Statistics and Application 统计学与应用, 2020, 9(5), 909-916
Published Online October 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.95095

Hans汉斯

基于关系结构的面板数据聚类方法研究

刘翠霞

重庆工商大学金融学院，重庆
Email: yueerliu_mail@163.com

收稿日期：2020年10月7日；录用日期：2020年10月22日；发布日期：2020年10月29日

摘　要

本文研究面板数据聚类方法，提出从面板数据变量之间影响与响应的结构关系上进行聚合分类的聚类方法，分别讨论了线性关系，非线性关系，基于轨迹特征和基于形状特征的多指标面板数据聚类方法。将相同结构关系的数据划分到相同的类中，不同关系结构的数据划分到不同的类中，使得类内有相同或相似结构关系与轨迹特征，类和类之间数据的结构关系与轨迹特征差异较大。

关键词

线性结构关系，非线性结构关系，形状相似性

Research on Clustering Method Based on Relationship Structure of Panel Data

Cuixia Liu

Department of Finance, Chongqing Technology and Business University, Chongqing
Email: yueerliu_mail@163.com

Received: Oct. 7th, 2020; accepted: Oct. 22nd, 2020; published: Oct. 29th, 2020

Abstract

This paper studies the panel data clustering method, and proposes a clustering method based on the structural relationship between the influence and response of the panel data variables. The linear relationship, the nonlinear relationship, the multi-index based on the trajectory feature and the shape feature are discussed respectively. This paper divides the data with the same structural relationship into the same class, and divides the data with different relationship structures into

文章引用: 刘翠霞. 基于关系结构的面板数据聚类方法研究[J]. 统计学与应用, 2020, 9(5): 909-916.
DOI: 10.12677/sa.2020.95095

different classes, so that the classes have the same or similar structural relationships and trajectory characteristics, and the structural relationships and trajectory characteristics of the data between classes and classes big different.

Keywords

Linear Structural Relationship, Nonlinear Structural Relationship, Shape Similarity

1. 问题提出

考虑变系数面板数据向量自回归模型[1] [2]：

$$y_{it}=\sum_{j=1}^{p}\beta_{i,j}y_{i,t-j}+\alpha_i+\varepsilon_{it},\left|\beta_i\right|<1 \tag{1}$$

其中，$\varepsilon_{it}\sim IID\left(0,\sigma_v^2\right),i=1,2,\cdots,N;t=1,2,\cdots,T$。

此模型中，需要估计的参数个数为$\left(p\times k+1\right)k\times N$个，当$k=3,N=5,p=3$时，待估计参数个数为150个，即便是每一个参数值与参数真值之间的差异很小，所有参数累计起来，造成总偏误比较大，模型拟合效果不一定理想。

此外，在面板数据建模分析中，不论运用基于哪种原理的估计方法，均是在系列假定条件与约束条件下，完成模型的参数估计问题，论证模型的有效性与一致性。然而，反映社会经济活动的大量宏(微)观数据，往往不一定符合模型所设定的假定条件，因此，运用论证再完整的模型，用来测度实际经济运行数据时，拟合经济系统中存在着相互嵌套、相互依赖又相互影响的经济变量之间的关系时，不一定得到期望的效果。鉴于此，在构造模型之前，先从数据变量关系结构出发，对数据进行合理的分类，一方面更符合经济理论与经济规律，如经济发展中，区域之间的差异往往是存在的，经济发达体和经济发展相对落后的区域变量之间的发展关系往往是不相同的；另一方面，降低了三维度面板数据计量模型的研究难度，尤其是在考虑时变系数，考虑动态影响关系时，有效降低模型估计参数个数，在建模分析前，对面板数据从关系结构上合理地分类，能有效解决面板数据模型中参数过度化问题。

本文将探讨面板数据关系聚类方法，分别讨论线性关系，非线性关系以及基于形状特征的关系聚类方法。

2. 基于参数关系面板数据聚类

2.1. 基于线性关系面板数据相似性度量

设面板数据模型数据生成过程为[3] [4]：

$$y_{it}^{(1)}=f\left(y_{it}^{(2)},y_{it}^{(3)},\cdots,y_{it}^{(m)}\right)+\varepsilon_{it}, \tag{2}$$

其中$y_{it}^{(2)},y_{it}^{(3)},\cdots,y_{it}^{(m)}$为影响变量，$y_{it}^{(1)}$为响应变量。

计算第i个截面个体中各个影响变量与响应变量之间的相关系数，记为$r_{il}^{(1)}$，

$$r_{il}^{(1)}=\frac{\operatorname{cov}\left(y_{it}^{(1)},y_{it}^{(l)}\right)}{\sqrt{\operatorname{var}\left(y_{it}^{(1)}\right)}\sqrt{\operatorname{var}\left(y_{it}^{(l)}\right)}},l=2,\cdots,m$$

其中 $y_{it}^{(2)},y_{it}^{(3)},\cdots,y_{it}^{(m)}$ 为影响变量， $y_{it}^{(1)}$ 为响应变量，有：

$$\operatorname{cov}\left(y_{it}^{(1)},y_{it}^{(l)}\right)=\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}^{(1)}-\overline{y}_{i}^{(1)}\right)\left(y_{it}^{(l)}-\overline{y}_{i}^{(l)}\right),$$

$$\overline{y}_{i}^{(1)}=\frac{1}{T}\sum_{t=1}^{T}y_{it}^{(1)},\ \overline{y}_{i}^{(l)}=\frac{1}{T}\sum_{t=1}^{T}y_{it}^{(l)}.$$

将第 i 个截面个体中各影响变量与响应变量构成的相关系数向量矩阵记为 $r_i^{(1)}=\left(r_{i2}^{(1)},r_{i3}^{(1)},\cdots,r_{im}^{(1)}\right)$。

设定每个截面个体为一个类 G_i，其中，

$$G_i=\left(\left(y_i^1\right)',\left(y_i^2\right)',\cdots,\left(y_i^m\right)'\right),i=1,2,\cdots,N,$$

$$y_i^k=\left(y_{i1}^k,\cdots,y_{iT}^k\right),t=1,2,\cdots,T.$$

现在构造反映结构关系的相似性度量指标为结构关系距离。

定义 1 线性结构关系距离：截面个体 i 和截面个体 j 之间的结构关系距离定义为：

$$d_{ij}=\sqrt{\left(r_i^{(1)}-r_j^{(1)}\right)\left(r_i^{(1)}-r_j^{(1)}\right)'},\tag{3}$$

其中， $r_i^{(1)}=\left(r_{i2}^{(1)},r_{i3}^{(1)},\cdots,r_{im}^{(1)}\right)$， $r_j^{(1)}=\left(r_{j2}^{(1)},r_{j3}^{(1)},\cdots,r_{jm}^{(1)}\right)$， $i,j=1,2,\cdots,N$ 。

上面定义的结构关系距离，满足距离的性质，若两个截面之间的结构关系完全相同，则结构关系距离为零，反之，自变量与因变量之间的关系差距越显著，则结构关系距离越大。

2.2. 非线性结构关系相似性度量

在面板数据模型数据生成过程为：

$$y_{it}^{(1)}=f\left(y_{it}^{(2)},y_{it}^{(3)},\cdots,y_{it}^{(m)}\right)+\varepsilon_{it},\tag{4}$$

其中 $y_{it}^{(2)},y_{it}^{(3)},\cdots,y_{it}^{(m)}$ 为影响变量， $y_{it}^{(1)}$ 为响应变量。3.5 节中，所刻画的数据生成过程中存在某种影响与响应关系是线性的，但是，自变量对因变量的影响关系可能是线性的，也可能是非线性的；可能是参数结构不变的，也可能存在跳跃性的机制转换也就是说，自变量的变化会引起因变量的变化是复杂的，本节探讨存在非线性结构关系的面板数据聚类问题。

在(4)中，若存在不同的截面个体 i,j， $i=1,2,\cdots,N$， $j=1,2,\cdots,N$， $i\neq j$，

假定 $f\left(y_{it}^{(2)},y_{it}^{(3)},\cdots,y_{it}^{(m)}\right)$ 为连续且处处可导、可微，有以下关系成立

$$\frac{\mathrm{d}y_{it}^{(1)}}{\mathrm{d}t}=\frac{\partial y_{it}^{(1)}}{\partial y_{it-l}^{(2)}}\frac{\mathrm{d}y_{it}^{(2)}}{\mathrm{d}t}+\frac{\partial y_{it}^{(1)}}{\partial y_{it-l}^{(3)}}\frac{\mathrm{d}y_{it}^{(3)}}{\mathrm{d}t}+\cdots+\frac{\partial y_{it}^{(1)}}{\partial y_{it-l}^{(m)}}\frac{\mathrm{d}y_{it}^{(m)}}{\mathrm{d}t},$$

其中 $\frac{\mathrm{d}y_{it}^{(i)}}{\mathrm{d}t},i=1,2,\cdots,m$ 为变量关于时间的导数， $\frac{\partial y_{it}^{(1)}}{\partial y_{it-l}^{(i)}},i=2,\cdots,m$ 为因变量关于各自变量的偏导数。

对于模型(4)，若满足第 i,j 个截面的数据关系结构相同或者相似，则：

$$y_{it}^{(1)}=f\left(y_{it}^{(2)},y_{it}^{(3)},\cdots,y_{it}^{(m)}\right)+\varepsilon_{it},i=1,2\cdots,N,$$

$$y_{jt}^{(1)} = f\left(y_{jt}^{(2)}, y_{jt}^{(3)}, \cdots, y_{jt}^{(m)}\right) + \varepsilon_{jt},\ j = 1,2,\cdots,N,$$

所刻画的两条曲线或者曲面是平行关系或者接近平行关系。即满足：

$$\frac{\partial y_{it}^{(1)}}{\partial y_{it-l}^{(k)}} \approx \frac{\partial y_{jt}^{(1)}}{\partial y_{jt-l}^{(k)}},\ k = 2,3,\cdots,m. \tag{5}$$

在实证研究中，建模分析前，模型的数据生成过程符合什么样的模型形式往往是未知的，因此，需要用样本信息特征来判定模型的初步特征，如，对于一个只含一个解释变量的二元面板模型，若第 i, j 个截面的数据关系结构相同或者相似，则有

$$\frac{\Delta y_{it}^{(1)}}{\Delta y_{it-l}^{(2)}} \approx \frac{\Delta y_{jt}^{(1)}}{\Delta y_{jt-l}^{(2)}}. \tag{6}$$

相应的，对于含有 m 个解释变量的面板数据模型，设定模型之间存在影响与响应的某种直曲线关系，若第 i, j 个截面的数据关系结构相同或者相似，则有

$$\frac{\Delta y_{it}^{(1)}}{\Delta y_{it-l}^{(2)}} \approx \frac{\Delta y_{jt}^{(1)}}{\Delta y_{jt-l}^{(2)}}, \frac{\Delta y_{it}^{(1)}}{\Delta y_{it-l}^{(3)}} \approx \frac{\Delta y_{jt}^{(1)}}{\Delta y_{jt-l}^{(3)}}, \cdots, \frac{\Delta y_{it}^{(1)}}{\Delta y_{it-l}^{(m)}} \approx \frac{\Delta y_{jt}^{(1)}}{\Delta y_{jt-l}^{(m)}}. \tag{7}$$

定义 2 非线性关系距离：将截面个体 i 和截面个体 j 之间的非线性关系距离定义为：

$$d_{ij} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\left(\frac{\Delta y_{it}^{(1)}}{\Delta y_{it-l}^{(2)}} - \frac{\Delta y_{jt}^{(1)}}{\Delta y_{jt-l}^{(2)}}\right)^2 + \left(\frac{\Delta y_{it}^{(1)}}{\Delta y_{it-l}^{(3)}} - \frac{\Delta y_{jt}^{(1)}}{\Delta y_{jt-l}^{(3)}}\right)^2 + \left(\frac{\Delta y_{it}^{(1)}}{\Delta y_{it-l}^{(m)}} - \frac{\Delta y_{jt}^{(1)}}{\Delta y_{jt-l}^{(m)}}\right)^2}, \tag{8}$$

公式(8)所刻画的距离，当两条曲线或者两个曲面之间是完全平行关系时，距离为零，当接近平行关系时，距离接近于零。

非线性关系距离满足以下两条性质：

性质一：对称性，即有 $d_{ij} = d_{ji}$。

性质二：不同的截面个体之间结构关系越接近，关系结构距离 d_{ij} 越小，二者之间关系结构关系差异越大，关系结构距离越大。当截面个体 i 和截面个体 j 之间为完全平行的关系时，$d_{ij} \to 0$。

2.3. 结构关系数据聚类

1) 初始聚类点确定

定义每一个截面个体为单独的一类，计算 N 个类之间的关系结构距离，构成初始关系结构距离矩阵，记为 $D_{(0)}$，

$$D_{(0)} = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1N} \\ 0 & 0 & d_{23} & \cdots & d_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_{(N-1)N} \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

其中，d_{ij} 为公式(3)或者(8)定义的线性或者非线性关系距离。

存在足够小的阈值 δ，当 $d_{ij} \le \delta$ 时，将类 i 和类 j 合并成新的一类，依次判定规则，将满足此规则构成的新的类集记为：$\{G_r, r = 1,2,\cdots,K\}$，此 K 个类构成了所有分类的初始分类聚合点。

2) 聚合规则

计算其余元素和$\{G_r, r=1,2,\cdots,K\}$中每个类的距离，将距离最小值对应的样本归并到相对应的类中，即

$$d_{rl}^2 = \min\left\{d_{i,j}^2, i \in G_i, i=1,2,\cdots,K, j \notin G_i, j=1,2,N-K\right\}. \tag{9}$$

将j对应的样本归并到类G_r中，重复判断，直到所有的元素按此规则被划分到相应的类中。

3. 基于轨迹特征的面板数据关系聚类

构造测度关系距离相似性的指标，设定数据之间的关系结构距离，对原始数据序列$\{y_{it}\}$分类，可操作性强，理解上也比较直观，但是，运用距离的测度指标，没有刻画出数据的轨迹特征，对于从轨迹上或者数据特征上，对数据进行分类，此方法有局限性，本节将考虑从轨迹形状模式出发，测度数据之间的轨迹形状相似性，对数据从轨迹形状上聚合分类，此分类方法适合于建立高频数据的建模分析。

3.1. 离散数据平滑处理

对于一组m维的存在内生影响关系的面板数据序列，现对于固定截面i，构成的N条存在内生关系的时间序列$\{y_{itm}\}, i=1,2,\cdots,N; t=1,2,\cdots,T$数据，在时间维度上，设定$m$个变量在各自时间维度上形成的数据轨迹呈现的是线性或者非线性特征，但是变量之间存在的影响相应关系为内生的线性影响关系。面板数据为考虑时间和空间上的离散数据集，从数据本身结构和特征出发，对数据进行分析，需要将离散的数据平滑化处理[5]，将离散数据转化为具有连续特征的函数性数据[6]。

通常，用插值或者平滑的方法对对离散数据$\{y_{it}\}$进行平滑处理，若$\{y_{it}\}$与平滑函数$\{f_i(t)\}$在t时刻数据完全相同，即

$$\left\|y_{it} - f_i(t)\right\| = 0,$$

此转换为插值。

相应地，若$\{y_{it}\}$与平滑函数$\{f_i(t)\}$在t时刻数据之间存在足够小，不为零的误差项$\{\varepsilon_{it}\}$，即

$$0 < \left\|y_{it} - f_i(t)\right\| < \varepsilon_{it},$$

称此转换过程为平滑。

本文采用平滑的定义对离散数据进行处理，通过函数性数据的拟合构造，对离散数据转化为函数性数据。常规离散数据函数性预处理方法有：线性平滑方法，基函数法，局部加权平滑法，粗糙惩罚法等，这里将针对本文要采用的基函数法进行简要的介绍。

3.2. 基函数确定

基函数定义：寻求能够有效拟合原始数据的函数$\varphi_k(t)(k=1,2,\cdots,K)$满足：

$$y_{it} = \sum_{k=1}^{K} \alpha_k \varphi_k(t), (k=1,2,\cdots,K)$$

其中，α_k为待估计的系数向量，$\varphi_k(t)(k=1,2,\cdots,K)$为基函数。

3.3. 基函数系数向量的估计

现假定基函数已经确定，在基函数确定，且在观测点t上，由基函数计算出的数据矩阵$\Phi = (\varphi_k(t))_{N\times K}$为满秩矩阵，$N$为截面个数，构造基于$N$个截面数据的残差平方和指标

$$SSE(y/\alpha)=(y-\Phi\alpha)'(y-\Phi\alpha), \tag{10}$$

其中， $y=(y_1,y_2,\cdots,y_N)'$,$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)'$ 运用 OLS 方法得到系数向量矩阵的估计值为

$$\alpha=(\Phi'\Phi)^{-1}\Phi'y.$$

3.4. 基于符号表示的相似性度量

考虑到本章节研究的数据为面板数据，若独立成 N 个截面数据序列来看，构成了 N 条时间维度上的时间序列数据，可以选用时间序列分析中的多项式混合模型，ARMA 模型，ARIMA 模型，Markov 模型，TARCH 模型，GARCH 模型等，具体的模型设定模式，要以数据本身的特征来最终确定。本部分假定模型为 ARMA 模型，即设定 $\{y_{it}\}$ 的数据生成过程为：

$$y_{it}=y_i(t)+u_{it}=\sum_{k=1}^{p}\phi_{ik}y_i(t-k)+\sum_{l=1}^{q}\theta_{il}\varepsilon_i(t-l). \tag{11}$$

传统聚类方法通常通过构造距离指标，构造相关系数指标来测度数据之间的相似性关系，若关系距离足够近，则被聚合到其中的一类中，否则，继续计算距离关系，继续聚合。这种基于数值模式的相似性度量指标，有其局限性，比如容易受尺度的影响的强噪声数据，基于数值模式的相似性度量方法不再适用。本节将提出基于符号表示的相似性度量方法，构造考虑轨迹和形状特征的，适合于面板数据建模分析的聚类方法。

构造基于形状特征的相似性度量指标时，需要以下几个定义。

极值集：对于不同的截面个体，$\{y_{it}\}$ 的光滑函数 $\{f_i(t)\}$，$\{y_{jt}\}$ 的光滑函数为 $\{f_j(t)\}$，若函数 $\{f_i(t)\}$，$\{f_j(t)\}$ 在时间维度 T 上连续且一阶可导，则 $\{f_i(t)\}$，$\{f_j(t)\}$ 在极小的领域内，一定存在相似的极值特征。记此极值特征为极值描述，用 $M(f(t))$ 表示，

$$M(f(t))=(m_1,m_2,\cdots,m_V),$$

其中，

$$m_v=\begin{cases}1, & \text{第}v\text{个顶点在 } f(t) \text{ 领域内达到极大值点}\\ 0, & \text{第}v\text{个顶点在 } f(t) \text{ 领域内达到极小值点}\end{cases}, v=1,2,\cdots,V$$

设 $M(f(t))$ 在一定领域内的顶点个数为 $\left|M(f(t))\right|$。

函数性数据等价关系定义：若两个函数性数据 $\{f_i(t)\}$，$\{f_j(t)\}$ 在任意相同领域 $(t,t+1)$ 内，均有 $M_i(f(t))=M_j(f(t))$，则称函数 $\{f_i(t)\}$，$\{f_j(t)\}$ 为 M 等价。

共享最大子集长度：记

$$\left|M(f_i(t)\cap f_j(t))\right|,$$

基于形状的相似性测度指标 D,

$$D(f_i(t),f_j(t))=\left|M(f_i(t))\right|+\left|f_j(t)\right|-\left|M(f_i(t)\cap f_j(t))\right|. \tag{12}$$

Salvatore Ingrassia (2003)已经证明该距离满足对称性，正定性，三角不等式关系，即：

$$D(f_i(t),f_j(t))\geq 0,$$

$$D(f_i(t),f_j(t))=D(f_j(t),f_i(t)),$$

$$D\left(f_i(t), f_j(t)\right) \le D\left(f_i(t), f_k(t)\right) + D\left(f_k(t), f_j(t)\right).$$

3.5. 基于形状特征的函数性数据聚类

完成对离散数据的连续性拟合处理，构造了基于形状特征的相似性度量指标，在此基础上，将从数据本身曲线特征上，完成对 N 条时间序列数据，也称为面板数据的分类处理。

针对我国经济发展的特征和由此导致的数据特征，本部分选用 K 均值聚类方法，完成对基于形状特征的函数性数据的聚合过程。

聚类过程分为以下几个步骤[7] [8]。

步骤一根据 N 组序列数据，估计未知函数 $y_i(t)$，实证分析时，要根据初步数据的曲线特征来进行。

步骤二：求解连续性数据序列的一阶导数 $y_i'(t)$，二阶导数，$y_i''(t)$ 从而求得 $y_i(t)$ 的极值集 $M\left(y_i(t)\right)$。

步骤三：计算 $y_i(t), i=1,2,\cdots,N$ 与“种子”样本序列的基于符号性数据的距离的最小值 $D^r\left(y_i(t)\right)$。其中 K 个的类核心“种子”样本序列为

$$\left\{y_i^k(t), k=1,2,\cdots,K\right\}, \tag{13}$$

其中，

$$D^r\left(y_i(t)\right) = \min\left\{D\left(y_i(t), y_i^k(t)\right), k=1,2,\cdots,K\right\},$$

$$D\left(y_i(t), y_i^k(t)\right) = \left|M\left(y_i(t)\right)\right| + \left|y_j^k(t)\right| - \left|M\left(y_i(t) \cap y_j^k\right)\right|$$

代表序列数据间的符号性相似距离。

步骤四：按照系统聚类原理完成聚合过程。

4. 小结

本文较系统研究了多指标面板数据关系聚类方法，在传统聚类的基础上，提出了一种基于数据变量之间影响关系相似性度量的结构关系聚类方法。结合面板数据聚类原理，针对面板数据变量之间的关系特征，构造了度量面板数据模型的相似性度量方法，包括线性结构关系的相似性度量指标，非线性结构关系的相似性度量指标，基于轨迹特征的相似性度量指标。完成了度量模型变量之间关系的相似性度量指标后，给出了针对不同变量关系结构的聚类过程。

基金项目

本文获得重庆市教委科学技术项目“面板数据聚类方法理论及应用研究”(KJ1600629)的资助，本文获得重庆市社会科学规划项目“基于面板数据的向量自回归模型的拓展研究”(2015PY25)的资助。

参考文献

[1] Ökonomie, R. (2009) Panel VAR Models with Spatial Dependence. *Economics Series*, No. 1, 237-275.

[2] Verdier, V. (2016) Estimation of Dynamic Panel Data Models with Cross-Sectional Dependence: Using Cluster Dependence for Efficiency. *Journal of Applied Econometrics*, **31**, 85-105. https://doi.org/10.1002/jae.2486

[3] 姜超. 多指标面板数据聚类的 SAS 实现[J].经济研究导刊, 2013(26):255-258.

[4] 刘翠霞, 史代敏. 基于关系聚类的动态面板数据模型及其应用研究[J]. 统计与信息论坛, 2015, 30(3): 10-16.

[5] 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.

[6] 任娟. 多指标面板数据融合聚类分析[J]. 数理统计与管理, 2016, 32(1): 57-67.

[7] 王双英, 王群伟, 曹泽. 多指标面板数据聚类方法及应用-以行业一次能源消费面板数据为例[J]. 数理统计与管

理, 2014(1): 42-49.

[8] 杨毅, 赵国浩, 秦爱民. 面板数据的有序聚类分析及其应用——以全球气候变化聚类分析为例[J]. 统计与信息论坛, 2012, 27(7): 13-18.