文章编号: 1001-0920 (2015) 08-1447-06 DOI: 10.13195/j.kzyjc.2014.0814

面板数据下的灰色指标关联聚类模型与应用

李雪梅, 党耀国, 王俊杰

(南京航空航天大学 经济与管理学院，南京 211106)

摘 要: 针对面板数据聚类研究存在的问题及现实需要, 构建面板数据下新的灰色指标关联聚类 (AGRA) 模型. 构造所有指标不同对象下时间序列的累加生成序列, 用生成序列的平均生成速率表征原序列的动态变化趋势; 单个指标所有对象的平均生成速率构成该指标的平均生成速率序列, 从而综合偏离、差离和分离的三重差异信息, 构建指标关联分析模型; 提出面板数据下 Mean-AGRA 灰色指标关联聚类算法, 并应用于我国区域生态环境评价指标的降维问题. 分析结果验证了所提出模型的实用性和有效性.

关键词: 灰色关联聚类；灰色关联分析；面板数据；AGRA 模型

中图分类号: N94 **文献标志码:** A

Grey relational clustering model for panel data clustering on indicators and its application

LI Xue-mei, *DANG Yao-guo*, *WANG Jun-jie*

(College of Economics and Management，Nanjing University of Aeronautics and Astronautics，Nanjing 211106, China．Correspondent：LI Xue-mei，E-mail：xuemeili85@163.com)

Abstract: In order to solve the problems in existing methods of clustering for panel data, grey clustering analysis based on grey relational analysis on accumulation sequences(AGRA) is put. The original data for certain index is accumulated generation firstly and the generation sequences are simulated. Then the dynamic similarity of change trend of the original panel data is characterized by the proximity of generation rate sequences. The AGRA model is comprehensive for deviation, difference and discrete degrees, and a grey clustering analysis is proposed. Based on the AGRA model, the factors in the regional ecological environment system are studied by using the grey clustering analysis. This application is presented to illustrate the effectiveness and practicality of the proposed model.

Keywords: grey relational cluster；grey relational analysis；panel date；AGRA model

0 引 言

灰色关联聚类是根据研究对象之间关联度的大小对其实现分类的一种方法. 二维数据下, 关于灰色关联模型 (GRA) 的大量研究成果[1-4]是灰关联聚类的基础, 可以根据实际问题的需要选取不同的 GRA 模型. 文献 [4] 指出, 现实中的大量实际问题和科学难题, 迫切需要运用关于面板数据等高维数据去研究, 因此, 面板数据的关联模型及聚类分析是非常有价值的研究方向. 在面板数据下的聚类方法中, 灰色关联聚类的优势是对样本数量和时间长度的要求不高. 面板数据下的灰关联聚类需解决两个问题, 一是面板数据的灰色关联分析: 文献 [5] 在三维空间讨论了指标的几何特征相似性; 文献 [6] 利用海塞矩阵定义凸度, 用数据的凸性表征样本之间的相似程度, 提出了三维灰色凸关联度; 文献 [7] 基于多指标面板数据的时空特征构建了灰色矩阵关联分析模型; 文献 [8] 利用面板数据的几何特征构建了灰色网格关联度. 二是根据关联度进行聚类, 现有针对面板数据的灰色关联聚类大多采用文献 [9] 提出的特征变量临界值聚类方法. 另外, 与灰色关联聚类相似, 传统聚类方法在面板数据下的

收稿日期: 2014-05-22；**修回日期:** 2014-08-01.

基金项目: 国家自然科学基金项目(71371098)；中央高校基本科研业务费专项基金项目(NC2012001)；江苏高校哲学社会科学重点研究基地重大项目(2012JDXM005)；江苏省普通高校研究生科研创新计划(CXZZ12_0174); 南京航空航天大学博士学位论文创新与创优基金项目(BCXJ12-11)；国家社会科学基金重大项目(14ZB151).

作者简介: 李雪梅(1985−), 女, 博士生, 从事灰色系统理论与应用的研究；党耀国(1964−), 男, 教授, 博士生导师, 从事灰色系统理论与应用等研究.

拓展多数是通过构造面板数据或其矩阵形式下新的距离或相似性度量[10-13], 有的也先对面板数据进行降维或转化再构建聚类模型[14].

现有文献中关于面板数据灰色关联分析和灰关联聚类的研究为数不多, 总结现有的面板数据灰色关联及聚类模型的研究, 可以发现以下值得改进的地方: 若考察指标在不同对象间的波动, 则不应仅限于相邻对象之间, 而应对所有对象全面考察, 以避免对象排列顺序不同引起关联结果的变化; 针对不等时长的面板数据的关联、评价与决策方法需要进一步研究; 直接衡量原始数据的接近性或相似性容易忽略数据需要经过数据变换才能显现的规律; 临界值关联聚类方法有时会出现聚类结果中有的类内指标关联度过小的现象.

鉴于以上问题, 本文先构建面板数据下指标AGRA模型, 进而构造灰色指标Mean-AGRA聚类模型. 非等时长的面板数据普遍存在, 基于Mean-AGRA的面板数据指标聚类模型具有广泛的应用领域, 本文以我国区域生态环境评价指标的降维过程为例来验证所提出模型的实用性和有效性.

1 面板数据指标AGRA关联分析模型的构建

定义1 $t\,(t=1,2,\cdots,t_i^s)$时刻对象$s\,(s=1,2,\cdots,N)$的指标$i\,(i=1,2,\cdots,m)$的值为$x_i^{(0)}(s,t)$, 则

$$X_i^{(0)}(s,t)=\begin{bmatrix} x_i^{(0)}(1,1) & x_i^{(0)}(1,2) & \cdots & x_i^{(0)}(1,t_i^1)\\ x_i^{(0)}(2,1) & x_i^{(0)}(2,2) & \cdots & x_i^{(0)}(2,t_i^2)\\ \vdots & \vdots & \ddots & \vdots\\ x_i^{(0)}(N,1) & x_i^{(0)}(N,2) & \cdots & x_i^{(0)}(N,t_i^N)\end{bmatrix}$$

为指标i的行为矩阵, 面板数据

$$X=(X_1^{(0)}(s,t),X_2^{(0)}(s,t),\cdots,X_i^{(0)}(s,t),\cdots,X_m^{(0)}(s,t))$$

为指标行为矩阵序列. 指标i在不同对象下的取值时长不一定相等.

通过累加生成可以看出灰作用量累积过程的发展趋势, 使原本看似没有规律的原始数据中蕴含的积分特征或规律显现出来[5], 且经过拟合累加生成序列再进行关联分析便可不受序列时长相等的限制.

指标i在对象s下的原始数据序列为$X_i^{(0)}(s,T)=(x_i^{(0)}(s,1),x_i^{(0)}(s,2),\cdots,x_i^{(0)}(s,t_i^s))$, 对此序列建立GM(1,1)模型[5], 得其累加序列的拟合方程为

$$\hat{x}_i^{(1)}(s,t)=c\mathrm{e}^{-at}+\frac{b}{a}. \tag{1}$$

其中

$$c=x_i^{(0)}(s,1)-\frac{b}{a};$$

$$x_i^{(1)}(s,k)=\sum_{q=1}^{k}x_i^{(0)}(s,q),\ k=1,2,\cdots,n;$$

$$Z_i^{(1)}(s,k)=\frac{1}{2}(x_i^{(1)}(s,k)+x_i^{(1)}(s,k-1));$$

$$(a,b)^{\mathrm{T}}=(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}Y;$$

$$Y=[\,x_i^{(0)}(s,2)\ \ x_i^{(0)}(s,3)\ \ \cdots\ \ x_i^{(0)}(s,t_i^s)\,]^{\mathrm{T}};$$

$$B=\begin{bmatrix} -z_i^{(1)}(s,2) & -z_i^{(1)}(s,3) & \cdots & -z_i^{(1)}(s,t_i^s)\\ 1 & 1 & \cdots & 1\end{bmatrix}^{\mathrm{T}}.$$

由$\hat{x}_i^{(1)}(s,t)$拟合方程及斜率含义可知, 拟合曲线t时刻切线斜率为

$$h_i^s(t)=\frac{\mathrm{d}\hat{x}_i^{(1)}(s,t)}{\mathrm{d}t}. \tag{2}$$

闭区间$[1,t_i^s]$内$\hat{x}_i^{(1)}(s,t)$拟合函数的均值为

$$\bar{x}_i^{(1)}(s,t)=\frac{1}{t_i^s-1}\int_1^{t_i^s}\hat{x}_i^{(1)}(s,t)\mathrm{d}t. \tag{3}$$

斜率是直线倾斜程度的量度, $X_i^{(1)}(s,T)$拟合曲线t时刻切线斜率是$X_i^{(1)}(s,T)$在t时刻的导数值, 即$X_i^{(1)}(s,T)$在t时刻变化的速度, 但没有消除数量级及量纲的影响, 再除以均值构造生成速率便达到了消除数量级及量纲的作用, 用生成速率的接近程度可以有效表示生成序列变化速度的相似性. 生成速率越接近, 原序列关联度越大.

定义2 设$t\in[a,b]$, 称

$$p_i^s(t)=\frac{h_i^s(t)}{\bar{x}_i^{(1)}(s,t)} \tag{4}$$

为$x_i^{(0)}(s,t)$在t时刻的生成速率, 则$X_i^{(0)}(s,T)\,(s=1,2,\cdots,N;\ i=1,2,\cdots,m)$的平均生成速率为

$$p_i^s=\frac{1}{t_i^s-1}\sum_{t=1}^{t_i^s}p_i^s(t)=\frac{1}{t_i^s-1}\sum_{t=1}^{t_i^s}\frac{h_i^s(t)}{\bar{x}_i^{(1)}(s,t)}. \tag{5}$$

定义3 $X_i^{(0)}(s,T)\,s=1,2,\cdots,N)$的平均生成速率构成的序列称为指标$i\,(i=1,2,\cdots,m)$的平均生成速率序列, 表示为$P_i=(p_i^1,p_i^2,\cdots,p_i^N)$.

通过GM(1,1)模型构建的指标平均生成速率序列对关联性及聚类误差的影响取决于两个方面, 一方面是GM(1,1)模型的模拟误差, 另一方面是平均生成速率序列的关联度的误差. 首先, GM(1,1)模型擅长处理小样本时间序列的模拟预测问题; 其次, 经过灰色累加生成一定程度上降低了干扰因素的影响, 对关联度计算及聚类结果的影响误差也可忽略.

下面用指标平均生成速率序列的偏离程度、差离程度和分离程度衡量指标的相似性, 从这3个角度可以较全面地体现原面板数据的时空特征. 用两个指标在不同对象下平均生成速率向量的夹角代表指标向量的偏离程度; 用两个指标在对应对象下的平均生成速率差值表示指标间的差离程度; 用指标在不同对

象下平均生成速率序列的离散系数反应指标的分离程度.

定义 4　指标 $i,j\,(i,j=1,2,\cdots,m)$ 的平均生成速率向量分别为

$$\vec{P}_i=(p_i^1,p_i^2,\cdots,p_i^N),$$
$$\vec{P}_j=(p_j^1,p_j^2,\cdots,p_j^N),$$

则指标 i,j 的偏离度为

$$\eta_{ij}=1-\cos(\vec{P}_i,\vec{P}_j)=1-\frac{\vec{P}_i\cdot\vec{P}_j}{|\vec{P}_i||\vec{P}_j|}=$$
$$1-\frac{\sum\limits_{s=1}^{N}p_i^s\cdot p_j^s}{\sqrt{\sum\limits_{s=1}^{N}(p_i^s)^2}\cdot\sqrt{\sum\limits_{s=1}^{N}(p_j^s)^2}}.\tag{6}$$

定义 5　指标 $i,j\,(i,j=1,2,\cdots,m)$ 的平均生成速率序列分别为

$$P_i=(p_i^1,p_i^2,\cdots,p_i^N),$$
$$P_j=(p_j^1,p_j^2,\cdots,p_j^N),$$

则关于对象 s 的指标 i,j 的差离值为

$$d_{ij}^s=|p_i^s-p_j^s|.\tag{7}$$

指标 i,j 的差离度为各个对象差离值的均值, 表示为

$$\xi_{ij}=\frac{1}{N}\sum_{s=1}^{N}d_{ij}^s.\tag{8}$$

定义 6　指标 $i,j\,(i,j=1,2,\cdots,m)$ 的平均生成速率序列分别为

$$P_i=(p_i^1,p_i^2,\cdots,p_i^N),$$
$$P_j=(p_j^1,p_j^2,\cdots,p_j^N),$$

则指标 i,j 的分离度为平均生成速率序列离散系数的差值, 表示为

$$\partial_{ij}=\Big|\frac{\frac{1}{N}\sqrt{\sum\limits_{s=1}^{N}(p_i^s-\bar{p}_i^s)^2}}{\bar{p}_i^s}-\frac{\frac{1}{N}\sqrt{\sum\limits_{s=1}^{N}(p_j^s-\bar{p}_j^s)^2}}{\bar{p}_j^s}\Big|.\tag{9}$$

为了综合原始面板数据变化趋势的偏离、差离和分离的三重异构信息, 先构造偏离关联度、差离关联度和分离关联度, 再构造三者融合的关联度.

定义 7　根据偏离度、差离度和分离度的含义, 分别定义偏离 (Deviate) 关联度、差离 (Difference) 关联度和分离 (Variance) 关联度如下:

$$\begin{cases}\mathrm{DE}_{ij}=\dfrac{1}{1+\eta_{ij}},\\ \mathrm{DI}_{ij}=\dfrac{1}{1+\xi_{ij}},\\ \mathrm{VA}_{ij}=\dfrac{1}{1+\partial_{ij}}.\end{cases}\tag{10}$$

3 种关联度是从不同的角度来度量指标的相似程度. 文章用平均生成速率向量的夹角来表征指标平均生成速率的偏离关联度, 代表了指标的平均生成速率构成的向量的偏离程度, 偏离关联度越大说明夹角越小, 向量变化方向越接近; 用指标平均生成速率的大小的差的绝对值来表征差离度, 进而构造差离关联度, 代表了指标变化速率的接近程度, 差离关联度越大说明指标的变化速率大小越接近; 用离散系数的差值来表征分离度, 进而构造分离关联度, 反映了指标在不同样本的变化速度的分散程度, 分离关联度越大说明指标在不同样本中的波动程度越接近.

定义 8　对指标的偏离关联度、差离关联度和分离关联度设置权重, 构造指标基于累加生成序列的关联度 (AGRA), 则指标 i 与 j 的 AGRA 关联度为

$$\zeta_{ij}=\gamma_{\mathrm{DE}}\mathrm{DE}_{ij}+\gamma_{\mathrm{DI}}\mathrm{DI}_{ij}+\gamma_{\mathrm{VA}}\mathrm{VA}_{ij},$$
$$\gamma_{\mathrm{DE}},\gamma_{\mathrm{DI}},\gamma_{\mathrm{VA}}\geqslant 0,\ \gamma_{\mathrm{DE}}+\gamma_{\mathrm{DI}}+\gamma_{\mathrm{VA}}=1,\tag{11}$$

其中 γ_{DE}、γ_{DI}、γ_{VA} 分别为偏离关联度、差离关联度和分离关联度的权重, 代表对 3 种信息的重视程度, 可根据实际问题的需要进行设置, 一般情况下可取 $\gamma_{\mathrm{DE}}=\gamma_{\mathrm{DI}}=\gamma_{\mathrm{VA}}=1/3$.

容易证明 AGRA 关联度满足规范性、无量纲化处理不变性、唯一性和相似性等.

由于研究中所用数据很多, 需要经过数据变换后才能显示真实规律, 且原始数据有关联关系的序列进行累加生成后也具有准指数规律. 因此, AGRA 关联度不仅适用于原始数据直接有关联关系的序列, 也适用于具有累积效应需要累加才能显现规律的序列. 另外, AGRA 关联度与对象排列顺序无关, 避免了对象排列顺序引起的指标关联序的变化. 运用拟合的累加生成序列也可以成功解决时长不等的问题.

2　基于 AGRA 模型的面板数据指标 Mean-AGRA 聚类方法

定义 9　指标 i,j 的 AGRA 关联度为 ζ_{ij}, 则

$$A=\begin{bmatrix}\zeta_{11} & \zeta_{12} & \cdots & \zeta_{1m}\\ & \zeta_{22} & \cdots & \zeta_{2m}\\ & & \ddots & \vdots\\ & & & \zeta_{mm}\end{bmatrix}.$$

称 A 为特征变量的扩展 AGRA 关联矩阵.

文献 [9] 中的临界值灰色关联聚类方法可以根据实际问题设定面板数据下指标聚类分析的关联度临界值 $r\in[0,1]$, 一般要求 $r>0.5$. 当 $\zeta_{ij}>r$, $i\neq j$ 时, 视 X_i 与 X_j 为同类特征. r 越接近于 1, 分类越细, 每一组分类中的变量越少; r 越小, 分类越粗.

定义 10 对指标进行 AGRA 分析, 得到指标的 AGRA 关联度, 则特征变量在临界值 r 下的分类称为特征变量的 r-AGRA 聚类.

r-AGRA 聚类方法与文献 [5-7] 类似, 有时会造成类内关联度较小的情况, 因此只能作为一种快速聚类方法适当选用.

下面构造面板数据指标 Mean-AGRA 聚类方法, 基本思想是: 首先将每个指标各成一类, 然后每次缩小一类, 每缩小一类, 类内平均关联度就要减小, 选择能使类内平均关联度最大的两类合并, 直至所有样本归为一类为止.

定义 11 设指标 $X_i\,(i=1,2,\cdots,p)$ 已成一类, 记为 $G_a=\{X_i, i=1,2,\cdots,p\}$, 则 G_a 类内关联度为

$$\zeta_a^{\rm mean}=\frac{2}{p(p-1)}\sum_{i=1,i<j}^{p}\zeta_{ij},$$

指标 X_l 与类 G_a 合并后的关联度为

$$\zeta_{al}^{\rm mean}=\frac{2}{p(p+1)}\Big(\sum_{i=1,i<j}^{p}\zeta_{ij}+\sum_{i=1,X_l\notin G_a}^{p}\zeta_{li}\Big),$$

类 $G_b=\{X_j, j=1,2,\cdots,q\}$ 与 $G_a=\{X_i, i=1,2,\cdots,p\}$ 的关联度为

$$\zeta_{ab}^{\rm mean}=\frac{2}{pq}\sum_{j=1,i<j}^{q}\sum_{i=1}^{p}\zeta_{ij},$$

$$X_i\in G_a,\ X_j\in G_b.$$

在此定义下的聚类方法称为平均关联度视角的 AGRA 聚类 (Mean-AGRA 聚类) 方法.

Mean-AGRA 聚类的算法如下.

Step 1: 由面板数据 AGRA 关联模型得到扩展 AGRA 关联矩阵 $A=\{\zeta_{ij}\}, 1\leqslant i,j\leqslant m$.

Step 2: 由 $\max\limits_{i\neq j}\zeta_{ij}=\zeta_{eu}$, 将 X_e 与 X_u 合并, 作为第 1 类 $G_1=\{X_e,X_u\}$.

Step 3: 分别计算未分组的指标与第 1 组合并后的关联度记为 $\zeta_{1l}^{\rm mean}$, 则

$$\zeta_{1l}^{\rm mean}=\frac{1}{3}\Big(\zeta_{eu}+\sum_{i=e,u;l\neq e,u}\zeta_{li}\Big).$$

同时计算未分组的指标间两两关联度的最大值, 即 $\max\limits_{i\neq j\neq e,u}\zeta_{ij}=\zeta_{hl}$, 再对这两个数值进行比较, 若 $\max\{\zeta_{hl},\zeta_{1v}^{\rm mean}\}=\zeta_{hl}$, 则将第 1 组之外的关联度最大的两个指标 X_h 与 X_l 合并, 作为 G_2, 否则将 X_v 与 G_1 合并, 若最大关联度不止一个, 则同时合并.

Step 4: 求类 $G_a=\{X_i, i=1,2,\cdots,p\}$ 与 $G_b=\{X_j, j=1,2,\cdots,q\}$ 的关联度

$$\zeta_{ab}^{\rm mean}=\frac{2}{pq}\sum_{j=1;i<j}^{q}\sum_{i=1}^{p}\zeta_{ij},$$

$$X_i\in G_a,\ X_j\in G_b,\ 1\leqslant a,\ b\leqslant m,$$

将关联度最大的两类合并.

Step 5: 重复 Step 3 和 Step 4 的过程, 直至所有指标归为一类, 聚类终止.

3 基于 Mean-AGRA 的面板数据指标聚类模型的构建步骤

本文构建 Mean-AGRA 的面板数据指标聚类模型的步骤如下.

Step 1: 由式 (1) 得到指标 $i\,(i=1,2,\cdots,m)$ 在对象 $s\,(s=1,2,\cdots,N)$ 下的累加生成序列的拟合方程;

Step 2: 由式 (2) ~ (5) 得到指标 $i\,(i=1,2,\cdots,m)$ 的平均生成速率序列;

Step 3: 由式 (6) ~ (9) 得到指标 $i,j\,(i,j=1,2,\cdots,m,\ i\neq j)$ 的偏离度、差离度和分离度;

Step 4: 由式 (10) 和 (11) 得到指标 $i,j\,(i,j=1,2,\cdots,m,\ i\neq j)$ 的 AGRA 关联度;

Step 5: 根据 Mean-AGRA 聚类算法的步骤得到指标的聚类过程.

4 实例分析

4.1 实例背景

生态环境的恶化已经成为困扰我国区域经济社会发展的重要问题, 为了有效地解决这一问题并为决策者提供决策有用的信息, 首先需要对生态环境进行全面、客观的评价, 而在具体评价过程中对指标体系进行筛选时, 往往容易出现指标重复现象. 因此, 如何避免这一现象是获得准确评价结果的关键, 而这一现象的规避可以采取本文提出的聚类模型来解决.

4.2 指标体系的构建与样本的选取

具体研究中, 在对我国区域生态环境评价指标进行聚类分析的基础上, 根据发展趋势的相似性程度对指标进行聚类, 并从中选取有代表性的指标构建相应的指标体系, 以作为进一步研究的依据, 以此可以达到避免指标重复并保证评价结果准确的目的. 下面以现有区域生态环境评价指标体系的研究[15]为基础, 从自然环境与社会环境两个二级指标出发构建指标体系, 根据空气质量、水环境、绿化环境、水污染、大气污染、废弃物污染、环境治理、城市基础建设、公共设施、人口因素这些三级指标选取了空气质量为优的天数占全年比例、人均水资源量、造林面积、绿化覆盖率、废水排放总量、二氧化硫排放总量、工业烟尘排放量、工业粉尘排放量、工业固体产生量、工业固体处置量、生活垃圾无害化处理率、环境污染治理投资总额、城市用水普及率、城市燃气普及率、人均城市道路面积、人均公园绿地面积、万人拥有

公共厕所、人均用水量、万人公共交通车辆、人口密度共20个量化指标作为初始指标体系; 时间跨度为2007～2012年; 样本为我国30个(由于西藏缺失数据较多, 暂未利用西藏的相关数据)省市自治区, 利用文中构建的AGRA模型和Mean-AGRA聚类方法对指标进行聚类分析.

4.3　聚类模型的构建

由于数据来源及统计口径的原因, 在这30个样本中, 2011年和2012年30个样本的工业烟尘排放量、2012年30个样本的环境污染治理投资总额、2007年海南绿化覆盖率、2008年内蒙古绿化覆盖率、2008年天津人均公园绿地面积存在数据缺失的现象, 本文采用AGRA模型, 通过拟合累加生成序列以把握时间序列的整体发展趋势, 可以有效处理这一问题. 另外, 本文构建的GM(1,1)模型的模拟误差均在允许范围以内. 根据本文第3节的步骤, 可以得到指标的聚类结果. 第1、第2类指标平均生成速率序列散点图如图1和2所示. 横坐标为30个省市, 纵坐标为平均生成速率的值, 一类图形代表一个指标的平均生成速率序列.

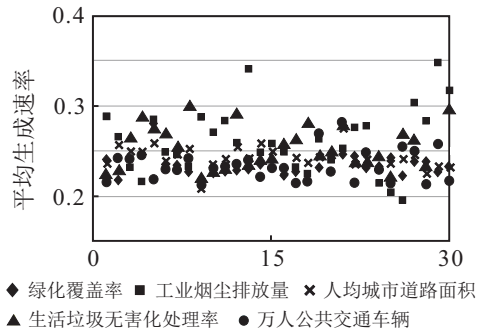

图 1　第 1 类指标平均生成速率序列散点图

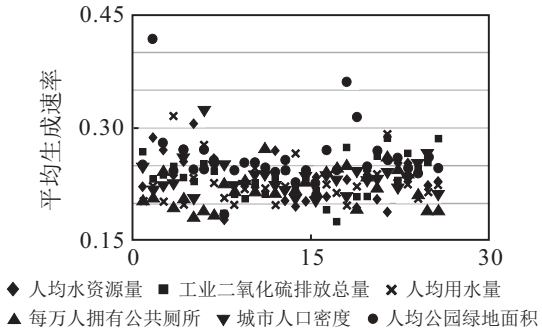

图 2　第 2 类指标平均生成速率序列散点图

4.4　结果分析

根据聚类结果及类内类间指标的密集程度, 本文认为初始指标体系可以分为10类, 其中工业烟尘排放量、生活垃圾无害化处理率、绿化覆盖率、人均城市道路面积、万人公共交通车辆归为第1类, 二氧化硫排放总量、人均用水量、人均水资源量、万人拥有公共厕所、人口密度、人均公园绿地面积、归为第2类, 城市用水普及率、城市燃气普及率归为第3类, 其余7个指标各成一类. 如图1所示, 第1类中指标的平均生成速率集中在0.2～0.3之间, 数值大小与分散程度都比较接近, 其中工业烟尘排放量与其他4个指标相比比较分散, 所以在第9步才将其归入这一类中. 如图2所示, 第2类中指标的平均生成速率介于0.15～0.45之间, 分散程度与第1类相比较大, 但类内取值相对接近. 同时可以发现, 指标相对密集的前两类集中代表了区域基础建设和公共设施水平, 说明代表区域基础建设与公共设施水平的指标发展趋势比较一致, 而环境污染和环境治理方面的指标与区域基础建设和公共设施水平的变化趋势相对差异程度较大. 以上的分析结果与现实情况也是相符的: 环境污染和环境治理属于“软件建设”的范畴, 而区域基础建设和公共设施水平则是“硬件建设”的范畴, 因此, 软硬件方面分在不同类别是合理的.

接下来确定10个代表性指标. 由于类内指标关联性较强, 选取哪个作为代表指标对问题的分析没有太大影响. 也可以参考这种方法选取: 前3类中的指标如果经过定性判断与独立成类的7个指标描述的对象相似, 则在类内选取代表性指标时不再考虑, 例如, 根据聚类结果, “造林面积”这一指标已经独立成类, 则第1类中不再选取同样反映绿化程度的绿化覆盖率作为代表性指标. 最终选取万人公共交通车辆、人口密度与城市用水普及率作为前3类的代表性指标, 与空气质量为优的天数占全年比例、工业固体产生量、废水排放总量、环境污染治理投资总额、工业固体处置量、造林面积、工业粉尘排放量7个独立成类的指标一起构成反映区域生态环境水平的指标体系. 最终确立的10个指标分属8个不同的二级指标, 说明其代表性, 并且从最初的20个指标降到10个, 也充分说明了降维的有效性. 因此, AGRA模型和Mean-AGRA聚类方法充分利用20个指标的30个样本在5年内的动态面板数据, 得到了指标的聚类结果, 从而达到了指标降维的效果.

5　结　论

针对现有灰色关联面板数据聚类及其他面板数据聚类存在的问题, 构建了面板数据AGRA模型, 通过生成速率序列的构造, 在对原始数据信息进行充分挖掘的同时, 达到了降维的效果, 不仅对等时长的面板数据有效, 也能有效处理非等时长的面板数据. 提出的新关联聚类原则避免了关联度较小的指标聚为一类的情况, 既拓展了灰色关联模型的应用范围, 又丰富了面板数据的研究. 将模型应用于我国区域生态

环境评价指标的降维问题中, 验证了本文所提出模型的实用性和有效性. 文中有效地结合了灰色准指数律挖掘原始数据的规律, 在以后的研究中也可以根据问题的实际背景考察其他规律; 文中 3 种关联度的权重确定方法除了主观确定也可以定量优化; 根据不同的聚类原则可以提出不同的聚类方法. 如何结合量化方法更合理地确定聚类数也是值得进一步研究的问题.

参考文献(References)

[1] 杜宏云, 施红星, 刘思峰, 等. 基于斜率判断的灰色周期关联度研究[J]. 中国管理科学, 2010, 18 (1): 128-132.
(Du H Y, Shi H X, Liu S F, et al. Study on the model of grey periodic incidence judged on slope[J]. Chinese J of Management Science, 2010, 18 (1): 128-132.)

[2] 孙玉刚, 党耀国. 灰色 T 型关联度的改进[J]. 系统工程理论与实践, 2008, 28 (4): 135-139.
(Sun Y G, Dang Y G. Improvement on grey T's correlation degree[J]. Systems Engineering-Theory & Practice, 2008, 28 (4): 135-139.)

[3] 郭昆, 张岐山. 基于灰关联分析的谱聚类[J]. 系统工程理论与实践, 2010, 30 (7): 1260-1265.
(Guo K, Zhang Q S. Spectral clustering based on grey relational analysis[J]. Systems Engineering-Theory & Practice, 2010, 30 (7): 1260-1265.)

[4] Liu S, Yang Y, Cao Y, et al. A summary on the research of GRA models[J]. Grey Systems: Theory and Application, 2013, 3(1): 7-15.

[5] 张可, 刘思峰. 灰色关联聚类在面板数据中的扩展及应用[J]. 系统工程理论与实践, 2010, 30(7): 1253-1259.
(Zhang K, Liu S F. Extended clusters of grey incidences for panel data and its application[J]. Systems Engineering-Theory & Practice, 2010, 30(7): 1253-1259.)

[6] 吴利丰, 刘思峰. 基于灰色凸关联度的面板数据聚类方法及应用[J]. 控制与决策, 2013, 28(7): 1033-1037.
(Wu L F, Liu S F. Panel data clustering method based on grey convex relation and its application[J]. Control and Decision, 2013, 28(7): 1033-1037.)

[7] 钱吴永, 王育红, 党耀国. 基于多指标面板数据的灰色矩阵关联模型及其应用[J]. 系统工程, 2013, 31(10): 70-74.
(Qian W Y, Wang Y H, Dang Y G. Grey matrix relational modeling and its application based on multivariate panel data[J]. Systems Engineering, 2013, 31(10): 70-74.)

[8] 刘震, 党耀国, 钱吴永, 等. 基于面板数据的灰色网格关联度模型[J]. 系统工程理论与实践, 2014, 34(4): 991-996.
(Liu Z, Dang Y G, Qian W Y, et al. Grey grid incidence model based on panel data[J]. Systems Engineering-Theory & Practice, 2014, 34(4): 991-996.)

[9] 刘思峰, 党耀国, 方志耕, 等. 灰色系统理论及其应用[M]. 第 3 版. 北京: 科学出版社, 2004.
(Liu S F, Dang Y G, Fang Z G, et al. Grey system theory and application[M]. 3th ed. Beijing: Science Press, 2004.)

[10] 朱建平, 陈民恳. 面板数据的聚类分析及其应用[J]. 统计研究, 2007, 24(4): 11-14.
(Zhu J P, Chen M K. The cluster analysis of panel data and its application[J]. Statistical Research, 2007, 24(4): 11-14.)

[11] 李因果, 戴翼, 何晓群. 基于自适应权重的面板数据聚类方法[J]. 系统工程理论与实践, 2013, 33(2): 388-395.
(Li Y G, Dai Y, He X Q. Clustering method for Panel data base on adaption weighting[J]. Systems Engineering-Theory & Practice, 2013, 33(2): 388-395.)

[12] Assmann C, Boysen-Hogrefe J. A Bayesian approach to model-based clustering for binary panel probit models[J]. Computational Statistics & Data Analysis, 2011, 55(1): 261-279.

[13] Juárez M A, Steel M F J. Model-based clustering of non-Gaussian panel data based on skew-t Distributions[J]. J of Business Economic Statistics, 2010, 28(1): 52-66.

[14] 徐华锋, 方志耕. 面板数据聚类分析的投影寻踪模型[J]. 统计与决策, 2010(4): 161-163.
(Xu H F, Fang Z G. Projection pursuit model of panel data clustering analysis[J]. Statistics and Decision, 2010(4): 161-163.)

[15] 罗上华, 马蔚纯, 王祥荣, 等. 城市环境保护规划与生态建设指标体系实证[J]. 生态学报, 2003, 23(1): 45-55.
(Luo S H, Ma W C, Wang X R, et al. A case study on indicator system of urban environmental protection and ecological construction[J]. Acta Ecologica Sinica, 2003, 23(1): 45-55.)

(责任编辑: 齐 霁)