则集(1)是一回事, 下面就把二者混为一谈.

由样本数据建立 Fuzzy 模型 R 的过程很简单: 只要把有样本数据支持的 $R(i, j)$取值为 1, 而其余 $R(i, j)=0$ 即可.

R 是 x 和 y 间的一个近似模型, 从它可以看出 x 和 y 的联系是否密切. 下面是两个 R 的例子.

$$R_1=\begin{bmatrix}8&1&0&0&0&0\\2&11&1&0&0&0\\0&3&20&5&0&0\\0&0&1&15&1&0\\0&0&0&3&16&4\\0&0&0&0&2&18\end{bmatrix},\quad R_2=\begin{bmatrix}7&2&1&0&0&1\\7&6&5&0&0&0\\0&2&8&6&2&0\\0&1&4&8&4&0\\0&0&0&3&10&9\\0&0&0&1&6&9\end{bmatrix}. \tag{4}$$

这里元素 $R(i,j)$的定义与前述稍有不同：$R(i,j)$="支持规则′IF $x=A_j$ THEN $y=B_i$′的样本数据个数". 例如，$R(3,4)=5$ 表明支持规则″IF $x=A_4$ THEN $y=B_3$″的样本数据共有5对等等. 一眼可以看出，R_1 中规则比较集中，R_2 中规则比较分散. 规则的集中程度，事实上反映了 y 与 x 的相关程度，规则愈集中则相关愈强. 所以，若设置一个反映 R 中规则集中程度的量 r，则 r 可作为反映 x 与 y 相关程度的一个指标.

问题是如何定义 r. 一个现成而明显的选择是以 R 中的规则数(即 R 中不为零的元素个数 n)来构造 r. 例如，对于 R_1，有 $n=16$；而对 R_2，$n=21$. 规则数目少，说明形成的规则比较集中，即相关较强；反之则相关较弱. 显然 R_1 对应着较强的相关. 不过，用 n 来构造 r 有如下不足之处：(1)对 n 很难做归一化处理. (2)即 $m>1$，即 y 受到不止一个 x 的影响时，更难以用 n 来反映每一个 x_j对 y 的影响(偏相关).

下面我们用 $R(i,j)$构造 r. 我们断言：设 y 与 x 有单值关系，则在矩阵 R 中，每一行中非零元素较少(即给定 B_i，与之对应的 A_j较少)时，x 与 y 相关较强，否则相关较弱. 其理由为：假如一行中非零元素较多，说明 x 在较大范围内变化时，y 不为所动，因此相关较弱. 但当某一行的元素 $R(i,j)$全为零时，则此行不能提供涉及相关程度的任何信息.

根据以上分析，x 与 y 间 Fuzzy 相关系数的一个可能的定义为

$$r'=\sum_i f(n_i)/p-q) \tag{5}$$

其中 n_i 表矩阵 R 第 i 行中非零元素的个数. 函数 $f(\cdot)$是一个从自然数集$\{0,1,\cdots\}$到$[0,1]$的映射. 根据前面的讨论，$f(\cdot)$应是 n_i 的减函数. 例如，可定义

$$f(n_i)=\begin{cases}1/n_i & n_i>0\\0 & n_i=0\end{cases} \tag{6}$$

其中 $n_i=0$ 属于未定情形，即此 n_i 实际上对 r' 未作贡献. 所以每有一个 $n_i=0$，就应使归一化因子(分母 p)中减去1. 故式中的 q 定义为：在 $i=1,\cdots p$ 中，$n_i=0$ 的个数.

把和(6)关于 r' 的定义用于前述 R_1，可得 $r'=[f(2)+f(3)+f(4)+f(2)+f(3)+f(2)]/6=[3\times(1/2)+3\times(1/3)]/6=0.416667$. 用于 R_2，就得到 $r'=[f(4)+f(3)+f(4)+f(4)+f(3)+f(3)]/6=[3\times(1/3)+3\times(1/4)]/6=0.291667$. 这两个 r' 确实描述了 R_1 和 R_2 在相关程度上的差异. 不过用(5)，(6)两式定义的 r' 略嫌苛刻，因为 R_1 从 Fuzzy 的观点来看相关较强，算得的 $r'=0.416667$ 太小，离完全相关($r'=1$)太远. 当然这可以通过修改(6)式 $f(\cdot)$的定义来加以调整.

然而仔细考察 R_1(或 R_2)就会明白，用 n_i 来构造 r 是不公平的. 因为一条规则可能只有一组样本数据来支持，也可能有许多组样本数据来支持. 因此，应该用样本数据的集中程度而非规则的集中程度来构造 r，即把支持每一条规则的样本数据数也考虑进来. 此

时，$f(\cdot)$的自变量将不再是 n_i，而是 R 的各行向量.

R 的每一行向量，反映了样本数据的集中程度. 当大多数数据都集中在少数元素上时，表明数据分布较集中，即相关程度较高(例如 R_1 的第 4 行[0 0 1 15 1 0])：若数据在各元素上分布较分散，就说明相关程度低(如 R_2 之第 2 行[7 6 5 0 0 0]). 以下用 $R(i)$表示矩阵 R 的第 i 行构成的行向量. 每一个 $R(i)$中样本数据分布的集中程度可以用多种方法表示. 例如，用统计学中熟悉的样本方差 $\sigma^2=\sum[x(k)-a]^2/N$($a$ 是 $x(k)$的均值)来表示. 对于向量 $R_1(4)=[0\ 0\ 1\ 15\ 1\ 0]$，可算出 $\sigma^2=5.9805$；对 $R_2(2)=[7\ 6\ 5\ 0\ 0\ 0]$，$\sigma^2=3.34664$. 前者大，说明 R_1 的第 4 行较之 R_2 的第 2 行有更强的相关. 不过用 σ 做相关指标也不方便，比如难于做归一化处理等.

我们采用如下的方式来计算每一个 $R(i)$中数据分布的集中程度. 首先定义 s_i，

$$s_i \triangleq \begin{cases} 0, & \text{当}R(i)\text{所有元素为}0 \\ 1+\sum_j |R(i,\ j)-R(i,\ jj)| \times R(i,\ j)/R(i,\ jj) & \text{其它} \end{cases} \tag{7}$$

其中求和是对 $j=1,\ \cdots,\ jj-1,\ jj+1,\ \cdots,\ p$ 进行的；$R(i,\ jj)$是 $R(i,\ j)$中的最大值，$R(i,\ jj)$可以看作为向量 $R(i)$的“数据分布中心”. 如存在多个 jj时可取其居中者. 例如，对于 $R_1(1)$，$jj=1$，$R_1(2)$的 $jj=2$，$R_1(3)$的 $jj=3$，$R_1(4)$的 $jj=4$，等等. 由定义(7)知，s_i 的最小非零值是 1，且 $s_i=1$ 表明 $R(i)$中只有 1 个非零元素；这对应了数据分布最集中(相关最强)的情形. 随着 $R(i)$中数据分布从 $R(i,\ jj)$这个中心元素向外延伸，$R(i)$的相关便减弱. 所以，当 $s>1$ 时，s 愈大，相关愈弱；s 愈小，相关愈强. 至于 $R(i)$所有元素为零时，样本数据未在 $R(i)$中出现，所以这一 $R(i)$不对总的相关系数做任何贡献：此时一方面 $s_i=0$，另一方面归一化因子(p)也要加 1(见下面(8)式). 最后，Fuzzy 相关系数 r 被定义为

$$r \triangleq \sum_i f(s_i)/(p-q) \in [0,\ 1] \tag{8}$$

其中：$q \triangleq s_1,\ \cdots,\ s_p$ 中，等于零的元素个数；$f(s_i)$定义为

$$f(s) \triangleq \begin{cases} 0 & \text{当}s=0\text{或}s>3 \\ 1/s^{1/2} & 1 \leqslant s < 2 \\ 1/s & 1.3 \leqslant s < 2 \\ 1/s^3 & 2 \leqslant s < 3 \end{cases} \tag{9}$$

这一定义正是根据上面的讨论得出的. 从式(8)可见，总相关系数 r 取决于每一行的相关水平. 把(8)和(9)定义的 Fuzzy 相关系数计算公式用于前述 R_1，有

$$s_1=1+1/8=1.125,\quad s_2=1+2/11+1/11=1.27273,$$
$$s_3=1+3/20+5/20=1.4,\quad s_4=1+1/15+1/15=1.13333,$$
$$s_5=1+3/16+4/16=1.570833,\quad s_6=1+2/18=1.11111$$

所以，Fuzzy 相关系数为

$$\begin{aligned} r &= [f(1.125)+f(1.27273)+f(1.4)+f(1.11111)+f(1.57083)+f(1.11111)]/(6-0) \\ &= (0.9428+0.8864+0.7143+0.9393+0.6366+0.9487)/6=0.844688 \end{aligned}$$

类似地，对 R_2 可得 $r=0.15714$.

讨论一. Fuzzy 相关系数 r 的定义，主要是函数 $f(\cdot)$的定义. 在保证 $f(0)=0$，

$f(1)=1$,及当 $s>1$ 时 $f(s)$是 s 的减函数诸条件的前提下,$f(\cdot)$当然可以有多种选择,很难找出什么“最优”的 $f(\cdot)$的定义.不过可以指出,$s>1$ 时,$f(s)-s$ 曲线愈平缓,r 对数据相关程度的灵敏度愈低,即结果愈“Fuzzy”;反之则结果趋于精确.因此,$f(\cdot)$如何定义,实与 r 的具体用途有关.特别是为建立 Fuzzy 模型而计算 r 时,r 的 Fuzzy 程度与模型的 Fuzzy 程度一致.

讨论二.由于每个论域都要划分为若干个小区间,我们自然面临两个问题:(1)一个论域划分为几个区间($p=?$);(2)如何划分.p 的选择影响结果的精确性,p 愈大,结果愈精确.因此可根据对精确性的要求来确定 p.至于如何划分,我们在后面的算例中用了两种方法,一种是把论域划分为 p 个等长的小区间(记为 Part−1),另一种是使每个小区间包含的样本数据对数相同(记为 Part−2).此外还可以采用聚类算法来确定论域.究竟以何种方法为好,只能看效果.

讨论三.从 r 的定义可见,我们的出发点是“如果样本数据中同一个 y 值对应的 x 值分布较集中,则 x 和 y 的相关程度较高;反之则较低”.这意味着(8),(9)式定义的 Fuzzy 相关系数 r 只适用于 x 和 y 有(Fuzzy)单值关系的场合.假如一个 y 可以对应多个 x(如 $y=\sin x$ 者),上述定义便不复适用.不过,可以修改上述关于 r 的定义,使之也适用于多值关系.比方说,在向量 $R(i)$中寻找一个以上的数据分布中心等.该问题本文不拟考虑.

讨论四.上述 Fuzzy 相关系数无法指明 y 与 x 间究竟是正相关还是负相关.

以上仅讨论了 y 只受一个 x 影响的情形($m=1$).下面给出根据观测样本$\{y(k),\ x_1(k),\ \cdots,\ x_m(k),\ k=1,\ \cdots,\ N\}$计算 Fuzzy 偏相关系数 $r_1,\ \cdots,\ r_m$ 的公式.限于篇幅,只给出计算 r_1 的公式,但其余 $m-1$ 个公式不难简单地类推获得.

步 1. 把论域划分为小区间,记为 $Y=(B_1,\ \cdots,\ B_p)$,$X_1=(A_{11},\ \cdots,\ A_{1p})$,$\cdots$,$X_m=(A_{m1},\ \cdots,\ A_{mp})$.再由样本数据计算出一个 $m+1$ 维矩阵 R

$$R \triangleq \{R(i,\ j_1,\ \cdots,\ j_m)\in\{0,\ 1,\ 2,\ \cdots\} \mid i,\ j_1,\ \cdots,\ j_m\in\{1,\ \cdots,\ p\}\} \tag{10}$$

其中的元素 $R(i,\ j_1,\ \cdots,\ j_m)$对应于一条 Fuzzy 规则

$$\text{IF } x_1=A_{j1} \text{ and}\cdots\text{and } x_m=A_{jm} \text{ THEN } y=B_i \tag{11}$$

且 $R(i,\ j_1,\ \cdots,\ j_m)$在数值上等于支持规则(11)的样本数据的组数.

步 2. Fuzzy 偏相关系数 r_1 定义为

$$r_1=\sum_i\sum_{j2}\cdots\sum_{jm} f(s(i,\ j_2,\ \cdots,\ j_m))/(p^m-q) \tag{12}$$

其中 $s(i,\ j_2,\ \cdots,\ j_m)$是个标量,它根据向量$[R(i,\ 1,\ j_2,\ \cdots,\ j_m),\ \cdots,\ R(i,\ p,\ j_2,\ \cdots,\ j_m)]$计算出,计算公式仿(7).函数 $f(\cdot)$的定义同(9),其中的 q 定义为:对所有 $i,\ j_2,\ \cdots,\ j_m=1,\ \cdots,\ p$,使 $s(i,\ j_2,\ \cdots,\ j_m)=0$ 的$(i,\ j_2,\ \cdots,\ j_m)$的个数.

3 若干实例

由于所提出的 Fuzzy 相关分析方法缺乏数学上的严格性,我们计算了大量的例子以考察此法的性能.下面给出部分计算结果.除特别指明的外,例子中的样本数据均由计算机根据仿真模型产生.其中的输入(x)是区间$[l,\ h]$上均匀分布的随机数,且输出(y)中叠加了在区间$[-a,\ a]$均匀分布的不相关噪声.

例 1 仿真模型为 $y=1/x$,数据区间$[l,\ h]=[0.1,\ 2]$,数据长度 $N=100$,每一论域

小区间数 $p=6$. 表 1 给出了不同噪声水平 a 下的 Fuzzy 相关系数 $r(F)$和用于比较的线性相关系数 $r(L)$.

表 1 不同噪声水平下的相关系数(论域划分 Part-1)

a	0	0.1	0.2	0.4
$r(F)$	0.994967	0.91918	0.760057	0.422866
$r(L)$	−0.776611	−0.77012	−0.776482	−0.772751

可见，$r(F)$在低噪声时较好地反映了 y 与 x 间的密切关系. 但噪声水平增大后 $r(F)$显著下降，因为此时对同一 y，x 的分布变得均匀了.

例 2 仿真模型为 $y=x_1/x_2$，$[l, h]=[0.1, 10]$，$N=200$，$p=6$. 结果示于表 2.

表 2 不同噪声水平下的相关系数(论域划分 Part-1)

a	0	0.1	0.2	0.4	0.8
$r_1(F)$	0.747651	0.745571	0.582984	0.425036	0.37038
$r_2(F)$	0.737114	0.703067	0.603453	0.472976	0.417601
$r_1(L)$	0.238858	0.239054	0.239178	0.239208	0.238431
$r_2(L)$	−0.541162	−0.541021	−0.540757	−0.539865	−0.536675

例 3 举一个动态系统建模的例子. 原始数据$\{y(k), u(k), k=1, \cdots, 296)\}$来自[3]. 假设 Fuzzy 模型形如 $y(k)=f(u(k-d_1), y(k-d_2))$，现欲用 Fuzzy 相关分析法确定最合适的 d_1 和 d_2 值. 分别用 r_1，r_2 表 $y(k)$—$u(k-d_1)$及 $y(k)$—$y(k-d_2)$间的相关系数. 取 $p=6$. 表 3 是计算结果.

表 3 例 3 之计算结果(论域划分 Part-2)

d_1	d_2	$r_1(F)$	$r_2(F)$	$r_1(L)$	$r_2(L)$
1	1	0.233647	0.89083		
2	1	0.412282	0.878879	−0.765083	0.976213
3	1	0.680837	0.921753	−0.852434	0.974311
4	1	0.830736	0.914437	−0.826464	0.939493
5	1	0.864612	0.906801	−0.557194	0.785173
6	1	0.881167	0.922332	0.180445	0.824994
7	1	0.763685	0.939771		
8	1	0.61688	0.946136		
9	1	0.389681	0.936308		
10	1	0.266908	0.905138		
6	2	0.89083	0.819761		
6	3	0.83451	0.451406		
6	4	0.809008	0.326365		
6	5	0.771322	0.28077		

从 Fuzzy 分析的结果来看，最合适的结构参数是 $d_1=6$，$d_2=1$；而线性相关分析则要求 $d_1=3$，$d_2=1$. 这也许说明 Fuzzy 模型的结构参数未必应与线性模型的结构一致.

4 结束语

我们提出的 Fuzzy 相关分析法从本质上讲是一种统计方法，它在线性相关分析法不适用的场合有可能提供合理的分析结果. 它目前的缺点是：(1)可调的因素过多(如 $f(\cdot)$

及 s 的定义等); (2)对 $y—x$ 间有多值关系的情形不适用. 我们下一步的目标是在这些方面加以改进.

参 考 文 献

1 Younger M S. A First Course in Linear Regression (2nd Edition). Boston: Duxbury Press, 1979

2 数学手册. 北京: 人民教育出版社, 1979

3 Box G E P, Jenkins G M. Time Series Analysis, Forecasting and Control. San Francisco: Holden Day, 1970

AN APPROACH TO FUZZY CORRELATION ANALYSIS

XU Chengwei

(Dept. Of Autocontrol, Kunming Polytechnic College)

Abstract

Linear correlation analysis approach can't be used to determine whether nonlinear correlation exists but to solve linear correlated or noncorrelated problems. This article presents an approximate algorithm—Fuzzy correlation analysis approach, which can be mainly used to measure the correlation levels between Fuzzy variants in Fuzzy modelling. This algorithm is based on the model of Fuzzy laws and is not mathematically strict; however, plenty of examples show it is good and practical.

Keywords: correlation analysis, Fuzzy, Fuzzy model