广西科学院学报
Journal of Guangxi Academy of Sciences

2009,25(1):46~49
Vol. 25,No. 1 February 2009

模糊聚类在西部省区经济发展状况分类中的应用*

The Application of Fuzzy Clustering in Classification of the Economic Development Situations of All Provinces in West Region

韦艳玲
WEI Yan-ling

(柳州职业技术学院信息工程系,广西柳州 545006)
(Department of Information Engineering, Liuzhou Vocational &Technical College, Liuzhou, Guangxi,545006, China)

摘要:依据2005年和2006年的有关经济统计数据,采用主成分分析与模糊聚类相结合的方法,对我国西部12个省区的经济发展状况进行模糊分类,初步划分出8种具有不同经济发展状况特征的类型。该方法比传统的聚类方法灵活,可以根据不同的要求获得不同的聚类结果,而且分析结果更贴近我国西部经济社会的实际情况。

关键词:模糊聚类 主成分分析 经济 分类
中图法分类号:F127 **文献标识码**:A **文章编号**:1002-7378(2009)01-0046-04

Abstract: Based on the concerned economic data of the year of 2005 and 2006, this paper uses fuzzy clustering analysis combining principal component analysis to preliminarily divide the economic development situations of all provinces in west region into eight types. Fuzzy clustering is more flexible than the traditional one. It may obtain the different clustering result according to the different request. Moreover the analysis result fits the actual situation of west region.

Key words: fuzzy clustering, principal component analysis, economy, classify

自从实施西部大开发战略以来,我国西部经济得到了快速发展[1]。但是,我国西部地区与东部、东北、中部地区在经济实力、工业化水平、城镇化水平等方面相比仍然存在一定差距。另外,西部地区内部各省区经济发展状况并不平衡。因此,对西部各省区经济发展状况进行合理分类,根据不同的具体情况制定针对性的发展政策,促进西部各省区经济发展,具有重要的现实意义。

由于经济发展状况的分类标准具有多元性,以及分类的界限无法进行精确的度量,所以西部各省区经济发展状况的各种类别之间的界限并不清晰和明确。有的学者分别用聚类和因子分析法对西部12个省区经济发展状况进行对比研究[2],有的学者用主成分分析法对我国中部六省与西部省份经济实力进行了比较[3]。但是,现有关于西部各省区经济发展状况的分类方法尚存在一些缺陷和不足,主要表现在,因子分析法和主成分分析法只能对各省区经济发展状况进行定量的排序,如何分类只是由主观判断来决定,而传统的聚类分析法是一种分类界限非常鲜明的硬性划分,缺少伸缩性。模糊聚类分析是当前在模糊数学中应用最多的几个方法之一,它可以根据不同的要求和分类标准获得不同的聚类结果,大大地提高了聚类的灵活性。由于具有较好的可伸缩性,模糊聚类是解决界限不清晰的聚类问题的较好办法。因此,采用模糊聚类对西部各省区经济发展状况进行分类更具合理性。

聚类分析中应尽可能使特征数减到最小。如果以经济指标原始数据为特征变量进行聚类,不仅因

收稿日期:2008-10-12
修回日期:2009-01-16
作者简介:韦艳玲(1970-),女,工程师,硕士,主要从事数据挖掘研究工作。
*广西自然科学基金项目(桂科自0481016)资助。

特征数多而计算量大,而且在一定程度上还存在经济指标数据反映的信息有所重叠等干扰因素。因此,本文采用主成分分析对西部各省区经济发展状况指标降维,排除干扰因素,再采用模糊聚类方法实现对西部各省区经济发展状况的分类。

1 评价指标及原始数据预处理

1.1 评价指标的选取

综合文献[2,3]的研究成果,选取7项指标来反映我国西部各省区经济发展状况,即地区生产总值($R1$)、人均地区生产总值($R2$)、固定资产投资($R3$)、居民消费价格指数($l4$)、城镇居民人均消费性支出($R5$)、人均财政收入($R6$)、海关进出口总额($R7$)。其中,地区生产总值和人均地区生产总值是用来反映地区经济实力的核心指标;固定资产投资是从资本形成方面反映地区经济实力,还可以反映地区经济发展的潜力和可持续性;居民消费价格指数和城镇居民人均消费性支出反映人民生活质量的高低;人均财政收入反映地区自我发展的能力;海关进出口总额反映地区对外贸易的实力。根据所研究问题的性质,在上述指标中,指标$l4$是逆指标,其他均为正指标。为了能对上述指标进行聚类分析,必须统一指标类型,即将逆指标转换为正指标。对于逆指标,我们直接求其倒数为正指标,即$R4 = 1/l4$。对于转换后的指标向量,为分析方便统一定义为:$R = (R1, R2, R3, R4, R5, R6, R7)$。

样本集用X表示,样本对象数为西部12个省区,即内蒙古、广西、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏、新疆,分别表示为$x1$、$x2$、$x3$、$x4$、$x5$、$x6$、$x7$、$x8$、$x9$、$x10$、$x11$、$x12$。

1.2 基于主成份分析法的数据预处理

由于第一次全国经济普查调整了国家和地区的生产总值及其产业构成,以及2005年以前部分资料的不可得性,所以我们不对2005年以前的数据进行分析。另外,2005年是"十五"规划的最后一年,而2006年是"十一五"规划的第一年。具有完整数据的这两年,能够客观真实地反映我国西部地区经济发展状况。

根据《中国统计年鉴——2006》和《中国统计年鉴——2007》,选取主要反映西部12个省区2005年和2006年的7个经济指标的平均值数据。即2005~2006年西部各省区的地区生产总值(当年价)、人均地区生产总值(当年价)、固定资产投资、城镇居民人均消费性支出、人均财政收入、海关进出口总额的平均值,2005~2006年西部各省区居民消费价格指数的几何平均值。

首先求出数据矩阵的特征值、特征值的方差贡献率和累积贡献率。由于各指标的量纲和单位不同,我们采用最小最大法对原始数据进行标准化处理。将数据输入MATLAB7.0进行主成分分析[4],发现前四个主成分的方差贡献率可达96.5569%,故可选取前四个主成分作为反映经济发展实力的综合指标。设$Y1$、$Y2$、$Y3$、$Y4$分别代表第一、第二、第三、第四主成份,则其线性组合为:

$$Y1 = -0.4563 \times R1 - 0.2196 \times R2 - 0.5179 \times R3 - 0.2984 \times R4 + 0.2918 \times R5 - 0.2026 \times R6 - 0.5101 \times R7;$$

$$Y2 = 0.2702 \times R1 - 0.5486 \times R2 + 0.1365 \times R3 - 0.6631 \times R4 - 0.4027 \times R5 - 0.0635 \times R6 + 0.0386 \times R7;$$

$$Y3 = -0.1955 \times R1 + 0.1268 \times R2 - 0.0822 \times R3 + 0.1352 \times R4 - 0.6773 \times R5 + 0.4994 \times R6 - 0.4611 \times R7;$$

$$Y4 = 0.0076 \times R1 - 0.3136 \times R2 + 0.0360 \times R3 - 0.0859 \times R4 + 0.4648 \times R5 + 0.8186 \times R6 + 0.0828 \times R7。$$

2 西部省区经济发展状况模糊聚类分析

在获得$Y1$、$Y2$、$Y3$、$Y4$作为模糊聚类的聚类变量后,就可以对样本集进行聚类。聚类分析采用等价闭包法,由模糊关系矩阵R求模糊等价关系矩阵$t(R)$,再求不同的λ值对应的λ截集,以得到不同的分类[5,6]。

2.1 建立模糊关系矩阵

以绝对值减数法建立相似关系矩阵R,其公式如下:

$$r_{ij} = \begin{cases} 1, & i = j, \\ 1 - C\sum_{k=1}^{m} |x_{ik} - x_{jk}|, & i \neq j, \end{cases}$$

式中,x_{ik}为第i行第k列的属性值,x_{jk}为第j行第k列的属性值,其中C为适当选取数,使$0 \leqslant r_{ij} \leqslant 1$。我们令属性个数为$m=4$、$i,j=1,2,\cdots,12$、$C=0.2$,计算得到的$R$矩阵如表1所示。

2.2 求模糊等价关系矩阵

通过褶积求R的传递闭包,即R自乘得$R \times R = R^2$,再自乘$R^2 \times R^2 = R^4$,直到$R^k = R^{2k}$为止,则模糊等价关系矩阵$t(R) = R^k = R^{2k}$,$k \in N$,模糊等价关系矩阵$t(R)$如表2所示。

表1 模糊关系矩阵 R

样本	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$	$x9$	$x10$	$x11$	$x12$
$x1$	1	0.69	0.71	0.63	0.52	0.65	0.62	0.65	0.57	0.63	0.78	0.66
$x2$	0.69	1	0.65	0.81	0.81	0.90	0.76	0.80	0.85	0.73	0.76	0.67
$x3$	0.71	0.65	1	0.54	0.57	0.71	0.67	0.74	0.61	0.63	0.71	0.61
$x4$	0.63	0.81	0.54	1	0.74	0.78	0.62	0.72	0.75	0.65	0.57	0.59
$x5$	0.52	0.81	0.57	0.74	1	0.87	0.86	0.82	0.95	0.86	0.75	0.68
$x6$	0.65	0.90	0.71	0.78	0.87	1	0.81	0.90	0.90	0.77	0.74	0.74
$x7$	0.62	0.76	0.67	0.62	0.86	0.81	1	0.71	0.86	0.78	0.83	0.58
$x8$	0.65	0.80	0.74	0.72	0.82	0.90	0.71	1	0.82	0.82	0.75	0.82
$x9$	0.57	0.85	0.61	0.75	0.95	0.90	0.86	0.82	1	0.85	0.80	0.69
$x10$	0.63	0.73	0.63	0.65	0.86	0.77	0.78	0.82	0.85	1	0.83	0.74
$x11$	0.78	0.76	0.71	0.57	0.75	0.74	0.83	0.75	0.80	0.83	1	0.65
$x12$	0.66	0.67	0.61	0.59	0.68	0.74	0.58	0.82	0.69	0.74	0.65	1

表2　模糊等价关系矩阵 $t(R)$

样本	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$	$x9$	$x10$	$x11$	$x12$
$x1$	1	0.78	0.74	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
$x2$	0.78	1	0.74	0.81	0.90	0.90	0.86	0.90	0.90	0.86	0.83	0.82
$x3$	0.74	0.74	1	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74
$x4$	0.78	0.81	0.74	1	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
$x5$	0.78	0.90	0.74	0.81	1	0.90	0.86	0.90	0.95	0.86	0.83	0.82
$x6$	0.78	0.90	0.74	0.81	0.90	1	0.86	0.90	0.90	0.86	0.83	0.82
$x7$	0.78	0.86	0.74	0.81	0.86	0.86	1	0.86	0.86	0.86	0.83	0.82
$x8$	0.78	0.90	0.74	0.81	0.90	0.90	0.86	1	0.90	0.86	0.83	0.82
$x9$	0.78	0.90	0.74	0.81	0.95	0.90	0.86	0.90	1	0.86	0.83	0.82
$x10$	0.78	0.86	0.74	0.81	0.86	0.86	0.86	0.86	0.86	1	0.83	0.82
$x11$	0.78	0.83	0.74	0.81	0.83	0.83	0.83	0.83	0.83	0.83	1	0.82
$x12$	0.78	0.82	0.74	0.81	0.82	0.82	0.82	0.82	0.82	0.82	0.82	1

2.3 求不同的 λ 值对应的 λ 截集

将 λ 依次按照 0.74，0.78，0.81，0.82，0.83，0.86，0.90，0.95 取值，分别求不同的 λ 截集。如当 $\lambda = 0.83$ 时，有 λ 截集为

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}。$$

当 $\lambda = 0.74$ 时，样本集 X 中的样本全为一类。

当 $\lambda = 0.78$ 时，X 分为{{$x1$,$x2$, $x4$,$x5$,$x6$,$x7$,$x8$,$x9$,$x10$,$x11$, $x12$},{$x3$}}两类。

当 $\lambda = 0.81$ 时，X 分为{{$x2$, $x4$,$x5$,$x6$,$x7$,$x8$,$x9$,$x10$,$x11$, $x12$},{$x1$},{$x3$}}三类。

当 $\lambda = 0.82$ 时，X 分为{{$x2$, $x5$,$x6$,$x7$,$x8$,$x9$,$x10$,$x11$, $x12$},{$x1$},{$x3$},{ $x4$}}四类。

当 $\lambda = 0.83$ 时，X 分为{{$x2$,$x5$,$x6$,$x7$,$x8$,$x9$,$x10$,$x11$},{$x1$},{$x3$},{$x4$},{$x12$}}五类。

当 $\lambda = 0.86$ 时，X 分为{{$x2$,$x5$,$x6$,$x7$,$x8$,$x9$,$x10$},{$x1$},{$x3$},{$x4$},{$x12$},{$x11$}}六类。

当 $\lambda = 0.90$ 时，X 分为{{$x2$,$x5$,$x6$,$x8$,$x9$},{$x1$}, {$x3$}, {$x4$}, {$x12$}, {$x11$}, {$x7$}, {$x10$}}八类。

当 $\lambda = 0.95$ 时，X 分为{{$x5$, $x9$},{$x1$},{$x3$},{$x4$},{$x12$},{$x11$},{$x7$},{$x10$}, {$x2$} ,{$x6$ },{$x8$}}11类。

从以上分析可以看出，当 λ 的值越大，则西部各省区的分类越细。例如：当 $\lambda >= 0.78$ 时，重庆($x3$)独立为一类；当 $\lambda >= 0.81$ 时，内蒙古($x1$)独立为一类；当 $\lambda >= 0.82$ 时，四川($x4$)独立为一类；当 $\lambda >= 0.83$ 时，新疆($x12$)独立为一类。

3　讨论

本文结合主成分分析，利用模糊聚类技术实现了对西部各省区经济发展状况的分类，初步划分了具有不同经济发展状况特征的西部各省区类型。相对于传统的聚类分析，本文的方法可以根据不同的

要求获得不同的聚类结果，这种软划分具有较好的可伸缩性，大大提高了聚类的灵活性。

在本文中，从 $\lambda=0.81$ 时开始，重庆和内蒙古就分别独立为一类。而如果把本文所用的原始数据按层次聚类分析方法进行聚类分析，则西部各省区可以分为三类，贵州、西藏、青海、宁夏为一类，四川、新疆为一类，其余各省区为一类，即重庆和内蒙古同属于一个类别[2]。但是，从经济发展的现实情况来看，重庆和内蒙古有各自的鲜明特点，并不能简单地划为同一个类别。重庆市作为西部地区第一个直辖市，由于国家的政策倾斜等因素，经济发展速度较快，城镇居民人均可支配收入和人均消费性支出均位居西部第一位。第三产业发展与城市发展具有密切的相互影响关系。就第三产业增加值占地区生产总值的比重来说，重庆位居全国第四位，仅次于北京、上海和西藏。考虑到西藏由于工业非常不发达导致第三产业增加值占地区生产总值的比重偏高的特殊情况，重庆应属于第三产业发展程度最高的西部省区。而且重庆是西南地区甚至是整个西部地区的“龙头”，具有辐射带动的独特作用。内蒙古农牧业较发达，不仅是我国重要的商品粮、油、糖生产基地，也是我国重要的畜牧业生产基地，其畜牧业综合生产能力居全国五大牧区之首。内蒙古也是我国的钢铁、煤炭生产基地，近年来，其钢铁、煤炭生产大增，而且其森林工业、农畜产品加工、电力、机械制造、化工、电子、纺织、制糖、造纸、轻工等也有重要地位。内蒙古的人均地区生产总值、人均固定资产投资、农村居民人均纯收入和农村居民人均生活消费支出均位居西部第一位。可以说，重庆是第三产业和城市型经济，侧重于经济发展，而内蒙古是农牧业、制造业和资源型经济，侧重于经济增长。可见，把重庆、内蒙古分别独立列为一类，更贴近我国经济社会的实际情况。

综上所述，模糊聚类分析有助于对西部各类省区的经济状况作进一步分析，并可以根据不同的具体情况制定针对性的发展政策，提供有效的激励或扶持措施，以便更合理地开发西部，并最终达到西部大发展的目的。

参考文献：

[1] 温家宝. 开拓创新，扎实工作，不断开创西部大开发的新局面 [N]. 人民日报，2005-02-05(2).

[2] 蒋志华，顾振海. 西部 12 省经济发展状况对比研究——基于聚类因子分析法的实证分析[J]. 经济体制改革，2006(12)：138-141.

[3] 麻晓刚，游达明. 我国中部六省与西部省份经济实力比较的实证分析[J]. 科技资讯，2006(23)：148-149.

[4] 苏金明，阮沈勇. MATLAB 工程数学[M]. 北京：电子工业出版社，2006.

[5] 李剑峰. 数据挖掘在公司财务分析中的应用[J]. 计算机工程与应用，2005(2)：217-219.

[6] 李鸿吉. 模糊数学基础及实用算法 [M]. 北京：科学出版社，2005.

（责任编辑：韦廷宗）

中美合作纳米线激光器研究获重大突破

湖南大学微纳技术研究中心邹炳锁教授领衔的纳米光子学小组与美国亚利桑那州立大学宁存政教授领衔的纳米光子学小组合作，将半导体激光芯片调谐范围扩大，成功演示出 500nm 绿光直至 700nm 红光，创下一个新的半导体激光器调谐范围的世界纪录，与原来调谐范围最长仅几十纳米相比实现了重大突破。

长期以来，如何提高半导体激光器的调谐范围从而充分发挥激光的作用，一直是国内外专家奋斗的目标；但是制约这一进步的主要因素就是一直无法攻克发光材料和基底材料的结构或应力配合问题，导致材料成分无法大幅调节，因此无法实现激光的大范围调谐。一般半导体激光器调谐范围最长仅几十纳米，制约了它在许多领域的应用。研究人员另辟蹊径，采用一维纳米结构生长技术，避免了材料中的结构配合问题，可以做出成分可大范围调节的纳米线，实现了从绿光、黄光、橙光到红光的单芯片上可调谐的激光发射，解决了这一国际难题。该项成果的材料将可应用于新光源、光通讯、分子和生物传感、太阳能电池等领域。

（据科学网）