文章编号: 1001-0920 (2013) 07-1033-04

基于灰色凸关联度的面板数据聚类方法及应用

吴利丰, 刘思峰

(南京航空航天大学 经济与管理学院, 南京 210016)

摘 要: 研究面板数据的聚类方法. 基于二维灰色凸关联度, 用二阶差商近似代替二阶导数, 利用黑塞矩阵的半正定性在三维空间中定义凸度; 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的概念, 讨论了三维灰色凸关联度的性质. 实例分析表明, 三维灰色凸关联度能够较好地反映面板数据的关联程度.

关键词: 灰色系统; 灰色关联度; 面板数据; 聚类分析

中图分类号: N94.5 **文献标志码:** A

Panel data clustering method based on grey convex relation and its application

WU Li-feng, LIU Si-feng

(College of Economics & Management, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. Correspondent: WU Li-feng, E-mail: wlf6666@126.com)

Abstract: The clustering method of panel data is studied. The 2nd order difference is used to substitute for the approximate 2nd derivative based on the grey convex relation for two dimension. The half positive characteristic of Hessian matrix is used to define the convexity degree. The similarity degree of objects is characterized by the convexity, the grey convex relation degree for three dimension is presented, and the properties of the grey convex relation degree for three dimension are discussed. An example is given to illustrate that the grey convex relation degree for three dimension can better reflect the correlation degree for three dimension panel data.

Key words: grey system; grey relational degree; panel data; clustering method

0 引 言

灰色关联分析是灰色系统理论中十分活跃的一个分支, 自提出以来便被广泛应用于系统的因素分析、方案决策和优势分析[1-2]. 众多学者以邓聚龙教授提出的灰色关联公理[3]为基础构造了若干新型灰色关联度模型[4], 有基于点关联系数的T型关联度[5]和B型关联度[6], 利用二序列曲线之间所夹面积来度量二序列曲线相似程度的绝对关联度[4]. 但是, 现有的灰色关联度大都主要适用于时间序列与截面数据的关联分析, 而对于面板数据的灰色关联分析则刚起步.

国内外对面板数据聚类分析的文献比较缺乏. 文献[7]将灰色绝对关联度扩展到三维面板数据中; 文献[8]分别评价了文献[9-11]的面板数据聚类方法, 并基于面板数据绝对量和动态量特征提出一种面板数据聚类方法; 文献[12]采用主成分分析方法对数据进行降维处理, 通过综合评价函数序列矩阵的相似指标对面板数据进行聚类. 然而, 文献[9-12]对统计量的预处理易造成数据信息丢失, 难以分析挖掘数据背后的深层次问题. 文献[13]将时间序列的局部变化特性与整体距离关系相结合, 将局部变化的信息融入相似测度的计算, 以实现序列的局部变化的形状特征提取, 提高了分类效果, 但是易丢失信息. 以往模型将所有指标同等对待, 只考虑不同类型指标的接近性, 未考虑指标发展趋势相似性和发展的协调性. 为此, 本文基于文献[14]的灰色凸关联度, 利用黑塞矩阵的半

收稿日期: 2012-03-05; **修回日期:** 2012-05-14.

基金项目: 国家自然科学基金项目(71173106, 70971064, 71171113); 江苏省高等学校优秀创新团队项目(P0702); 南京航空航天大学创新群体项目(Y0553); 特聘教授科研创新基金项目(1009-260812); 南京航空航天大学博士学位论文创新与创优基金项目(BCXJ12-13); 江苏省普通高校研究生科研创新计划项目(CXLX12_0176); 中央高校基本科研业务费专项资金项目.

作者简介: 吴利丰(1983−), 男, 博士生, 从事灰色系统理论的研究; 刘思峰(1955−), 男, 教授, 博士生导师, 从事数量经济学、灰色系统理论等研究.

正定性在三维空间中定义凸度, 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的概念, 讨论了三维灰色凸关联度的性质. 实例分析表明, 所提出的三维灰色凸关联度能够较好地反映面板数据的关联程度.

1 三维灰色凸关联度

对于二维数据, 正离散系统行为序列 $X=(x(1),x(2),\cdots,x(n))$, 若有

$$1\leqslant k<n-1,\ x(k+1)\geqslant\frac{x(k)+x(k+2)}{2},$$

则称 X 在 $t=k+1$ 处是凸的.

对于三维数据, 给出以下定义.

定义 1 面板数据中第 i 个对象在第 s 个指标 t 时刻的值记为 $x_i(s,t),s=1,2,\cdots,m,t=1,2,\cdots,n,3\leqslant n,3\leqslant m$, 则称

$$X_i=\begin{bmatrix}x_i(1,1)&x_i(1,2)&\cdots&x_i(1,n)\\x_i(2,1)&x_i(2,2)&\cdots&x_i(2,n)\\\vdots&\ddots&\vdots&\vdots\\x_i(m,1)&x_i(m,2)&\cdots&x_i(m,n)\end{bmatrix},$$

$$i=1,2,\cdots,L$$

为对象 i 的行为矩阵.

定义 2 设行为矩阵

$$X_i=\begin{bmatrix}x_i(1,1)&x_i(1,2)&\cdots&x_i(1,n)\\x_i(2,1)&x_i(2,2)&\cdots&x_i(2,n)\\\vdots&\vdots&\ddots&\vdots\\x_i(m,1)&x_i(m,2)&\cdots&x_i(m,n)\end{bmatrix},$$

$$i=1,2,\cdots,L,$$

则称

$$H(X_i(s,t))=\begin{bmatrix}\dfrac{\partial^2x_i(s,t)}{\partial s^2}&\dfrac{\partial^2x_i(s,t)}{\partial s\partial t}\\\dfrac{\partial^2x_i(s,t)}{\partial t\partial s}&\dfrac{\partial^2x_i(s,t)}{\partial t^2}\end{bmatrix},$$

$$i=1,2,\cdots,L$$

为 X_i 在点 (s,t) 处的黑塞矩阵, 其中

$$\frac{\partial x(s,t)}{\partial s}=x_i(s+1,t)-x_i(s,t),$$

$$\frac{\partial^2x(s,t)}{\partial s\partial t}=x_i(s+1,t+1)-x_i(s,t+1)-(x_i(s+1,t)-x_i(s,t)),$$

$$\frac{\partial^2x(s,t)}{\partial t^2}=(x_i(s,t+1)-x_i(s,t))-(x_i(s,t)-x_i(s,t-1)).$$

如果黑塞矩阵 $H(X_i(s,t))$ 是半正定矩阵, 即 $H(X_i(s,t))$ 的所有主子式大于或等于零, 则称 X_i 在点 (s,t) 处是凸的.

称 $\dfrac{\partial^2x_i(s,t)}{\partial s^2}$, $\dfrac{\partial^2x_i(s,t)}{\partial t^2}$ 和 $\dfrac{\partial^2x_i(s,t)}{\partial s^2}\dfrac{\partial^2x_i(s,t)}{\partial t^2}-\dfrac{\partial^2x_i(s,t)}{\partial t\partial s}\dfrac{\partial^2x_i(s,t)}{\partial s\partial t}$ 为点 (s,t) 的凸度, 凸度可以看成具有 3 个分量的向量. 如果凸度的 3 个数都为正, 则说明这一点高于周围其他点; 如果两点的凸度相等, 则这两点分别高于周围点的程度相等; 如果 2 个对象行为矩阵在某点的凸度越接近, 则说明 2 个对象行为矩阵在某点的关联度越大.

记

$$(X_i)_{11}=\frac{\partial^2x_i(s,t)}{\partial s^2},$$

$$(X_i)_{12}=\frac{\partial^2x_i(s,t)}{\partial s^2}\frac{\partial^2x_i(s,t)}{\partial t^2}-\frac{\partial^2x_i(s,t)}{\partial t\partial s}\frac{\partial^2x_i(s,t)}{\partial s\partial t},$$

$$(X_i)_{22}=\frac{\partial^2x_i(s,t)}{\partial t^2},$$

则可以看出, $(X_i)_{11}$ 给出了某时期内指标间的凸性, $(X_i)_{22}$ 反映了指标随时间变化的速度, $(X_i)_{12}$ 反映了指标发展的协调水平.

定义 3 设正的对象行为矩阵

$$X_i=\begin{bmatrix}x_i(1,1)&x_i(1,2)&\cdots&x_i(1,n)\\x_i(2,1)&x_i(2,2)&\cdots&x_i(2,n)\\\vdots&\vdots&\ddots&\vdots\\x_i(m,1)&x_i(m,2)&\cdots&x_i(m,n)\end{bmatrix},$$

$$i=1,2,\cdots,L,$$

则称

$$\gamma_{ij}^{11}(s,t)=\frac{1}{1+|(X_i)_{11}-(X_j)_{11}|},$$

$$\gamma_{ij}^{22}(s,t)=\frac{1}{1+|(X_i)_{22}-(X_j)_{22}|},$$

$$\gamma_{ij}^{12}(s,t)=\frac{1}{1+|(X_i)_{12}-(X_j)_{12}|}$$

为 X_i 与 X_j 在点 (s,t) 的三维灰色凸关联系数 $(i,j=1,2,\cdots,L)$, 称

$$\gamma_{ij}=\frac{1}{3(n-2)\times m}\sum_{t=2}^{n-1}\sum_{s=1}^{m}\gamma_{ij}^{22}(s,t)+\frac{1}{3n\times(m-2)}\sum_{t=1}^{n}\sum_{s=2}^{m-1}\gamma_{ij}^{11}(s,t)+\frac{1}{3(m-2)\times(n-2)}\sum_{t=2}^{n-1}\sum_{s=2}^{m-1}\gamma_{ij}^{12}(s,t)$$

为 X_i 与 X_j 的三维灰色凸关联度. 若对于相关对象行为矩阵 X_i 和 X_j 有 $\gamma_{ik}>\gamma_{jk}$, 则称对象 X_i 与 X_k 的关联优于对象 X_j 与 X_k 的关联, 记为 $X_i\succ X_j$, 称"$\succ$"为由三维灰色凸关联度导出的灰色关联序.

基于三维灰色凸关联度的面板数据聚类方法步骤如下.

Step 1: 根据定义 1, 将面板数据转化为对象行为矩阵.

Step 2: 对原始数据进行均值化处理, 消除量纲影响, 令

$$x_i'(s,t)=\frac{x_i(s,t)}{\bar{x}(s,t)},\ \bar{x}(s,t)=\frac{1}{L}\sum_{i=1}^{L}x_i(s,t).$$

对象行为矩阵变为

$$X_i'=\begin{bmatrix} x_i'(1,1) & x_i'(1,2) & \cdots & x_i'(1,n)\\ x_i'(2,1) & x_i'(2,2) & \cdots & x_i'(2,n)\\ \vdots & \vdots & \ddots & \vdots\\ x_i'(m,1) & x_i'(m,2) & \cdots & x_i'(m,n)\end{bmatrix},$$

$$i=1,2,\cdots,L.$$

Step 3: 先按定义 2 计算每个点的凸度, 再根据定义 3 计算任意 2 个对象的三维灰色关联度, 并构造对象的关联矩阵

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1L}\\ & \gamma_{22k} & \cdots & \gamma_{2L}\\ & & \ddots & \vdots\\ & & & \gamma_{LL}\end{bmatrix}.$$

实际应用中可根据问题设定聚类分析的关联度临界值 $r\in[0.5,1]$. 当 $\gamma_{ij}>r$ 时, 视 X_i 与 X_j 为同类指标, 遍历关联矩阵便得到聚类结果.

Step 4: 设定关联度临界值 r, 根据关联矩阵对指标进行聚类分析.

2　三维灰色凸关联度的性质

定理 1　三维灰色凸关联度具有以下性质:

1) 规范性, $0<\gamma_{ij}\leqslant 1$;

2) 偶对称性, 即 $\gamma_{ij}=\gamma_{ji}$, 且 X_i 与 X_j 在点 (s,t) 的关联系数与 X_i 和 X_j 在点 (t,s) 的关联系数相等;

3) 接近性, 即相对凹凸程度越接近, γ_{ij} 越大;

4) 可比性, 唯一性;

5) 干扰因素独立性[14];

6) 数乘变换一致性, 数乘变换保序性[15].

证明　对于性质 1), 三维灰色凸关联系数

$$0<\gamma_{ij}^{11}(s,t)\leqslant 1,\ 0<\gamma_{ij}^{22}(s,t)\leqslant 1,$$

$$0<\gamma_{ij}^{12}(s,t)\leqslant 1,$$

显然这些关联系数的均值 γ_{ij} 满足 $0<\gamma_{ij}\leqslant 1$. 性质 2) ~ 性质 5) 显然成立. 对于性质 6), 因为数据采用均值化处理, 对于发生数乘变换的对象行为矩阵, 其元素经过均值化处理后每个点的关联系数不变, 所以满足数乘变换一致性和数乘变换保序性. □

注 1　X_i 与 X_j 在点 (s,t) 的关联系数与 X_i 和 X_j 在点 (t,s) 的关联系数相等表明, 三维坐标的转化不影响关联系数的大小; 性质 5) 说明不同量纲的数据不影响关联度的大小.

因为凸关联度需采用 3 个数描述凸性, 因此本文的三维灰色凸关联度适用于每个维度上至少有 3 个数据的面板数据.

3　实例分析

选择文献 [16] 的数据 (见表 1), 依次设长沙、株洲、益阳、衡阳、邵阳为 X_1,X_2,X_3,X_4,X_5.

表 1　湖南省 5 市历年每万人中科技活动人员数、R&D 占 GDP 比例和第二产业增加值

年份	对象	长沙	株洲	益阳	衡阳	邵阳
	每万人中科技活动人员数	56.2	33.7	5.8	13.9	5.8
2003	R&D 占 GDP 比例 / %	1.6	1.3	0.1	0.3	0.2
	第二产业增加值 / 亿元	444.9	185.5	60.9	149.3	71.1
	每万人中科技活动人员数	72.7	29.2	4.4	17.2	4.1
2004	R&D 占 GDP 比例 / %	1.97	0.76	0.38	0.16	0.3
	第二产业增加值 / 亿元	541.7	218.4	71.1	181.9	84.5
	每万人中科技活动人员数	79.8	43.2	4.9	18.5	5.3
2005	R&D 占 GDP 比例 / %	1.69	0.79	0.19	0.35	0.33
	第二产业增加值 / 亿元	642.1	264.6	84.2	224.3	101.2
	每万人中科技活动人员数	83	50.6	6	17.9	4.8
2006	R&D 占 GDP 比例 / %	1.73	0.78	0.2	0.39	0.18
	第二产业增加值 / 亿元	791	311.8	102.2	272.1	122.4
	每万人中科技活动人员数	87.9	43.8	6	18.1	5.5
2007	R&D 占 GDP 比例 / %	1.88	1.12	0.34	0.47	0.18
	第二产业增加值 / 亿元	984.8	396.3	131.2	328.8	145.4
	每万人中科技活动人员数	100.2	46.1	5.1	22.1	5.8
2008	R&D 占 GDP 比例 / %	1.93	1.17	0.5	0.55	0.2
	第二产业增加值 / 亿元	1 567.4	497.4	170.7	412.4	182.2

对象行为矩阵为

$$X_1=\begin{bmatrix} 56.2 & 1.60 & 444.9\\ 72.7 & 1.97 & 541.7\\ 79.8 & 1.69 & 642.1\\ 83.0 & 1.73 & 791.0\\ 87.9 & 1.88 & 984.8\\ 100.2 & 1.93 & 1\,567.4\end{bmatrix},$$

$$X_2=\begin{bmatrix} 33.7 & 1.30 & 185.5\\ 29.2 & 0.76 & 218.4\\ 43.2 & 0.79 & 264.6\\ 50.6 & 0.78 & 311.8\\ 43.8 & 1.12 & 396.3\\ 46.1 & 1.17 & 497.4\end{bmatrix},\ X_3=\begin{bmatrix} 5.8 & 0.1 & 60.9\\ 4.4 & 0.38 & 71.1\\ 4.9 & 0.19 & 84.2\\ 6 & 0.2 & 102.2\\ 6 & 0.34 & 131.2\\ 5.1 & 0.5 & 170.7\end{bmatrix},$$

$$X_4=\begin{bmatrix} 13.9 & 0.3 & 149.3\\ 17.2 & 0.16 & 181.9\\ 18.5 & 0.35 & 224.3\\ 17.9 & 0.39 & 272.1\\ 18.1 & 0.47 & 328.8\\ 22.1 & 0.55 & 412.4\end{bmatrix},\ X_5=\begin{bmatrix} 5.8 & 0.2 & 71.1\\ 4.1 & 0.3 & 84.5\\ 5.3 & 0.33 & 101.2\\ 4.8 & 0.18 & 122.4\\ 5.5 & 0.18 & 145.4\\ 5.8 & 0.2 & 182.2\end{bmatrix}.$$

取 $r=0.6$, 按上述方法计算三维灰色凸关联度, 得到对象关联矩阵为

$$\begin{bmatrix} 1 & 0.643 & 0.591 & 0.769 & 0.368 \\ & 1 & 0.497 & 0.656 & 0.290 \\ & & 1 & 0.588 & 0.894 \\ & & & 1 & 0.334 \\ & & & & 1 \end{bmatrix}.$$

聚类结果如下: 长沙、株洲、衡阳为一类, 邵阳和益阳为一类, 与现实情况基本相符. 实际上, 长沙作为省会, 株洲作为湖南重要的工业基地, 衡阳作为湘南地区的经济中心和综合交通枢纽, 这些地区的经济实力较强. 从科技投入、创新能力和科技第二产业的协调程度看, 长沙、株洲和衡阳属于第一阶梯, 科技投入较高, 从而促进了第二产业的较快发展; 而邵阳主要受交通瓶颈制约, 经济发展较慢; 益阳是环洞庭湖城市群的主要城市之一, 石长铁路和长益高速公路开通以后, 尽管与长株潭区域的经济社会联系愈来愈紧密, 但由于经济基础差, 经济发展水平仍然较低. 邵阳和益阳两地属于科技投入相对不足、第二产业相对不发达的地区. 利用文献 [7] 的聚类方法的聚类结果为: 长沙、衡阳为一类, 邵阳、株洲和益阳为一类, 与实际情况有差别, 说明本文的聚类方法能够反映面板数据的关联程度.

4 结 论

针对面板数据的聚类问题, 从三维空间中的凸度入手提出了三维灰色凸关联度, 讨论了三维灰色凸关联度的性质. 三维灰色凸关联度适合于面板数据的指标聚类, 能够从动态角度描述事物的类别. 所提出的方法思路较为清晰, 在以矩阵序列为研究对象的计算机控制、图像处理领域具有较好的应用价值.

参考文献(References)

[1] 王嚞华, 朱建军, 方志耕. 基于灰色关联度的多阶段语言评价信息集结方法[J]. 控制与决策, 2013, 28(1): 109-114.
(Wang H H, Zhu J J, Fang Z G. Aggregation of multi-stage linguistic evaluation information based on grey incidence degree[J]. Control and Decision, 2013, 28(1): 109-114.)

[2] 宋捷, 党耀国, 花增木. 基于灰色聚类的群决策方法研究[J]. 控制与决策, 2010, 25(10): 1593-1597.
(Song J, Dang Y G, Hua Z M. Study on group decision-making method based on grey cluster model[J]. Control and Decision, 2010, 25(10): 1593-1597.)

[3] 邓聚龙. 灰理论基础[M]. 武汉: 华中科技大学出版社, 2002: 22-41.
(Deng J L. The foundation of grey theory[M]. Wuhan: Press of Huazhong University of Science & Technology, 2002: 22-41.)

[4] 刘思峰, 党耀国, 方志耕. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2004: 51-79.
(Liu S F, Dang Y G, Fang Z G. The grey system theory and application[M]. Beijing: Sciences Press, 2004: 51-79.)

[5] 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]. System Engineering Theory & Practice, 2008, 28(4): 135-139.)

[6] 王清印, 崔援民, 赵秀恒, 等. 预测与决策的不确定性数学模型[M]. 北京: 冶金工业出版社, 2001: 143-170.
(Wang Q Y, Cui Y M, Zhao X H, et al. The uncertainty mathematical model of forecasting and decision making[M]. Beijing: Metallurgical Industry Press, 2001: 143-170.)

[7] 张可, 刘思峰. 灰色关联聚类在面板数据的扩展及应用[J]. 系统工程理论与实践, 2010, 30(7): 1253-1259.
(Zhang K, Liu S F. Extended clusters of grey incidence for panel data and its application[J]. System Engineering Theory & Practice, 2010, 30(4): 1253-1259.)

[8] 李因果, 何晓群. 面板数据聚类方法及应用[J]. 统计研究, 2010, 27(9): 73-77.
(Li Y G, He X Q. Panel data clustering method and application[J]. Statistical Research, 2010, 27(9): 73-77.)

[9] Bonzo D C, Hermosilla A Y. Clustering panel data via perturbed adaptive simulated annealing and genetic algorithms[J]. Advances in Complex Systems, 2002, 5(4): 339-360.

[10] 郑兵云. 多指标面板数据的聚类分析及其应用[J]. 数理统计与管理, 2008, 27(3): 265-270.
(Zheng B Y. The clustering analysis of multivariable panel data and its application[J]. Application of Statistics and Management, 2008, 27(2): 265-270.)

[11] 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.)

[12] 肖泽磊, 李帮义, 刘思峰. 基于多维面板数据的聚类方法探析及实证研究[J]. 数理统计与管理, 2009, 28(5): 831-838.
(Xiao Z L, Li B Y, Liu S F. The discussion on the clustering way based on the multi-dimensional panel data and empirical analysis[J]. Application of Statistics and Management, 2009, 28(5): 831-838.)

[13] 任娟, 陈圻. 基于形状的多指标面板数据聚类方法及其应用[J]. 数理统计与管理, 2011, 26(10): 27-33.
(Ren J, Chen Q. Clustering and its empirical study bases on shape for multivariable panel data[J]. Application of Statistics and Management, 2011, 26(10): 27-33.)

(下转第1045页)