一种基于时序数据的动态聚类分析方法*

程建权　黄经南

(武汉测绘科技大学城市建设学院,武汉市珞喻路 129号, 430079)

摘　要　本文基于灰色关联度和模糊聚类分析原理及多目标决策技术,提出了一种基于时间序列数据的动态聚类分析方法,并以复杂的城市化进程为例加以应用,得出了一些有意义的结果。实践证明,该方法适合于多指标、多时段的(复杂)动态系统建模。

关键词　时间序列;灰色关联度;模糊聚类分析;多目标决策;动态聚类
分类号　TU983; O159

聚类分析是系统工程及城市规划等领域常用的一种结构分析方法,是根据系统内部结构的相似性而将若干对象划分为几种类型供进一步分析决策,如模糊聚类分析、灰色聚类分析、系统聚类分析、k-均值聚类分析、最优聚类分析等。这些方法的共同特点是以若干指标因子来反映系统的内部结构,采用不同的方法来量化对象间的相似性程度,并设计不同的标准进行聚类。但它们都没有考虑变量的时间因素(即动态变化过程的相似性),是一种静态的聚类分析,在应用中有很大的局限性。本文将基于灰色关联度和模糊聚类分析原理及多目标决策技术,提出了一种基于时间序列数据的动态聚类分析方法。

1　基本原理

设聚类论域为 $S=[S_1, S_2, \cdots, S_m]$,针对某一聚类目标 D,反映系统内部结构的指标元素集为 $Y=[Y_1, Y_2, \cdots, Y_n]$,时间长度为 T,则用于反映系统动态变化过程的时间序列数据为 X_{ijk},其中 $i=1,2,\cdots,m$; $j=1,2,\cdots,n$; $k=1,2,\cdots,t$。

1.1　动态变化过程相似性量化

灰色系统理论所创立的关联度分析方法实质上是按照待分析系统特征参量,或者说指标间的几何形状、发展趋势接近的程度来量化系统间动态变化过程的相似或相异程度(灰色关联度),是动态关联的一种系统分析方法。针对任一指标 Y_j ($j=1,2,\cdots,n$),本文应用此基本原理来量化聚类对象 S_i ($i=1,2,\cdots,m$)间动态变化过程的相似程度。任取一指标 Y_j,对应 Y_j 的时间序列数据为 X_{ijk} ($i=1,2,\cdots,m$; $k=1,2,\cdots,t$),简写为 X_{ik}。则相对于 S_i(即选取 X_{ik} 为参考序列), S_l ($l=1,2,\cdots,m$)与 S_i 间的相似程度可定量化为:

$$r_{il}=\sum_{k=1}^{T}\gamma_{il(k)}$$

$$\gamma_{il(k)}=(\Delta_{\min}+p\cdot\Delta_{\max})/(\Delta_{il(k)}+\Delta_{\max})$$

(p 为分辨系数, $0<p<1$)　　(1)

$$\Delta_{il(k)}=|x_{lk}-x_{ik}|,\ \Delta_{\min}=\min_k\min_l|x_{lk}-x_{ik}|,$$

$$\Delta_{\max}=\max_k\max_l|x_{lk}-x_{ik}|$$

当 $i=1,2,\cdots,m$ 时,可得关联矩阵 $R_{m\times m}$:

$$R=r_{il}=\begin{bmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & & \vdots \\ r_{m1} & \cdots & r_{mm} \end{bmatrix}$$

其中, $r_{ii}=1$; $r_{il}\neq r_{li}$,即不满足对称性,这是由于动态关联度 r_{il} 是相对于不同的参考序列 x_{ik} 而计算的。为此,令

$$r'_{il}=r'_{li}=(r_{il}+r_{li})/2　　(2)$$

则 $r'_{ii}=1$, $r'_{il}=r'_{li}$,矩阵 $R=(r'_{il})_{m\times m}$ 满足自反性与对称性,表达了聚类论域 S 在指标 Y_j 上的一种动态相似关系,反映了系统动态变化过程的局部相似性。

1.2　相似性综合

基于以上算法,用 $R_j=(r_{il}^{(j)})_{m\times m}$ 表示聚类论域 S 在指标 Y_j 上的相似矩阵,则针对某一聚类目标 D 的整体相似性是建立在指标集 $Y=[Y_1, Y_2, \cdots, Y_n]$ 的系统综合上,综合的方式取决于指标 Y_j 间的相互关系以及对聚类目标 D 的影响程度。基于多维价值并合原理,本文设计了以下模型。

1)线性加权模式

收稿日期: 1997-10-06。程建权,男, 31岁,讲师,现从事城市信息系统与城市系统建模研究。
* 荷兰政府 DSO援助计划资助项目。

$$R = (r_{il})_{m\times m} = w_1R_1 + w_2R_2 + \cdots + w_nR_n = \sum_{j=1}^{n} w_jR_j \quad (3)$$

其中, w_j 表示不同指标 Y_j (j= 1, 2,… , n)在聚类目标 D中的相对重要性程度,即权重,且满足归一性, $0 \leqslant w_j \leqslant 1$, $\sum_{j=1}^{n} w_j = 1$,当取 $w_j = 1/n$ (j = 1, 2,… , n)时,即转化为算术平均值模型。

此模式需均衡地考虑全部指标,且指标间相互独立,其相似性可互相补偿,反映了“好坏搭配”的特性。这类模式常使分布较均匀,由此而得到的分类的分辨率较低,分类数偏少。

2)乘法模式

$$R = (r_{il})_{m\times m}$$

$$r_{il} = \sqrt[n]{r_{il}^{(1)}\times r_{il}^{(2)}\times \cdots \times r_{il}^{(n)}} = \sqrt[n]{\prod_{j=1}^{n} r_{il}^{(j)}} \quad (4)$$

其中, $R_1 = (r_{il}^{(1)})_{m\times m}$, $R_2 = (r_{il}^{(2)})_{m\times m}$, $R_n = (r_{il}^{(n)})_{m\times m}$,即几何平均值模型。

此模式中所有指标需全面考虑,且需全部同时满足,反映了“不可偏废”的特征。相对于模式(1),这类模式常使分布较离散,分类的分辨率较高,分类数偏多。

3)代换 (取大)模式

$$R = (r_{il})_{m\times m}, r_{il} = \max\{r_{il}^{(1)}, r_{il}^{(2)}, \cdots, r_{il}^{(n)}\} \quad (5)$$

此类模式属主要因素突出型,即单项相似,整体就相似的情况,反映了“一好遮百丑”的特征。由于仅取最大值,故这种模式常导致部分信息丢失,所对应分类的灵敏度较高。

4)加乘混合模式

$$R = (r_{il})_{m\times m}$$

$$r_{il} = \lambda\sum_{j=1}^{n} w_j r_{il}^{(j)} + (1-\lambda)\sqrt[n]{\prod_{j=1}^{n} r_{il}^{(j)}} \quad 0 < \lambda < 1 \quad (6)$$

此类模式既要求全部指标同时满足,且指标间又可相互补偿。对应分类较细,但难以反映不同指标权重值对分类结果的影响,即灵敏度分析较为困难。

5)取大乘法混合模式

$$R = (r_{il})_{m\times m},$$

$$r_{il} = \lambda\max\{r_{il}^{(1)}, r_{il}^{(2)}, \cdots, r_{il}^{(n)}\} + (1-\lambda)\sqrt[n]{\prod_{j=1}^{n} r_{il}^{(j)}} \quad 0 < \lambda < 1 \quad (7)$$

此类模式既全面考虑又兼顾重点指标。对应分类较细,但没有考虑指标间的相对重要性。

用户还可根据聚类性质和实际需要,开发定义其它模型。由以上比较可知,不同模型具有不同的特征和一定的适用范围,用户应根据指标性质和聚类要求细心选择。一般来说,以模型 1)~ 3)为主,当判断模糊时,可试用混合模型 4)和 5)。容易证明,基于以上模型得出的矩阵 R既满足自反性又满足对称性,为相似矩阵,表示论域 S相对于聚类目标 D的系统整体相似性,其实质是定义在论域 S上的一种模糊相似关系。

1.3 聚类

模糊相似矩阵 R反映了论域 S中各聚类对象间在动态变化过程中的亲疏、远近关系,一般不满足传递性,至多通过有限次 $\log_2 m$ 的求平方 (传递闭包)运算,可求出对应的模糊等价矩阵,记之为 R^*。对于任意的阈值 λ ($0<\lambda<1$),进行对应的 λ-截矩阵运算,所得的 R_λ^* 为对应的布尔等价矩阵,可唯一确定一种分类。这样,当 λ 在一定的范围内变化时,即得到聚类论域 S的不同分类 (动态聚类图),用户可根据精度要求和不同分类的实际意义确定 λ 水平值,也可根据下列的 F统计量方法确定最优的 λ 值,即

$$F = \frac{\sum_{i=1}^{c} n_i(\bar{y}_i - \bar{y}_0)^2/(c-1)}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)/(n-c)} \quad (8)$$

其中, $y_k = (\sum_{j=1}^{n} r_{jk} - 1)/(n-1)$ (k= 1, 2,… , n); c是分类数; n 为总样本数; n_i 为第 i类样本数; $\bar{y}_i$是第 i类样本平均值; $\bar{y}_0$为全体样本平均值;当 $F > F_\alpha$时,类与类之间差异显著,分类最优,若同时有几个分类满足 $F > F_{0.05}$时,取差值中最大的。

1.4 灵敏度分析

在以上所设计的相似性综合模型中,模型 1)和 4)中均涉及到权重向量 w,反映了不同指标在聚类目标中的相对重要性,是一种主观判断,可应用比较矩阵法 (CMM)、环比评分法 (CCM)、模糊区间法 (FIM)、重要性排序法 (IUM)、模糊子集法 (FSM)、层次分析法 (AHP)等或采用文献 [4]提出的主客观并合的方法来计算权重。显然,聚类分析结果随相似性综合模型和对应的权重向量 w而变化,为避免这种主观判断所产生分类的不确定性,还需进行灵敏度分析,即以相似性综合模型和权重向量 w为控制参数,征求不同专家意见,多次调整判断矩阵,选用不同综合模型,多次聚类。若分类结果基本稳定,则模型可靠性较高,分类计算完成。

2 应用实例

在我国,城镇化过程是一个复杂的动态系统,受多因素、多指标的综合影响。传统的以单一的相对指标——非农业人口数/总人口数来直接量化城市化水平,较难系统地反映城镇化的复杂过程。本文基于系统工程思想,选取主要反映城市化不同侧面的多因素指标参与建模。为方便起见,本文选取以下5个指标,以检验该方法的可操作性,时间长度为10年(1984~1993年):Y_1——第二、第三产业产值/国民生产总值;Y_2——年投资总额;Y_3——非农业人口数/总人口数;Y_4——第二、第三产业就业人数/就业总人数;Y_5——城镇土地使用规模。

此例的聚类论域是由武汉市郊县新洲县的18个镇构成的集合,即 $S=$ {潘塘镇,三店镇,…,双柳镇},用于量化城镇化进程的指标集为 $Y=[Y_1, Y_2, \cdots, Y_5]$,时间长度 $T=10$,用 $x_{ijk}^{(0)}$ ($i=1,2,\cdots,18$; $j=1,2,\cdots,5$; $k=1,2,\cdots,10$)表示原始数据(由新洲县统计部门提供)。

2.1 建模过程

1)数据处理(标准化)

灰色关联分析中常用的标准化方法有初值化法、均值化法、区间值化法等。初值化方法较适合于量化动态变化过程,即记

$$x_{ijk}=x_{ijk}^{(0)}/x_{ij1}^{(0)} \qquad k=1,2,\cdots,10$$

2)计算不同指标的动态相似矩阵

指标 Y_3 的原始数据见表1,基于灰色关联度分析所得的关联矩阵见表2。R 不满足对称性,应用公式(2)可将其转化为相似矩阵 R_3。依此类推,可分别得到5个指标的相似矩阵 R_1、R_2、R_3、R_4、R_5。

表1 对应指标 Y_3 的原始时间序列数据

Tab.1 The Original Time Series Data Corresponding to Factor Y_3

镇名	1984年	1985年	1986年	1987年	1988年	1989年	1990年	1991年	1992年	1993年
潘塘	0.071	0.086	0.081	0.086	0.088	0.088	0.063	0.069	0.071	0.072
三店	0.050	0.060	0.056	0.040	0.042	0.042	0.042	0.044	0.044	0.044
凤凰	0.039	0.046	0.044	0.032	0.034	0.034	0.035	0.037	0.038	0.039
徐古	0.059	0.071	0.067	0.073	0.073	0.073	0.070	0.070	0.070	0.070
李集	0.061	0.074	0.070	0.079	0.082	0.079	0.065	0.068	0.071	0.074
新集	0.069	0.082	0.078	0.075	0.074	0.074	0.056	0.058	0.050	0.078
城关	0.372	0.447	0.423	0.455	0.424	0.417	0.447	0.456	0.489	0.507
张店	0.017	0.021	0.020	0.030	0.034	0.032	0.031	0.030	0.031	0.033
旧街	0.062	0.075	0.071	0.075	0.074	0.074	0.055	0.054	0.055	0.053
仓埠	0.317	0.381	0.361	0.345	0.364	0.357	0.350	0.345	0.370	0.373
孔埠	0.046	0.055	0.052	0.058	0.065	0.066	0.054	0.055	0.055	0.054
辛冲	0.039	0.047	0.044	0.048	0.046	0.047	0.046	0.046	0.046	0.046
周埠	0.033	0.040	0.038	0.040	0.043	0.044	0.042	0.041	0.042	0.051
汪集	0.096	0.116	0.110	0.113	0.115	0.116	0.111	0.113	0.114	0.115
阳逻	0.247	0.297	0.281	0.308	0.318	0.329	0.338	0.354	0.379	0.385
金台	0.016	0.020	0.019	0.022	0.027	0.030	0.028	0.030	0.030	0.033
大埠	0.073	0.088	0.083	0.110	0.115	0.120	0.115	0.113	0.113	0.117
双柳	0.048	0.058	0.055	0.104	0.092	0.095	0.098	0.099	0.098	0.097

3)相似性综合

此5个指标针对于城市化进程既相互独立又互相弥补,故较易用线性加权模式来综合。本文采用AHP法计算权重向量 $w=(w_1, w_2, w_3, w_4, w_5)$,其中的判断矩阵是对专家定性思维过程的定量化,可应用DELPHI法或BRAIN STORMING(头脑风暴)法来收集不同专家的意见,以避免较大的主观性。对应的判断矩阵及权重值如表3所示,由公式(3)得到18个镇间城镇化过程整体相似性矩阵 R 如表4所示。

表 2 对应指标 Y_3的灰色关联矩阵 R(部分数据)

Tab. 2 The Grey Relative Matrix R Corresponding to the Factor Y_3

镇名	潘塘	三店	凤凰	徐古	李集	新集	城关	张店	旧街	仓埠
潘塘	1.00	0.82	0.88	0.88	0.87	0.86	0.82	0.61	0.93	0.86
三店	0.84	1.00	0.94	0.76	0.77	0.86	0.75	0.59	0.88	0.80
凤凰	0.89	0.94	1.00	0.79	0.80	0.85	0.78	0.60	0.85	0.84
徐古	0.86	0.70	0.73	1.00	0.91	0.78	0.90	0.60	0.81	0.89
李集	0.86	0.71	0.76	0.92	1.00	0.78	0.84	0.61	0.81	0.89
新集	0.88	0.86	0.85	0.82	0.82	1.00	0.82	0.61	0.90	0.88
城关	0.80	0.69	0.73	0.91	0.84	0.78	1.00	0.61	0.79	0.88
张店	0.60	0.55	0.57	0.63	0.63	0.58	0.64	1.00	0.59	0.61
旧街	0.93	0.87	0.84	0.84	0.83	0.89	0.82	0.60	1.00	0.85
仓埠	0.86	0.76	0.81	0.91	0.90	0.86	0.89	0.61	0.83	1.00

表 3 判断矩阵及权重 w

Tab. 3 The Judge Matrix and the Corresponding Weight Vector w

指标	Y_1	Y_2	Y_3	Y_4	Y_5	权重
Y_1	1	2	3	3	8	0.426
Y_2		1	2	2	5	0.250
Y_3			1	1	3	0.137
Y_4				1	3	0.137
Y_5					1	0.050

检验值 CR= 0.027

表 4 18个镇间城镇化过程整体相似性矩阵 R(部分数据)

Tab. 4 The Overall Similarity Matrix in Urbanization Process Among the 18 Towns

镇名	潘塘	三店	凤凰	徐古	李集	新集	城关	张店	旧街	仓埠
潘塘	1.000	0.900	0.924	0.908	0.926	0.923	0.856	0.897	0.931	0.816
三店	0.900	1.000	0.917	0.846	0.928	0.930	0.843	0.895	0.922	0.815
凤凰	0.924	0.917	1.000	0.879	0.910	0.920	0.831	0.891	0.912	0.795
徐古	0.908	0.846	0.879	1.000	0.889	0.870	0.837	0.860	0.875	0.795
李集	0.926	0.928	0.910	0.889	1.000	0.943	0.865	0.919	0.933	0.836
新集	0.923	0.930	0.920	0.870	0.943	1.000	0.849	0.911	0.954	0.831
城关	0.856	0.843	0.831	0.837	0.865	0.849	1.000	0.824	0.838	0.886
张店	0.897	0.895	0.891	0.860	0.919	0.911	0.824	1.000	0.915	0.784
旧街	0.931	0.922	0.912	0.875	0.933	0.954	0.838	0.915	1.000	0.828
仓埠	0.816	0.815	0.795	0.795	0.836	0.831	0.886	0.784	0.828	1.000

4)模糊聚类和灵敏度分析

表 4所示的相似矩阵 R 经过 4次传递闭包运算,得到对应的模糊等价矩阵 R^*。当 λ= 0.84至 0.96时(步长为 0.01),动态聚类图为图 1所示,当 $\lambda < 0.84$时,全体样本分为一类,而当 $\lambda > 0.96$时,每个样本为一类,分类数随 λ 增加而增加。当调整判断矩阵,权重向量 w改变为 w= (0.49, 0.17, 0.13, 0.11, 0.10)以及选用模型 4)进行灵敏度分析时,其分类结果基本一致,说明了图 1的可靠性。

2.2 结果分析

根据图 1结果和公式(8),检验不同 λ 所对应分类的显著性,当 λ= 0.91时,$F > F_{0.05}$,且差异显著,则 18个镇可划分为 5大类,即{阳逻镇},{城关镇},{仓埠镇},{双柳镇},{其它镇}。这个结果与武汉市郊县新洲县近些年的实际情况非常相

符。阳逻镇是武汉地区开放开发的前沿阵地和新的经济增长点,它不仅是国家重要的能源基地,而且化工、棉纺、建材、机械、汽配等基础工业已形成规模,是全县的支柱产业,工业的快速发展有力地促进了该地区的城市化进程;城关镇具有优厚的基础设施;双柳镇交通发达、区位条件优越;仓埠镇资源丰富、交通便利,在开放政策和市场经济指导下,其城市化进程相对较快。以上的分类结果将有助于了解新洲县城市化发展的动态过程及其空间分布,为其制定未来城镇发展格局提供决策依据。

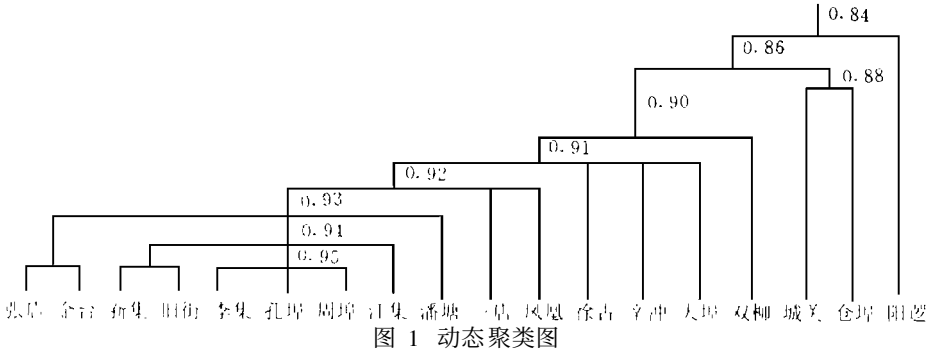

图 1 动态聚类图

Fig. 1 The Dynamic Cluster Figure

3 结 论

本文所提出的方法是基于灰色关联度和模糊聚类分析原理,并综合了多目标决策技术。这种方法具有以下特点: 1)不追求大的样本量; 2)不要求待分析序列具有某种特殊的分布; 3)计算过程简单; 4)综合了序列分布的客观信息与指标的主观判断; 5)所设计的相似性综合模型可适用于不同的条件。但在具体应用过程中,指标体系的设计至关重要,是动态聚类分析的核心。相似性综合模型的选择应取决于待分析动态系统的内部特征。如何结合空间定位,基于时空数据库,向时空综合聚类分析发展,仍有待于系统工程和 GIS领域的进一步研究和完善。

参 考 文 献

1 邓聚龙.灰色系统理论教程.武汉: 华中理工大学出版社, 1990
2 汪培庄.模糊集合论及其应用.上海: 上海科技出版社, 1983
3 李万绪.基于灰色关联度的聚类分析方法.系统工程, 1990, 8(3): 37~44
4 程建权.多指标综合评价中一种计算权重的改进方法.系统工程理论与实践, 1994, 8(11): 51~57
5 王发曾,袁中金.省域新设城市综合研究.郑州: 河南大学出版社, 1995

A Method of Dynamic Cluster Analysis Based on Time Series Data

Cheng Jianquan *Huang Jingnan*
(School of Urban Studies, WTUSM, 129 Luoyu Road, Wuhan, China, 430079)

Abstract Based on the principles of grey relative degree, fuzzy cluster analysis and multi-objective decision making technique, the paper presents a method of dynamic cluster analysis through time series data. Taking complicated urbanization process for an example, the method is applied with some significant results. It is proved that the method could be employed to model any complicated dynamic system with multi-factors.

Key words time series; grey relative degree; fuzzy cluster analysis; multi-objective decision making; dynamic cluster