

目 录

绪论 社会科学中的数学	(i)
第一章 概 论	(1)
第一节 综合评价问题	(1)
第二节 综合评价方法的产生、发展和研究现状	(2)
第二章 常规综合评价方法	(5)
第一节 几个评价实例	(5)
第二节 评价指标的选取	(10)
第三节 无量纲化方法	(22)
第四节 权的确定	(39)
第五节 常见的综合方法	(45)
第三章 综合评价的主成分方法与因子分析法	(53)
第一节 主成分分析	(53)
第二节 主成分分析方法的统计依据	(57)
第三节 几个实例	(63)
第四节 因子分析法	(70)
第五节 因子分析的实例和计算方法	(72)
第四章 综合评价的聚类分析与判别分析方法	(79)
第一节 综合评价的系统聚类法	(79)
第二节 综合评价的动态聚类法	(89)
第三节 有序样本的综合评价聚类方法	(93)
第四节 综合评价的距离判别分析	(99)
第五节 贝叶斯 (Bayes) 判别	(106)
第六节 费歇判别	(111)
第七节 逐步判别	(116)
第五章 其他综合评价方法	(123)
第一节 距离综合评价方法	(123)
第二节 灰色关联度评价法	(129)

第三节	DEA 方法	(140)
第六章	模糊综合评价	(167)
第一节	模糊综合评价的基本程序	(167)
第二节	模糊单因素评价	(172)
第三节	单因素模糊评价的综合	(182)
第四节	模糊综合评价结果向量的分析	(185)
第五节	多级模糊综合评价	(196)
第七章	多维标度法	(208)
第一节	相似性度量	(208)
第二节	托格森 (Torgerson) 方法	(215)
第三节	K-L 方法	(225)
第四节	谢帕尔德方法	(228)
第五节	克拉斯卡尔方法	(234)
第六节	最小维数分析法 (MDA 方法)	(239)
第七节	综合评价方法优选探讨	(247)
参考文献	(250)
名词索引	(252)

第一章 概 论

第一节 综合评价问题

在日常生活、工作中,我们经常会遇到综合评价问题.设想你是一个消费者,需要购买一台电脑,你会怎样做?显然,你会先了解一下市场上出售的各种品牌、各种档次的电脑情况,比如,质量、价格等,然后再依据自身的经济状况和对电脑的使用要求从各种电脑中选择一种你认为总体上比较合适的电脑.又若你是一个大公司的总经理,你的公司有许多分公司,了解各分公司的经营状况是非常必要的.也许你还想在年末奖励一些经营好的公司,而对一些经营较差的公司进行改造,那么你的依据是什么?恐怕对各分公司的综合评价结果是唯一的依据.再若你有一大笔资金,有人为你提供了几种投资方案,你将如何决策?对各种方案的综合考虑与比较将有助于你的决策.你若是一名探矿者,在某个地方探测到了各方面的数据,那么,此处是否有矿?丰度如何?你必须对各方面情况进行综合分析,必要的时候与历史数据比较以作出判断等等.事实上,上面所要解决的问题在很大程度上都涉及到所谓的综合评价问题.综合评价作为专业名词其实具有较为通俗的含义,简单地说,就是对客观事物以不同侧面所得的数据作出总的评价.综合评价涉及到日常生活中的方方面面,小到商品,大到社会经济发展状况及卫星运行状况等.因此,对综合评价方法的研究具有很现实的意义.

综合评价的研究对象通常是自然、社会、经济等领域中的同类事物(横向)或同一事物在不同时期的表现(纵向).具体的综合评价一般表现为以下几类问题:

第一类问题是对所研究事物进行分类. 俗话讲,物以类聚,人

以群分,把多个事物中具有相同或相近属性的事物归为一类,有利于对客观事物进行科学的管理.比如依据经济发展状况对我国各地区分类,有利于国家制定有关政策,促进我国经济稳定协调发展.

第二类综合评价问题表现为对上述分类的序化,即在第一类问题基础上对各小类按优劣排出顺序.比如对我国各地区按经济状况分类后再进一步明确:哪些地区经济发展状况好,哪些地区经济发展状况不佳,等等,这将为客观经济管理提供信息.

第三类综合评价问题表现为对某一事物作出整体评价.当然也必须有参考系,否则无法作出评价.如果已经有一些同类事物的评价结果(即了解其综合表现情况),就称其为有训练的样本,这样,只需将所评对象与这些有训练样本进行比较,用训练样本的先验信息对该对象作出评价.这在地质勘探、天气预报等方面有较为广泛的应用.即使对于每一个评价对象,通过综合评价和比较,可以找到自身的差距,也便于及时采取措施,对症下药.

第二节 综合评价方法的产生、 发展和研究现状

综合评价的依据就是指标,而指标按不同的标志可分为实物指标和价值指标,相对指标和绝对指标,单项指标和综合指标等.

我们往往用最终结果来衡量事物发展的情况,而结果则用实物指标来反映,比如生产了多少吨粮食,多少台拖拉机,多少头牛等.这种衡量方式显然是非常粗略的,因为粮食有等级品种之分,牛有大小轻重之分,拖拉机有型号功率之分,并且由于度量单位不同而无法汇总相比.为了解决这样的问题,就产生了价值综合指标,通过引进价格这一共同度量因素解决不同实物指标的可综合问题.比如非常典型的总产值指标,用总产值进行评价,导致大家都盲目地追求高产值.随着社会经济的进一步发展,管理的重心从单纯的追求高产出而转向注重效益,即追求以尽量少的投入而得

到较高的产出.用价值综合指标进行评价就不能满足这一要求.效益表现在多个方面,比如能耗,劳动生产率,资金使用效益,等等.为了从效益角度对事物进行综合评价,就产生指标体系评价法,即用不同的指标对事物发展的多个方面分别予以反映.这种指标体系法虽能全面反映某一个事物的发展状况,但在不同事物间比较时又遇到了困难.因为各个指标的同时使用,经常会发生不同指标之间相互矛盾的情况,因而不能对被评价事物作时间和空间上的整体对比.比如在比较甲、乙两个企业同一时期经济效益的优劣时,往往会遇到这样的情况,甲企业有几项经济效益指标好于乙企业,同时,乙企业又有另外几项经济效益指标好于甲企业.这时就无法判断甲、乙两个企业的经济效益到底谁好谁差.同样在分析比较同一企业不同时期经济效益的发展变化时,也常常会遇到类似的情况.正是由于指标体系的这一不足,人们又发展了多指标综合评价方法,即把反映被评价事物的多个指标的信息综合起来,得到一个综合指标,由此来反映被评价事物的整体情况,并进行横向和纵向的比较.这样既有全面性,又有综合性.近年来,围绕着多指标综合评价,其他领域的相关知识不断渗入,使得多指标综合评价方法不断丰富,有关这方面的研究也不断深入.主要表现在以下几个方面:

第一,评价所用的指标是多种多样的,评价的问题也不是单一的,20世纪60年代产生的模糊数学在综合评价中得到了较为成功的应用,产生了特别适合于对主观或定性指标进行评价的模糊综合评价方法.

第二,多指标综合评价中比较难以解决的是综合时各指标间信息的重复问题,近几十年来迅速发展的多元统计分析为解决这一问题提供了可能性,因而产生了主成分、因子评价法,另外判别分析、聚类分析也为解决第一、第二类及第三类综合评价问题提供了较好的定量方法.

第三,由于评价对象的多样性及评价的决策作用,多目标决策方法也溶入到综合评价中来,比如功效系数法,AHP法等,开阔了

评价方法的思路.

第四,运筹学的新发展产生了将投入和产出指标分离开来评价部门间相对有效性的数据包络分析方法,在对非单纯盈利部门进行评价时取得了很好的效果.

第五,信息论,灰色系统理论等也渗透到综合评价中来,产生了熵值法、灰色关联度评价法等.

第六,多维标度分析及空间统计学的发展提高了统计分析技术上的整合能力,使多目标综合评价方法的应用更加深入.

到目前为止,已经出现了多种的综合评价方法,但这并不意味着综合评价方法和理论已十分完善,还有不少的问题正在不断研究和完善之中,或者还有待于进一步的解决.比如,综合评价方法很多,我们在实际中如何选用?针对同一问题,不同的方法会得到不同的结果,如何解释?如何辨别不同方法对不同问题的优劣?如何衡量综合评价结果的客观准确性?等等问题还需要我们进一步探索和研究,使综合评价方法和理论不断丰富、完善.

第二章 常规综合评价方法

常规方法是指不涉及模糊数学、运筹学、多元统计分析等其他学科的方法;另一方面,它也是各类文献资料中经常见到的方法。

通过这一章的学习,可以了解综合评价的各类问题的共性与特性,常规方法的优点与不足。

第一节 几个评价实例

这一节就是介绍几个有代表性的例子,目的是引出综合评价的基本问题,然后分别给以讨论,了解综合评价的完整过程。

一、综合国力评价

从本世纪 60 年代开始,一些学者尝试对综合国力进行定量分析的研究。I. P. 考尔是第一个对综合国力进行定量测算的学者,他把度量国力状况的指标,选取为人口、国土面积、钢消费量、能源消费量、国民生产总值、总军事实力等六项(见表 2-1),将各国占世

表 2-1 综合国力评价指标和权数

序号 i	1	2	3	4	5	6
指标 x_i	人口	面积	钢消费量	能源消费量	国民生产总值	总军事实力
权数 w_i	200	200	100	100	200	200

界总数的比重作为处理对象,按事先确定的权数加权平均,其结果作为该国综合国力的总得分,由此来进行各国的比较。用公式可表

示为

$$y_j = \frac{\sum_{i=1}^6 w_i \frac{x_{ij}}{X_i}}{\sum_{i=1}^6 w_i} \quad (2-1)$$

$(j = 1, 2, \dots, n)$

式中符号含义为

- n : 参评国家个数;
- x_{ij} : 第 j 国第 i 项指标值;
- X_i : 第 i 项指标世界总计值;
- y_j : 第 j 国综合国力总得分.

当然该方法有它的历史局限性. 80 年代初美国乔治敦大学战略与国际研究中心主任 R. S. 克莱茵提出了测算综合国力的“国力方程”, 80 年代中期日本经济企划厅综合计划局提出了用综合国力指数表示综合国力的大小, 虽在一些方面作了改进, 但基本的思路是相似的, 即: 综合国力 → 构成要素分解 → 指标选择 → 指标值转换 → 权数确定 → 多指标综合 → 比较结果排序.

二、经济效益综合评价

经济效益评价历来是人们关注的问题, 也是现代经济管理中一个比较重要的研究课题. 这方面的文献介绍了不少评价方法.

1. 综合经济效益指数法

该方法是 1983 年刘亮等在第三次全国统计科学讨论会上提出来的. 综合经济效益指数法, 是在制定一套合理的经济效益指标体系的基础上, 把某一年限各项经济效益的数值 (或者是几年的平均值) 作为统一的基数, 然后把报告期各参评单位的每一项指标实际值与该指标的基数值作比较, 计算出该指标的指数值, 最后将各单位各项经济效益指数值加权平均就可求得各单位的综合经济效益指数. 用公式表示如下:

$$y_j = \frac{1}{\sum_{i=1}^p w_i} \left(\frac{x_{1j}}{X_1} w_1 + \dots + \frac{x_{pj}}{X_p} w_p \right) \quad (2-2)$$

$$(j = 1, 2, \dots, n)$$

式中符号含义为

n : 参评单位个数,

x_{ij} : 第 j 个参评单位第 i 项指标值,

X_i : 第 i 项指标的基数值,

w_i : 第 i 项指标的权数,

y_j : 第 j 个参评单位的综合经济效益指数.

得到各单位综合经济效益指数后,就可以由此来评价各单位经济效益水平的高低.刘亮等用全国的数据进行了计算,以1979年各项指标的实际值为基数,测算了1980,1981,1982年的综合经济效益指数.后来国家计委“关于建立和完善计划指标综合体系的暂行规定”中,采用了这种办法,计算了国民经济和社会发展综合指标的各年综合分,计算中,以1980年为基期,1985年综合分为108分,2000年综合分为190分.

下面我们用实例来说明,为了便于看清方法,我们采用简单的数据,表2-2是三个地区经济效益的各项指标实际值.假设各指标的基数值分别如表2-3中所示,由该基数值求得的各指标的指数值也列于表2-3中.

表 2-2 三地区经济效益指标值(虚拟)

指标 x_i	i	单 位	指标实际值 x_{ij}		
			甲地区 $j=1$	乙地区 $j=2$	丙地区 $j=3$
产品销售率	1	%	75	85	60
百元产值实现利税	2	元	25	28	13
可比产品成本降低率	3	%	3	2	-1
全员劳动生产率	4	千元/人	9	12	6
万元产值能耗	5	吨	45	25	18

表 2-3 三地区经济效益指标指数值

指 标	基 数 值	指标的指数值		
		甲	乙	丙
x_1	80	0.94	1.06	0.75
x_2	25	1	1.12	0.52
x_3	5	0.6	0.4	-0.2
x_4	10	0.9	1.2	0.6
x_5'	5	0.44	0.8	1.11

由于第 5 个指标(万元产值能耗)的大小与经济效益呈反方向,所以将该指标取倒数变为每吨能耗的产值(单位:百元),记为 x_5' ,甲、乙、丙三个地区的指标值分别为 2.22, 4, 5.55, 分别除以基数值 5 就可得到 x_5' 的指数值,如表 2-3 中最后一行所示。

设 5 个指标权数分别为 2, 3, 1, 2, 2, 则由公式(2-2)可得三地区的综合经济效益指数分别为

甲地区

$$y_1 = \frac{0.94 \times 2 + 1 \times 3 + 0.6 \times 1 + 0.9 \times 2 + 0.44 \times 2}{2 + 3 + 1 + 2 + 2} = 0.816$$

乙地区

$$y_2 = \frac{1.06 \times 2 + 1.12 \times 3 + 0.4 \times 1 + 1.2 \times 2 + 0.8 \times 2}{2 + 3 + 1 + 2 + 2} = 0.988$$

丙地区

$$y_3 = \frac{0.75 \times 2 + 0.52 \times 3 + (-0.2) \times 1 + 0.6 \times 2 + 1.11 \times 2}{2 + 3 + 1 + 2 + 2} = 0.628$$

可以看出,乙地区经济效益最好,甲地区次之,丙地区最差。

2. 改进的功效系数法

在 1982 年我国的经济效益问题大讨论中,庞皓、谢胜智等提出用“改进的功效系数法”来计算综合经济效果。基本思想是:先确定反映经济效益的几种指标,然后将异度量的各指标值分别转化

为无量纲的相对数,即功效分数,再用加权平均将功效分数综合起来得到总的经济效益分数,由它表示经济效益的水平,用于各单位之间的比较.用公式可表示为

$$d_{ij} = \frac{x_{ij} - x_i^{(s)}}{x_i^{(h)} - x_i^{(s)}}, \quad i = 1, 2, \dots, p \quad (2-3)$$

$$j = 1, 2, \dots, n$$

$$y_j = \sum_{i=1}^p w_i d_{ij} / \sum_{i=1}^p w_i \quad (2-4)$$

或者

$$y_j = \left(\prod_{i=1}^p d_{ij}^{w_i} \right)^{\frac{1}{\sum_{i=1}^p w_i}} \quad (2-5)$$

符号含义如下:

p : 选取指标个数,

$x_i^{(s)}$:第 i 项指标的不允许值,

$x_i^{(h)}$:第 i 项指标的满意值,

d_{ij} : 第 j 个单位的第 i 项指标的功效分数.

其他几个符号含义同公式(2-2).某项指标的不允许值是指该项指标在参评各单位中不应该出现的最坏值,满意值即该项指标在参评单位中可能达到的最好值.对功效分数用公式(2-4)或公式(2-5)综合具有不同的评价效果,这个我们将在第五节合成方法中讨论.总的功效分数高,则综合经济效益好,反之则差.下面用表2-2中数据进行计算.表2-4是由公式(2-3)计算的三个地区各项指标的功效分数.

表 2-4 三地区经济效益指标功效分数

指标 x_i	满意值 $x_i^{(h)}$	不容许值 $x_i^{(s)}$	功效分数 d_{ij}		
			甲 $j=1$	乙 $j=2$	丙 $j=3$
x_1	100	50	80	88	68
x_2	35	15	80	86	56
x_3	5	0	84	76	52
x_4	15	5	76	88	64
x_5	6	40	54	78	86

设 5 个指标权数分别为 2, 3, 1, 2, 2 采用几何平均法(公式(2-5))可得各地区总的功效分数分别为

$$\text{甲地区: } y_1 = \sqrt[10]{80^2 \times 80^3 \times 84 \times 76^2 \times 54^2} = 73.56$$

$$\text{乙地区: } y_2 = \sqrt[10]{80^2 \times 86^3 \times 76 \times 88^2 \times 78^2} = 84.07$$

$$\text{丙地区: } y_3 = \sqrt[10]{68^2 \times 56^3 \times 52 \times 64^2 \times 86^2} = 64.67$$

由此可以看出:乙地区综合经济效益较好,甲地区次之,丙地区最差.

第二节 评价指标的选取

从上一节的实例可以看出,评价指标的选取是否合适,直接影响到综合评价的结论,指标是不是选取得越多就越全面呢?太多了,事实上是重复性的指标,会有干扰;太少了,可能所选的指标缺乏足够的代表性,会产生片面性.每一项指标都是从一个方面反映了评价对象的某些信息,如何正确地、科学地使用这种信息,就是综合评价要处理的问题.

很明显,评价指标的选取与具体问题所涉及的专业知识有关,也与我们能考察获取的手段有关.例如评价参加高考的学生,是否能录取,考试科目太多了,学生受不了.口试可以了解到学生的反应能力快慢,但实际上是无法进行的.尽管如此,仍然有一些原则,一些数学方法可以帮助我们选择.下面分两段来介绍.

一、选取评价指标的一些原则

选取评价指标要遵循的原则,通常有以下几条,这些供我们在解决实际问题时参考.

1. 目的明确

所选用的指标目的很明确.从评价的内容来看,该指标确实能反映有关的内容,反映多与少是另一类问题.决不能将与评价对象、评价内容无关的指标也选择进来.比如要评价一个企业的活力

如何,就要选择与企业自身发展能力有关的指标,例如劳动生产率,市场占有率,产品的优势,等等.所以选取指标目的明确是非常重要的.

2. 比较全面

选择的指标要尽可能覆盖评价的内容,如果有所遗漏,评价就会出偏差.当然,要做到全面是不容易的,但要努力、尽量去做.比如评价科技实力时,既要考虑科技人才、科技知识结构等重要因素,还要考虑科技投入、科技人员流动情况等有关的因素,这样才能比较全面.本书中的一些例子,有时为了便于说明方法,只选用了几个指标,希望不要产生误解,认为只要有几个指标就可以进行综合评价了.

比较全面的另一种说法就是有代表性,所选的指标确能反映要评价的内容,虽然不是全部,但代表了某一侧面.

3. 切实可行

用通俗一些说法,就是可操作性.有些指标虽然很合适,但无法得到,就不切实可行,缺乏可操作性.例如一个学生的才能,没有办法可以直接测量,只能通过做题、面试、科研,几方面去考察.综合评价在一定意义下就是凭借一些可以直接观察、测量的指标去推断不可观察、测量的性能.

以上几条原则还需在实际中灵活考虑和运用.下面我们通过例子来了解指标选取的思路.

例 2.1 外国直接投资经济效益的系统评价^[13].

改革开放以来,我国在吸引外资方面取得了相当的成效.外国直接投资是主要方式之一,因而评价外国直接投资的经济效果具有重要的理论和实践意义.

要对外国直接投资的经济效果进行综合评价,首先得弄清外国直接投资经济效益的基本含义和这种经济效益在经济发展中的形成过程.外国直接投资的经济效果一般说来表现在以下四个方面:

(1)通过有效利用直接投资引入的资金、技术和管理经验等资

源,可以增强受资国运用国内外生产经营资源的能力。(2)运用生产经营资源的能力增强后,能够提高受资国生产经营的总体水平。(3)经过上述能力增强和水平提高,在合理政策导向的条件下,可以改善受资国的经济结构和促进经济稳定迅速地增长。(4)受资国的经济稳定迅速地增长反过来又会进一步吸引更多的外国直接投资,从而实现良性循环。

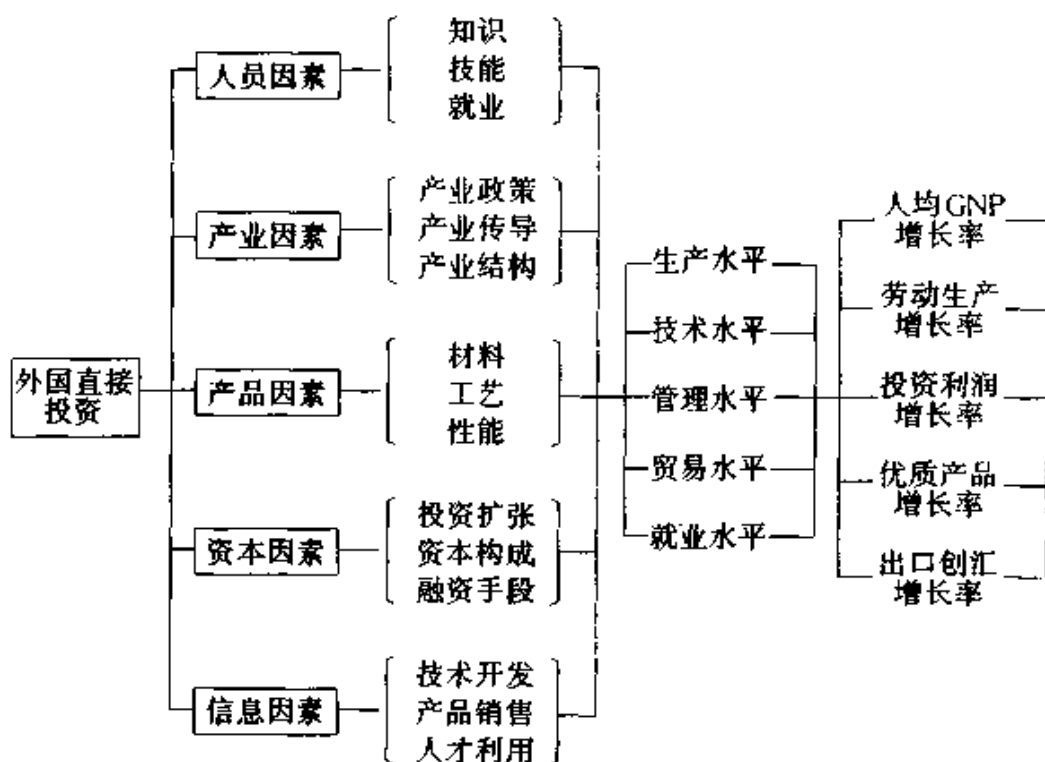


图 2-1 外国直接投资经济效益的形成过程

外国直接投资效果在经济发展中的形成过程如图 2-1 所示。概括地讲,这个过程是通过外国直接投资影响了经济增长的五种资源因素,从而提高五个方面的水平,最终体现在五个增长率上。在弄清外国直接投资的作用机理后,再分层确定评价的指标体系(见图 2-2),第一层次(一级指标)确定为外国直接投资影响的六大方面:调整产业结构,提高技术水平、经营管理水平、资本形成水平,改善贸易结构,提高就业水平。然后根据一级指标的要求,考虑到与直接投资经济效益的关联程度、不同地区利用外资的现状,与

实际部门的工作人员一同筛选确定出 24 个二级指标,建立起评价的指标体系。

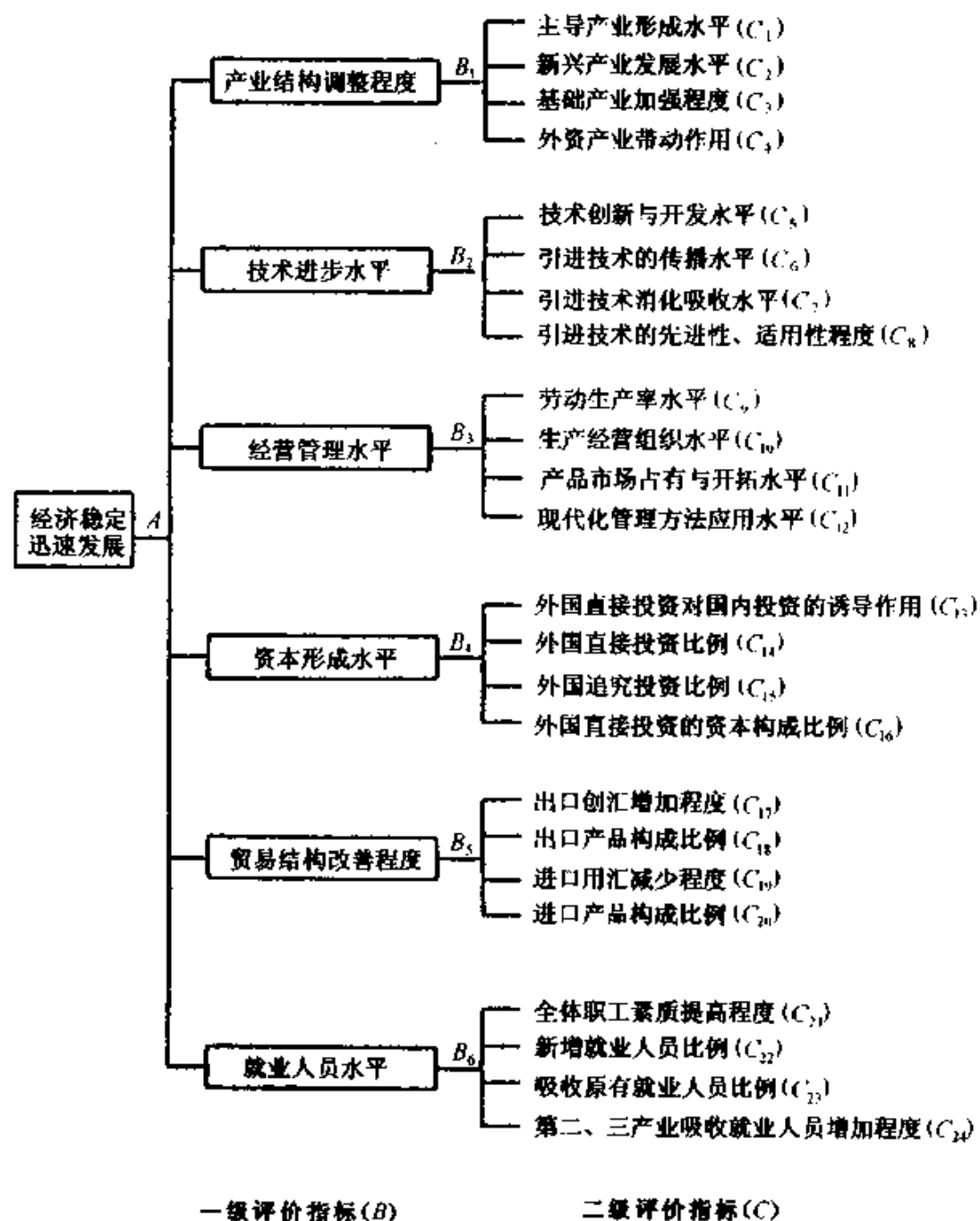


图 2-2 外国直接投资经济效果系统评价的指标体系

从该例可以看出,选取评价指标的前提是对被评事物发展的内在机理要比较清楚,因而多为评价者和有关专家共同确定,最终

的结果带有一定的主观性。

例 2.2 我国各地区普通高等教育发展水平的综合评价^[14].

近年来,我国普通高等教育得到了迅速发展,为国家培养了大批人才.但由于各地区经济发展水平不均衡,加之高等院校原有布局使各地区高教发展的起点不一致,因而各地区高教发展水平就存在一定的差异.对我国各地区普通高等教育发展水平进行综合评价,有利于管理和决策部门从宏观上把握各地区普通高教发展现状,更好地指导和规划高教事业的健康发展.

我们从高等教育的五个方面选取10项相对指标(见图2-3).

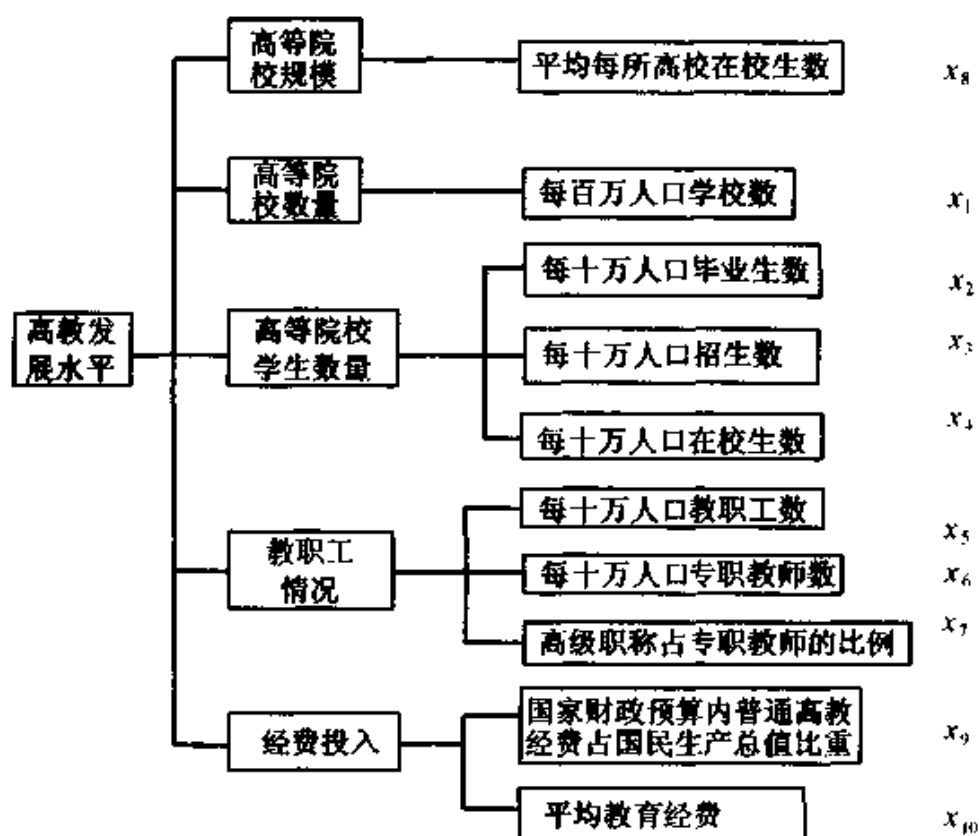


图 2-3 高教发展水平评价指标体系

需要说明一点,高等教育的质量是高等教育发展水平的一个重要标志,它表现在严格而高效的教学管理和高质量的教学水平,

最终体现为高素质的毕业生. 鉴于高等教育质量难以量化, 从可操作性考虑, 这里改用教职工情况予以近似地反映. 另外, 这 10 个评价指标中的某些指标之间(比如反映高等院校学生数量的三个指标之间)可能存在较强的相关性, 下面我们将用定量方法予以筛选.

二、定量指标筛选方法

在按一些原则确定指标体系后, 这些量都是可以观察、测量的. 在这个基础上, 就可以用统计分析中的一些方法来选出一部分, 它们有很好的代表性, 使我们综合评价时, 工作就更容易些.

面对一大堆指标(如上面举的例子, 都有 20 个左右的指标), 这里有重复反映某些内容的, 都考虑会过于偏重某一侧面. 但若删去的过多或不当, 就会不全面, 丢失重要信息. 所以这是一个矛盾, 又要尽可能全面、又希望指标数量不要过多. 解决这个矛盾我们常用下面几种统计方法.

1. 条件广义方差极小^[17]

从统计分析的眼光来看, 给定 p 个指标 x_1, \dots, x_p 的 n 组观察数据, 就称为给了 n 个样本, 相应的全部数据用矩阵 X 表示, 即

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{matrix} \leftarrow \text{第一个样本} \\ \leftarrow \text{第二个样本} \\ \vdots \\ \leftarrow \text{第 } n \text{ 个样本} \end{matrix}$$

每一行代表一个样本的观察值, X 是 $n \times p$ 的矩阵, 利用 X 的数据, 可以算出变量 x_i 的均值、方差与 x_i, x_j 之间的协方差, 相应的表达式是:

$$\text{均值} \quad \bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai}, \quad i = 1, 2, \dots, p$$

$$\text{方差} \quad s_{ii} = \frac{1}{n} \sum_{a=1}^n (x_{ai} - \bar{x}_i)^2, \quad i = 1, 2, \dots, p$$

$$\text{协方差 } s_{ij} = \frac{1}{n} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j), \quad i \neq j$$

$$i, j = 1, 2, \dots, p$$

由 s_{ii}, s_{ij} 形成的矩阵

$$S = (s_{ij})_{p \times p} \quad (2-6)$$

称为 x_1, \dots, x_p 这些指标的方差、协方差矩阵(样本的), 或简称为样本的协方差阵. 用 S 的行列式值

$$|S|$$

反映这 p 个指标变化的状况, 称它为广义方差, 因为 $p=1$ 时 $|S| = |s_{11}| =$ 变量 x_1 的方差, 所以它可以看成是方差的推广. 可以证明, 当 x_1, \dots, x_p 相互独立时, 广义方差 $|S|$ 达到最大值; 当 x_1, \dots, x_p 线性相关时, 广义方差 $|S|$ 的值是 0. 因此, 当 x_1, \dots, x_p 既不独立, 又不线性相关时, 广义方差的大小反映了它们内部的相关性.

现在来考虑条件广义方差, 将 (2-6) 式分块表示, 也就是将 x_1, \dots, x_p 这 p 个指标分成两部分, (x_1, \dots, x_{p_1}) 和 (x_{p_1+1}, \dots, x_p) , 分别记为 $x_{(1)}$ 与 $x_{(2)}$, 即

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} x_{(1)} \\ x_{(2)} \end{pmatrix} \begin{matrix} p_1 \times 1 \\ p_2 \times 1 \end{matrix}, \quad p_1 + p_2 = p$$

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}$$

这样表示后, s_{11}, s_{22} 分别表示 $x_{(1)}$ 与 $x_{(2)}$ 的协方差阵. 给定 $x_{(1)}$ 之后, $x_{(2)}$ 对 $x_{(1)}$ 的条件协方差阵, 从数学上可以推导得到(在正态分布的前提下)

$$S(x_{(2)} | x_{(1)}) = s_{22} - s_{21} s_{11}^{-1} s_{12} \quad (2-7)$$

(2-7) 表示当已知 $x_{(1)}$ 时, $x_{(2)}$ 的变化状况. 可以想到, 若已知

$x_{(1)}$ 后, $x_{(2)}$ 的变化很小, 那么 $x_{(2)}$ 这部分指标就可以删去, 表示 $x_{(2)}$ 所能反映的信息, 在 $x_{(1)}$ 中几乎都可得到, 因此就产生条件广义方差最小的删去方法. 方法如下:

将 x_1, \dots, x_p 分成两部分, (x_1, \dots, x_{p-1}) 看成 $x_{(1)}$, x_p 看成 $x_{(2)}$, 用(2-7)就可算出 $S(x_{(2)}|x_{(1)})$, 此时是一个数值, 它是识别 x_p 是否应删去的量, 记为 t_p . 类似地, 对 x_i , 可以将 x_i 看成 $x_{(2)}$, 余下的 $p-1$ 个看成 $x_{(1)}$, 用(2-7)算出一个数, 记为 t_i . 于是得到 t_1, t_2, \dots, t_p 这 p 个值, 比较它们的大小, 最小的一个是可以考虑删去的, 这与所选的临界值有关, 这个临界值 C 就是自己选的, 认为小于这个 C 就可删去, 大于这个 C 不宜删去. 给定 C 之后, 逐个检查

$$t_i < C, i = 1, 2, \dots, p$$

是否成立, 有就删, 删去后对留下的变量, 可以完全重复上面的过程, 因此, 这样可以进行到没有可删的为止, 这就选得了既有代表性, 又不重复的指标集.

从(2-7)式可以看到, 如有经验, 不必逐个考虑, 完全可以将指标分组, 按组来考虑, 方法、步骤与上面所说的相同.

2. 极大不相关^[18]

容易想到, 如果 x_1 与其他的 x_2, \dots, x_p 是独立的, 那就表明 x_1 是无法由其他指标来代替的, 因此保留的指标应该是相关性越小越好, 在这个想法指引下, 就导出极大不相关方法. 首先利用(2-6)式, 求出(样本的)相关阵 R ,

$$R = (r_{ij}) \quad (2-8)$$

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \quad i, j = 1, 2, \dots, p$$

r_{ij} 称为 x_i 与 x_j 的相关系数, 它反映了 x_i 与 x_j 的线性相关程度. 现在要考虑的是一个变量 x_i 与余下的 $p-1$ 个变量之间的线性相关程度, 称为复相关系数. 通常记为

$$\rho_{x_i|x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_p}$$

这个符号太复杂,现在简化为 ρ_i ,但要注意它的意义. ρ_i 可以由下面的公式来计算.先将 R 分块,例如要计算 ρ_p ,就将 R 写成

$$R = \begin{bmatrix} R_{-p} & r_p \\ r_p^T & 1 \end{bmatrix}_1^{p-1} \quad (R_{-p} \text{ 表示除去 } x_p \text{ 的相关阵})$$

(注意 R 中的主对角元素 $r_{ii}=1, i=1,2,\dots,p$),于是

$$\rho_p^2 = r_p^T R_{-p}^{-1} r_p \quad (2-9)$$

类似地要计算 ρ_i^2 时,将 R 中的第 i 行、第 i 列经过置换,放在矩阵的最后一行,最后一列,此时

$$R \xrightarrow{\text{置换后}} \begin{bmatrix} R_{-i} & r_i \\ r_i^T & 1 \end{bmatrix}$$

于是 ρ_i^2 的计算公式为

$$\rho_i^2 = r_i^T R_{-i}^{-1} r_i, i = 1, 2, \dots, p$$

算得 $\rho_1^2, \dots, \rho_p^2$ 后,其中值最大的一个,表示它与其余变量相关性最大,指定临界值 D 之后,当 $\rho_i^2 > D$ 时,就可以删去 x_i .

下面我们仍用例 2.2 来作一次删去的示范,说明这个方法.

例 2.3(续例 2.2) 我国各地区高教发展水平的十项指标值如表 2-5 所示,其中 x_1, x_2, \dots, x_{10} 的含义见图 2-3. 试用极大不相关法进行筛选.

由表 2-5 中数据可求得十个指标中的每一个与其余九个指标的复相关系数如下(记 x_i 与其余九个指标的复相关系数为 $\rho_i^{(1)}$):

$$\begin{array}{ll} \rho_1^{(1)} = 0.99786, & \rho_2^{(1)} = 0.99692 \\ \rho_3^{(1)} = 0.99923, & \rho_4^{(1)} = 0.99946 \\ \rho_5^{(1)} = 0.99926, & \rho_6^{(1)} = 0.99952 \\ \rho_7^{(1)} = 0.93591, & \rho_8^{(1)} = 0.94687 \\ \rho_9^{(1)} = 0.92324, & \rho_{10}^{(1)} = 0.80920 \end{array}$$

表 2-5 高教发展水平指标值

序号	地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	北京	5.96	310	461	1557	931	319	44.36	2615	2.20	13631
2	上海	3.39	234	308	1035	498	161	35.02	3052	0.90	12665
3	天津	2.35	157	229	713	295	109	38.40	3031	0.86	9385
4	陕西	1.35	81	111	364	150	58	30.45	2699	1.22	7881
5	辽宁	1.50	88	128	421	144	58	34.30	2808	0.54	7733
6	吉林	1.67	86	120	370	153	58	33.53	2215	0.76	7480
7	黑龙江	1.17	63	93	296	117	44	35.22	2528	0.58	8570
8	湖北	1.05	67	92	297	115	43	32.89	2835	0.66	7262
9	江苏	0.95	64	94	287	102	39	31.54	3008	0.39	7786
10	广东	0.69	39	71	205	61	24	34.50	2988	0.37	11355
11	四川	0.56	40	57	177	61	23	32.62	3149	0.55	7693
12	山东	0.57	58	64	181	57	22	32.95	3202	0.28	6805
13	甘肃	0.71	42	62	190	66	26	28.13	2657	0.73	7282
14	湖南	0.74	42	61	194	61	24	33.06	2618	0.47	6477
15	浙江	0.86	42	71	204	66	26	29.94	2363	0.25	7704
16	新疆	1.29	47	73	265	114	46	25.93	2060	0.37	5719
17	福建	1.04	53	71	218	63	26	29.01	2099	0.29	7106
18	山西	0.85	53	65	218	76	30	25.63	2555	0.43	5580
19	河北	0.81	43	66	188	61	23	29.82	2313	0.31	5704
20	安徽	0.59	35	47	146	46	20	32.83	2488	0.33	5628
21	云南	0.66	36	40	130	44	19	28.55	1974	0.48	9106
22	江西	0.77	43	63	194	67	23	28.81	2515	0.34	4085
23	海南	0.70	33	51	165	47	18	27.34	2344	0.28	7928
24	内蒙古	0.84	43	48	171	65	29	27.65	2032	0.32	5581
25	西藏	1.69	26	45	137	75	33	12.10	810	1.00	14199
26	河南	0.55	32	46	130	44	17	28.41	2341	0.30	5714
27	广西	0.60	28	43	129	39	17	31.93	2146	0.24	5139
28	宁夏	1.39	48	62	208	77	34	22.70	1500	0.42	5377
29	贵州	0.64	23	32	93	37	16	28.12	1469	0.34	5415
30	青海	1.48	38	46	151	63	30	17.87	1024	0.38	7368

可见,指标 x_6 与其余指标间的复相关系数最大,因而它最能被其余指标替代,故先将 x_6 剔除掉,再计算余下的九个指标中的

每一个与其余八个指标的复相关系数,记为 $\rho_i^{(2)}$:

$$\rho_1^{(2)} = 0.99671, \quad \rho_2^{(2)} = 0.99647$$

$$\rho_3^{(2)} = 0.99923, \quad \rho_4^{(2)} = 0.99945$$

$$\rho_5^{(2)} = 0.99497, \quad \rho_7^{(2)} = 0.92327$$

$$\rho_8^{(2)} = 0.94490, \quad \rho_9^{(2)} = 0.90513$$

$$\rho_{10}^{(2)} = 0.76681$$

由此看出,我们应剔除指标 x_4 ,同理再计算余下八个指标的复相关系数,结果如下:

$$\rho_1^{(3)} = 0.99657, \quad \rho_2^{(3)} = 0.99582$$

$$\rho_3^{(3)} = 0.99839, \quad \rho_5^{(3)} = 0.99292$$

$$\rho_7^{(3)} = 0.92182, \quad \rho_8^{(3)} = 0.94102$$

$$\rho_9^{(3)} = 0.90321, \quad \rho_{10}^{(3)} = 0.75579$$

指标 x_3 应被剔除,在余下的七个指标中,如果再计算下去,应剔除的将是 x_1 ,但考虑到指标 x_1 (每百万人口学校数)反映着高教发展水平的五个方面之一——高等院校数量,如果剔除的话,这个侧面将不能被反映,所以剔除到此为止,我们将余下的七个指标作为评价指标.从图 2-3 看出,剔除的 x_4, x_3 与留下的 x_2 反映同一个侧面, x_6 与 x_5, x_7 反映同一个侧面,因而 x_4, x_3, x_6 在一定程度上可以被其他指标代替,这与定性分析的结果也是吻合的.

3. 选取典型指标

如果开始考虑的指标过多,则可以将这些指标先进行聚类,而后在每一类中选取若干个典型指标.关于聚类分析我们将在第四章介绍.在每一类中选取典型指标可以用上述方法 1 或 2.这两种方法的计算量都相当大,下边介绍一种用单相关系数选取典型指标的方法,该方法较为粗略,但其计算简单,在实际中可依据具体情况选用.

假设反映事物同一侧面的或聚为同一类的指标有 n 个,分别为 a_1, a_2, \dots, a_n . 第一步计算 n 个指标之间的相关系数矩阵 R

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

第二步计算每一指标与其他 $n-1$ 个指标的决定系数(相关系数的平方)的平均值 \bar{r}_i^2

$$\bar{r}_i^2 = \frac{1}{n-1} \left(\sum_{j=1}^n r_{ij}^2 - 1 \right), i = 1, 2, \dots, n \quad (2-10)$$

则 \bar{r}_i^2 粗略地反映了 a_i 与其他 $n-1$ 个指标的相关程度. 第三步比较 \bar{r}_i^2 的大小, 若有

$$\bar{r}_k^2 = \max_{1 \leq i \leq n} \bar{r}_i^2 \quad (2-11)$$

则可选取 a_k 作为 a_1, a_2, \dots, a_n 的典型指标, 需要的话, 还可以在余下的 $n-1$ 个指标中继续选取. 这里之所以要用相关系数的平方是为了防止相关系数可能为负, 因而无法直接相加求平均. 如果相关系数均为正, 则也可直接用相关系数. 比如, 例 2.2 中的指标 x_5, x_6, x_7 , 反映了高教发展水平的同一个侧面, 其相关阵为

$$R = \begin{pmatrix} 1 & 0.99859 & 0.55988 \\ 0.99859 & 1 & 0.55001 \\ 0.55988 & 0.55001 & 1 \end{pmatrix}_{3 \times 3}$$

直接用相关系数由(2-10)式可求得

$$\bar{r}_5 = \frac{1}{3-1} (1 + 0.99859 + 0.55988 - 1) = 0.77924$$

同理可得

$$\bar{r}_6 = 0.7743, \quad \bar{r}_7 = 0.55495$$

\bar{r}_5 最大, 故应选 x_5 为 x_5, x_6, x_7 的典型指标. 如果再要选一个, 应在 x_6 和 x_7 之间选取, 而 x_6 和 x_7 之间相关性为 0.55001, 无法再用上边的方法, 但从相关阵可以看出 x_6 与 x_5 的相关性(0.99859)要大于 x_7 与 x_5 的相关性(0.55988), 已经选取了 x_5 , 它已将 x_6 基本代替, 因而应选 x_7 . 即如果从 x_5, x_6, x_7 中选两个

指标,应选 x_5 和 x_7 .

第三节 无量纲化方法

在本章第一节综合国力评价实例中,I.P.考尔选取了人口、面积、钢消费量等6个指标.这6个指标显然是异量纲的,而且数值差异较大,直接将它们加权平均是不合适的,也没有实际意义.考尔将各指标值与世界总计相比较,把指标值转化为无量纲的相对数——比重,同时数值大小规范在 $[0,1]$ 内.这种去掉指标量纲的过程,我们称为数据的无量纲化(也称为数据的规格化),它是指标综合的前提.如果我们把指标无量纲化以后的数值称为指标评价值,那么无量纲化过程就是指标实际值转化为指标评价值的过程,无量纲化方法也就是指如何实现这种转化.从数学角度讲就是要确定指标评价值依赖于指标实际值的一种函数关系式.我们把无量纲化方法从几何的角度归结为三类:直线型无量纲化方法、折线型无量纲化方法、曲线型无量纲化方法.

一、直线型无量纲化方法

先来分析考尔所用的方法,不妨用 x 表示指标实际值,用 y 表示指标评价值, x_{ij} 表示第 j 国的第 i 项指标实际值, X_i 表示第 i 项指标世界的总数,则令

$$y_{ij} = \frac{x_{ij}}{X_i}, \quad (2-12)$$

指标评价值与实际值之间是一种线性关系(如图2-4).指标评价值随实际值等比例变化.常用的直线型无量纲化方法有以下几种:

1. 阈值法

阈值也称临界值,是衡量事物发展变化的一些特殊指标值,比如极大值、极小值、满意值、不允许值等.阈值法是用指标实际值与阈值相比以得到指标评价值的无量纲化方法,主要公式及特点等如表2-6所示,其中 n 为参评单位的个数.

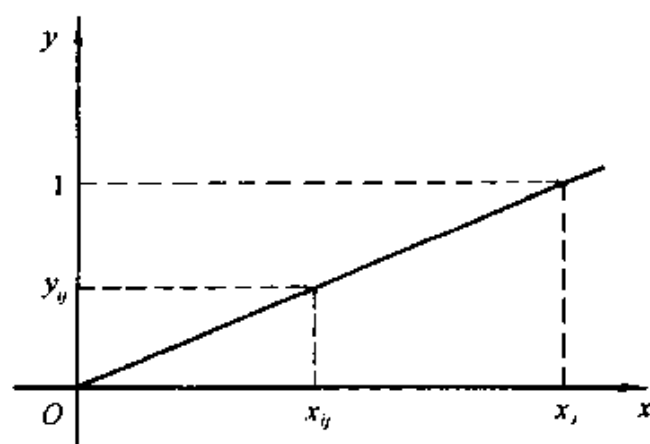


图 2-4

实际在多数情况下,所有的指标值均大于 0,故以上按所有 $x_i > 0$ 处理,事实上若有部分 $x_i \leq 0$,公式仍然适用.最后两个公式实质是一样的,只不过评价值的范围有所变化.我们可以用同样的方法将前几个公式的评价值范围变到我们希望的范围.公式 2 和公式 3 适合于处理指标体系中的逆指标,比如经济效益评价中的成本类指标,社会效益评价中的资源耗费类指标等.当然我们也可以事先将其转化为正指标,然后再与其他指标一起无量纲化,这在后边还会谈到.在弄清这些公式的特点之后,我们就可以灵活选用或构造新的公式.需要再提及的是,如果我们所选取的评价指标都是正指标(或都是逆指标),则在一次综合评价中对所有指标应采取同一种无量纲化公式.如果在评价中既有正指标,又有逆指标,且不作逆指标的转化处理的话,则应对正、逆指标分别采取两种相互对应的无量纲化公式.所谓相互对应的公式是指两种公式得到的评价值范围应是一致的,这样才可以进行综合.比如表 2-6 中公式 1 和公式 2,公式 3 和公式 4 都是相互对应的.将公式 3 稍作变形就可得到与公式 5 对应的公式.

在前面提到的综合国力评价中,R. S. 克莱茵提出的综合国力方程就是利用公式 1 进行指标数据的无量纲化的.比如,克莱茵将各国的国民生产总值 GNP 与当时 GNP 最大的美国相比,得到各国 GNP 的评价值.为了通俗,他又将其乘以 100 转化为百分

表 2-6 几种阈值法参照表

序号	公式	影响评价因素	评价范围	几何图形	特点
1	$y_i = \frac{x_i}{\max_{1 \leq i \leq n} x_i}$	$x_i, \max_{1 \leq i \leq n} x_i$	$[\frac{\min x_i}{\max x_i}, 1]$		评价值随指标值增大,若指标值均为正,则评价值不可能为零,指标最大值的评价值为 1.
2	$y_i = \frac{\max x_i + \min x_i - x_i}{\max x_i}$	$x_i > 0$ $x_i, \max x_i$ $\min x_i$	$[\frac{\min x_i}{\max x_i}, 1]$		评价值随指标值增大而减小,适合于对逆指标进行无量纲处理,即无量纲化和指标转化同时进行.
3	$y_i = \frac{\max x_i - x_i}{\max x_i - \min x_i}$	$x_i, \max x_i$ $\min x_i$	$[0, 1]$		同上
4	$y_i = \frac{x_i - \min x_i}{\max x_i - \min x_i}$	同上	$[0, 1]$		评价值随指标值增大而增大,指标最小值的评价值为零,指标最大值的评价值为 1.
5	$y_i = \frac{x_i - \min x_i}{\max x_i - \min x_i} \cdot k + q$	$x_i, \min x_i$ $\max x_i, k, q$	$[q, k + q]$		评价值随指标值增大而增大,指标最小值的评价值为 q,指标最大值的评价值为 k + q.

制.经济效益评价中的功效系数法是公式 4 的应用,即最大最小值分别取为满意值和不允许值.改进的功效系数法是公式 5 的应用,即 k 取 40, q 取 60.在综合经济效益指数中将各指标数值与基期相比进行无量纲化,实际也是一种阈值法.另外实际中也有将指标实际值除以该指标的第一个样本值或均值,分别称为初值化和均值化,实质也是一种阈值法.

2. 标准化方法

统计学理论告诉我们,要对多组不同量纲的数据进行比较,可以先将它们分别标准化,转化成无量纲的标准化数据.而综合评价就是要将多组不同的数据进行综合,因而可以借助于标准化方法来消除数据量纲的影响.标准化公式为

$$y_i = \frac{x_i - \bar{x}}{s} \quad (2-13)$$

上式中

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2-14)$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (2-15)$$

指标实际值与评价值的关系如图 2-5 所示.

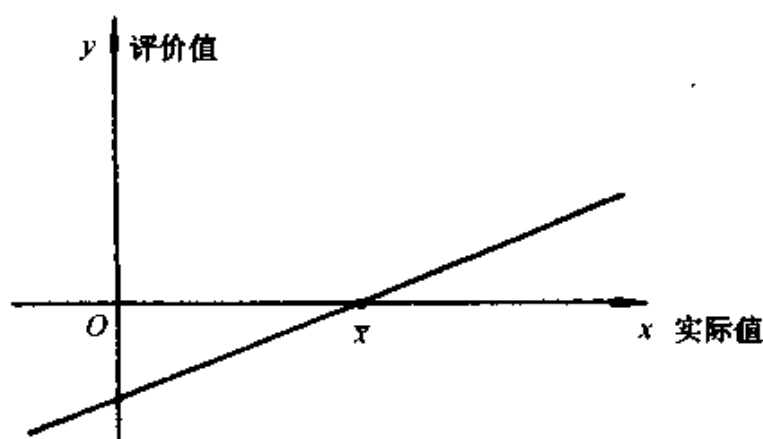


图 2-5

可以看出,无论指标实际值如何,指标的评价值总是分布在零的两侧.指标实际值比平均值大的,其评价值为正,反之为负,实际值距平均值越远则其评价值距零越远.这种方法与阈值法最大的不同在于:第一,它利用了原始数据的所有信息;第二,它要求样本数据较多;第三,它的评价值结果超出 $[0,1]$ 区间,有正有负.为了更符合习惯,我们可以将其转化为百分数形式,比如用公式(见图 2-6):

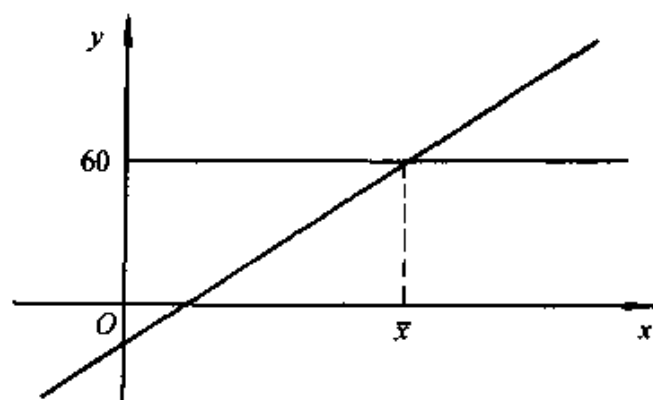


图 2-6

$$\begin{aligned}
 y_i &= 60 + \frac{x_i - \bar{x}}{10s} \times 100 \\
 &= 60 + \frac{x_i - \bar{x}}{s} \times 10
 \end{aligned}
 \tag{2-16}$$

均值转化为 60,超过均值的转化为 60 以上,反之在 60 以下.这种“百分数”还不同于一般的百分数,因为个别极端数值的转化值可能超出 $[0,100]$ 区间.另外,也有的将均值转化为 50.

在第三、四章介绍的综合评价的多元统计方法中,大多是用标准化方法进行数据的无量纲化的.

例 2.4 某次考试中有关的统计结果及甲、乙两考生的成绩如表 2-7 所示.

表 2-7

科 目	原始分数		全体考生		无量纲化结果	
	甲	乙	均值 \bar{x}	标准差 s	甲	乙
数学	78	82	80	8	57.5	62.5
物理	45	41	42	4	67.5	57.5
化学	72	74	74	6	56.7	60

这里就涉及一个对甲、乙两考生理科学习水平的综合评价问题,评价指标为三科目的考试成绩.我们可以利用公式(2-16)对甲、乙两名考生的成绩进行无量纲化,以甲生数学成绩为例

$$y_{\text{甲数}} = 60 + \frac{78 - 80}{10 \times 8} \times 100$$

$$= 57.5$$

计算结果见表 2-7 最右边一栏.还有两个问题说明一下,第一个问题是:有必要对考试分数进行无量纲化吗?通常的看法是考试成绩是一种分数,本身没有单位,而且都是百分制,因此不必要无量纲化,直接相加求总分即可.事实上,这种看法是有误解的.由于不同科目试题的难易程度、份量都不一定相同,其分数的“含金量”并不相同,因而不能直接相加.本例中物理较难,因而分数普遍偏低,数学则相反,因而物理中的 1 分就比数学中的 1 分具有更高的“含金量”.无量纲化以后,各科分数都以 60 为中心而分布,具有了可比性,因而可以相加.例中甲、乙两考生的原始总分分别为 195 和 197,乙比甲高两分.无量纲化以后甲、乙两考生的分数分别为 188.3 和 180.甲比乙高.很显然,后一个结果才是客观合理的.虽然乙在数学和化学上共比甲高出 6 分,而甲在物理上仅比乙高出 4 分,但这 4 分的“含金量”要比那 6 分的“含金量”高.近年来,我国的高考实行用标准分录取就是源于这个思想.后面我们将对高考标准分再行讨论.第二个问题是,本例实际上是应该对所有考生进行综合评价,因而均值和标准差是由全体考生计算而来.我们这里为了计算方便,便于说明想法,仅对其中的两名考生进行了计

算.

3. 比重法

比重法是将指标实际值转化为它在指标值总和中所占的比重,主要公式有:

$$y_i = \frac{x_i}{\sum_{i=1}^n x_i} \quad (2-17)$$

或

$$y_i = \frac{x_i}{\sqrt{\sum_{i=1}^n x_i^2}} \quad (2-18)$$

公式(2-17)适合指标值均为正数的情况,且评价值之和满足

$$\sum_{i=1}^n y_i = 1 \quad (2-19)$$

公式(2-18)适合于指标值有负值的情况,一般情况下,指标评价值不满足(2-19)式,而是满足 $\sum_{i=1}^n y_i^2 = 1$. 考尔在综合国力评价中就是利用比重法消除量纲影响的.

以上介绍了三种常用的直线型无量纲化方法,这些方法的最大特点是简单、直观. 直线型无量纲化方法实质是假定指标评价值与实际值成线性关系,评价值随实际值等比例变化,也就是说指标值在不同区间内变化对被评事物的综合水平影响是一样的(如图2-7所示),即在事物发展的前期和后期指标值相同的变化量引起评价值的变化量是相同的. 而这一点与事物发展变化的实际情况往往并不符合,这也就是直线型无量纲化方法的最大缺陷. 比如我们要评价一个学生,学习成绩是一个指标,如果我们将学习成绩用直线型无量纲化方法转化为评价值,再与其他指标综合,这就意味着学习成绩从40分增加到50分和由90分增加到100分对评价起的作用是相同的(二者导致评价值的增加量是相同的),这显然与实际情况不符. 实际中后者要比前者难得多,需要比前者付出更多的努力,因此,从这个角度讲,应该给后者以较多的评价值增

加量.为了解决这个问题很自然我们就会想到用折线或曲线来代替直线,这就是我们下面要介绍的折线型和曲线型无量纲化方法.

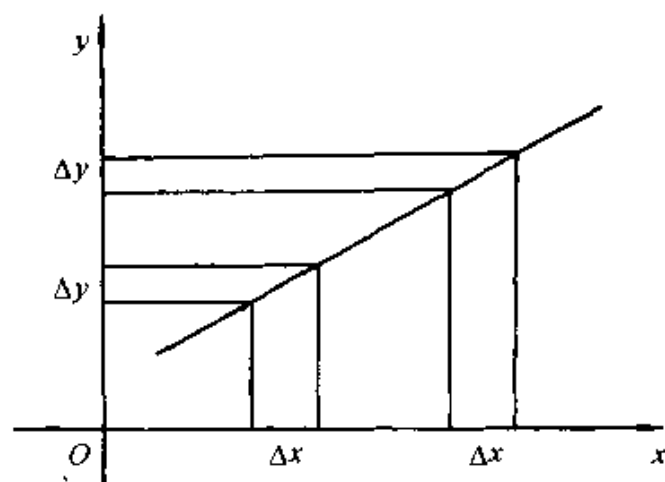


图 2-7

二、折线型无量纲化方法

折线型无量纲化方法适合于事物发展呈现阶段性,指标值在不同阶段变化对事物总体水平影响是不相同的.构造折线型无量纲化方法与直线型不同之处在于必须找出事物发展的转折点的指标值并确定其评价值.常用的有下面三种类型.

1. 凸折线型

采用凸折线型无量纲化公式,指标值在前期的变化被赋以较多的评价值增加量,如图 2-8 所示,(a)适合于正指标,(b)适合于逆指标.

比如用阈值法可构造如下折线型公式(如图 2-9 所示):

$$y_i = \begin{cases} \frac{x_i}{x_m} y_m, & 0 \leq x_i \leq x_m \\ y_m + \frac{x_i - x_m}{\max_i x_i - x_m} (1 - y_m), & x_i > x_m \end{cases} \quad (2-20)$$

式(2-18)中 x_m 为转折点指标值, y_m 为 x_m 的评价值.

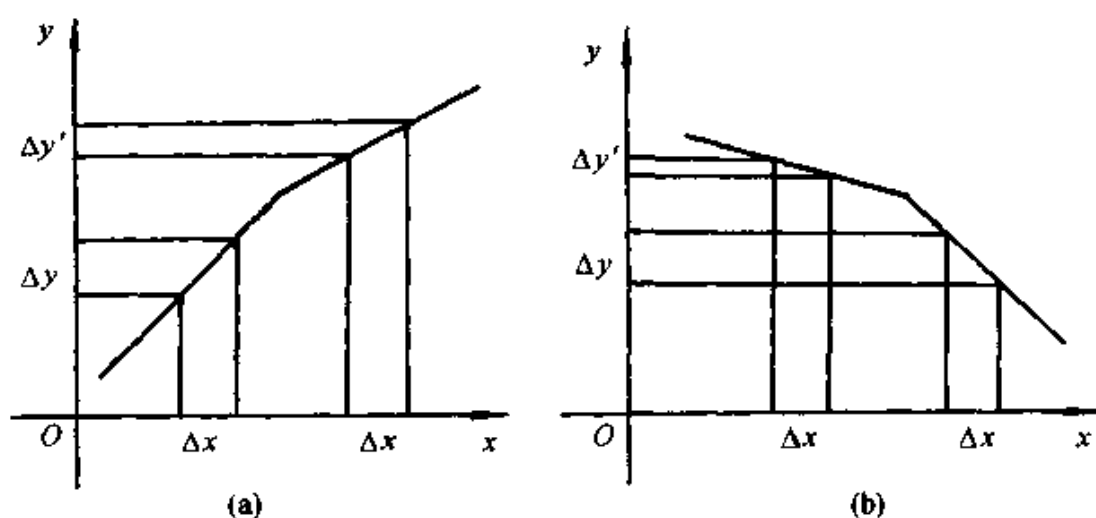


图 2-8

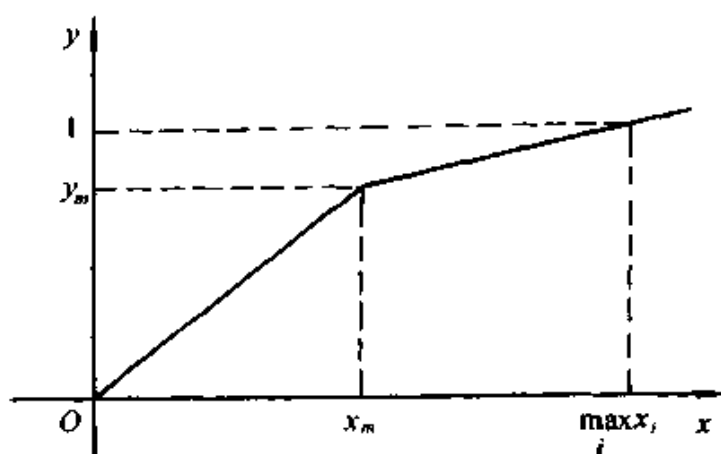


图 2-9

2. 凹折线型

与凸折线型不同,凹折线型无量纲化公式对指标后期变化赋予较多评价值增加值,指标后期变化对事物发展总体水平影响较大.如图 2-10 所示.

在式(2-18)中将 y_m 取小一些即可得到凹折线型无量纲化公式.

3. 三折线型

常用的三折线型无量纲化公式有图 2-11 所示两种形式.

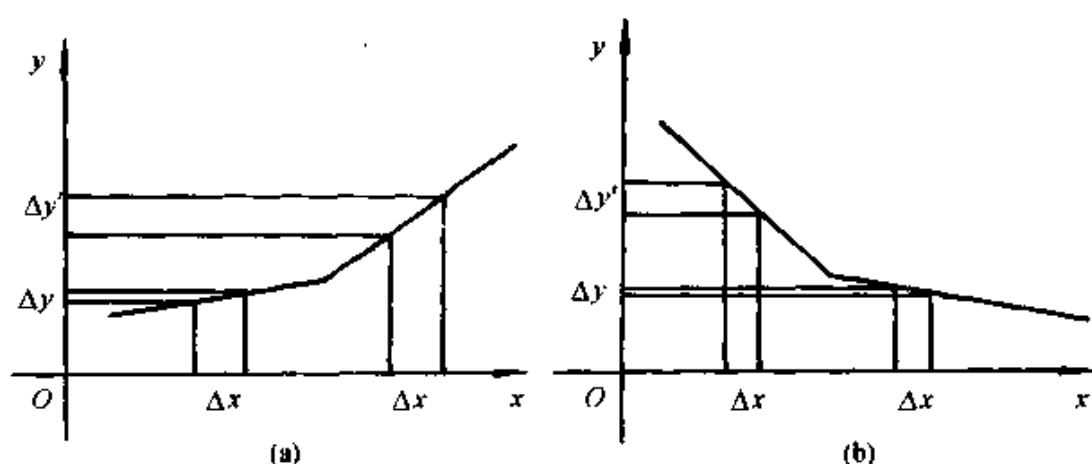


图 2-10

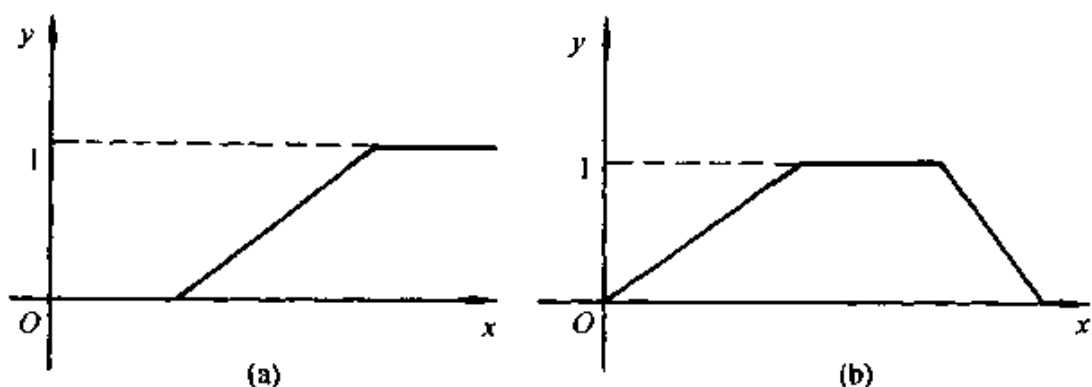


图 2-11

(a)适合于某些事物要求指标值在某区间内变化,若超出这个区间则指标值的变化对事物的总体水平几乎没有什么影响.比如评价人的健康状况,以饮食量作为一个衡量指标,显然该指标值在一个范围内变化才有意义,太小或太大对一个人的健康状况没有什么影响.(b)适合于适度指标的无量纲化,即指标值过大或过小都会对事物产生不利影响.比如评价综合国力,以人口作为一个指标.相对于有限的各种资源来讲,一国人口只有在适度的范围内才会有利于其综合国力的提高,过多或过少都会削弱其综合国力,对人口这个适度指标采用(b)型无量纲化公式是比较合适的.

从理论上讲,折线型无量纲化方法比直线型无量纲化方法更

符合事物发展的实际情况,但应用的前提是评价者必须对被评事物有较为深刻的理解和认识,合理地确定出指标值的转折点及其评价值.

三、曲线型无量纲化方法

有些事物发展阶段性的分界点不很明显,而前中后各期发展情况又截然不同,也就是说指标值变化对事物总体水平的影响是逐渐变化的,而非突变的.在这种情况下,曲线型无量纲化公式更为合适.常用的曲线型无量纲化公式如表 2-8 所示.

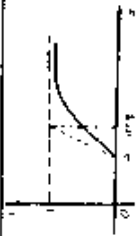
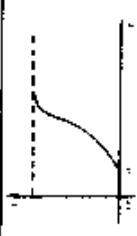
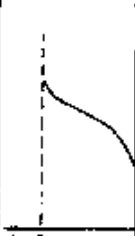
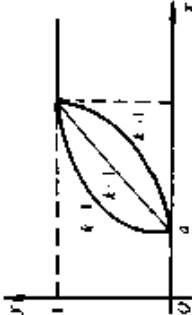
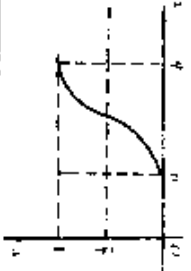
例 2.5 无量纲化在高考录取工作中的应用.

高考就是对考生进行综合评价,各科目成绩即为评价指标,综合评价的结果就是录取的依据.以往的作法就是直接将各科目原始成绩相加,总分就是综合评价值,然后依据各考生的总分从高到低录取.从例 2.4 中我们已经可以看出这样的作法是不合理的.这里的核心问题是:各科目的分数不是同质的,不具有可比性,因而不能直接相加.必须首先对不同科目的分数进行无量纲化,即转化为标准分数,然后才能相加.

化为标准分数有几种不同的方式,比如线性标准分数、正态化标准分数、中位数标准分数等.由公式(2-16)得到的标准分数就是线性标准分数,用公式(2-16)进行无量纲化的前提是不同科目的分数都要呈现正态分布,否则,不同科目分数经公式(2-16)转化后还不具可比性.例如,如果一个科目分数呈正态分布,另一科目分数呈偏态分布,那么相同的线性标准分数可能对应不同的百分等级¹⁾,难以做到准确的比较.通常,各科目考试成绩在考生人数较多时,从理论上讲应该呈现正态分布,但是实际考试中,由于题目难度不同及评分误差等随机因素影响,很难使考试分数的实

1)百分等级:在一批分数中,小于某分数的分数个数占整批分数个数的百分比,称为该分数在这批分数中的百分等级.比如,某班共 50 人,某人成绩为 80 分,80 分以下有 30 人,则该考生的百分等级为 60($30 \div 50 = 60\%$).

表 2-8 曲线型无量纲化公式一览表

名称	图形	解析式	特点
升半Γ型		$y = \begin{cases} 0 & 0 \leq x \leq a \\ 1 - e^{-k(x-a)} & x > a \end{cases}$ $k > 0$	指标评价价值随实际值变化,到后期逐渐缓慢直至几乎不变,适合于指标值在后期变化对事物发展总体水平影响较小的情况。
升半正态型		$y = \begin{cases} 0 & 0 \leq x \leq a \\ 1 - e^{-k(x-a)^2} & x > a \end{cases}$ $k > 0$	指标评价价值随实际值中期变化较快,而后期相对较慢,适合于指标中期值变化对事物发展总体水平影响较大的情况。
升半柯西型		$y = \begin{cases} 0 & 0 \leq x \leq a \\ \frac{k(x-a)^2}{1+k(x-a)^2} & x > a \end{cases}$ $k > 0$	同上
升半凹凸型		$y = \begin{cases} 0 & 0 \leq x \leq a \\ a(x-a)^k & a \leq x \leq a + \frac{1}{\sqrt[k]{ka}} \\ 1 & x \geq a + \frac{1}{\sqrt[k]{ka}} \end{cases}$	指标评价价值随指标实际值的变化逐渐加快或逐渐减慢。
升半岭型		$y = \begin{cases} 0 & a \leq x \leq a \\ \frac{1}{2} - \frac{1}{2} \sin \frac{\pi}{b-a} (x - \frac{a+b}{2}) & a < x \leq b \\ 1 & x > b \end{cases}$	指标评价价值随指标实际值中期变化快,前后期较慢,且呈对称情况。

际分布与正态分布完全一致. 因此, 实践中并不采用公式(2-16)进行无量纲化, 而是将各科目原始分数分别进行正态化变换, 以使得源于不同分布的分数可以比较, 然后再相加.

正态化变换是一种非线性变换, 不论原始分数的分布呈何种形态, 通过这种变换, 可以使变换后的分数呈现标准正态分布, 因而变换后的分数称为正态化标准分数. 利用这种分数可以明确知道, 原分数在整批分数中所处的位置. 对原始分数进行正态化变换得到正态化标准分数的步骤是:

(1) 对原始分数排序, 求出每个原始分数 x 对应的百分比

$$P_x = \frac{x \text{ 分以下的考生人数}}{\text{考生人数}} \quad (2-21)$$

(2) 查标准正态分布表, 得到每个分数的正态化标准分数 Z_x 如图 2-12 所示, 阴影部分面积为 P_x , 即

$$\int_{-\infty}^{Z_x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = P_x$$

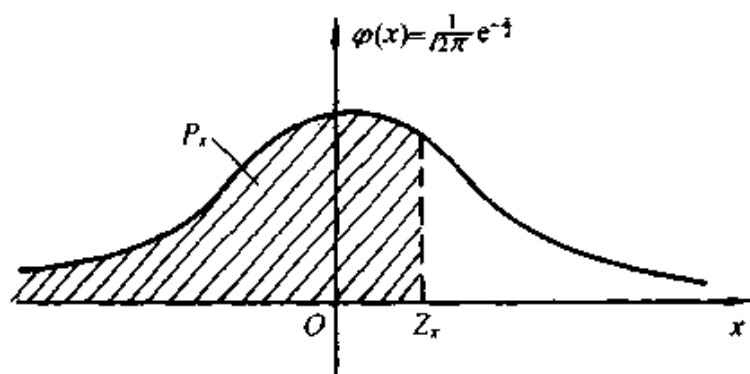


图 2-12 $N(0,1)$ 分布

比如, 某考生某科原始分数所对应的百分位数值为 0.5, 则查表可知其正态化标准分为 0, 另一考生某科原始分数所对应的百分位数值为 0.8413, 则查表可知其正态化标准分为 1. 反过来由正态化标准分数也可查得原始分数所对应的百分比.

与原始分数相比, 正态化标准分数服从标准正态分布, 故而形式简单、整齐, 它能准确保留原始数据的相对位置信息, 合理调整

分数间距使不同科目间具有可比性. 由于正态化标准分数绝对值较小, 有正还有负, 且多为小数, 实践中在不同科目相加前再作一次线性变换, 得到标准 T 分数

$$T = 500 + 100Z \quad (2-22)$$

这样可以将分数的档次拉开, 便于比较. 各科目的分数都经过正态化变换和(2-22)式的变换得到标准 T 分数后, 就可以进一步形成综合分. 实际中具体的操作方法可参阅文献[9].

无量纲化方法在使用时, 要尽量选择适合于讨论对象性质的方法, 不要不加思考, 随便选用一种, 当然也可以选用几种, 然后分析不同的无量纲化方法对结论会产生多大的影响. 实际工作表明, 不是越复杂的方法就越合适, 关键在于是否切合实际的要求, 在这个前提下, 应该说越简便、越方便施用, 越会受欢迎.

四、逆指标和适度指标的处理

对于选出的评价指标, 如果是正指标, 即指标值越大越好, 上面介绍的种种方法是适用的. 对于逆指标和适度指标, 上面介绍的方法有些是不适用的, 比较常用的方法是首先将它们转换为正指标, 然后再无量纲化.

逆指标(记为 x , 其 n 个样本值为 x_1, \dots, x_n)转换为正指标, 可选用的简单变换是(x'_i 表示转换后的指标):

对 $i = 1, 2, \dots, n$, 取 $x'_i = \frac{1}{x_i}$ (假定 $x_i > 0, i = 1, 2, \dots, n$) 或

$$x'_i = \frac{1}{k + \max_{1 \leq i \leq n} |x_i| + x_i} \quad (x_i \text{ 可以是负值}, i = 1, 2, \dots, n)$$

其中 k 是常数, 是选定的, 且 $k > 0$.

对适度指标值 x_1, \dots, x_n , 假定最合适的值是 a , 离 a 偏差越大越不好, 因此

$$|a - x_i|$$

就反映了 x_i 不好的程度, 它就相当于一个逆指标, 于是

$$x'_i = \frac{1}{1 + |a - x_i|}, i = 1, 2, \dots, n$$

就是一个正指标.

如果适度指标的偏差在正、负方向的作用是不对称的,那么用 $|a - x_i|$ 来衡量就会有问题,可以用 $(a - x_i)$ 或 $(a - x_i)^3$ 乘以适当的系数来调整,例如用

$$3(a - x_i) + 4|a - x_i|$$

在 $x_i < a$ 时,得 $7|a - x_i|$; 当 $x_i > a$ 时,得 $|a - x_i|$. 反映了偏小时影响坏得多,偏大时影响要小一些. 用这些方法就可以比较合适地处理适度指标的转换,对转换后的指标再进行无量纲化处理.

五、定性指标的量化方法

在综合评价时,会遇到一些定性的指标,定性指标的信息不加利用,又很可惜,直接使用,又有困难,通常总希望能给以量化,使量化后的指标可与其他定量指标一起使用.

定性指标中有两类:名义指标与顺序指标. 名义指标,实际上只是一种分类的表示. 例如性别:男、女;企业所在地:北京、上海、广州、沈阳. 这类指标只能有代码,无法真正量化. 顺序指标,如优、良、中、劣;甲等、乙等;等等,这类指标是可以量化的,所以这一段主要是指顺序指标量化的方法.

如果我们已将全部对象按某一种性质排出了顺序,我们用 $a \succ b$ 表示 a 优于 b , a 排在 b 的后面,全部对象共有 n 个,用 a_1, \dots, a_n 表示,并且不妨假定

$$a_1 \prec a_2 \prec a_3 \prec \dots \prec a_n$$

现在的问题是如何对每个 a_i 赋以一个数值 x_i , x_i 能反映这一前后的顺序. 设想这个顺序是反映了某一个难以测量到的量,例如一个人感觉到的疼痛的程度,从无感觉的痛到有一点痛,到中等疼痛,一直到痛得受不了,比如分成 n 种,记为 $a_1 \prec a_2 \prec a_3 \prec \dots \prec a_n$. 这个疼痛的量是无法测量的,只能比较而排出顺序,设想这个量 x 是客观存在的,可以认为它遵从正态分布 $N(0, 1)$, 于是 a_1, a_2, \dots, a_n 分别反映了 x 在不同范围内人的感觉,设 x_i 是相应

于 a_i 的值,由于 a_i 在全体 n 个对象中占第 i 位,即小于等于它的成员有 i/n ,因此可以想到,若取 y_i 为正态 $N(0,1)$ 的 i/n 分位数,即

$$P(x < y_i) = i/n, i = 1, 2, \dots, n-1$$

那么 y_1, y_2, \dots, y_{n-1} 将 $(-\infty, \infty)$ 分成了 n 段:

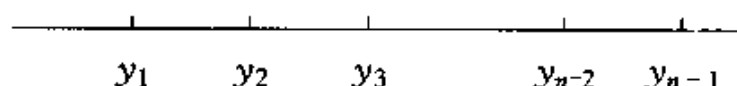


图 2-13

很明显, a_i 表示它相应的 x_i 值应在 (y_{i-1}, y_i) 这个区间之内,在 (y_{i-1}, y_i) 中选哪一个为代表才好呢?自然要考虑概率分布,比较简便可以操作的方法就是选中位数,即 x_i 满足

$$P(x < x_i) = \frac{i-1}{n} + \frac{1}{2} \frac{1}{n} = \frac{i-0.5}{n}$$

$$i = 1, 2, \dots, n$$

其中 x 是 $N(0,1)$ 的分布.于是利用正态概率表,很快就可以查出相应的各个 x_i ,这样就把顺序变量定量化了.

把这个方法稍作推广,就可以处理等级数据的量化.例如一个班上优、良、中、差四级学生的人数如表 2-9.

表 2-9

等级成绩 y_i	差 y_1	中 y_2	良 y_3	优 y_4
人数 f_i	2	10	28	10
占全班百分数	0.04	0.20	0.56	0.20
从差到 y_i 的累计百分数	0.04	0.24	0.80	1.00

差、中、良、优各自对应的量化值 x_i 该如何确定呢?设想班上的成绩是呈正态分布的,因此可以假定未观察到的成绩 $x \sim N(0,1)$,现在

$$\begin{aligned}
y_1:0.04, & \quad P(x < x_1) = \frac{1}{2}(0.04) = 0.02 \\
y_2:0.24, & \quad P(x < x_2) = 0.04 + \frac{1}{2}0.20 = 0.14 \\
y_3:0.80, & \quad P(x < x_3) = 0.24 + \frac{1}{2}0.56 = 0.52 \\
y_4:1.00, & \quad P(x < x_4) = 0.80 + \frac{1}{2}0.20 = 0.90
\end{aligned}$$

查正态分布表,就得到

$$0.02 \text{ 对应的 } x_1 = -2.055$$

$$0.14 \text{ 对应的 } x_2 = -1.080$$

$$0.52 \text{ 对应的 } x_3 = 0.052$$

$$0.90 \text{ 对应的 } x_4 = 1.283$$

这样就把等级的成绩改为“标准分”的成绩.

将上述方法用一般化的公式来描述,若用统计的术语来叙述,使公式更易于理解和表示. 设 u_α 使

$$\int_{-\infty}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \alpha$$

则称 u_α 是标准正态分布 $N(0,1)$ 的 α 分位数,因此上面例中的 x_i , 可以用 α 分位数表示:

$$x_1 = u_{0.02}, \quad x_2 = u_{0.14}$$

$$x_3 = u_{0.52}, \quad x_4 = u_{0.90}$$

一般说来,若有 k 类 a_1, a_2, \dots, a_k , a_1 最差, a_2 比 a_1 好, \dots , 依次递升, a_k 最好, a_i 类各占的比例和累计的比例为

类	a_1	a_2	a_3	$\dots a_k$
各类比例	p_1	p_2	p_3	$\dots p_k$
累计(由 a_1 起始)的比例	p_1	$p_1 + p_2$	$p_1 + p_2 + p_3 \dots 1$	

于是 a_i 对应的 x_i 应有性质:

$$P(x < x_i) = \sum_{j=1}^{i-1} p_j + \frac{1}{2} p_i, i = 1, 2, \dots, k$$

用 $N(0,1)$ 的 α 分位数 u_α 来表示,就得

$$\left\{ \begin{array}{l} x_1 = u_{p_1}/2 \\ x_2 = u_{p_1 + \frac{1}{2}p_2} \\ \vdots \\ x_i = u_{\sum_{j=1}^{i-1} p_j + \frac{1}{2}p_i} \\ \vdots \\ x_k = u_{\sum_{j=1}^{k-1} p_j + \frac{1}{2}p_k} \end{array} \right. \quad (2-23)$$

这就给出了一般化的表达式.

第四节 权的确定

一、权的定义

用若干个指标 x_1, x_2, \dots, x_p 进行综合评价时,其对评价对象的作用,从评价的目标来看,并不是同等重要的.例如要比较不同地区居民生活开支的情况,通常称为生活水平,各种代表性物品的价格 x_i 就是反映这方面内容的,很明显,对一般居民而言,粮食、副食品的价格就很重要,生活开支中这一部分的比例较大,对文化、旅游这一类开支就少,这一类价格的作用就不大,所以选定了代表性物品后,常常对不同物品的价格赋以不同的权,然后来进行综合,权的数值大就认为重要,数值小就认为不重要.

从权的属性来看,可以分为以下几类:

1. 从含信息的多少来考虑.有关的信息多,权数就大,有关的信息少,就将权的数值取得小.
2. 从指标的区分对象能力来考虑.所谓综合评价,就是将评价对象给以区别,并排出先后的次序,所以一个指标从区别这些对象的性质来看,能力强的就应重视,能力弱的,就不应重视.这种权数又称为敏感性权.

当然,从不同的角度考虑,可以有种种办法,这样去罗列就相当繁琐.例如指标数值的质量如何,也就是数据的可信度的程度差异,质量好的权是否应该大一些,质量差的权就小一些.又如从统

计的观点看,相关性大的指标反映的实质上是同一个内容,不相关的指标反映了真正的不同内容,所以在赋权时要考虑这些差别.

总之,指标的权就是体现在综合评价时,对该指标重视的程度.这是定性的描述什么是权,下面来讨论具体的如何给出权的方法.

二、确定权的方法

对实际问题选定被综合的指标后,确定各指标的权的值,常有以下几种方法,这些方法都是利用专家或个人的知识或经验,所以有时称为主观赋权法,但这些专家的判断本身也是从长期实际中来的,不是随意设想的,应该说有客观的基础;另一些方法是从指标的统计性质来考虑,它是由调查所得的数据决定,不需征求专家们的意见,所以有的书上称为客观赋权方法.

在这些方法中,德尔菲(Delphi)方法是经常被采用的,其他方法相对来说,就用得不多,这里都列举在下面,以供比较.

1. 德尔菲法

德尔菲法又称为专家法,其特点在于集中专家的经验与意见,确定各指标的权数,并在不断的反馈和修改中得到比较满意的结果.基本的步骤如下:

(1)选择专家.这是很重要的一步,选得好不好将直接影响到结果的准确性.一般情况下,选本专业领域中既有实际工作经验又有较深理论修养的专家 10~30 人左右,并需征得专家本人的同意.

(2)将待定权数的 p 个指标和有关资料以及统一的确定权数的规则发给选定的各位专家,请他们独立地给出各指标的权数值.

(3)回收结果并计算各指标权数的均值与标准差.

(4)将计算的结果及补充资料返还给各位专家,要求所有的专家在新的基础上重新确定权数.

(5)重复上述第(3)和(4)步,直至各指标权数与其均值的离差

不超过预先给定的标准为止,也就是各专家的意见基本趋于一致,以此时各指标权数的均值作为该指标的权数.

此外,为了使判断更加准确,令评价者了解已确定的权数把握性的大小,还可以运用“带有信任度的德尔菲法”,该方法需在上述第(5)步每位专家给出最后权数值的同时,标出各自所给权数值的信任度,并求出平均信任度.这样,如果某一指标权数的信任度较高时,就可以有较大的把握使用它,反之,只能暂时使用或设法改进.

文献[15]在对上海证券交易所 30 指数的二届成分股进行聚类分析时,指标的选择及赋权采用了德尔菲法.所选专家为国内五家较大规模的证券公司研究人员和操盘手、大学的教学研究人员共 25 人,对 10 个指标赋以权重,经过三轮反馈调整,结果如下:

地区因素	0.05	价格	0.05
换手率	0.15	市盈率	0.10
总股本	0.10	流通股	0.15
每股收益	0.15	每股净资产	0.10
净资产收益率	0.10	振幅	0.05

德尔菲法是调查、征集意见、汇总分析、反馈、再调查……一个反复的过程,专家们是处于互相不知道的隔离状态,每个人的信息是他自己的知识、经验、专长以及调查机构反馈给他的汇总情况的集中体现,这就便于集中智慧.所以不少方法也都或多或少借用这一想法,反复比较、协调,求得较好的结果和比较一致的意见.

2. 相邻指标比较法

这一方法往往与德尔菲法结合使用,在发给专家的征询意见表时,为了便于专家考虑,先将所选的指标按一定考虑排好顺序:

$$x_1, x_2, x_3, \dots, x_k$$

然后让专家填一张表(见表 2-10).

表 2-10

指标(1)	参考指标(2)	相对重要性 $\frac{(1)}{(2)} = g_i$	权 ω'_i	归一化权 ω_i
x_1	x_1	$1(g_1)$	1	ω_1
x_2	x_1	$1.2(g_2)$	1.2	ω_2
x_3	x_2	\vdots	\vdots	\vdots
\vdots	\vdots			
x_k	x_{k-1}	$2.5(g_k)$	ω'_k	ω_k
合计	—	—	$\sum \omega'_i$	1

把 x_2 与 x_1 相比, x_3 与 x_2 相比, \dots , x_k 与 x_{k-1} 相比, 相比的重要性的值列在第 3 列. ω'_i 从第 3 列的值可以算出, 它表示各个指标与 x_1 相比的重要性. 由于 x_i 总是与上一个 x_{i-1} 相比, 因此 g_i 表示 x_i 与 x_{i-1} 相比的重要性

$$\omega'_i = \prod_{j=2}^i g_j = g_i g_{i-1} g_{i-2} \cdots g_2, i = 2, 3, \dots, k$$

而 ω_i 只是将 ω'_i 归一化, 注意 $\omega'_1 = g_1 = 1$,

$$\omega_i = \omega'_i / \sum_{j=1}^k \omega'_j, i = 1, 2, \dots, k$$

相邻指标相比, 一是为了方便, 一次比较两个指标就可以了; 二是为了求 ω'_i 方便, 只需将前面的 g_j 的值乘以 g_i .

当然也可以只给专家指标 x_1, \dots, x_k , 让专家自己去排一个次序逐步比较. 从上面的介绍就可以看出, 这个比较顺序, 并不要求后一个比前一个重要 (即 $g_i \geq 1$), 关键是两者便于比较就好. 让专家自己去排一个比较的次序, 其目的也是为此. 当然, 有时为了便于汇总各个专家的意见, 可以指定一个指标为 x_1 , 它是各个专家比较的基准.

3. 统计方法

从收集到的指标的数据来看, 数据本身提供的信息中是否能

确定合适的权呢？常见的有用方差的倒数为权、变异系数为权和复相关系数的倒数为权这几种。这些方法的起因是统计中关于综合预测有如下一条定理：

定理 2.1 设 y_1, \dots, y_k 分别为真值 θ 的预测值，它们是相互独立的，并且 $Ey_i = \theta, \text{Var}(y_i) = \sigma_i^2$ ，则它们的加权平均预测值 $\sum_{i=1}^k \omega_i y_i$ ，在 $\omega_i = \sigma_i^{-2} / \sum_{j=1}^k \sigma_j^{-2}, i = 1, 2, \dots, k$ 时，相应的方差 $\text{Var}(\sum_{i=1}^k \omega_i y_i)$ 达到最小值。

证明 利用 y_i 的独立性，

$$\text{Var}\left(\sum_{i=1}^k \omega_i y_i\right) = \sum_{i=1}^k \omega_i^2 \sigma_i^2 \triangleq f(\omega_1, \dots, \omega_k)$$

注意到约束条 $\sum_{i=1}^k \omega_i = 1, \omega_i > 0, i = 1, 2, \dots, k$ ，因此，这是一个条件极值问题。用拉氏乘法，令

$$F(\omega_1, \dots, \omega_k) = \sum_{i=1}^k \omega_i^2 \sigma_i^2 - \lambda \left(\sum_{i=1}^k \omega_i - 1 \right)$$

后，就有

$$0 = \frac{\partial F}{\partial \omega_i} = 2\omega_i \sigma_i^2 - \lambda, i = 1, 2, \dots, k$$

因此

$$\omega_i = \lambda \sigma_i^{-2}, i = 1, 2, \dots, k$$

注意到 $\sum_{i=1}^k \omega_i = 1$ ，就求得 $\lambda = \left(\sum_{j=1}^k \sigma_j^{-2} \right)^{-1}$ ，这样

$$\omega_i = \left(\frac{1}{\sigma_i^2} \right) / \sum_{j=1}^k \left(\frac{1}{\sigma_j^2} \right), i = 1, 2, \dots, k$$

这条定理告诉我们，用方差的倒数作为权来综合各种相互独立的预测，效果好（方差达到最小）。然而，综合评价与预测还不一样，所以方差小的这个准则并不适用。于是就派生出变异系数法、复相关系数法。一组数据的变异系数是它的标准差除以均值的绝对值，即对数据

$$z_1, \dots, z_n$$

记 $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$, $s_z = \left(\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2 \right)^{1/2}$, 则

$$v_z = s_z / |\bar{z}| \quad (2-24)$$

就是 z_1, \dots, z_n 的变异系数.

于是对选的指标 x_1, \dots, x_k , 利用被评价对象的数据, 各个指标都有各自的变异系数. 为了方便, 用 v_i 表示 x_i 的变异系数, $i = 1, 2, \dots, k$, 此时, x_i 相应的权就是 $v_i / \sum_{j=1}^k v_j$. 这种加权的方法是为了突出各指标的相对变化幅度, 从评价的目的来看, 就是区别被评价的对象, v_i 的值大表示 x_i 在不同的对象身上变化大, 区别对象能力强, 所以应给以重视.

另一种是考虑复相关系数, 每一个被选的指标 x_i , 用其余的指标对它的相关程度——复相关系数 $\rho_{x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_k}$ 来考虑时, 复相关系数简记为 ρ_i , 它反映了非 x_i 的那些指标能替代 x_i 的能力. 当 $\rho_i = 1$ 时, x_i 可以去掉, 因为用非 x_i 的值就可定出 x_i 的值; 当 ρ_i 很小时, 非 x_i 的值并不能代替它, 所以用 $|\rho_i|^{-1}$ 作为权是合适的, 它就是复相关系数倒数的绝对值.

从各种角度去考虑评价问题, 就会引出各种各样的权, 那么对各种各样的权, 又如何能综合成一个合适的权呢? 这就是权的综合问题. 这一问题的一个简单的处理方法, 就是把各种权相乘, 产生一个新的权. 例如对指标 x_1, \dots, x_k 由两种考虑导出 $\omega_i^{(1)}$ 与 $\omega_i^{(2)}$ 两种权系数, 那么令

$$\omega_i = \frac{\omega_i^{(1)} \omega_i^{(2)}}{\sum_{j=1}^k \omega_j^{(1)} \omega_j^{(2)}}, i = 1, 2, \dots, k \quad (2-25)$$

就是新的合成的权.

下面用一个例子来说明这一方法的应用.

例 2.6 (续例 2.3) 对评价高教发展水平的七个评价指标

进行客观赋权.

计算过程略,结果见表 2-11. $\omega_i^{(1)}$ 和 $\omega_i^{(2)}$ 分别表示由变异系数法和复相关系数法所确定的权数, ω_i 表示由(2-25)合成的权数.

表 2-11

	x_1	x_2	x_3	x_7	x_8	x_9	x_{10}
\bar{x}_i	1.247	66.467	126.5	30.12	2381.267	0.553	7645.267
S_i	1.079	62.353	177.2	5.904	591.149	0.397	2488.248
v_i	1.156	1.066	0.714	5.102	4.028	1.393	3.073
$\omega_i^{(1)}$	0.070	0.064	0.043	0.309	0.244	0.084	0.186
ρ_i	0.995	0.989	0.993	0.891	0.925	0.901	0.741
$\omega_i^{(2)}$	0.131	0.132	0.131	0.146	0.141	0.144	0.176
ω_i	0.062	0.058	0.038	0.306	0.232	0.083	0.222

可以看出 $\omega_i^{(1)}$ 相差较大, $\omega_i^{(2)}$ 则比较均衡, 由于用乘法求组合权, 故 ω_i 相差较大.

第五节 常见的综合方法

前几节讨论了综合评价的一些准备工作, 或者说是数据的预处理, 这一节介绍一些常见的综合方法, 对这些方法的优劣作一些简短的讨论.

一、各种平均值

最常见的综合评价方法都是与平均值有关的, 加权算术平均、加权几何平均、算术平均与几何平均联合使用等这一类方法. 我们分别来作一些讨论.

1. 加权算术平均

这是最常见的方法, 对每个指标 x_i 给定权 ω_i , 然后用加权的

算术平均值来综合,也即评价值 y 的表达式为(考察的指标为 x_1, \dots, x_k)

$$y = \sum_{i=1}^k \omega_i x_i, \omega_i \geqslant 0, i = 1, 2, \dots, k, \sum_{i=1}^k \omega_i = 1 \quad (2-26)$$

当 $\omega_i = \frac{1}{k}, i = 1, 2, \dots, k$ 时,它就是通常的算术平均值,若对 x_1, \dots, x_k 中的最大值、最小值都赋以数值为 0 的权,这就是“去掉一个最高分,去掉一个最低分”的办法,对剩下的权都相等,实质上还是一种加权的算术平均.所以(2-26)概括了很多方法.

关于加权算术平均的效果,下面这条定理给我们很多启示.

定理 2.2 设指标 x_1, \dots, x_k 的相关系数矩阵 $R = (r_{ij})$, 显然有 $r_{ii} = 1, r_{ij} = \text{Cov}(x_i, x_j) / (\text{Var}(x_i) \text{Var}(x_j))^{\frac{1}{2}}, i, j = 1, 2, \dots, k$. 则任给两个 x_1, \dots, x_k 的加权算术平均

$$y_1 = \sum_{i=1}^k a_i x_i, \quad y_2 = \sum_{i=1}^k b_i x_i$$

$$a_i \geqslant 0, b_i \geqslant 0, i = 1, 2, \dots, k$$

且有

$$\sum_{i=1}^k a_i = \sum_{i=1}^k b_i = 1$$

总有 y_1, y_2 的相关系数 $\rho \geqslant \min_{1 \leqslant i, j \leqslant k} r_{ij}$.

证明 因为讨论的是相关系数,注意到相关系数对于线性变换是不变的,这样就可以假定 x_i 的方差都相等,并且都是 1,此时协方差就是相关系数,即 $r_{ij} = \text{Cov}(x_i, x_j)$ 对 $i, j = 1, 2, \dots, k$ 都成立.于是

$$\begin{aligned} \text{Var}(y_1) &= \sum_{i,j=1}^k a_i a_j \text{Cov}(x_i, x_j) \\ &= \sum_{i,j=1}^k r_{ij} a_i a_j \leqslant \sum_{i,j=1}^k a_i a_j = 1 \end{aligned}$$

同理 $\text{Var}(y_2) \leqslant 1$, 因此

$$\rho = \frac{\text{Cov}(y_1, y_2)}{\sqrt{\text{Var}(y_1) \text{Var}(y_2)}} \geqslant \text{Cov}(y_1, y_2)$$

$$\begin{aligned}
&= \sum_{i,j=1}^k r_{ij} a_i b_j \geq \left(\min_{1 \leq i,j \leq k} r_{ij} \right) \sum_{i,j=1}^k a_i b_j \\
&= \min_{1 \leq i,j \leq k} r_{ij}
\end{aligned}$$

这条定理告诉我们,如果指标 x_1, \dots, x_k 彼此之间相关系数很大, $r_{ij} \geq 0$, 那么任何两个加权算术平均之间的相关性也非常大, 所以加权就没有意义, 它与普通算术平均值是差不多的, 因为它们之间相关性很强. 平均值的上述性质很早就被发现了, 但人们还往往认为加权平均比简单的平均好, 这是习惯势力的影响. 在综合评价的问题中, 人们希望用正向指标来评价, 所以选出的指标往往相关性很大, 而且是正相关, 因此考虑加权的意义就不很大.

2. 加权几何平均.

如果所选的评价指标 x_i 是比例型的, 如受教育人数占全民人数的比例、贫困人口的比例等, 因为这个比值随单位的选择不同是不会改变的, 它是无量纲的, 因此相加, 或加权相加再平均也可以. 另一种的比例实际上是比值型的, 它有量纲, 如劳动生产率, 用人均产值算, 量纲是元/人·年, 也可以为万元/人·年, 随量纲的不同, 这个数值就很不同, 把这一类数据作算术平均、加权算术平均综合时, 量纲的改变对综合值的影响是明显的. 在这种情况下, 用几何平均或加权几何平均就可消除这种影响, 其理由从下面的论述中就可以看出.

设 x_1, \dots, x_k 是 k 个被综合的指标, 各自量纲的变化可以将 $x_i \rightarrow c_i x_i$, c_i 是量纲变化引出的比值常数, 用算术平均来评价时, 甲企业和乙企业的 x_i 值分别用 u_i 和 v_i 表示, 于是数值

$$\frac{1}{k} \sum_{i=1}^k u_i, \frac{1}{k} \sum_{i=1}^k v_i$$

的大小就评出了甲、乙的优劣, 但是

$$\frac{1}{k} \sum_{i=1}^k u_i > \frac{1}{k} \sum_{i=1}^k v_i$$

时, 如果换一下量纲来计算, 就是比较

$$\frac{1}{k} \sum_{i=1}^k c_i u_i, \frac{1}{k} \sum_{i=1}^k c_i v_i$$

能否还与上述不等式保持一致,这就难以保证.但是对于几何平均,这是一定能保证的,因为量纲变换引出的 $c_i > 0, i = 1, 2, \dots, k$, 所以

$$\left(\prod_{i=1}^k c_i u_i \right)^{\frac{1}{k}} \geq \left(\prod_{i=1}^k c_i v_i \right)^{\frac{1}{k}}$$

与

$$\left(\prod_{i=1}^k u_i \right)^{\frac{1}{k}} \geq \left(\prod_{i=1}^k v_i \right)^{\frac{1}{k}}$$

是一定保持一致的,所以用几何平均来综合比算术平均好.对于加权的算术平均与加权的几何平均也是如此.对于指标 $x_i > 0, i = 1, 2, \dots, k$, 一般几何平均的加权形式是

$$\omega_i \geq 0, i = 1, 2, \dots, k \text{ 且 } \sum_{i=1}^k \omega_i = 1$$

加权几何平均值 g 是

$$g = \prod_{i=1}^k x_i^{\omega_i} \quad (2-27)$$

g 称为 x_1, \dots, x_k 的几何加权平均, ω_i 均相等为 $\frac{1}{k}$ 时,就是常见的几何平均.

(2-27)可以有各种变化,例如某一指标 x_i 的值越大越不好,如文盲率、婴儿死亡率等,于是可以将 x_i^{-1} 作为综合的对象,于是(2-27)就成为

$$g^* = \prod_{i=1}^{k_1} x_i^{\omega_i} / \prod_{i=k_1+1}^k x_i^{\omega_i} \quad (2-28)$$

这表示 x_1, \dots, x_{k_1} 均为正向指标, x_{k_1+1}, \dots, x_k 是逆向指标.

例如美国卫生组织(American Social Health Association)评价经济发展基本需要的程度时,就是一些指标的几何平均,现已被大家接受,称为 ASHA 指标,它的定义是

$$\text{ASHA 指标} = \frac{\text{就业率} \times \text{识字率} \times \frac{\text{平均期望寿命}}{70} \times \text{人均 GNP 增长率}}{\text{人口出生率} \times \text{婴儿死亡率}}$$

很明显,这个综合指标的意义是什么.

几何平均在一定意义下就是各种性质都具有的程度,只有各种性质均衡增长时,几何平均值才会增大,例如 ASHA 指标中,单是识字率增长,整个 ASHA 指标的值不会增长多少,但若分子的四项都增长时,它就有明显的增长.所以用它来反映变化、发展的状况,的确是具有“综合性”的.与算术平均数相比,就明显看出差别,算术平均值受个别极端值(太大,太小)的影响非常敏感,几何平均就不是这样,因此在评奖中是否可以采用几何均值而不是算术均值呢?这是应该思考的.

从上面的讨论就可以想到,对某些问题,最适宜的方法也许是一部分指标用加权算术平均,一部分指标用加权几何平均,然后再综合,这样就可以派生出许多方法,为我们综合评价提供了较宽的选择范围.例如西方国家的经济业绩指数 EPI 的定义是

$$\text{EPI} = \frac{\text{实际 GNP 增长率}}{\text{通货膨胀率} + \text{失业率}}$$

这是一个典型的把加法、乘法、正向指标、反向指标同时使用的实例,它对我们考虑一些综合指标时,会有帮助.

下面我们通过继续讨论高教发展水平的例子来看这些方法的使用.

例 2.7 (续例 2.3,例 2.6) 对我国各地区高教发展水平进行综合评价.

按例 2.3 的结果选取七项指标进行综合评价:

x_1 :每 100 万人口学校数;

x_2 :每 10 万人口毕业生数;

x_3 :每 10 万人口教职工数;

x_4 :高级职称占专职教师的比例;

x_5 :平均每所高校在校生数;

表 2-12

样本序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	4.368505	3.905746	4.540068	2.411978	0.395388	4.159534	2.405601			
2	1.986190	2.686871	2.096501	0.829927	1.134627	0.874563	2.017376			
3	1.022140	1.451958	0.950903	1.402447	1.099103	0.773720	0.699180			
4	0.095169	0.233083	0.132619	0.055840	0.537484	1.681310	0.094739			
5	0.234215	0.345348	0.098758	0.707970	0.721871	-0.033026	0.035259			
6	0.391800	0.313272	0.149549	0.577544	-0.281260	0.521612	-0.066419			
7	-0.071686	-0.055598	-0.053612	0.863804	0.248217	0.067817	0.371640			
8	-0.182922	0.008554	-0.064898	0.469138	0.767545	0.269504	-0.154031			
9	-0.275619	-0.039560	-0.138262	0.240470	1.060196	-0.411189	0.056559			
10	-0.516632	-0.440506	-0.369639	0.741847	1.026363	-0.461610	1.490901			
11	-0.637138	-0.424468	-0.369639	0.423404	1.298714	-0.007815	0.019183			
12	-0.627868	-0.135787	-0.392212	0.479302	1.388370	-0.688508	-0.337694			
13	-0.498092	-0.392392	-0.341422	-0.337132	0.466436	0.445980	-0.145993			
14	-0.470283	-0.392392	-0.369639	0.497934	0.400463	-0.209502	-0.469514			
15	-0.359047	-0.392392	-0.341422	-0.030546	-0.030900	-0.764140	0.023604			
16	0.039551	0.312203	-0.070542	-0.709777	-0.543462	-0.461610	-0.774146			
17	-0.192192	-0.215976	-0.358352	-0.188073	-0.477488	-0.663297	-0.216726			
18	-0.368316	-0.215976	-0.284989	-0.760593	0.293891	-0.310345	-0.830008			
19	-0.405395	-0.376354	-0.369639	-0.050872	-0.115481	-0.612875	-0.780174			
20	-0.609329	-0.504657	-0.454289	0.458976	0.180552	-0.562454	-0.810718			
21	-0.544441	-0.488619	-0.465576	-0.265990	-0.688941	-0.184291	-0.587053			
22	-0.442474	-0.376354	-0.335779	-0.221950	0.226226	-0.537243	-1.430833			
23	-0.507362	-0.536733	-0.448646	-0.470945	-0.063041	-0.688508	0.113627			
24	-0.377586	-0.376354	-0.347065	-0.418436	-0.590827	-0.587664	-0.829607			
25	0.410339	-0.648998	-0.290632	-3.052365	-2.657989	1.126672	2.633874			
26	-0.646408	0.552771	-0.465576	-0.289704	-0.068116	-0.638086	-0.776155			
27	-0.600059	-0.616922	-0.493792	0.306529	-0.397982	-0.789351	-1.007242			
28	0.132248	-0.296165	-0.279345	-1.256889	-1.490770	-0.335556	-0.911592			
29	-0.562980	-0.697111	-0.505079	-0.338825	-1.543210	-0.537243	-0.896320			
30	0.215675	-0.456544	-0.358352	-2.075016	-2.295981	-0.436399	-0.111431			

x_9 :国家财政预算内普通高教经费占国民生产总值比重;

x_{10} :生均教育经费.

考虑到样本个数较多,我们选用标准化方法(2-13)进行数据的无量纲化,无量纲化后结果见表2-12.我们分别以例2.6确定的三种权数进行合成,考虑到指标已经过筛选,选用加法合成(2-26)可得综合评价结果(见表2-13).

表 2-13 综合评价结果表

样本 序号	(1)		(2)		(3)	
	评价值	位次	评价值	位次	评价值	位次
1	2.391661	1	3.110864	1	2.375057	1
2	1.383921	2	1.649333	2	1.394694	2
3	1.101941	3	1.043078	3	1.086193	3
4	0.334897	8	0.403584	4	0.326142	8
5	0.441071	5	0.295268	5	0.427395	5
6	0.195359	12	0.220430	7	0.187491	12
7	0.390943	6	0.212332	8	0.399839	6
8	0.310855	9	0.156940	9	0.296532	9
9	0.280631	11	0.075380	10	0.273831	11
10	0.636770	4	0.273904	6	0.685811	4
11	0.362164	7	0.059040	11	0.356717	7
12	0.295861	10	-0.044924	12	0.275734	10
13	-0.054809	14	-0.106389	13	-0.056707	14
14	0.072187	13	-0.145318	14	0.057989	13
15	-0.142107	16	-0.258296	16	-0.132309	16
16	-0.554673	26	-0.427791	23	-0.570992	26
17	-0.313407	20	-0.328931	18	-0.309310	19
18	-0.395519	23	-0.374226	21	-0.419761	22
19	-0.309065	19	-0.400174	22	-0.326779	20
20	-0.107198	15	-0.336872	19	-0.128076	15
21	-0.246059	17	-0.255674	15	-0.206354	17
22	-0.394288	22	-0.480892	24	-0.438507	23
23	-0.287053	18	-0.352795	20	-0.269782	18
24	-0.542475	25	-0.519109	26	-0.555857	25
25	-1.030913	29	-0.263662	17	-0.897119	28
26	-0.404976	24	-0.498608	25	-0.418719	24
27	-0.359164	21	-0.526487	27	-0.378795	21
28	-0.970783	28	-0.659921	28	-0.979746	29
29	-0.798600	27	-0.733008	29	-0.799560	27
30	-1.287173	30	-0.787076	30	-1.255055	30

注:(1)为用变异系数法确定的权数的合成结果.

(2)为用复相关系数法确定的权数的合成结果.

(3)为用(1)、(2)两种权数组合所得权数的合成结果.

从表 2-13 可以看出:第一,利用变异系数法和复相关系数法所确定权数的合成结果在排序有一定的差异,这体现了不同权数的作用.从表 2-11 可以看出,这两种权数有着较大的差异,前者在 x_7, x_8, x_{10} 上权数较大,而后者除在 x_{10} 上稍大以外其余基本均衡,因而就导致了不同的评价结果.第二,用变异系数法确定权数的合成结果体现了较好的区分度,而用复相关系数法确定权数的合成结果则体现了各指标信息的合理利用.可以看出,评价值的区分度主要体现在第 4~30 号样本上,前者评价值分布在 $-1.3 \sim 0.7$ 之间,而后者评价值分布在 $-0.8 \sim 0.5$ 之间,相对而言,前者评价值具有较好的区分度.第三,从表 2-11 可以看出组合权数与变异系数法所确定权数没有太大的差异,因而两种权数的评价排序结果几乎完全相同.第四,我们认为,就这个评价问题而言,我们的目的并不在于择优,而是在于客观地反映情况,因而用复相关系数法确定权数的结果较为客观,如果再能将专家的主观赋权结果结合起来,那将会取得更好的效果.第五,通过本例,我们可以体会到权数的作用.

第三章 综合评价的主成分方法与因子分析法

综合评价从数学的眼光来看,就是建立一种从高维空间到低维空间的映射,这种映射能保持样本在高维空间的某种“结构”,其中最明显的是与“序”有关的结构,因为综合评价的目的往往与排序是分不开的.这一章的重点是讨论因子分析法、主成分分析法在综合评价中的应用与特点.

第一节 主成分分析

首先,我们看一个实例,用主成分分析解决问题的过程,然后讨论相应的数学问题是什么,在统计中分析的方法是怎样的.这是一个企业经济效益的综合评价问题.

为了评价企业的经济效益,先要设定评价的指标,也就是评价企业经济效益的指标体系,这一专题不是我们这本书讨论的内容,我们这里只讨论指标体系选定后如何评价.

评价指标体系实际上是发展的,有的选四个指标:

总产值/消耗,净产值/工资,盈利/工资,销售收入/成本.也有用下面的八个指标:

x_1 : 固定资产利税率;

x_2 : 资金利税率;

x_3 : 销售收入利税率;

x_4 : 资金利润率;

x_5 : 固定资产产值率;

x_6 : 流动资金周转天数;

x_7 : 万元产值能耗;

x_8 : 全员劳动生产率.

考虑到可持续发展,有人建议将环保有关的指标也应列入.下面我们举的例子是对 15 家水泥生产企业按 1984 年的数据进行的 평가,其原始数据见表 3-1,用的是上面所说的八个指标: $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$.

表 3-1 各厂的指标值

<div> <div>指标</div> <div>指标值</div> </div> <div> 厂家 与代号 </div>	x_1 %	x_2 %	x_3 %	x_4 %	x_5 %	x_6 (天)	x_7 (吨)	x_8 (万元/人·年)
1 琉璃河	16.68	27.75	31.84	18.40	53.25	55	28.83	1.75
2 邯郸	19.70	27.56	32.94	19.20	59.82	55	32.92	2.87
3 大同	15.20	23.40	32.98	16.24	46.78	65	41.69	1.53
4 哈尔滨	7.25	8.97	21.30	4.76	34.39	62	39.28	1.63
5 华新	29.45	56.49	40.74	43.68	75.32	69	26.68	2.14
6 湘乡	32.93	42.78	49.98	33.87	66.46	50	32.87	2.60
7 柳州	25.39	37.85	36.76	27.56	68.18	63	35.79	2.43
8 峨眉	15.05	19.49	27.21	14.21	56.13	76	35.76	1.75
9 耀县	19.82	28.78	33.41	20.17	59.25	71	39.13	1.83
10 永登	21.13	35.20	39.16	26.52	52.47	62	35.08	1.73
11 工源	16.75	28.72	29.62	19.23	55.76	58	30.08	1.52
12 抚顺	15.83	28.03	26.40	17.43	61.19	61	32.75	1.60
13 大连	16.53	29.73	32.49	20.63	50.41	69	37.57	1.31
14 江南	22.24	54.59	31.05	37.00	67.95	63	32.33	1.57
15 江油	12.92	20.82	25.12	12.54	51.07	66	39.18	1.83

注意到这些指标的特性,有些是越大越好的正向指标,如 $x_1, x_2, x_3, x_4, x_5, x_8$;但是 x_6 与 x_7 是反向指标,数值越大就越不好,这时就考虑它们各自的倒数来参与综合评价(其原因在第二章已叙述过).

现在来进行统计分析.从表 3-1 上的数据可以看出,共有 15 个样本,每一个样本有八个指标.对各个指标先作数据的标准化处理,求出每个指标的均值,分别用 $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_8$ 表示,再求各自的

标准差,然后将原始数据标准化,即各个指标的样本值 x_{ai} (第 α 个企业的第 i 个指标值)减去 \bar{x}_i 后再除以第 i 个指标的标准差,这些在第二章中均已介绍.于是 x_{ai} 就转换成 \tilde{x}_{ai} , \tilde{x}_{ai} 的特点是各个指标的均值为 0,标准差是 1,因此这 8 个指标的协方差阵就是相关矩阵,经计算,得到相关矩阵是

$$R = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ 1 & 0.849 & 0.925 & 0.902 & 0.850 & 0.325 & 0.491 & 0.586 \\ & 1 & 0.693 & 0.988 & 0.860 & 0.117 & 0.610 & 0.525 \\ & & 1 & 0.776 & 0.615 & 0.367 & 0.349 & 0.522 \\ & & & 1 & 0.856 & 0.129 & 0.607 & 0.317 \\ & & & & 1 & 0.099 & 0.620 & 0.976 \\ & & & & & 1 & 0.284 & 0.504 \\ & & & & & & 1 & 0.194 \\ & & & & & & & 1 \end{bmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{matrix}$$

由于 R 是对称矩阵,我们只需写出它的上三角部分就可以了.进一步计算 R 矩阵的特征值及特征向量,这类计算一般的软件中已有专门的模块,调用就可以算出.全部特征根之和是 R 中对角元素之和,就是数值 8.计算显示依大小次序排列后,前三个特征根之和就占 8 个指标的 89.865%,因此选用前三个就可以综合这 8 个指标的信息.这三个因子分别用 y_1, y_2, y_3 表示,就得到

$$\begin{aligned} y_1 &= 0.43128x_1 + 0.40519x_2 + 0.31894x_3 + 0.41899x_4 \\ &\quad + 0.39940x_5 + 0.15494x_6 + 0.29191x_7 + 0.24879x_8 \\ y_2 &= 0.05226x_1 - 0.28635x_2 + 0.15310x_3 - 0.24014x_4 \\ &\quad - 0.17611x_5 + 0.68000x_6 - 0.13622x_7 + 0.56449x_8 \\ y_3 &= -0.0000x_1 + 0.02041x_2 - 0.32459x_3 - 0.04617x_4 \\ &\quad - 0.01325x_5 + 0.42736x_6 + 0.77781x_7 - 0.23352x_8 \end{aligned}$$

将各厂相应的 y_i 值算出,这三个 y_i 的值反映了工厂的三个方面的状况,从 y_i 表达式中 x_i 的系数就可以看出 y_i 主要是体现哪些 x_i 的内容,注意到上述表达式中 x_i 是指标准化后的变量,因此系

数的大小比较是有意义的,这样去看,就得到:

y_1 : 反映盈利能力 x_1, x_2, x_3, x_4, x_5 的这些指标;

y_2 : 资金、人力利用方面的 x_6, x_8 ;

y_3 : 产值、能耗方面的 x_7 .

将有关数据列在表 3-2,相应的名次也一并列出,以供比较.

表 3-2 名次表

水泥 厂名	盈利能力方面		资金、人力利用方面		产值能耗方面		综合评价	
	Y_1	名次	Y_2	名次	Y_3	名次	E 值	名次
琉璃河	-0.089	7	0.635	4	1.929	1	0.239	6
邯郸	0.75	5	2.155	2	0.092	7	0.832	5
大同	-1.857	13	-0.094	6	-0.994	14	-1.288	13
哈尔滨	-4.207	15	0.922	3	0.242	6	-2.482	15
华新	4.184	1	-1.744	15	0.262	5	2.374	2
湘乡	3.826	2	2.189	1	-0.537	12	2.7365	1
柳州	1.744	4	0.483	5	-0.981	13	1.089	3
峨眉	-1.736	12	-0.855	12	-0.450	10	-1.287	12
耀县	-0.434	8	-0.534	11	-1.175	15	-0.479	10
永登	0.534	6	-0.072	7	-0.345	9	0.2113	7
工源	-0.470	9	-0.075	8	1.275	2	-0.185	8
抚顺	-0.715	10	-0.422	10	0.974	3	-0.421	9
大连	-1.217	11	-1.163	13	-0.514	11	-1.016	11
江南	1.889	3	-1.488	14	0.489	4	0.992	4
江油	-2.173	14	0.015	6	-0.270	8	-1.394	14

从 y_1, y_2, y_3 综合起来看,给出的是表 3-2 综合评价的 E 值, E 值是将 y_1, y_2, y_3 加权得来的,按它们各自的贡献率作为权重,具体方法以及贡献率的定义,在以下各节会作介绍.这样我们就解决了一个实际的评价问题.

将这个过程与第二章常规方法相比,明显要复杂些,问题在于这种复杂的结果是否有更合理的结论.

在第二节我们详细说明这一方法的统计依据,还指出它的优缺点与进一步发展的可能性;然后举例说明这种方法的应用,这就

是第三节的例子,第四节介绍进一步的因子分析方法,第五节介绍因子分析的算法和实例,实例说明了因子分析不仅能给出评价,而且还能指出影响评价名次的原因,为进一步改善指出努力方向,这是因子分析很具特色的优点.

第二节 主成分分析方法的统计依据

这一节从统计分析的角度来说明主成分分析的方法,然后将它用于综合评价后会有些什么不合适的疑点,再将有关的各种看法给以介绍.

我们考虑的统计问题可以陈述如下:

设有 p 个指标 x_1, x_2, \dots, x_p , 这 p 个指标反映了客观对象的各个特性, 因此每个对象观察到的 p 个指标值就是一个样本值, 它是一个 p 维的向量. 如果观察了 n 个对象, 就有 n 个 p 维向量, 共有 np 个数据, 用矩阵 X 表示就有

$$X_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

每一行就是一个样本的观察值. 统计问题是: 已知数据矩阵 X , 能否找到反映 p 个指标 x_1, x_2, \dots, x_p 的线性函数 $\sum_{i=1}^p a_i x_i$, 它能最好地反映 x_1, x_2, \dots, x_p 的变化状况. 也就是说, 把 p 个变量在 n 个样本上的差异, 能否用它们的一个线性函数的差异来综合表示, 如果行, 这个线性函数就是一个代表性很好的指标, 它就是这 p 个变量的主要成分 (principle component), 找出这个主要成分的方法就称为主成分分析方法.

将这个问题先用概率的语言描述成一个数学问题, 解这个数学问题后, 就得到了统计分析的方法. 把 p 个指标 x_1, \dots, x_p 看成随机变量, 它们的期望值和协方差矩阵是

$$Ex = E \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} Ex_1 \\ Ex_2 \\ \vdots \\ Ex_p \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

$$V = (v_{ij}) = (\text{Cov}(x_i, x_j)) = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{pmatrix}$$

v_{ii} 就是第 i 个变量的方差, 因此这 p 个变量总的变化状况可以用

$\sum_{i=1}^p v_{ii}$ 来反映. 现在考虑它们的线性函数 $\sum_{i=1}^p a_i x_i$, 记它为 y , 于是 y

的方差 $\text{Var}(y) = \text{Var}\left(\sum_{i=1}^p a_i x_i\right) = \sum_{i=1}^p \sum_{j=1}^p a_i a_j v_{ij} = a^T V a$, $a =$

$\begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}$. 要寻找最能反映这些 x_i 变化的, 就要求 $\text{Var}(y)$ 尽可能大,

从 $\text{Var}(y) = a^T V a$ 可以看出, 对向量 a 的长度要作一些限制, 否

则 $\text{Var}(y)$ 可以无限增大而没有意义, 自然限制 $a^T a = \sum_{i=1}^p a_i^2 = 1$.

因此数学问题就是: 已知协方差矩阵 V , 求满足约束条件 $a^T a = 1$ 的 a , 使 $a^T V a$ 达到最大值.

现在来解这个问题, 这是一个条件极值问题, 用拉氏乘子法就可以处理. 令

$$f(a) = a^T V a - 2\lambda(a^T a - 1)$$

于是

$$\frac{\partial f}{\partial a} = 2Va - 2\lambda a$$

因此由 $\frac{\partial f}{\partial a} = 0$ 导出方程

$$Va = \lambda a, a^T a = 1$$

这表示 a 是矩阵 V 的特征向量, 利用上式, 就知道

$$\text{Var}(a^T x) = a^T V a = \lambda a^T a = \lambda$$

也即特征根 λ 就是 $a^T x$ 的方差, 因此只要求出最大特征根所对应的特征向量 a , 寻找主成分的问题就解决了.

这就告诉我们, 主成分的求法是:

1. 先求出样本的协方差矩阵 V ;

2. 求 V 的最大特征根 λ 和相应的特征向量 $a = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}$, 于是

$a_1 x_1 + \cdots + a_p x_p$ 就是所要的主成分分量.

这样求得的主成分, 它能反映原来 p 个变量变化状况的多少呢? 很自然用 $\text{Var}(a^T x) = \lambda$ 去和原来的 $\sum_{i=1}^p v_{ii}$ 相比, 即看

$$\lambda / \sum_{i=1}^p v_{ii}$$

的值是多少, 这个比值称为主成分 $a^T x$ 的贡献率, 贡献率越大就越表明 $a^T x$ “综合”的能力强.

有人会问, 方程

$$V a = \lambda a$$

是特征方程, 它有许多解, 除了最大的特征根之外, 其余的特征根和特征向量是否也有统计意义呢? 这个问题的答案是肯定的. 将协方差阵 V 的特征根按大小排列, 共有 p 个, 记为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

因为 V 非负定, 所以特征根全是非负的, 它们相应的特征向量用 r_1, r_2, \cdots, r_p 表示, 于是有

$$V r_i = \lambda_i r_i, i = 1, 2, \cdots, p$$

且

$$r_i^T r_i = 1, r_i^T r_j = 0, (i \neq j), i, j = 1, \cdots, p$$

注意到

$$\text{Var}(r_i^T x) = r_i^T V r_i = \lambda_i r_i^T r_i = \lambda_i$$

$$\text{Cov}(r_i^T x, r_j^T x) = r_i^T V r_j = \lambda_j r_i^T r_j = 0 (i \neq j)$$

这表明令 $y_i = r_i^T x = \sum_{j=1}^p r_{ij} x_j$, $r_i = \begin{bmatrix} r_{i1} \\ \vdots \\ r_{ip} \end{bmatrix}$, $i = 1, \dots, p$ 后, y_1, \dots, y_p

是彼此不相关的,而且 y_i 的方差就是 λ_i . 从根与系数的关系可以得到

$$\sum_{i=1}^p \lambda_i = \sum_{i=1}^p v_{ii}$$

这表示 y_1, \dots, y_p 的全部方差正好是 x_1, \dots, x_p 的全部方差, y_1, \dots, y_p 全面反映了 x_1, \dots, x_p 的变化状况,所以 y_1, \dots, y_p 称为 x_1, \dots, x_p 的全部主成分分量. 将 x_1, \dots, x_p 转换成 y_1, \dots, y_p 的优点是 $y_1, y_2, y_3, \dots, y_n$ 等前几个主成分会集中 x_1, \dots, x_p 全部方差的 90% 或 85% 以上,所以讨论少数几个主成分就可以了,而且它们彼此不相关,反映了完全不同的方面. 这些正是主成分分析的优点.

综合评价就是选用第一个主成分分量作为评价指标,它是原来指标 x_1, \dots, x_p 的线性函数,如果最大特征根 λ_1 在 $\sum_{i=1}^p v_{ii}$ (或 $\sum_{i=1}^p \lambda_i$) 中占的百分比超过 80%, 这个综合指标就能较好地反映选用的评价指标 x_1, \dots, x_p ; 如果 λ_1 在 $\sum_{i=1}^p v_{ii}$ 中占的百分比不超过 80%, 这个综合指标就不是很理想; 如果不到 50%, 那就很不好, 实际上这个主成分并不能反映这些 x_1, \dots, x_p 的变化状况, 这时就要考虑第二个、第三个主成分分量, 把它们放在一起分析、比较, 再作出评价. 所以用主成分分析, 并不总是成功的, 有时就会遇到上面这一类问题.

用主成分分析时, 把第一主成分作为综合评价的代表值, 尽管它的方差 λ_1 占全部的 85% 以上, 很有代表性, 但是

$$y = \sum_{i=1}^p a_i x_i = a_1 x_1 + a_2 x_2 + \dots + a_p x_p$$

这一表达式中,系数 a_i 的值或符号可能与实际意义会不相符合,这时就可以考虑对主成分分量的坐标系作一些旋转,这就引向了因子分析的内容,这些内容在本章以后的有关章节中会详细说明.

有一些误解,这里也可以附带说明一下.当 x_1, x_2, \dots, x_p 都是正向指标时,从 y 的表达式

$$y = a_1x_1 + a_2x_2 + \dots + a_px_p$$

可以认为 a_i 相当于对 x_i 附以的权重,因此有的人认为 $a_i \geq 0$,

$\sum_{i=1}^p a_i = 1$, 应该自然成立,如果不成立,就感到有问题,似乎方法有毛病.这是一种误解.因为原则上凡是与 a_1, \dots, a_p 成比例的向量,都是最大特征根 λ_1 相应的特征向量,因此将 a_i 换成 $a_i / \sum_{i=1}^p a_i$

后,其效果是一样的,用 a_i 作为系数来综合评价排序,和用 $\tilde{a}_i =$

$a_i / \sum_{i=1}^p a_i$ 作为系数来综合评价序不会有什么不同.所以这是一种

误解,反过来说,主成分分析可以从客观的数据来定出合理区分对象的权系数,可以清楚显示各个指标 x_i 在综合评价的作用.

另一个问题是第一主成分的 a_i 可能会出现负值.若 a_i 相应的 x_i 是反向指标,很明显,令 $-x_i$ 作为指标参评,代替 x_i , a_i 就是正值.问题在于 x_i 有时是正向指标,为什么 a_i 会小于 0? 这时,应考察 x_1, \dots, x_p 的相关矩阵,与 x_i 相关性很大的指标 $x_{j1}, x_{j2}, \dots, x_{jp}$ 等,是否相应的 a_i 是正的而且都相当大,这表示这一类指标在参与综合评价时,过于重复产生影响,所以这时应从 x_i 中删去一些指标,重新考虑综合才好.有时 a_i 中会有很多是负的,这时注意将所有 a_i 的符号同时改向,就会只有少数负号,用后者来评价就可

可以了,其依据和上面谈到的加权的看法相同,当 $a = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}$ 是特

征向量时, $-a = \begin{pmatrix} -a_1 \\ \vdots \\ -a_p \end{pmatrix}$ 也是特征向量.

有了这些准备,我们就可以从样本资料矩阵 $X = (x_{ai}), a = 1, 2, \dots, n, i = 1, 2, \dots, p$ 出发,求出指标 x_1, \dots, x_p 的样本协方差矩阵:

$$S = (s_{ij}), s_{ij} = \frac{1}{n} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j)$$

$$\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai}, i, j = 1, 2, \dots, p$$

从 S 求它的特征根与特征向量,用

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p \geq 0$$

$$r_1, r_2, r_3, \dots, r_p, r_i = \begin{pmatrix} r_{i1} \\ \vdots \\ r_{ip} \end{pmatrix}, i = 1, \dots, p$$

分别表示特征根与相应的特征向量,看 $\lambda_1 / \sum_{i=1}^p \lambda_i$ 的值是否足够大,如果足够大就用 r_1 的各个分量作为系数得到综合评价的指标 y :

$$y = r_{11}x_1 + r_{12}x_2 + \dots + r_{1p}x_p, r_1 = \begin{pmatrix} r_{11} \\ \vdots \\ r_{1p} \end{pmatrix}$$

求出各个样本的 y 值,按 y 值的大小对样本给出评价的顺序.下一节我们按照这一步骤用一个例子来说明它的实现过程.

最后,在结束这一节之前,还需谈两点要注意的事情:

1. 我们处理问题时,不一定总是从原始的指标出发,求协方差阵或相关阵,有时可以将原始指标作一些变换,直接处理它们各自的函数,例如取对数,特别是一些百分比数据,因为比例值都在 0 与 1 之间,它的变化值不可能大,取对数后可以在 $(-\infty, 0)$ 之间变化,容易比较出差别来;当 $0 < x < 1$ 时,有时取

$$y = \ln \frac{x}{1-x}$$

后, y 的变化范围在 $(-\infty, \infty)$ 之间,更容易处理.所以作为主成分

分析的指标可以是原始指标的函数,这样求出的主成分就出现非线性函数的形式,下节有一个例子可以说明这一点.

2. 前面已提到过,有时第一个主成分分量 y_1 方差所占的比例,也就是它的贡献率不是很大,需要几个分量合起来才能占到 85% 以上,这时应该如何综合呢? 上节的例子已经出现了类似的问题. 遇到这种情况,可以有种种处理的方法,考虑因子分析,作旋转(见本章后面几节)以及加权等等都是可以的,但需针对实际问题来选择,不能死套一种方法.

在有些场合,需要将专家们的意见和实际资料提供的信息,作出一个综合的结论. 专家的意见是主观的,实际资料是客观的,这两者往往不是一致的,这一类问题处理时更应该具体分析,种种评价的方法只是一种参考.

第三节 几个实例

这一节我们完整地先介绍一个实例,讨论的内容是安徽省 16 个地区宏观经济发展的情况,资料来源见文献[36].

选用的评价指标是八个:

x_1 : 固定资产利税率;

x_2 : 流动资金利税率;

x_3 : 销售收入利税率;

x_4 : 净产值的利税率;

x_5 : 总产值的利税率;

x_6 : 人均的利税率;

x_7 : 全员劳动生产率;

x_8 : 物耗利损率.

16 个地区的原始数据见表 3-3.

将表中的数据记为 $x_{\alpha i}$, α 表示地区号, i 表示指标号,于是

$$X = (x_{\alpha i})$$

是 16×8 的矩阵, 求出均值向量及协方差矩阵:

$$\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai}, S = (s_{ij}), n = 16$$

$$s_{ij} = \frac{1}{n} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j), i, j = 1, 2, \dots, 8$$

表 3-3 原始数据表

地 指 标 区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
①合肥市	0.461	0.334	0.164	0.534	0.146	0.359	2.450	0.202
②芜湖市	0.537	0.372	0.177	0.549	0.147	0.335	2.281	0.201
③蚌埠市	0.607	0.419	0.204	0.712	0.192	0.469	2.440	0.263
④淮南市	0.021	0.044	0.019	0.071	0.018	0.027	1.454	0.025
⑤马鞍山市	0.298	0.520	0.182	0.608	0.177	0.408	2.313	0.249
⑥淮北市	0.054	0.168	0.060	0.325	0.074	0.081	1.104	0.095
⑦铜陵市	0.152	0.206	0.101	0.344	0.086	0.174	2.028	0.114
⑧安庆市	0.327	0.364	0.128	0.537	0.119	0.270	2.273	0.153
⑨黄山市	0.185	0.164	0.084	0.302	0.083	0.128	1.546	0.114
⑩阜阳地区	0.467	0.336	0.154	0.522	0.139	0.277	1.987	0.190
⑪宿县地区	0.232	0.211	0.091	0.391	0.083	0.139	1.685	0.105
⑫滁县地区	0.514	0.401	0.149	0.539	0.129	0.299	2.324	0.169
⑬六安地区	0.132	0.183	0.080	0.284	0.069	0.084	1.224	0.091
⑭宣城地区	0.202	0.275	0.124	0.415	0.113	0.178	1.575	0.156
⑮巢湖地区	0.246	0.291	0.118	0.430	0.104	0.144	1.392	0.136
⑯池州地区	0.153	0.185	0.086	0.330	0.082	0.119	1.452	0.110
均值 \bar{x}	0.287	0.280	0.120	0.430	0.110	0.218	1.845	0.148
方差 s^2	0.0326	0.0149	0.0025	0.0243	0.0020	0.0168	0.2153	0.0039
标准差 s	0.180	0.122	0.050	0.156	0.044	0.130	0.464	0.062

为了消除各指标之间因度量单位不同引起的差异, 就考虑将数据标准化之后的协方差矩阵, 也就是原数据的相关矩阵, 记它为 R , 现算出 R 的值是

$$R = \begin{bmatrix} 1.0 & & & & & & & & \\ 0.79470 & 1.0 & & & & & & & \\ 0.90238 & 0.93677 & 1.0 & & & & & & \\ 0.87596 & 0.94517 & 0.96418 & 1.0 & & & & & \\ 0.85711 & 0.94333 & 0.98509 & 0.97449 & 1.0 & & & & \\ 0.88386 & 0.91742 & 0.96172 & 0.93856 & 0.96536 & 1.0 & & & \\ 0.83027 & 0.78948 & 0.82625 & 0.79037 & 0.79134 & 0.90716 & 1.0 & & \\ 0.84280 & 0.93372 & 0.98138 & 0.95733 & 0.99771 & 0.96491 & 0.78619 & 1.0 \end{bmatrix}$$

从 R 矩阵可以看出,这些指标的相关性非常高,大部分相关系数都在 0.85 以上,而且都是正相关。

求 R 的特征根,得 8 个特征根的值和它们各自相应的贡献率(见表 3-4)。

表 3-4

λ_i	7.322787	0.348345	0.186249	0.086731	0.035174	0.017495	0.003174	0.000044
贡献率%	91.535	4.35	2.33	1.08	0.44	0.22	0.04	0.005
累计 贡献率%	91.535	95.885	98.215	99.295	99.735	99.955	99.995	100

从贡献率看出,选用第一个主成分就已足够好了,它相应的特征向量是

$$(0.37700, 0.35114, 0.36541, 0.36012, 0.36359, 0.36414, 0.32353, 0.36119)$$

因此综合评价函数

$$y = 0.37700x_1^* + 0.35114x_2^* + 0.36541x_3^* + 0.36012x_4^* + 0.36359x_5^* + 0.36414x_6^* + 0.32353x_7^* + 0.36119x_8^*$$

这里变量用的是标准化后的变量,也即

$$x_i^* = \frac{x_i - \bar{x}_i}{\sqrt{s_{ii}}}, \quad i = 1, 2, \dots, 8$$

所以将上述公式代入 y 的表达式后,才能得到用原始指标 x_1, \dots, x_8 表示的综合评价函数。从 y 的表达式看,对标准化变量 x_1^* ,

\cdots, x_8^* 而言, 相应的系数差别很小, 几乎就是一样, 这又一次印证了上章我们叙述过的定理, 相关性很强的指标, 任一加权平均与算术平均数的相关性很大. 但还原到原来的变量 x_1, \cdots, x_8 时, x_i 的系数就与 $\sqrt{s_{ii}}$ 成反比, 标准差 $\sqrt{s_{ii}}$ 越大的相应的系数就越小, 注意到 $\frac{1}{\sqrt{s_{ii}}}$ 在某种意义下是反映了指标 x_i 的“精度”, 这就表示精度高的, 它的系数就大, 就更应重视, 这是合理的.

今算得

$$y = 8.7889 + 2.0944x_1 + 2.8782x_2 + 7.3082x_3 + 2.3085x_4 \\ + 8.2634x_5 + 2.8011x_6 + 0.6973x_7 + 5.8256x_8$$

因此就可以算出各地区相应的综合评价值. 从系数的大小来看, 从大到小排列, 得 x_5, x_3 和 x_8 , 最小的是 x_7 , 就这些指标来看也合于实际, 这样就得到了综合评价的模型.

利用上述 y , 算出各地区的评价值及相应的名次见(表 3-5).

表 3-5 各地区名次表

区号	评价值	名次	区号	评价值	名次
1	9.3	4	9	4.9	12
2	9.5	3	10	8.5	6
3	11.6	1	11	5.5	11
4	1.9	16	12	8.9	5
5	10.4	2	13	4.2	14
6	3.9	15	14	6.5	8
7	5.7	10	15	6.2	9
8	8.0	7	16	4.87	13

结论是:

蚌埠最好, 依次为

马鞍山、芜湖、合肥、滁县

淮南最差, 然后稍好一些是

淮北、六安、宣城、池州

这个排序很明显与工业发展的状况有关.

下面举的例子就不再详细论述全部的处理过程,重点突出这个例子的特殊之处.

例 3.1^[2] 我国农民家庭消费结构的分析.

以全国 30 个省、自治区、直辖市 1993 年的资料为依据,选择了八个指标:

- x_1 : 食品;
- x_2 : 衣着;
- x_3 : 居住;
- x_4 : 家庭设备及有关服务;
- x_5 : 医疗保健;
- x_6 : 交通通讯;
- x_7 : 文教、娱乐消费;
- x_8 : 其他非商品及服务消费.

这些 x_i 都是指该项支出占全部消费的百分比,因此自然有 $x_i \geq 0$,

$\sum_{i=1}^8 x_i = 1$. 这表示这些指标本身是线性相关的, $x = \begin{bmatrix} x_1 \\ \vdots \\ x_8 \end{bmatrix}$ 的协

方差矩阵是退化的(即它的行列式值为 0),有关的资料见表 3-6.

表 3-6 各省市农民家庭消费状况

地 区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
北京	48.16	10.58	11.29	9.26	4.62	3.04	11.61	1.44
天津	53.31	9.54	15.98	6.10	3.39	2.54	7.59	1.63
河北	58.39	7.50	13.86	5.16	6.32	1.63	5.90	1.23
山西	57.36	11.20	11.89	5.80	3.57	2.00	6.74	1.44
内蒙古	58.40	8.14	14.04	4.56	3.76	1.83	8.15	1.12
辽宁	55.14	10.47	12.34	4.79	4.30	1.98	8.91	2.07
吉林	60.63	9.70	9.99	4.03	4.44	1.94	8.15	1.12
黑龙江	61.4	8.88	15.25	3.88	3.38	1.56	5.04	0.97
上海	46.43	7.14	16.24	11.75	2.28	2.98	9.99	3.19
江苏	50.21	7.40	17.89	8.58	3.32	2.81	7.76	2.03
浙江	50.16	7.46	17.35	7.63	3.97	3.05	6.83	3.54

续表

地 区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
安徽	63.90	6.95	10.09	5.41	3.31	1.87	6.84	1.62
福建	60.60	4.99	13.44	5.04	2.43	2.90	7.46	3.13
江西	61.35	5.94	13.36	5.00	3.45	2.41	7.27	1.22
山东	57.36	8.31	13.64	5.60	3.41	2.04	8.32	1.31
河南	59.21	8.15	13.27	5.32	4.01	1.55	6.94	1.54
湖北	61.85	6.46	10.75	5.43	3.07	1.51	9.88	1.05
湖南	61.10	5.56	13.23	4.91	2.98	1.98	9.08	1.14
广东	52.77	3.73	16.98	6.69	3.17	4.28	9.41	2.97
广西	64.33	4.31	10.70	5.24	2.77	2.01	9.89	0.75
海南	67.17	4.67	7.06	4.87	2.55	2.66	9.08	1.94
四川	63.29	6.33	12.20	5.23	3.13	1.95	6.99	0.88
贵州	70.98	6.46	8.37	4.32	1.87	1.49	4.92	1.60
云南	61.20	6.92	13.20	6.50	3.22	1.65	5.96	1.35
西藏	66.53	8.78	18.21	3.40	0.49	0.98	0.62	1.00
陕西	57.13	7.75	15.39	5.50	4.58	1.33	7.08	1.24
甘肃	55.36	7.35	19.24	5.13	4.01	1.62	6.02	1.28
青海	55.50	9.54	20.86	4.43	3.84	1.75	3.01	1.08
宁夏	58.75	8.87	12.0	5.57	4.67	2.04	7.11	0.95
新疆	52.15	12.94	14.70	5.37	4.22	2.77	6.24	1.61

直接计算由这些数据给出的协方差矩阵,标准化之后是相关矩阵,由大到小排列后,它的前四个特征根为

$$\lambda_1 = 3.2272, \quad \lambda_2 = 1.8598$$

$$\lambda_3 = 1.4147, \quad \lambda_4 = 0.5751$$

这四个加起来的贡献率为 88.46%, 单用 λ_1 还不足 50%, 所以降维的效果不明显。

对原始数据作对数中心化变换(具体方法见前面第二章), 对新的指标求主成分, 得到第一主成分的贡献率就已达 94.42%, 降维效果非常明显. 相应的主成分为

$$y = 0.7307 \lg x_1 + 0.0677 \lg x_2 + 0.2553 \lg x_3 - 0.0382 \lg x_4 \\ - 0.2058 \lg x_5 - 0.36 \lg x_6 + 0.0217 \lg x_7 - 0.4715 \lg x_8$$

$$= \lg \frac{x_1^{0.7307} x_2^{0.0677} x_3^{0.2553} x_7^{0.0217}}{x_4^{0.0382} x_5^{0.2058} x_6^{0.36} x_8^{0.4715}}$$

因此用 y 的值来评价各地区的消费结构的差异是可以的. y 的表达式显示:

(i) 农民家庭中, 消费的主要项目是食品、住房、医疗保健、交通通讯与非商品消费; 在衣着、文教娱乐、家庭设备等方面消费是不多的.

(ii) 食品与住房是同步的; 医疗和交通是同步的. 实际情况也是如此. 与 1990 年相比, 农民家庭中食品和住房均下降 0.74 与 3.46 个百分点, 而医疗和交通分别上升 0.28 和 0.82 个百分点. 这就告诉我们, 上述评价函数中, x_i 所处的位置与关系和实际是相符的. 我们比较一下, 如直接用原始指标, 第一主成分相应的

$$\begin{aligned} \tilde{y} = & 0.479x_1 + 0.054x_2 - 0.155x_3 - 0.4786x_4 \\ & - 0.1119x_5 - 0.4731x_6 - 0.3245x_7 - 0.419x_8 \end{aligned}$$

它表示食品与住房是不同步的, x_1, x_4, x_6, x_8, x_7 是重要的消费内容. 它与实际情况就不太相符, 这涉及到成分数据 (即百分比数据) 的特性, 因为 $x_i \geq 0$, $\sum_{i=1}^8 x_i = 1$, 因此某一个 x_i 上升时, 其他的 x_j 就会受影响, 就要下降, 所以似乎内部一定是互相制约的负相关, 因此内部的正相关性不易得到反映. 经过对数变换后

$$\begin{aligned} & \lg x_1, \lg x_2, \lg x_3, \lg x_4 \\ & \lg x_5, \lg x_6, \lg x_7, \lg x_8 \end{aligned}$$

它们是线性无关的, 所以更易于表现出它们内部的真实关系, 本例就是一个很好的说明.

本例的第二个特点是如此求出的 y , 不仅可以用来评价, 还可以用于分析农民家庭消费结构的内在联系, 这对我们也是有启发作用的.

第四节 因子分析法

因子分析法是主成分分析的一种自然的延伸,因子分析的特点是它的解不具有唯一性,正因为是不唯一的,我们可以从中选择适合所考虑的具体问题的解,但是什么是适合,什么是不适合,就会有各种不同的看法,因而意见也易于分歧.这一节我们着重介绍因子分析的原理和算法,下一节举例来说明它在综合评价中的应用和应用的的具体算法.

我们沿用前面的记号, x_1, \dots, x_p 表示原始变量,用向量 x 表示, x 看成是随机向量,它的期望值是向量 μ ,它的协方差矩阵是 V ,即

$$V = (v_{ij}) = (\text{Cov}(x_i, x_j))$$

由于 V 是非负定的矩阵,因此可以用正交阵 Γ 将 V 对角化,对角元素均为 V 的特征根, Γ 的元素用 σ_{ij} 表示,写成

$$\Gamma = (\sigma_{ij}) = (\sigma_1 \quad \sigma_2 \quad \cdots \quad \sigma_p)$$

注意 σ_i 都是 $p \times 1$ 的向量, Γ 的正交性表现为

$$\sigma_i^T \sigma_i = 1, \sigma_i^T \sigma_j = 0 (i \neq j), i, j = 1, 2, \dots, p$$

用 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 表示 V 相应的特征根,于是对角化定理就是说

$$\Gamma^T V \Gamma = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$$

或者

$$V = \Gamma \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} \Gamma^T = \sum_{i=1}^p \lambda_i \sigma_i \sigma_i^T$$

令

$$y = \Gamma^T x = \begin{bmatrix} \sigma_1^T x \\ \vdots \\ \sigma_p^T x \end{bmatrix}$$

后,自然有

$$\begin{aligned} \text{Var}(y) &= (\text{Cov}(\sigma_i^T x, \sigma_j^T x)) \\ &= \text{Var}(\Gamma^T x) = \Gamma^T V \Gamma = \begin{bmatrix} \lambda_1 & 0 \\ & \ddots \\ 0 & \lambda_p \end{bmatrix}. \end{aligned}$$

这表示从 x 的协方差阵 V , 可以找到 x 的线性变换 Γ^T , 使 $y = \Gamma^T x$ 的各个分量不相关, 各自的方差正好是 V 的特征根, Γ 矩阵就是特征向量组成的矩阵. 主成分分析就是将 x 转换成 y , y 的各个分量是不相关的, 方差集中到前几个主成分上. 注意到 $y = \Gamma^T x$, Γ 是正交阵, 因此

$$\begin{aligned} x = \Gamma y &= \Gamma \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ & \ddots \\ 0 & \sqrt{\lambda_p} \end{bmatrix} \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 \\ & \ddots \\ 0 & 1/\sqrt{\lambda_p} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \\ &\stackrel{\text{记}}{=} \Gamma \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ & \ddots \\ 0 & \sqrt{\lambda_p} \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_p \end{bmatrix} \\ &\stackrel{\text{记}}{=} Az \end{aligned}$$

易见 z 的各个分量是互不相关的, 而且 $\text{Var}(z_i) = 1$. 从表达式 $x = Az$ 可以看出, 如果方差集中在少数几个 z_1, \dots, z_k 中, $\lambda_{k+1} = \dots = \lambda_p$ 都几乎是 0, 于是

$$x = A_1 \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix} + \epsilon$$

因此可以设想 x 是由少数因子的线性函数生成的, 这些因子彼此是不相关的, 而且方差都是 1, 用 f_1, \dots, f_k 表示这些因子, 就得到

$$\begin{matrix} x & = & A & f & + & \epsilon \\ p \times 1 & & p \times k & k \times 1 & & p \times 1 \end{matrix}$$

$$\text{Var}(f) = I_k, f \text{ 与 } \epsilon \text{ 不相关}$$

问题是 A 与 f 都是未知的, 如何从 x 的观察数据去求出 A 与 f 的估计值. 从上面的讨论就可以看出主成分分析为因子的寻找提供了一个途径.

从因子结构的表达式来看

$$x = Af + \epsilon = AP^{-1}Pf + \epsilon$$

总是成立的, 如果选 P 是正交阵, 令 $g = Pf$, 于是

$$x = AP^{-1}g + \epsilon \stackrel{\text{记}}{=} Bg + \epsilon$$

也是一种解, 此时 g 是由 f 经正交变换得来的, g 的各分量仍然不相关, 各自的方差都是 1, 而且 g 与 ϵ 不相关, 这样的由 f 到 g 称为正交旋转. 如果 P 不是正交阵, 而是一般的非退化矩阵, 这样得到的 g 称为由 f 经斜旋转得来的, 此时 g 不再保持 f 原有的性质:

各分量不相关, 各自的方差都是 1.

然而经过旋转以后的因子, 有时它的实际意义比较明显, 这已被许多实际例子所证实.

有关这些旋转的计算公式及软件, 在一般的多元统计分析教材和软件中都可找到, 这里就不逐一介绍了, 下一节用一些实例来给以说明.

第五节 因子分析的实例和计算方法

这一节用我国 44 个城市第三产业发展水平的综合评价作为例子, 介绍因子分析的实际计算方法.

选用下面 20 个指标:

x_1 : 人口数;

x_2 : GDP (国内生产总值);

x_3 : 第三产业增加值;

x_4 : 货运总量;

- x_5 : 批、零、贸商品销售总额;
 x_6 : 外贸收购总额;
 x_7 : 年末银行贷款余额;
 x_8 : 社会零售物价指数;
 x_9 : 实际利用外资;
 x_{10} : 万名职工拥有科技人员的人数;
 x_{11} : 旅游外汇收入;
 x_{12} : 第三产业的就业比例;
 x_{13} : 邮电业务总量;
 x_{14} : 职工人均工资;
 x_{15} : 人均居住面积;
 x_{16} : 用水普及率;
 x_{17} : 煤气普及率;
 x_{18} : 人均道路面积;
 x_{19} : 人均公共绿地面积;
 x_{20} : 政策体制.

全部原始资料可以从《城市统计年鉴 1993~1994》查到,这里就不将表列出了.

用主成分分析法求相关矩阵的特征根,其结果见表 3-7.

表 3-7

主成分分量序号	特征值	贡献率%	累计贡献率%
1	7.9913	40	40
2	3.7347	18.7	58.7
3	1.7313	8.7	67.4
4	1.4953	7.5	74.9
5	0.9837	4.9	79.8
6	0.8124	4.1	83.9
7	0.7183	3.6	87.5
8	0.6261	3.1	90.6

这表示单独用一个或少数二三个主成分分量是不能反映这 20 个指标的变化状况,若用到七八个,就可以达到 85% ~ 90%,用 f_1, \dots, f_8 分别表示前八个主分量,并且规格化使它们各自的方差都是 1,于是

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \end{bmatrix} = A_{20 \times 8} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_8 \end{bmatrix} + \epsilon, \text{Cov}(f, \epsilon) = 0$$

矩阵 $A_{20 \times 8}$ 称为因子荷载矩阵,因为 A 矩阵的每一行反映了指标对因子 f_1, \dots, f_8 依赖的状况,例如

$$x_1 = a_{11}f_1 + a_{12}f_2 + \dots + a_{18}f_8 + \epsilon_1$$

这表示 x_1 与 f_1, \dots, f_8 的关系,注意到 f_i 彼此不相关,并且各自的方差为 1,因此 x_1 的方差

$$\text{Var}(x_1) = a_{11}^2 + a_{12}^2 + \dots + a_{18}^2 + \text{Var}(\epsilon_1)$$

a_{1i}^2 反映了因子 f_i 荷载了 x_1 的方差的量,所以 a_{ij} 称为荷载矩阵是有实际意义的,对其余的 x_i 也是类似的.

实际上 x 的协方差矩阵

$$\begin{aligned} \text{Var}(x) &= \text{Var}(Af) + \text{Var}(\epsilon) \\ &= A \text{Var}(f) A^T + \text{Var}(\epsilon) \\ &= AA^T + \text{Var}(\epsilon) \end{aligned}$$

所以误差 ϵ 的协方差矩阵

$$\text{Var}(\epsilon) = \text{Var}(x) - AA^T$$

这就告诉我们,求出因子荷载阵 A 之后, $\text{Var}(x) - AA^T$ 应该与一个对角阵相近,它的非对角元素应近似于 0,这也就提供了选择因子个数是否合适的一项依据.

因子荷载的值还明确显示了因子与原指标之间的关系,因为

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \dots + a_{i8}f_8 + \epsilon_i, i = 1, 2, \dots, 20$$

由于 $f_1, \dots, f_8, \epsilon_i$ 彼此不相关, $\text{Var}(f_i) = 1$, 因此 x_i 与 f_j 的协方差就是 a_{ij} , 即

$$\text{Cov}(x_i, f_j) = a_{ij}$$

于是 x_i 与 f_j 的相关系数

$$\rho_{x_i, f_j} = \frac{\text{Cov}(x_i, f_j)}{\sqrt{\text{Var}(x_i)\text{Var}(f_j)}} = \frac{a_{ij}}{\sqrt{\text{Var}(x_i)}}$$

如果原始数据已标准化, 则 $\text{Var}(x_i) = 1$, 因此

$$\rho_{x_i, f_j} = a_{ij}$$

这就告诉我们 a_{ij} 的大小、正负号都反映了 x_i 与 f_j 的相关状况. 如果 $a_{ij}, j=1, 2, \dots, 8$ 之间有的接近于 0, 有的接近于 ± 1 , 这表示因子 f_j 主要是反映了那些接近 ± 1 的原始指标. 这样就产生两个值得注意的推论:

(i) 可以利用因子荷载的值反映原始指标与因子的联系, 利用因子将原始指标给以分类, 然后可以判断这个因子的实际意义是什么.

这一推论使因子分析的作用大大加强, 因为有一些潜在因素, 原来是无法度量与观察的, 往往通过这一方法可以发现, 度量它的大小.

(ii) 将因子作正交旋转(或斜旋转), 使荷载矩阵中每一行的数值尽可能两极化(接近 0 或接近 ± 1), 这样得到的因子便于发现它的实际意义. 基于这个想法, 就引导出极大方差旋转的方法, 英文名词是 Varimax 旋转, 在一些软件中已有专门的程序.

表 3-7 给出前五个主成分分量作为公共因子(此时总的贡献率为 79.8%, 将近 80%, 所以也就可以作综合评价了), 表 3-9 把未旋转的荷载矩阵 A_0 和经 Varimax 旋转后的 A 作一比较, 就可以看出差别, 表 3-8 列出全部特征根和相应的贡献率.

从旋转后的因子荷载矩阵来看, 五个因子的意义比较明显, 也就是将 20 个指标分成五大类:

1. 第三产业的基本经济因子:

人口数、GDP、三产增加值、货运总量、
 $x_1 \quad x_2 \quad x_3 \quad x_4$

表 3-8 因子特征值及在总方差中比重

因子	特征值	在总方差中 的比重(%)	累计比重(%)	因子	特征值	在总方差中 的比重(%)	累计比重(%)
1	7.9913	40.0	40.0	11	0.2928	1.5	96.8
2	3.7347	18.7	58.7	12	0.1811	0.9	97.7
3	1.7313	8.7	67.4	13	0.1670	0.8	98.5
4	1.4953	7.5	74.9	14	0.1300	0.6	99.1
5	0.9857	4.9	79.8	15	0.0789	0.4	99.5
6	0.8124	4.1	83.9	16	0.0451	0.2	99.7
7	0.7183	3.6	87.5	17	0.0362	0.1	99.8
8	0.6261	3.1	90.6	18	0.0284	0.1	99.9
9	0.5910	3.0	93.6	19	0.0153	0.1	100.0
10	0.3366	1.7	95.3	20	0.0028	0.0	100.0

表 3-9 因子载荷矩阵和旋转后因子载荷矩阵

变 量	因子载荷矩阵 A_0					旋转后因子载荷矩阵 A				
	1	2	3	4	5	1	2	3	4	5
x_1 POP	0.59	-0.52	0.24	0.01	0.34	0.49	-0.18	-0.22	0.24	0.63
x_2 GDP	0.97	-0.14	0.07	0.06	0.06	0.94	-0.02	-0.06	0.08	0.30
x_3 TRP	0.98	-0.09	-0.02	0.07	0.02	0.96	-0.02	0.01	0.00	0.23
x_4 FT	0.63	-0.30	-0.06	-0.16	0.49	0.51	-0.19	0.01	-0.12	0.67
x_5 SPWT	0.95	-0.14	-0.04	0.08	0.01	0.94	-0.06	-0.04	0.00	0.23
x_6 TPC	0.80	0.11	0.13	-0.03	-0.17	0.78	-0.00	0.28	0.16	0.04
x_7 LBB	0.95	-0.11	-0.06	0.07	-0.05	0.94	-0.07	-0.01	-0.01	0.16
x_8 RPI	0.16	0.51	0.15	-0.59	-0.01	0.06	0.06	0.80	0.04	0.01
x_9 FCA	0.89	0.26	0.00	-0.12	-0.08	0.86	0.08	0.36	-0.03	0.07
x_{10} NST	0.06	-0.60	0.33	0.53	0.04	0.10	0.01	-0.73	0.41	0.21
x_{11} TR	0.84	0.10	-0.17	0.23	-0.26	0.90	0.08	-0.03	-0.10	-0.16
x_{12} TRET	0.19	0.31	-0.80	0.15	0.01	0.26	0.03	-0.00	-0.81	-0.26
x_{13} PTS	0.95	0.08	-0.09	0.09	-0.14	0.96	0.04	0.08	-0.05	0.02
x_{14} AS	0.47	0.71	-0.11	-0.16	-0.03	0.45	0.35	0.60	-0.26	-0.14
x_{15} PCLS	-0.09	0.59	0.03	0.57	-0.11	0.01	0.71	-0.10	-0.12	-0.40
x_{16} PUMH	0.19	-0.27	0.56	0.03	-0.52	0.25	-0.12	-0.09	0.75	-0.22
x_{17} PUGH	0.16	0.51	0.27	0.36	0.44	0.11	0.77	0.07	-0.06	0.27
x_{18} PCRA	-0.17	0.65	0.48	0.22	0.06	-0.18	0.76	0.26	0.22	-0.11
x_{19} PCGA	-0.05	0.74	-0.05	0.29	0.08	-0.03	0.68	0.20	-0.28	-0.23
x_{20} P	0.33	0.59	0.46	-0.31	0.11	0.22	0.42	0.69	0.24	0.15

批零贸易商品销售总额、年末银行贷款余额、

x_5 x_7

实际利用外资额、旅游外汇收入、三产就业比率、

x_9 x_{11} x_{12}

邮电业务总量、外贸收购总额；

x_{13} x_6

2. 基础环境因子：

人均居住面积、用水普及率、煤气普及率、

x_{15} x_{16} x_{17}

人均道路面积、人均公共绿地面积；

x_{18} x_{19}

3. 政策性因子：

社会零售物价指数、职工人均工资、政策体制；

x_8 x_{14} x_{20}

4. 人员素质因子：

万名职工科技人员数 x_{10} 。

5. 补充因子：

它与第一、二类因子部分相重，弥补一、二类因子的不足部分。

在这个基础上，选择哪一个来综合评价都不理想，应根据这 5 个因子的值再综合，自然考虑将五个因子在各个城市所得的值给以加权平均，作为综合评价的依据。问题是权怎么定？按贡献率的大小来加权平均，就是一个比较容易想到的方法，看来也有一定道理，于是有综合评价值

$$E = w_1 f_1 + w_2 f_2 + w_3 f_3 + w_4 f_4 + w_5 f_5$$

$$w_i = \lambda_i / \sum_{i=1}^{20} \lambda_i, i = 1, 2, 3, 4, 5$$

f_i 是旋转后的 x_j 的线性函数（即 A 矩阵各列相应的线性函数）。

按上述 E 评价的结论见表 3-10 给出的值。这 44 个名次与实际还是比较相符的。

不仅如此，我们还可按四类因子得分的状况对几个重要城市第三产业结构、基础作一分析比较，具体数值见表 3-11。

表 3-10 44 个主要城市第三产业发展水平(E 值)

序号	城市	E	序号	城市	E	序号	城市	E	序号	城市	E	序号	城市	E
1	上海	4.77	10	厦门	0.43	19	武汉	-0.19	28	南通	-0.49	37	西安	-0.92
2	北京	3.06	11	宁波	0.37	20	石家庄	-0.20	29	秦皇岛	-0.51	38	连云港	-0.93
3	广州	2.34	12	烟台	0.36	21	福州	-0.22	30	郑州	-0.54	39	南昌	-0.93
4	深圳	2.08	13	杭州	0.31	22	成都	-0.22	31	乌鲁木齐	-0.56	40	贵阳	-1.22
5	天津	0.90	14	沈阳	0.28	23	湛江	-0.26	32	长春	-0.57	41	兰州	-1.25
6	珠海	0.82	15	南京	0.26	24	北海	-0.26	33	昆明	-0.58	42	银川	-1.30
7	大连	0.78	16	海口	0.23	25	济南	-0.29	34	长沙	-0.66	43	呼和浩特	-1.44
8	威海	0.46	17	汕头	-0.05	26	哈尔滨	-0.43	35	太原	-0.66	44	西宁	-1.69
9	青岛	0.44	18	重庆	-0.81	27	温州	-0.46	36	南宁	-0.87			

表 3-11 重要城市样本第三产业因素结构分析

序号	城市	基本经济因子	基础环境因子	政策性可控因子	劳动力质量因子	总 计
1	上海	5.007	-0.187	-0.002	-0.047	4.771
2	北京	2.912	0.013	-0.109	0.24	3.056
3	广州	2.496	0.043	-0.055	-0.140	2.344
4	深圳	1.272	0.845	-0.017	-0.025	2.075
5	天津	0.994	-0.159	0.063	0.005	0.903
21	福州	-0.258	0.093	0.004	-0.058	-0.219
22	成都	0.048	-0.418	0.042	0.108	-0.220
23	湛江	-0.268	0.112	0.083	-0.189	-0.262
24	北海	-0.821	0.511	0.195	-0.147	-0.265
25	济南	-0.300	-0.189	0.095	0.109	-0.285
40	贵阳	-1.004	0.026	-0.241	-0.004	-1.225
41	兰州	-0.852	-0.365	-0.076	0.042	-1.251
42	银川	-1.156	-0.221	-0.011	0.089	-1.299
43	呼和浩特	-1.051	-0.189	-0.197	-0.001	-1.438
44	西宁	-1.19	-0.398	-0.104	0.007	-1.685

从这个例子可以看出,用因子分析作综合评价不仅可以给出排名顺序,还可以进一步探索影响排名次序的因素,从而找到进一步改善努力的方向,这是一般评价方法无法代替的。

第四章 综合评价的聚类分析与判别分析方法

本书第一章中对综合评价的问题分了三个类型,这是按照综合评价的目的进行的.其中第一、三类评价问题就是本章所要重点解决的综合评价问题.比如,在地质勘探中,我们要根据地质结构数据,物探、化探指标来判别矿化类型、矿质结构与矿物贮藏丰瘠程度,等等;在医学诊断中,要根据病人体温、白血球数目以及其他病理指标诊断病人的病变类型,等等.这些综合评价问题都是以“有训练样本”为前提条件.这种综合评价问题可利用本章所介绍的判别分析法加以解决.如果是“无训练样本”的综合评价问题,如对某地区工业企业的经营状况进行综合评价,对某些医院以及医院内部的不同科室的经营管理进行综合评价就必须先对总体进行聚类,然后再进行判别,给出当前样本的综合评价结果.这些都是这一章要解决的问题.

第一节 综合评价的系统聚类法

聚类分析是数值分类学的基本内容,是对统计样本进行定量分类的一种多元统计分析方法.将这种方法应用于综合评价,一方面可以对分类评价问题给出直接的评价结果,另一方面,也为其他综合评价方法如后续判别分析提供训练样本,形成综合评价的框架结构以便提高综合评价的效果.从本节开始我们分别介绍系统聚类与动态聚类和有序样品聚类的方法.为此我们先引入一些基本概念.

一、常用统计距离与聚类距离的不同定义和性质

1. 距离公理:设 x_1, x_2, \dots, x_n 为 n 个样本,第 i 个样本 x_i 与第 j 个样本 x_j 之间建立了一个函数关系式 $d_{ij} = d(x_i, x_j)$,如果它满足:

(i)非负性: $d_{ij} \geq 0$,对一切 i, j 成立.

(ii) 规范性: $d_{ij} = 0$, 当且仅当向量 $x_i = x_j$.

(iii) 对称性: $d_{ij} = d_{ji}$, 对一切 i, j 成立.

(iv) 三角不等式: $d_{ij} \leq d_{ik} + d_{kj}$, 对一切 i, j, k 都成立. 则称 d_{ij} 为样本 x_i 与 x_j 之间的距离.

常见的统计距离有:

(1) 闵可夫斯基(Minkowski)距离

设 x_i 均为 p 维向量, 且 $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}, i = 1, 2, \dots, n$

称

$$d_{ij}(q) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^q \right]^{\frac{1}{q}}, i, j = 1, 2, \dots, n$$

为闵可夫斯基距离.

当 $q = 1, 2$, 及 $q \rightarrow \infty$ 时, 分别得到绝对距离, 欧氏距离及切比雪夫距离.

(2) B 模距离

对于任给的正定矩阵 B , 由下式确定距离

$$d_{ij} = \left[(x_i - x_j)^T B (x_i - x_j) \right]^{\frac{1}{2}} \\ i, j = 1, 2, \dots, n$$

所得的距离称为 x_i 与 x_j 的 B 模距离.

当 $B = I$ (即单位矩阵) 时, B 模距离是常见的欧氏距离. 设 X 为

一 p 维随机向量, 即有 $X = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$, 方差 $D(x_i) = \sigma_i^2, i = 1, 2, \dots, p$. 则

当 $B = \text{diag} \left\{ \frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots, \frac{1}{\sigma_p^2} \right\}$ 时, 得精度加权距离; 当 $B = [\text{Cov}(x)]^{-1} \triangleq \Sigma^{-1}$ 时, 得到多元统计中经常使用的马哈拉诺比斯(Mahalanobis)距离, 简称马氏距离.

二、系统聚类的一般原则

系统聚类(Hierarchical Clustering Methods)有两个基本的思路.一个是将被评价对象每一个单元(或样本)看成一个类,通过建立相似性度量,逐步将类由多变少.另一个思路是相反地将全部被评价对象看成一类,通过建立相似性度量将类由少变多.就其应用于综合评价问题应遵循如下一般原则:

(i)首先将被评价的 n 个个体看成 n 个类,这时类间距离与样品间距离是相等的.

(ii)按照被评价对象的评价指标体系的特征,选择适当的“距离”作为不相似性度量,并求出最小类间距离.

(iii)将最小距离的类并为一类,并求出新类与其余类之间的距离,并选出最小类间距离.

(iv)重复(iii)步骤,直至所有类归为一类.

(v)在所取“距离”意义下,画出按相似性或相近程度联结的谱系图.

(vi)按综合评价的精度要求,选择阈值,确定聚类结果并给出综合评价的结果.

三、八种距离聚类方法及步骤

现在我们用 G_1, G_2, \dots, G_m 表示 m 个类,用 d_{ij} 表示样品 i 与 j 的距离,用 D_{pq} 表示 G_p 与 G_q 的距离,则 $D_{pq} = \min_{i \in G_p, j \in G_q} d_{ij}$, $p, q = 1, 2, \dots, m$.

(1)最短距离法聚类方法

(i)选择适当的距离函数,计算样品两两间距离 d_{ij} ,形成样品距离对称阵,记为 $D_{(0)}$ (这时 $D_{pq} = d_{pq}$).

(ii)选择 D_0 中最小元素,无妨设为 D_{pq} ,则将 G_p 与 G_q 合并为新类 G_r ; $G_r = G_p \cup G_q$.

(iii)计算新类与其他类间的距离:

$$\begin{aligned}
 D_{rk} &= \min_{i \in G_r, j \in G_k} d_{ij} \\
 &= \min \left\{ \min_{i \in G_p, j \in G_k} d_{ij}, \min_{i \in G_q, j \in G_k} d_{ij} \right\} \\
 &= \min \{ D_{pk}, D_{qk} \},
 \end{aligned}$$

并用 D_{rk} 代替 D_0 中 p, q 行与 p, q 列, 得距离对称矩阵 $D_{(1)}$.

(iv) 对于 $D_{(1)}$, 重复(ii), 直到所有的类并为一类为止. 注意: 如果在 $D_{(k)}$ 中重复(ii)时, 最小元不止一个, 则应将对应类同时合并; 在给定聚类阈值 T 时, 当 $D_{rk} \leq T$ 时, 聚类过程可以结束; 如果选择相似性度量是“距离”, 则方法与步骤同上; 如果是用相似系数作为相似性度量, 则只要将以上过程中的“min”改为“max”即可.

(2) 最长距离法

在(1)所述的方法中, 只要将 D_{pq} 定义为 $D_{pq} = \max_{i \in G_p, j \in G_q} d_{ij}$, 相应合并类后定义新类与剩余类的距离为 $D_{rk} = \max \{ D_{pk}, D_{qk} \}$ (其中的 $D_r = G_p \cup G_q$). 其余步骤与方法同(1).

(3) 中间距离法

该方法与(1)所不同的是在初始距离矩阵 $D_{(0)}$ 中, 用元素 d_{ij}^2 代替 d_{ij} , 记这时所得矩阵为 $D_{(0)}^2$, 定义新类 $G_r = G_p \cup G_q$ 与剩余类 G_k 的距离为

$$D_{kr}^2 = \frac{1}{2} D_{kp}^2 + \frac{1}{2} D_{kq}^2 - \frac{1}{4} D_{pq}^2$$

剩余步骤同(1).

(4) 重心法

该法类间距离采用两类的重心(即均值)之间的距离作为相异性度量, 设 G_p, G_q 的重心为 \bar{x}_p 与 \bar{x}_q , 则定义

$$D_{pq} = d_{\bar{x}_p \bar{x}_q}, \bar{x}_j \text{ 为 } G_j \text{ 类向量的均值向量}, j = 1, 2, \dots, m$$

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 - \frac{n_p}{n_r} \cdot \frac{n_q}{n_r} \cdot D_{pq}^2$$

其中 $G_r = G_p \cup G_q$, n_p, n_q, n_r 分别是 G_p, G_q, G_r 中样品数目. 其余步骤同(1).

(5)类平均法

如果两类 G_p, G_q 之间距离按如下定义:

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{i \in G_p, j \in G_q} d_{ij}^2$$

$$D_{rk}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2$$

其中 n_r, n_p, n_q 同(4). 其余步骤同(4).

(6)可变类平均法

该法是在(5)中, 将新类 $G_r = G_p \cup G_q$ 与剩余类的距离定义为

$$D_{kr}^2 = \frac{n_p}{n_r} (1 - \beta) D_{kp}^2 + \frac{n_q}{n_r} (1 - \beta) D_{kq}^2 + \beta D_{pq}^2$$

时, 对应的系统聚类方法称为类平均法, 其中常数 $\beta < 1$.

(7)可变法

如果将合并新类与剩余类间距离式定义为

$$D_{kr}^2 = \frac{1-\beta}{2} [D_{kp}^2 + D_{kq}^2] + \beta D_{pq}^2$$

时(其中 $\beta < 1$), 对应的系统聚类法叫做可变法.

(8)离差平方和法

该方法的思想与方差分析类似. 从数值分类学的角度看, 类分的适当时应该是类内样品离差的平方和尽量小, 类间离差平方和较大. 因此, 可将整体类内离差的平方和极小原则做为系统聚类的准则.

设有 G_1, G_2, \dots, G_m 个类, α_{ij} 表示第 i 类中第 j 个样本在 p 个评价指标下的标志值. 令

$$S = \sum_{k=1}^m S_k = \sum_{k=1}^m \sum_{i=1}^{n_k} (\alpha_{ik} - \bar{\alpha}_k)^T (\alpha_{ik} - \bar{\alpha}_k)$$

其中 $\bar{\alpha}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \alpha_{ik}$. 定义两类 G_p, G_q 合并后的离差增量为平方距离:

$$D_{pq}^2 \triangleq S_r - S_p - S_q = \frac{n_p \cdot n_q}{n_p + n_q} (\bar{x}_p - \bar{x}_q)^T (\bar{x}_p - \bar{x}_q)$$

其中 n_p, n_q 同(7)中定义, $\bar{x}_p = \frac{1}{n_p} \sum_{k=1}^{n_p} \alpha_{pk}, \bar{x}_q = \frac{1}{n_q} \sum_{k=1}^{n_q} \alpha_{qk}$. 则有递推

关系式:

$$D_{kr}^2 \triangleq \frac{n_k + n_p}{n_r + n_k} D_{kp}^2 + \frac{n_k + n_q}{n_r + n_k} D_{kq}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2$$

以上 8 种不同类型的类间距离定义, 易证它们都符合距离公理. 完全可以做为样本之间的相异程度的度量建立系统聚类过程, 所不同的只是看是否适合所要进行的聚类样本的特点. 1967 年 Lance 和 Williams 给出了合并前、后类间统一的递推公式为

$$D_{kr}^2 = \alpha_p D_{kp}^2 + \alpha_q D_{kq}^2 + \beta D_{pq}^2 + \gamma |D_{kp}^2 - D_{kq}^2|,$$

其中系数 $\alpha_p, \alpha_q, \beta$ 与 γ 对不同的聚类方法有不同的取值, 现将其列表 4-1.

表 4-1

方法名称	γ	α_p	α_q	β
最短距离法	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
最长距离法	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0
中间距离法	0	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{4} \leq \beta \leq 0$
重心法	0	n_p/n_r	n_q/n_r	$-\alpha_p \alpha_q$
类平均法	0	n_p/n_r	n_q/n_r	0
可变类平均法	0	$(1-\beta)n_p/n_r$	$(1-\beta)n_q/n_r$	<1
可变法	0	$(1-\beta)/2$	$(1-\beta)/2$	<1
离差平方和法	0	$(n_k + n_p)/(n_k + n_r)$	$(n_k + n_q)/(n_r + n_k)$	$-n_k/(n_r + n_k)$

四、系统聚类法应用于综合评价的实例

例 4.1^[12] 城镇居民消费水平通常用以下八项指标来描述
 x_1 : 人均粮食支出(元/人);

- x_2 :人均副食支出(元/人);
 x_3 :人均烟、酒、茶支出(元/人);
 x_4 :人均其他副食支出(元/人);
 x_5 :人均衣着商品支出(元/人);
 x_6 :人均日用品支出(元/人);
 x_7 :人均燃料支出(元/人);
 x_8 :人均非商品支出(元/人).

原始数据见表 4-2. 为了对城镇居民的消费状况有一个深入的了解,我们先从评价指标的分析入手,表 4-2 实质上给出了一个 8 维的 30 个样本点的样本信息阵. 为此我们先对指标进行聚类,并用相似系数作为相似性度量,可以得到相关系数矩阵见表 4-3.

表 4-2 30 个城市消费结构原始数据表

样本	次序 编号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
北京	1	7.78	48.44	8.00	20.51	22.12	15.73	1.15	16.61
天津	2	10.85	44.68	7.32	14.51	17.13	12.08	1.26	11.57
河北	3	9.09	28.12	7.40	9.62	17.26	11.12	2.49	12.65
山西	4	8.35	23.53	7.51	8.62	17.42	10.00	1.04	11.21
内蒙古	5	9.25	23.75	6.61	9.19	17.77	10.48	1.72	10.51
辽宁	6	7.90	39.77	8.49	12.94	19.27	11.05	2.04	13.29
吉林	7	8.19	30.50	4.72	9.78	16.28	7.60	2.52	10.32
黑龙江	8	7.73	29.20	5.42	9.43	19.29	8.49	2.52	10.00
上海	9	8.28	64.34	8.00	22.22	20.06	15.52	0.72	22.89
江苏	10	7.21	45.79	7.66	10.36	16.56	12.86	2.25	11.69
浙江	11	7.68	50.37	11.35	13.30	19.25	14.59	2.75	14.87
安徽	12	8.14	37.75	9.61	8.49	13.15	9.76	1.28	11.28
福建	13	10.60	52.41	7.70	9.98	12.53	11.70	2.31	14.69
江西	14	6.25	35.02	4.72	6.28	10.03	7.15	1.93	10.39
山东	15	8.82	33.70	7.59	10.98	18.82	14.73	1.78	10.10
河南	16	9.42	27.93	8.20	8.14	16.17	9.42	1.55	9.76

续表

样本	次序 编号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
湖北	17	8.67	36.05	7.31	7.75	16.67	11.68	2.38	12.88
湖南	18	6.77	38.69	6.01	8.82	14.79	11.44	1.74	13.23
广东	19	12.47	76.39	5.52	11.24	14.52	22.00	5.46	25.50
广西	20	7.27	52.65	3.84	9.16	13.03	15.26	1.98	14.57
海南	21	13.45	55.85	5.50	7.45	9.55	9.52	2.21	16.30
四川	22	7.18	40.91	7.32	8.94	17.60	12.75	1.14	14.80
贵州	23	7.67	35.71	8.04	8.31	15.13	7.76	1.41	13.25
云南	24	9.98	37.69	7.01	8.94	16.15	11.08	0.83	11.67
西藏	25	7.94	39.65	20.97	20.82	22.52	12.41	1.75	7.90
陕西	26	9.41	28.20	5.77	10.80	16.36	11.56	1.53	12.17
甘肃	27	9.16	27.98	9.01	9.32	15.99	9.10	1.82	11.35
青海	28	10.06	28.64	10.52	10.05	16.18	8.39	1.96	10.81
宁夏	29	8.70	28.12	7.21	10.53	19.45	13.30	1.66	11.96
新疆	30	6.93	29.85	4.54	9.49	16.62	10.65	1.88	13.61

利用系统聚类法的一般准则,将八个指标看成八个类 G_1, G_2, \dots, G_8 ; 取表 4-3 相似系数最大的两类 G_2 与 G_8 合并为一个新类 G_9 (其中相似系数 $r_{2,8}=0.835$), 再计算 G_9 与各类的相关系数, 再找最大的相关系数, 每次缩小一类就得谱系聚类图 4-1。我们从图 4-1 的情况可以将消费结构的评定划分为以下三个方面。一方面是由人均副食支出, 日用品支出, 及非商品服务性支出为主的消费领域, 这一领域是消费结构中起主导作用的方面; 其次一个领域是由购买烟、酒、茶, 衣着及其他副食方面的支出。最后是粮食与燃料两个指标构成的消费领域。从三个消费领域的构成看第一个领域的消费比重较大, 是研究消费结构的重点内容, 它是影响消费结构分析中恩格尔系数的主要因素。第二个消费领域, 具有较大的消费弹性。第三个领域是较稳定的一部分。

表 4-3 八指标相似系数矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_2	0.3335						
x_3	-0.0547	-0.229					
x_4	-0.0614	0.3988	0.5333				
x_5	-0.2896	-0.1564	0.4966	0.6984			
x_6	0.1962	0.7165	0.0328	0.4782	0.2837		
x_7	0.3484	0.4131	-0.1383	-0.1714	-0.2097	0.4082	
x_8	0.3190	0.8350	0.2583	0.3125	-0.0815	0.7099	0.3982

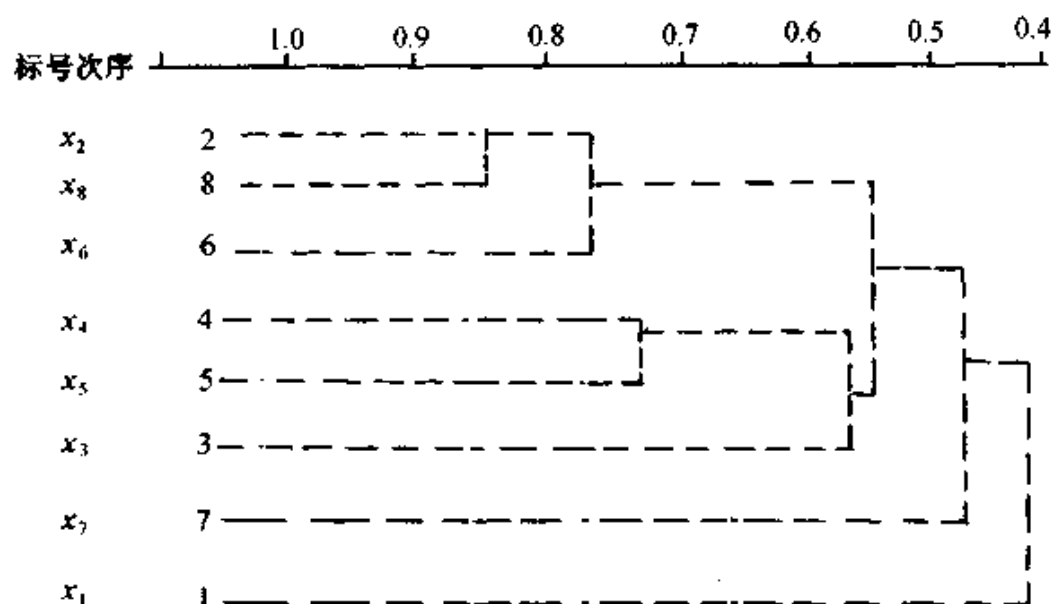


图 4-1 城镇居民消费指标聚类图

下面我们利用上节讲述的平均距离法对我国 30 个省市的消费结构进行聚类给出综合评价结果, 聚类图见图 4-2. 从图 4-2 看出可分成四类:

$$\text{I} = \{1, 2, 10, 11\}$$

$$\text{II} = \{3, 4, 5, 7, 8, 16, 26, 27, 28, 29, 30\}$$

$$\text{III} = \{6, 12, 14, 15, 17, 18, 22, 23, 24\}$$

$$\text{IV} = \{13, 20, 21\}$$

9, 19, 25 不易归类, 可做进一步的判别分析. 从表 4-2 的数据与

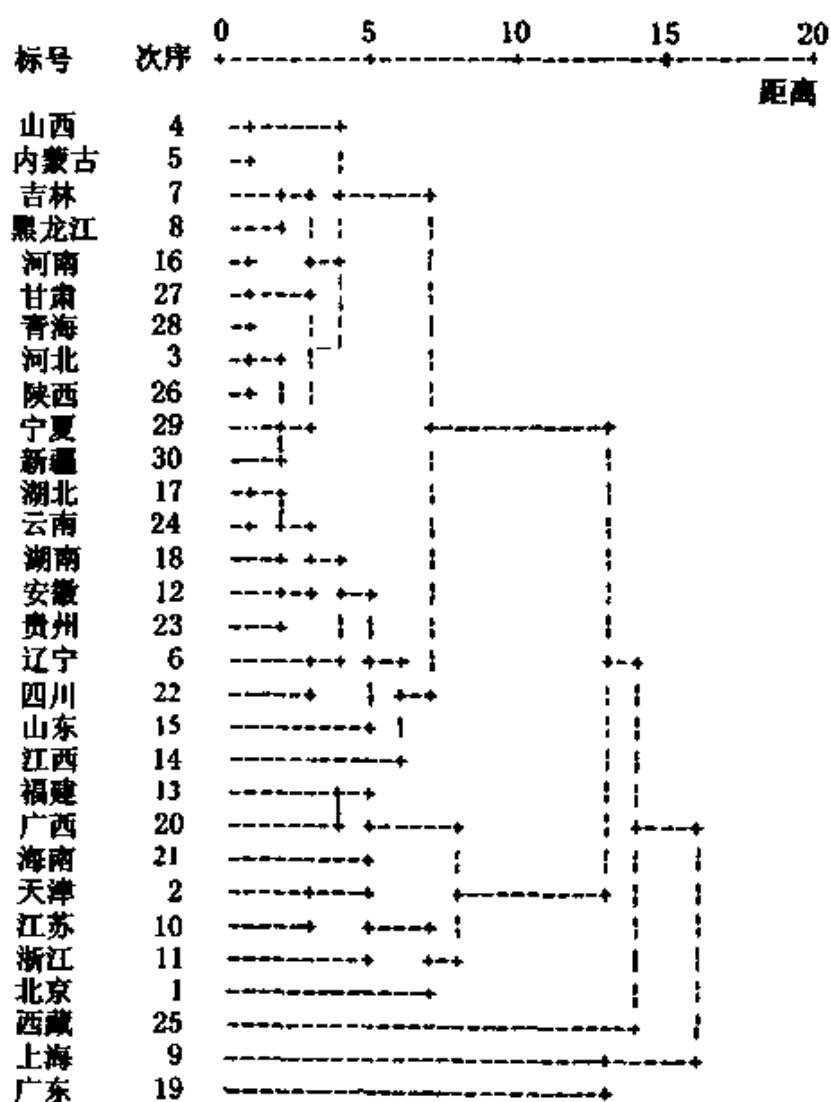


图 4-2 30 个省市消费结构聚类图

消费结构指标聚类的分析结果不难看出,北京、天津、江苏、浙江四省市在第一消费领域具有较大的比重,特别是副食品支出比重较大,说明这四省市人民生活水平较高,人们饮食结构较为丰富,这些是这一类的特性.第Ⅱ类 11 个省市在所有消费领域都有较低的支出水平,它反映了我国消费结构的主流,从这一消费层看主要的差异在副食品的支出差异上.这里有两个结论是明显的,一是饮食文化的层次较明显,这可以归结出我国消费领域饮食现代化、城市化的潜力巨大,由此可见第三产业发展仍有较大潜力.二是从消费

结构看,我国仍处于温饱型层面,非商品性消费支出比例不大,因而提高人们生活质量的问题仍是全社会关注的重要问题,这也符合我国实际情况.第Ⅲ类9个省区,属较发达省份,不过第Ⅱ类中存在的问题仍在这里可以见到.第Ⅳ类比第Ⅲ类要高些(生活水平),比第一类低些.最后,上海、广东、西藏三省市的特点比较明显,其中上海、广东的副食品支出、非商品支出都显著地高,西藏的消费结构非常具有特点,这些结果为地方政府制定产业结构发展战略具有十分重要的意义.综上所述,我国居民消费结构具有显著地层次结构与地方特色,大力发展第三产业,根本性地提高我国人民的生活水平,还有艰巨的工作要做.改革人民饮食结构,是提高生活质量的一个重要环节.

第二节 综合评价的动态聚类法

动态聚类法与系统聚类相比的突出特点是动态聚类对样本分类的确定性要求较弱,利用数值分析中的迭代思想使样本的聚类趋于一致.

在动态聚类中有两个问题是必须得到较好解决的.一个是初始分类中如何较好选择聚点问题;另一个问题是如何去修改初始分类使得最终分类比较合理.这两个问题都涉及一些基本的多元统计知识,我们以下只给出结果,对这些理论有兴趣的同志可参看文献[3,11].

一、初始分类中聚点的选择

1. 经验选点法:如果对评价对象有较好的经验知识或训练样本,则可以凭经验知识确定目前样本的分类,并确定每一类的代表样本作为聚点.如对某班同学的学习成绩进行聚类,一般分为不及格、及格、良好、优良、优秀等.我们可以选取各类代表样本做为聚点,如取45,65,75,85,90然后进入下一步聚类分析.

2. 随机选点法:当样本量 n 很大时,可以先随机抽选 p 个

样本,然后用系统聚类法将其分类,以每类的重心作为初始抽选样本聚点.如刚才的例子,我们可以以 60,80 为分界将三部分 $[0,59]$ $[60,79]$ $[80,100]$ 样本分别按系统聚类法聚类,取每类重心做为样本聚点,然后进入下一步.

3. 极小极大原则:要将 n 个样品分成 k 类,先取 x_{i_1}, x_{i_2} ,使得 $d(x_{i_1}, x_{i_2}) = d_{i_1 i_2} = \max_{i,j} \{d_{i,j} \mid i, j = 1, 2, \dots, n\}$.若已取 t 个聚点 $x_{i_1}, x_{i_2}, \dots, x_{i_t}$,则第 $t+1$ 个聚点为

$$\min_{r=1,2,\dots,t} d(x_{i_{t+1}}, x_{i_r}) = \max_{j \neq i_1, i_2, \dots, i_t} \{ \min_{r=1,2,\dots,t} d(x_j, x_{i_r}) \}$$

如此下去,直到选取第 k 个聚点为止.然后进入下一步.

4. 界值确定法:对于事先确定的正数 d ,把所有样品的重心作为第一聚点,然后把每一个样品输入,如果输入样本与已确定的聚点的距离大于 d ,则该样本作为一个新聚点,否则不计入聚点集.如此直到所有样本输完为止.

5. 密度法:人为确定两正数 $d_1, d_2 (d_2 > d_1)$,以每个样本点为球心,以 d_1 为半径,落在该球内的样本数称为该样本的密度.选取最大密度样本点作为第一个聚点,再选次大密度点,求该点与第一聚点的距离,如果该距离大于 d_2 ,选该点为第二聚点,否则不选入该点为聚点.依次按密度大小选下去.这样得到若干个两两距离都大于 d_2 的聚点集,然后进入下一步.

二、聚点选取后的初始分类原则

将所有样本逐个输入,计算该样本点到所有聚点的距离,将该样本点归入距离最小的聚点所在的类.

三、对初始分类的修改方法

1. 按批修改原则:

(i)选取一批聚点后,给出样本之间距离的定义.

(ii)按就近原则将所有样本归类.

(iii)计算每一类重心,以重心作为新一批聚点,再按就近原则

归类,当所有新的重心形成的聚点与上次聚点重合时,过程终止,动态聚类结束,否则回到(ii).

2. k -Means 方法的原则:

(i)确定三个数 k, C, R .

(ii)取前 k 个样品作为凝聚点,计算 k 个聚点间距离,如小于 C ,则将相应两聚点合并,用两聚点重心作为新聚点.重复以上步骤,直到所有聚点间距离大于 C 为止.

(iii)将余下 $n - k$ 个样本逐个输入,如样本与聚点之间的距离最小的大于 R ,则将该样本作为新聚点.如小于 R 则将该样本归入最近聚点所在的类.重新计算该类重心,并以此重心为新的聚点,回到(ii)步.

(iv)将所有样品输入,按(iii)中办法归类,直至新的分类与上次完全相同,聚类过程结束,否则重复(iv).

四、动态聚类法应用于综合评价的例子

例 4.2^[10] 从 12 个不同地区测量某树种的平均发芽率 y_1 与发芽势 y_2 ,其数据资料见表 4-4,试对该树种的培植问题给出综合评定.

解 我们选用欧氏距离,将这 12 个地区的种子按批修改法聚类.

(i)根据经验,取 12 个样本中的 $x_5 = (0.688, 0.605)^T$ 和 $x_{12} = (0.777, 0.723)^T$ 为聚点.

(ii)初始分类,顺序计算 $x_i, i \neq 5, 12$ 与聚点 x_5 与 x_{12} 的距离,若 $d(x_i, x_5) < d(x_i, x_{12})$,则归入 x_5 所在类,否则归入 x_{12} 所在类.现将结果列在表 4-5,得到两类:

$$\pi_1^{(0)} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_{10}, x_{11}\}$$

$$\pi_2^{(0)} = \{x_7, x_9, x_{12}\}$$

(iii)修改初始分类,计算 $\pi_1^{(0)}, \pi_2^{(0)}$ 的重心

$$\bar{X}_1^{(0)} = (0.647, 0.450)^T, \bar{X}_2^{(0)} = (0.823, 0.704)^T$$

现将 $\bar{X}_1^{(0)}$ 与 $\bar{X}_2^{(0)}$ 作为新的聚点, 将 $x_i (i=1, 2, \dots, 12)$ 与 $\bar{X}_1^{(0)}$ 和 $\bar{X}_2^{(0)}$ 的欧氏距离求出, 按就近原则将 $x_1 \sim x_{12}$ 分为两个新类:

$$\pi_1^{(1)} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_{10}\}$$

$$\pi_2^{(1)} = \{x_7, x_9, x_{11}, x_{12}\}$$

现再计算 $\pi_1^{(1)}$ 与 $\pi_2^{(1)}$ 的重心:

$$\bar{X}_1^{(1)} = (0.636, 0.433)^T, \bar{X}_2^{(1)} = (0.802, 0.672)^T$$

它们与 $\bar{X}_1^{(0)}$ 与 $\bar{X}_2^{(0)}$ 不同, 故应作为新的聚点, 再将 $x_i (i=1, 2, \dots, 12)$ 逐个计算到两聚点的距离, 按就近原则聚为两类:

$$\pi_1^{(2)} = \{x_1, x_2, x_3, x_4, x_6, x_8, x_{10}\},$$

$$\pi_2^{(2)} = \{x_5, x_7, x_9, x_{11}, x_{12}\}.$$

继续进行第三次修改, 修改后分类结果与第二次完全相同, 分类停止. 第三次修改结果列在表 4-5. 得最后分类为

$$\pi_1^{(3)} = \{x_1 \sim x_4, x_6, x_8, x_{10}\}$$

$$\pi_2^{(3)} = \{x_5, x_7, x_9, x_{11}, x_{12}\}$$

(iv) 综合评价: 从聚类分析的结果看, 显然 $\pi_1^{(3)}$ 中的样本表现出这些地区对该种树木生长有某些不适应性. 多数发芽率都不过 0.65, 其中虽然 1 号地区与 4 号地区表现出较高的发芽率, 但是发芽势较低, 表现了生长环境某些不适应性. 而 $\pi_2^{(3)}$ 类中样本就表现出了较好的性态. 如果培植此种树木应选择相应于 $\pi_2^{(3)}$ 中地区作为基地.

表 4-4 12 个地区某种树发芽情况

地区号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
	1	2	3	4	5	6	7	8	9	10	11	12
y_1	0.707	0.600	0.693	0.717	0.688	0.533	0.877	0.513	0.815	0.633	0.740	0.777
y_2	0.385	0.433	0.505	0.343	0.605	0.380	0.713	0.353	0.675	0.465	0.580	0.723

表 4-5 12 个样本点初始分类表及第三次分类表

距离 聚点	样本	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
$x_5 =$ $(0.688, 0.605)^T$		0.221	0.192	0.100	0.259	0.00	0.274	0.217	0.307	0.145	0.152	0.055	0.148
$x_{12} =$ $(0.777, 0.723)^T$		0.345	0.339	0.235	0.385	0.161	0.421	0.100	0.455	0.063	0.295	0.148	0.00
所属类		$\pi_1^{(0)}$	$\pi_1^{(0)}$	$\pi_1^{(0)}$	$\pi_1^{(0)}$	$\pi_1^{(0)}$	$\pi_1^{(0)}$	$\pi_2^{(0)}$	$\pi_1^{(0)}$	$\pi_2^{(0)}$	$\pi_1^{(0)}$	$\pi_1^{(0)}$	$\pi_2^{(0)}$
$\bar{X}_1^{(2)} =$ $(0.628, 0.409)^T$		0.084	0.032	0.114	0.110	0.205	0.100	0.392	0.126	0.326	0.055	0.205	0.348
$\bar{X}_2^{(2)} =$ $(0.779, 0.659)^T$		0.283	0.288	0.176	0.322	0.105	0.371	0.110	0.405	0.045	0.243	0.084	0.063
所属类		$\pi_1^{(3)}$	$\pi_1^{(3)}$	$\pi_1^{(3)}$	$\pi_1^{(3)}$	$\pi_2^{(3)}$	$\pi_1^{(3)}$	$\pi_2^{(3)}$	$\pi_1^{(3)}$	$\pi_2^{(3)}$	$\pi_1^{(3)}$	$\pi_2^{(3)}$	$\pi_2^{(3)}$

第三节 有序样本的综合评价聚类方法

对于聚类问题,有一种特殊的类型,它的聚类不能破坏对象之间原有的自然顺序.例如奥运会的 100 米男子的记录,从记录的数据来看,它有一个自然的顺序:第一届、第二届、……,因此聚类就是将这些记录按时间先后顺序分为几个阶段,每个阶段内部是相近的,不同的阶段有明显的差别.以百米成绩为例,可以分为 10 秒以上,10 秒附近,10 秒以内这几个阶段.又如一个地质剖面,按它的化学元素所占的比例可以将一层一层聚类,显然这种聚类不能破坏剖面上各层的自然顺序,也就是将剖面分割成几段,每一段由若干个层面组成.所以这种聚类的特点是将有序的对象按它的序分割成几段,段内差异小,段与段之间差异大,因此这种聚类称为分割.自然的序往往是以时间为参照,或以位置为参考,总之是按某个量的大小来排序.下面就介绍有序样本的聚类.

所谓有序样本的聚类问题,就是要寻求一种分割,使得分割后所形成的样本“段”的段内差异尽可能地小,各段之间的差异尽可能地大,这种分割称为相应给定分段数下的最优分割.为了寻求这种最优分割,我们先引入几个概念.

一、几个基本概念

设给了有序样本 x_1, x_2, \dots, x_n , 每个样本 x_i 都是 p 维向量. 我们定义, $G_{ij} \triangleq \{x_i, x_{i+1}, \dots, x_j\}$, $i < j$ 是一类, 类 G_{ij} 的直径 $D_{(ij)} \triangleq \sum_{k=i}^j (x_k - \bar{x}_{ij})^T (x_k - \bar{x}_{ij})$, 其中 $\bar{x}_{ij} = \frac{1}{j-i+1} \sum_{k=i}^j x_k$, 即段内均值.

(i) 目标函数

将 n 个样本分成 k 类记为 $P(n, k)$, 它使得 n 个样本在原序结构上分成 k 段 $\{x_1, x_2, \dots, x_{i_2-1}\}, \{x_{i_2}, x_{i_2+1}, \dots, x_{i_3-1}\}, \dots, \{x_{i_k}, \dots, x_n\}$, 其中 $1 < i_2 < \dots < i_k \leq n$, 定义在 $P(n, k)$ 给定下的目标函数为(注意 $i_1 = 1$)

$$E(P(n, k)) \triangleq \sum_{l=1}^k D(i_l, i_{l+1} - 1)$$

显然从以上定义和方差分析的概念可知类内直径实质上就是类内变差的度量, 目标函数就是总体分类后全部 k 个类的类内变差总和. 分类愈合理, 则 $E[P(n, k)]$ 值愈小. 因此我们所求的分类应是使得该分类的目标函数达到极小.

(ii) 最优分割即目标函数的最优解的特性

$$E_0[P(n, k)] = \min_{k \leq j \leq n} \{E_0[P(j-1, k-1)] + D(j, n)\}$$

其中

$$\begin{aligned} E_0[P(n, k)] &= \min_{P(n, k)} E[P(n, k)] \\ &= \text{最优 } k \text{ 分割相应的目标函数值} \end{aligned}$$

(iii) 最优解 $E_0[P(n, k)]$ 的求法

当 n, k 固定时, 首先寻找 i_k (最后一段第一样本), 使得

$E_0[P(i_k - 1, k - 1)] + D(i_k, n)$ 达到极小,即

$$\begin{aligned} & E_0[P(i_k - 1, k - 1)] + D(i_k, n) \\ &= \min_{k \leq j_k \leq n} \{E_0[P(j_k - 1, k - 1)] + D(j_k, n)\} \end{aligned}$$

然后再来寻找 i_{k-1} ,使得

$$\begin{aligned} & E_0[P(i_{k-1} - 1, k - 2)] + D(i_{k-1}, i_k) \\ &= \min_{k-1 \leq j_{k-1} \leq i_k} \{E_0[P(j_{k-1} - 1, k - 2)] + D(j_{k-1}, i_k)\} \end{aligned}$$

依此类推,找出 $i_{k-2}, i_{k-3}, \dots, i_2, i_1 (=1)$,故得最优 k 分割.

(iv)最优分割数 k 的确定

关于 k 的取值通常有两种办法,一是按经验人为地确定 k 值,或按最优二分割、三分割方法,直到能说明问题为止;^[3]二是给定一个阈值 $\delta > 0$,使得 $|E_0[P(n, k - 1)] - E_0[P(n, k)]| < \delta$ 时,则满足该式的 k 值即为最小分类数 k .

二、有序样本综合聚类评价的方法与步骤

以下我们通过具体实例来说明这一方法的步骤.

例 4.3^[11] 根据某地云杉生长过程得出140年之内每10年直径生长量 x_1 ,与每10年树高生长量 x_2 ,数据见表4-6.

表 4-6 云杉的直径和树高每10年生长量

树龄	10	20	30	40	50	60	70	80	90	100	110	120	130	140
x_1	0.2	1.4	2.8	3.0	2.8	2.6	2.7	1.5	2.4	1.4	1.8	1.5	1.7	1.0
x_2	0.8	1.5	1.7	2.2	1.6	1.8	1.5	1.0	1.5	1.3	1.2	1.1	0.1	0.7

第一步 计算14个样本(为了方便将树龄按从左到右将样本编号为1,2, ..., 14).一切可能的样本段直径

$$d_{i,j} = \sum_{k=i}^j \sum_{l=1}^p [x_{kl} - \bar{x}_l(i,j)]^2$$

这里 $p=2, i, j=1, 2, \dots, 14$,结果列于下表4-7.

第二步 计算最小目标函数 $E[P(m, t)]$.

表 4-7 14 个样品一切可能的样品段的直径 $d_{i,j}$

$\begin{matrix} j \\ i \end{matrix}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0													
2	0.965	0												
3	3.833	1.000	0											
4	6.160	1.780	0.145	0										
5	6.884	1.930	0.233	0.200	0									
6	7.193	1.940	0.288	0.267	0.040	0								
7	7.477	2.023	0.380	0.375	0.067	0.050	0							
8	8.224	3.409	2.227	2.156	1.448	1.213	0.845	0						
9	8.291	3.420	2.266	2.180	1.448	1.230	0.947	0.530	0					
10	8.845	4.389	3.495	3.294	2.308	1.896	1.428	0.733	0.520	0				
11	8.993	4.725	3.940	3.664	2.523	2.022	1.472	0.738	0.553	0.085	0			
12	9.396	5.391	4.734	4.349	3.004	2.366	1.682	0.816	0.695	0.107	0.050	0		
13	11.125	7.368	6.765	6.245	4.582	3.789	2.877	1.862	1.804	1.028	0.787	0.520	0	
14	12.394	8.975	8.489	7.787	5.820	4.802	3.660	2.397	2.382	1.356	1.128	0.767	0.425	0

表 4-8 最小目标函数 $E_0(P(m, t))$

样品数 m 分割数 t	2	3	4	5	6	7	8	9	10	11	12	13	14
2	0.00	0.965 (2)	1.110 (2)	1.198 (2)	1.252 (2)	1.345 (2)	3.191 (2)	3.230 (2)	4.388 (1)	4.725 (1)	5.390 (1)	7.369 (1)	8.975 (1)
3		0.00	0.145 (2)	0.233 (2)	0.287 (2)	0.380 (2)	1.345 (7)	1.875 (7)	2.078 (7)	2.082 (7)	2.161 (7)	3.206 (7)	3.742 (7)
4			0.00	0.145 (4)	0.185 (4)	0.211 (4)	0.380 (7)	0.910 (7)	1.113 (7)	1.117 (7)	1.196 (7)	2.161 (12)	2.586 (12)

(表的以下部分略去)

首先计算分割数 $t=2$, 当 $m=2$ 时, 显然 $E_0[P(2,2)]=0$; 当 $m=3, t=2$ 时, 前三个样品分成二段, 有两种分法: $\{1\}, \{2,3\}$ 及 $\{1,2\}, \{3\}$ 这时

$$\begin{aligned} E_0[P(3,2)] &= \min_{2 \leq j \leq 3} (d_{1,j-1} + d_{j,3}) \\ &= \min[(d_{1,1} + d_{2,3}), (d_{1,2} + d_{3,3})] \\ &= \min[(0 + 1.0)(0.965 + 0)] = 0.965 \end{aligned}$$

因此对于 $m=3, t=2$ 时, 最优二分割点在第二个样本之后, 并用小括号列在表 4-8 相应的位置上. 同理对于 $m=4$, 有

$$\begin{aligned} E_0[P(4,2)] &= \min_{2 \leq j \leq 4} (d_{1,j-1} + d_{j,4}) \\ &= \min[(0 + 1.780), (0.965 + 0.145), (3.833 + 0)] \\ &= 1.110 \end{aligned}$$

即分割点在第二个样本之后, 并填入表 4-8, 同理得 $m=5, 6, \dots, 14$ 的最优二分割的 $E_0[P(m,2)]$, 填入表 4-8 的第一行.

接着计算第二行, 即分割数 $t=3$. 当 $m=3$ 时, 显然

$$E_0[P(3,3)]=0$$

当 $m=4$ 时, 有

$$\begin{aligned} E_0[P(4,3)] &= \min\{[E_0(P(2,2)) + d_{3,4}], [E_0(P(3,2)) + d_{4,4}]\} \\ &= \min\{(0 + 0.145), (0.965 + 0)\} \\ &= 0.145 \end{aligned}$$

可见前 4 个样品的最优分割是 $\{1\}, \{2\}, \{3,4\}$, 第二个分割点在第二个样本之后, 并将 0.145 及第二分割点填在表 4-8 第二行相应位置. 同理可得第三行, 第四行, …… , 第十三行, 并将所有数据填入表 4-8.

第三步 分段. 先找 14 个样本三分段的最优三分割点. 即找 j_3 满足

$$\begin{aligned} E_0[P(14,3)] &= E_0[P(j_3 - 1, 2)] + d_{j_3, 14} \\ &= \min_{j \leq i_3 \leq 14} [E_0(P(i_3 - 1, 2)) + d_{i_3, n}] \end{aligned}$$

从表中可得三分割为 $\{1,2\}, \{3 \sim 7\}, \{8 \sim 14\}$.

同样可得 14 个样本点的最优四分割为

$$\{1,2\}, \{3 \sim 7\}, \{8 \sim 12\}, \{13,14\}$$

第四步 综合评价分析.

从表 4-6 的原始数据我们可以看到如果就从云杉生长作为木材原料来讲分为三段是较合理的. 如果是作为研究植物生长周期或其他生化问题, 也许是分为四个阶段比较好, 在这里我们不谈这个问题. 显然, 最优三分割提供的三个时期, 反映了杉木成长的三个特殊时期, 前 20 年生长缓慢, 30 到 70 年生长迅速是云杉的成才期, 这些为我们开发利用杉木提供了科学依据.

第四节 综合评价的距离判别分析

判别分析是根据样本的评价指标的观察值来推断该样本的所属总体, 它以有训练样本为前提. 比如本章一开始所讲到的, 在地质找矿中我们要根据异常点的地质结构、化探和物探的各项化验指标来判断该异常点属于哪一种矿化类型. 在国民经济宏观管理中, 人们要通过产业间各项经济指标的观察值, 判定当前经济领域中宏观经济运行是否正常, 以便制定各项经济宏观政策. 这些都涉及判别分析问题, 它给综合评价第三类问题的解决提供了科学的数值分析方法.

综合评价的距离判别分析, 用统计的语言说, 就是已知有 q 个总体 G_1, G_2, \dots, G_q , 它们的分布函数为 $F_1(x), \dots, F_q(x)$, 其中 $F_i(x), i=1, 2, \dots, q$ 都是与 p 个评价指标相对应的 p 维函数. 因此这 q 个总体的统计规律便由这 q 个分布函数所决定. 对于当前一个样本 x_1, x_2, \dots, x_n , 其中每个 $x_i (i=1, 2, \dots, n)$ 都是 p 维向量, 也就是 p 个评价指标的观察值. 现在我们要判定这 n 个样本分属于哪些总体. 当然我们希望这个判别准则应该是某种规则下的最优判别. 如错判的概率最小, 或错判损失最小等. 下面我们来介绍几个常用的方法.

一、距离判别分析在两总体下的综合评价应用

设有两个总体即 $q = 2$ 时, 无妨设两总体协差阵相同都是 $\Sigma (i > 0)$, 且两总体 G_1 与 G_2 的均值向量为 μ_1, μ_2 . 那么对于一个样本 $x_{p \times 1}$ 要判断它属于哪个总体, 应看它与哪个总体最相似, 或来自哪个总体的概率最大. 如假设检验就是以极大似然为基础的, 本节是以“距离”作为相似性度量, 关于统计距离在本章第一节已有全面讲述, 这里我们采用 B 模下的马氏距离. 定义样本 x 到 G_1 与 G_2 的距离为 $d(x, \mu_1)$ 与 $d(x, \mu_2)$, 则建立判别规则应该是

$$\begin{cases} x \in G_1, \text{当 } d(x, \mu_1) < d(x, \mu_2) \\ x \in G_2, \text{当 } d(x, \mu_2) < d(x, \mu_1) \end{cases} \quad (4-1)$$

将(4-1)式简化一下就有

$$\begin{aligned} d^2(x, \mu_1) - d^2(x, \mu_2) &= (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) \\ &\quad - (x - \mu_2)^T \Sigma^{-1} (x - \mu_2) \\ &= -2[x - (\mu_1 + \mu_2)/2]^T \Sigma^{-1} (\mu_1 - \mu_2) \\ &\triangleq -2W(x) \end{aligned}$$

即

$$W(x) = (x - \bar{\mu})^T \Sigma^{-1} (\mu_1 - \mu_2)$$

其中 $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$. 这时判别准则(4-1)式就成为

$$\begin{cases} x \in G_1, & W(x) > 0 \\ x \in G_2, & W(x) < 0 \end{cases} \quad (4-2)$$

显然 $W(x)$ 是 x 的线性函数, 故也称为线性判别函数. 这样给出一个样本观察值向量 x , 代入(4-2)很快就可以做出综合判定, 给出样本的综合评价. 但在这里有必要对评价效果给出说明, 正如统计假设检验中两类错误的问题, 这种距离判别法也会出错.

(i) 以马氏距离建立的最小距离判别分析是以总体有显著差异为前提的. 为了说明问题, 我们设 $G_1 \sim N(\mu_1, \delta^2)$, $G_2 \sim N(\mu_2, \delta^2)$, 即两总体为一维正态总体. 如果已知样本 x 是来自总体 G_1

的,但若它的取值落在 $\bar{\mu}$ 的右侧则由(4-2)的准则应判为 G_2 的总体,即出现误判.虽然在总体有显著差异条件下(图4-3)这种误判概率很小,但当总体差异不很显著时(图4-4),则很大.在实际生活中,综合评价结果对于两总体的错判情况往往具有很大的差异,为了控制犯某种错误的概率,可以将判别限(在(4-2)中是 $\bar{\mu}$)向另一方倾斜(图4-5).可以证明在正态总体下(4-2)式所定义的线性判别函数对于两种误判的概率是相同的.

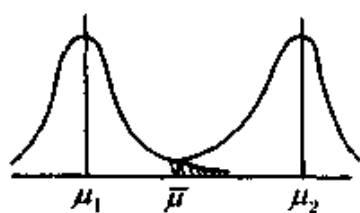


图4-3

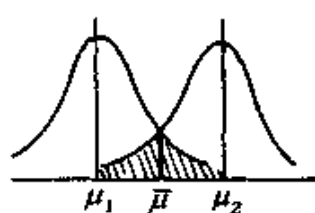


图4-4

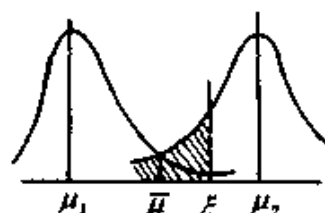


图4-5

(ii)为了减少 $\bar{\mu}$ 点附近判错的情况,我们可以通过划定待判区域,减少误判的发生.如取 $a < \bar{\mu} < b$,将判别准则定为

$$\begin{cases} x \in G_1, & \text{当 } x \leq a \\ x \in G_2, & \text{当 } x \geq b \\ \text{待判}, & \text{当 } a < x < b \end{cases}$$

(iii)当总体分布类型未知时,从(4-1)、(4-2)可知,只要总体二阶矩存在即可进行.如果总体二阶矩也未知,则可以利用样本来估计总体的一、二阶矩.设 $x_i^{(1)}, i=1,2,\dots,n_1, x_i^{(2)}, i=1,2,\dots,n_2$ 分别为总体 G_1 与 G_2 的样本,由多元统计分析知道,令

$$\bar{x}^{(1)} = \frac{1}{n} \sum_{i=1}^{n_1} x_i^{(1)}, \quad \bar{x}^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} x_i^{(2)}$$

$$A_1 = \sum_{i=1}^{n_1} (x_i^{(1)} - \bar{x}^{(1)})(x_i^{(1)} - \bar{x}^{(1)})^T$$

$$A_2 = \sum_{i=1}^{n_2} (x_i^{(2)} - \bar{x}^{(2)})(x_i^{(2)} - \bar{x}^{(2)})^T$$

则 Σ 可以用

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2} (A_1 + A_2)$$

来估计. 于是可以建立判别函数

$$W(x) = \left[x - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)}) \right]^T \hat{\Sigma}^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) \quad (4-3)$$

剩余的判别法则同(4-2).

(iv) 若 G_1 与 G_2 的协方差阵有显著差异, 则(4-2)就成为

$$\begin{aligned} W(x) &= d^2(x, \mu_1) - d^2(x, \mu_2) \\ &= (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \end{aligned} \quad (4-4)$$

该式确定的是 x 的二次函数.

二、多总体距离判别的综合评价方法

假设有 q 个总体 G_1, G_2, \dots, G_q . 它们有公共的正定协方差阵和不同的均值向量 $\mu_i, i = 1, 2, \dots, q$. 建立判别函数, 我们仍采用 B 模下的马氏距离

$$W_{ij}(x) = \left[x - \frac{1}{2}(\mu_i + \mu_j) \right]^T \Sigma^{-1} (\mu_i - \mu_j) \quad i, j = 1, 2, \dots, q$$

现在建立判别规则:

$$x \in G_i, \quad \text{如果 } x \in D_i, \quad i = 1, 2, \dots, q \quad (4-5)$$

其中

$$D_i = \{x \mid W_{ij}(x) > 0, \text{ 对一切 } j \neq i \text{ 成立}\} \quad i = 1, 2, \dots, q$$

对于 D_i 的边界点归入 D_i 的下指标最小的区域. 显然以上规则构成了对 p 维空间的一个剖分.

其次, 如果总体 G_1, G_2, \dots, G_q 的二阶矩不知道, 则可以利用样本矩来估计总体矩. 设从 q 个总体中分别抽取 n_1, n_2, \dots, n_q 个样本, 定义

$$\bar{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad i = 1, 2, \dots, q$$

x_{ij} 为来自第 i 个总体中的第 j 个样本.

$$S_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{\mu}_i)(x_{ij} - \bar{\mu}_i)^T, \quad i = 1, 2, \dots, q$$

$$\hat{\Sigma} = \frac{1}{n - q} \sum_{i=1}^q S_i \quad (4-6)$$

$$n = n_1 + \dots + n_q$$

则相应的判别规则为

$$W_{ij}(x) = \left[x - \frac{1}{2}(\bar{\mu}_i + \bar{\mu}_j) \right]^T \hat{\Sigma}^{-1} (\bar{\mu}_i - \bar{\mu}_j)$$

$$x \in G_i, \text{ 若 } x \in D_i$$

其中

$$D_i = \{x \mid W_{ij}(x) > 0, \text{ 对一切 } j \neq i\} \quad i = 1, 2, \dots, q \quad (4-7)$$

最后, 如果 q 个总体协差阵不同, 如设 μ_i, Σ_i 分别为 G_i 的期望值向量与协差阵, $i = 1, 2, \dots, q$. 则直接由下式确定判别准则:

$$\begin{cases} x \in G_i, \text{ 若 } d^2(x, \mu_i) = \min_{1 \leq j \leq q} d^2(x, \mu_j) \\ i = 1, 2, \dots, q \\ \text{对于 } d^2(x, \mu_i) = d^2(x, \mu_j) \text{ 时, } x \text{ 归 } \mu \text{ 的下指标最小的一类} \end{cases} \quad (4-8)$$

显然在(4-7)与(4-8)式下, 分别构成了 p 维空间的一个剖分, 因而判别规则完备.

三、距离判别分析下综合评价的应用案例

例 4.4^[11] 在某地区的沙岩中, 采集了 56 个样本, 根据它们的化学成分, 可将每个样本对应于如下三总体:

G_1 : wilhilm 沙岩;

G_2 : 低 mulinia 沙岩;

G_3 : 上沙岩.

对每个样本考虑五个含量指标: x_1 : 钒; x_2 : (铁) $^{\frac{1}{2}}$; x_3 : (铍) $^{\frac{1}{2}}$; x_4 :

(饱和烃)⁻¹; x_5 : 芳香烃. 前三个为痕量元素, 后两个变量是从气相层析曲线段确定的. $n_1=7, n_2=11, n_3=38$ 是通过 56 个样本的聚类形成的对应于以上三总体的样本个数. 从该样本中可以得到如下估计量

$$\hat{\mu}'_1 = (3.229, 6.589, 0.303, 0.150, 11.540)$$

$$\hat{\mu}'_2 = (4.445, 5.667, 0.344, 0.157, 5.484)$$

$$\hat{\mu}'_3 = (7.226, 4.634, 0.598, 0.223, 5.768)$$

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 + n_3 - 3} A$$

$$A = \sum_{i=1}^3 \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_i)(x_{ij} - \hat{\mu}_i)^T$$

$$= \begin{bmatrix} 187.575 & & & & * \\ 1.957 & 41.789 & & & \\ -4.031 & 2.128 & 3.580 & & \\ 1.092 & -0.143 & -0.284 & 0.077 & \\ 79.672 & -28.243 & 2.559 & -0.996 & 338.023 \end{bmatrix}$$

由多元统计分析知道 $\hat{\Sigma}$ 是三总体等协差阵的无偏估计.

现对于该地区的任一样本 x , 可以建立如下判别函数与判别规则:

$$\begin{cases} x \in G_i, \text{ 若 } x \in D_i = \{x \mid W_{ij}(x) > 0, \text{ 对一切 } j \neq i\}, \\ \quad i = 1, 2, 3 \\ W_{ij}(x) = \left[x - \frac{1}{2}(\hat{\mu}_i + \hat{\mu}_j) \right]^T \cdot \hat{\Sigma}^{-1}(\hat{\mu}_i - \hat{\mu}_j), \\ \quad i, j = 1, 2, 3 \end{cases}$$

这只要求出 $\hat{\Sigma}^{-1}$ 即可进行判别, 从而通过判别对 x 的总体沙岩型给出综合评定.

例 4.5^[3] 生物统计学家经常提出由个体的“大小”及“形状”因子来进行判别精神病与非精神病人, 即 $x = (x_1, x_2)^T$. 现从正常人与非正常人(即精神病患者)中各取 25 个样本得如下结果:

$$\bar{x}_1 = \frac{1}{25} \sum_{i=1}^{25} x_{1i} = (20.80, 12.32)^T$$

$$\bar{x}_2 = \frac{1}{25} \sum_{i=1}^{25} x_{2i} = (12.80, 36.40)^T$$

$$S_1 = \frac{1}{24} A_1 = \begin{pmatrix} 6.90 & -5.27 \\ -5.27 & 40.89 \end{pmatrix}$$

$$S_2 = \frac{1}{24} A_2 = \begin{pmatrix} 36.75 & 13.92 \\ 13.92 & 281.92 \end{pmatrix}$$

$$\hat{\Sigma} = \frac{A_1 + A_2}{n_1 + n_2 - 2} = \begin{pmatrix} 21.83 & 4.33 \\ 4.33 & 164.40 \end{pmatrix}$$

由以上条件我们可以求得

$$\hat{\Sigma}^{-1} = \frac{1}{3570.1} \begin{pmatrix} 164.40 & -4.33 \\ -4.33 & 21.83 \end{pmatrix}$$

$$\bar{x}_1 - \bar{x}_2 = (8.00, -24.08)^T$$

$$\bar{x}_1 + \bar{x}_2 = (33.60, 48.72)^T$$

代入判别函数

$$W(x) = \left[x - \frac{1}{2}(\bar{\mu}_1 + \bar{\mu}_2) \right]^T \hat{\Sigma}^{-1}(\bar{\mu}_1 - \bar{\mu}_2)$$

得

$$\begin{aligned} W(x) &= \frac{1}{3570.1} (1419x_1 - 560x_2 - 10198) \\ &\approx \frac{280}{3570.1} (5x_1 - 2x_2 - 36) \end{aligned}$$

令 $W^*(x) = 5x_1 - 2x_2 - 36$, 显然 $W(x)$ 与 $W^*(x)$ 的判别法则等效, 即 $W^*(x) > 0 \Leftrightarrow W(x) > 0$.

现在对新的样本 x 比如经检查, 若 $x = (28, 36)^T$ 代入 $W^*(x)$ 式有

$$W^*(x) = 22 > 0$$

故当前样本 $x = (28, 36)^T$ 是正常人.

注意: 该例题是通过病理指标选做主因子分析, 然后按前两个

主因子来解释“大小”和“形状”的. 一般情况下, 我们也可以直接构造该病理指标下的距离判别分析, 只不过运算复杂一些.

第五节 贝叶斯(Bayes)判别

前面介绍的各种综合评判问题, 都是从技术上给出综合评价的方法, 并没有考虑综合评价产生偏误要造成损失的问题. 如对外贸易中生产线的整套贸易, 如果将重要技术指标的合格与否产生误判, 则引起的经济损失是巨大的. 再比如药品检验中, 如果将有毒药品检验为无毒药品, 则后果是不堪设想的. 因此引进考虑误判损失的综合评价方法是十分有益的. 贝叶斯综合评价方法正是利用贝叶斯统计推断的方法, 通过建立损失函数求得综合评价问题的贝叶斯解, 给出综合评价结果. 下面我们先从贝叶斯的理论模型谈起, 在这里我们只给出结果, 对理论有兴趣的读者请参阅文献[3, 11].

一、Bayes 综合评价模型简介

设有 q 个总体 G_1, G_2, \dots, G_q , 分别是具有 p 维分布密度函数 $p_1(x), p_2(x), \dots, p_q(x)$ 的 p 元总体. D_1, D_2, \dots, D_q 是对 p 维空间 R^p 的一个剖分, 即它们互不相交, 所有 q 个区域构成 p 维空间 R^p .

设 $L(i, j)$ 表示样本来自 G_i 而判为属于 G_j 的损失. 这一误判发生的概率为

$$P(j|i) = \int_{D_j} p_i(x) dx \quad (j \neq i)$$
$$i, j = 1, 2, \dots, q \quad (4-9)$$

设 q 个总体出现的先验概率为 $\pi_1, \pi_2, \dots, \pi_q$. 则利用 Bayes 统计推断, 通过划分 D_1, D_2, \dots, D_q 来判别样本 x 归属的平均损失(expected cost of misclassification)ECM(\mathcal{D})为

$$\text{ECM}(\mathcal{D}) \triangleq g(D_1, D_2, \dots, D_q) = \sum_{k=1}^q \pi_k \sum_{j \neq k}^q L(k, j) P(j | k) \quad (4-10)$$

所谓贝叶斯方法就是选择剖分 \mathcal{D}^* , 使得

$$\text{ECM}(\mathcal{D}^*) = \min_{\mathcal{D}} \text{ECM}(\mathcal{D}) \quad (4-11)$$

满足(4-11)的剖分 \mathcal{D}^* 就是该判别问题的贝叶斯解. 特别地如果总体先验分布 $\pi_1, \pi_2, \dots, \pi_q$ 与总体分布密度 $p_1(x), \dots, p_q(x)$ 以及损失函数 $L(i, j), i, j = 1, 2, \dots, q$ 都是给定的, 则称 $h_l(x) =$

$\sum_{\substack{i=1 \\ i \neq l}}^q \pi_i p_i(x) L(i, l)$ 为后验损失 ($l = 1, 2, \dots, q$). 这时(4-11)式相应的解等价于

$$\text{ECM}(\mathcal{D}^*) = \{D_1^*, D_2^*, \dots, D_q^*\}$$

其中

$$\left. \begin{aligned} D_l^* &= \{x: h_l(x) < h_i(x) \quad i = 1, 2, \dots, q, i \neq l\} \\ l &= 1, 2, \dots, q \end{aligned} \right\} \quad (4-12)$$

由(4-12)给出的评价结果正是我们所考虑的加上损失后的综合评价的贝叶斯解. 下面先从两总体谈起.

二、两总体下的 Bayes 综合评价模型

如果综合评价的目标仅是“是”与“非”的评价过程, 一般总可以归结为一个两总体的问题. 如果我们已知总体 G_1, G_2 的分布密度为 $p_1(x)$ 与 $p_2(x)$. 定义损失函数为 $L(1, 2), L(2, 1)$, 总体的先验概率 π_1, π_2 为已知, 则利用上段的论述, 贝叶斯解 \mathcal{D}^* 为

$$\begin{aligned} \mathcal{D}^* &= (D_1^*, D_2^*) \\ D_1^* &= \left\{x: \frac{p_1(x)}{p_2(x)} \geq \frac{L(1, 2)}{L(2, 1)} \cdot \frac{\pi_2}{\pi_1}\right\} \\ D_2^* &= \left\{x: \frac{p_1(x)}{p_2(x)} < \frac{L(1, 2)}{L(2, 1)} \cdot \frac{\pi_2}{\pi_1}\right\} \end{aligned} \quad (4-13)$$

式(4-13)的证明略, 有兴趣的读者可参考文献[3]. 这样就得到贝

叶斯判别规则:

$$\begin{cases} x \in G_1, \text{若 } x \text{ 使得 } \frac{p_1(x)}{p_2(x)} \geq \frac{L(1,2)}{L(2,1)} \frac{\pi_2}{\pi_1} \\ x \in G_2, \text{若 } x \text{ 使得 } \frac{p_1(x)}{p_2(x)} < \frac{L(1,2)}{L(2,1)} \frac{\pi_2}{\pi_1} \end{cases}$$

特别地,如果两总体 G_1, G_2 分别为正态总体 $N(\mu^{(i)}, \Sigma_i) i=1, 2$, 则有一般的判别公式:

(i) 若 $\Sigma_1 = \Sigma_2 = \Sigma > 0$, 则

$$\begin{cases} D_1^* = \{x: W(x) \geq \beta\} \\ D_2^* = \{x: W(x) < \beta\} \\ W(x) = \left[x - \frac{1}{2}(\mu^{(1)} + \mu^{(2)}) \right]^T \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \quad (4-14) \\ \beta = \ln \frac{L(1,2)\pi_2}{L(2,1)\pi_1} \end{cases}$$

如果总体参数 $\mu^{(1)}, \mu^{(2)}, \Sigma_1, \Sigma_2$ 未知, 可用样本估计值代替总体的值.

(ii) 若 $\Sigma_1 \neq \Sigma_2, \Sigma_1 > 0, \Sigma_2 > 0$, 则在(4-14)式中, 只要将相应 $W(x)$ 与 β 作适当修改, 评价规则不变.

$$\begin{cases} D_1^* = \{x: W(x) \geq \beta\} \\ D_2^* = \{x: W(x) < \beta\} \\ W(x) = -\frac{1}{2}x^T(\Sigma_1^{-1} - \Sigma_2^{-1})x + (\mu^{(1)T}\Sigma_1^{-1} - \mu^{(2)T}\Sigma_2^{-1})x \\ \beta = \ln \frac{L(1,2)\pi_2}{L(2,1)\pi_1} + \frac{1}{2} \ln \left| \frac{\Sigma_1}{\Sigma_2} \right| \\ \quad + \frac{1}{2}(\mu^{(1)T}\Sigma_1^{-1}\mu^{(1)} - \mu^{(2)T}\Sigma_2^{-1}\mu^{(2)}) \end{cases} \quad (4-15)$$

同(i), 如果总体的参数未知, 可用样本估计值去代替.

三、多总体下的贝叶斯综合评价方法

在这里我们只以正态总体为例. 设 G_1, G_2, \dots, G_q 都是正态总体 $N_p(\mu_i, \Sigma_i), i=1, 2, \dots, q, (\Sigma_i > 0)$ 无妨设误判损失为 1, 则利用多元正态密度函数

$$f_i(x) = (2\pi)^{-\frac{p}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\}$$

$$i = 1, 2, \dots, q$$

以及 q 个总体的先验概率为 $\pi_1, \pi_2, \dots, \pi_q$, 可得判别规则为

$$\begin{cases} x \in G_i, \text{若样本 } x \text{ 使得 } h_i(x) = \min_{1 \leq j \leq q} h_j(x) \\ \text{当 } x \text{ 属于几个 } G_i \text{ 时, 判归下标最小的一个} \end{cases} \quad (4-16)$$

因为

$$L(i, j) = 1$$

故

$$\begin{aligned} h_i(x) &= \sum_{\substack{j=1 \\ j \neq i}}^q p_j f_j(x) \quad (i = 1, 2, \dots, q) \\ &= \sum_{j=1}^q p_j f_j(x) - p_i f_i(x) \end{aligned}$$

这时 $\min_{1 \leq j \leq q} h_j(x) \Leftrightarrow \max_i p_i f_i(x)$. 这正是最大后验规则. 在正态总体就成为

$$\begin{aligned} \max_j \ln[p_j f_j(x)] &= \max_j \left[\ln \pi_j - \frac{1}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_j| \right. \\ &\quad \left. - \frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) \right] \end{aligned}$$

记

$$d_k(x) = \ln p_k - \frac{1}{2} [\ln |\Sigma_k| + (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)]$$

$$k = 1, 2, \dots, q$$

显然 $d_k(x)$ 综合了广义方差 $|\Sigma_k|$, 先验概率 p_k , 以及 x 到 μ_k 的马氏距离对判别的“贡献”. 如果误判损失不为 1 则还有损失约束. 故称 $d_k(x)$ 为样本 x 在第 k 个总体的二次判别得分. (4-16) 式的

判别规则就成为

$$x \in G_j, \quad d_j(x) = \max_{1 \leq i \leq q} \{d_i(x)\} \quad (4-17)$$

若 x 属于几个 G_i , 则判归下标最小者. 在实际应用时, 多数情况下 μ_i 与 $\Sigma_i, i=1, 2, \dots, q$ 是未知的. 可以利用样本先对总体参数做出估计.

三、例题

例 4.6^[10] 对三种鸢尾花分别抽取 $n_1 = n_2 = n_3 = 50$ 的样本, 对这种植物的两项指标, $x_1 =$ 花瓣宽, $x_2 =$ 花萼宽进行测量, 所得结果估计三总体的均值向量与协方差矩阵为

$$G_1: \mu_1 = (3.46, 0.25)^T, \quad \hat{\Sigma}_1 = \begin{pmatrix} 0.0661 & 0.0298 \\ 0.0298 & 0.0601 \end{pmatrix}$$

$$G_2: \mu_2 = (2.77, 1.30)^T, \quad \hat{\Sigma}_2 = \begin{pmatrix} 0.0711 & 0.0333 \\ 0.0333 & 0.0400 \end{pmatrix}$$

$$G_3: \mu_3 = (2.79, 2.01)^T, \quad \hat{\Sigma}_3 = \begin{pmatrix} 0.0588 & 0.0323 \\ 0.0323 & 0.0810 \end{pmatrix}$$

试判定在损失为 1 的误判函数定义下, 样本 $(2.8, 1.5)^T$ 应属哪个总体.

解 假设三总体具有相同协方差阵, 且都服从二维正态分布. 这时利用相同协方差阵多总体的协方差矩阵的估计

$$\begin{aligned} \hat{S} &= \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + (n_3 - 1)S_3}{n_1 + n_2 + n_3 - 3} \\ &= \frac{49}{147}(\hat{\Sigma}_1 + \hat{\Sigma}_2 + \hat{\Sigma}_3) = \begin{pmatrix} 0.0653 & 0.0318 \\ 0.0318 & 0.0604 \end{pmatrix} \end{aligned}$$

就得

$$\hat{S}^{-1} = \frac{1}{0.0029} \begin{pmatrix} 0.0604 & -0.0318 \\ -0.0318 & 0.0653 \end{pmatrix}$$

现对先验信息作为无信息先验处理, 利用贝叶斯假设应服从离散均匀分布. 利用(4-17)式有三总体二次判别:

$$d_1(x) = 0.2009x_1 + (-0.0937)x_2 - 0.3358$$

$$d_2(x) = 0.1259x_1 + 0.0032x_2 - 0.1723$$

$$d_3(x) = 0.1045x_1 + 0.0425x_2 - 0.1885$$

将样本 $x_0 = (2.8, 1.5)^T$ 代入上式三者得分

$$d_1(x_0) = 0.0861$$

$$d_2(x_0) = 0.1754$$

$$d_3(x_0) = 0.1679$$

利用判别规则(4-17)式,应判 $x_0 \in G_2$. 对样本的综合评定给出了结果.

第六节 费歇判别

在综合评价的实践中,针对不同的评价问题和评价指标的不同特性,要选择适当的评价方法,以提高评价效用. 对于某些评价问题,由于人们的认识已有相当的深度,指标之间的关系以及评价指标对于评价目标的反映都是比较清楚的,这类评价问题用第二章常规方法就可以得到较好的评价效果. 对于另一些相对认识不够深入的领域,由于人们对设置的评价指标以及评价指标所反映的关系认识非常浅薄,只有通过评价客体的数量关系,才有可能帮助人们从变量规律的认识,达到对事物质的认知. 这就使得以数值计算与数值分析为基础的多元统计分析显示出在综合评价中应用的能力. 本章前几节就是从统计距离入手对综合评价的第一、第三类问题的评价方法给出了初步性的介绍. 这一节是以方差分析的思想为原则,介绍判别分析在另一视角下,是如何进行综合评价的.

一、费歇(Fisher)判别分析的准则

设有 q 个总体 G_1, G_2, \dots, G_q , 相应的均值向量和协方差阵为 $\mu_1, \mu_2, \dots, \mu_q; \Sigma_1, \Sigma_2, \dots, \Sigma_q$. 正像距离判别和 Bayes 判别一样, 我们也希望建立一个线性判别函数, 即对于样本 x 有常数向量 c ,

形成判别函数 $C(x) = c^T \cdot x$, 则有 $x \in G_i$ 时令 $c(x)$ 的期望和方差分别为 e_i 与 v_i^2 , 即

$$e_i = E(c(x) | G_i) = c^T \mu_i, i = 1, 2, \dots, q$$

$$v_i^2 = \text{Var}(c(x) | G_i) = c^T \Sigma_i c, i = 1, 2, \dots, q$$

令

$$B_0 \triangleq \sum_{i=1}^q \left(e_i - \frac{1}{q} \sum_{i=1}^q e_i \right)^2$$

$$E_0 \triangleq \sum_{i=1}^q v_i^2 = \sum_{i=1}^q c^T \Sigma_i c \quad (4-18)$$

从方差分析的思想上看, (4-18) 式中 B_0 相当于组间差, E_0 相当于组内差. 如果判别很有效, 应使组内差尽量小, 组间差尽量大, 亦即 $\frac{B_0}{E_0}$ 的值应极大化. 这就是费歇判别分析的准则. 我们定义 $\mu(c) = \frac{B_0}{E_0}$, 称 $\mu(c)$ 为判别效率.

如果我们令 $B = \mu^T \left(I_q - \frac{1}{q} J \right) \mu$, $E = \sum_{i=1}^q (\Sigma_i)$, 则

$$\mu(c) = \frac{c^T B c}{c^T E c} \quad (4-19)$$

其中 I_q 为 q 阶单位方阵, J 为 q 阶元素全为 1 的方阵, $\mu = (\mu_1, \mu_2, \dots, \mu_q)_{q \times p}^T$. 从 (4-19) 式, 我们可以看到要使效率 $\mu(c)$ 达到最大, 也就是求 B 相对于 E 的最大特征根, 从多元分析中我们已知这时的向量 c 正是这个特征根相对应的特征向量. 故有如下费歇判别准则.

定理 5.1 设 $\lambda_1 \geq \lambda_2 \geq \dots \lambda_r \geq 0$ 为 $E^{-1}B$ 的 r 个非零特征根, $r \leq \min(q-1, p)$, c_1, c_2, \dots, c_r 为相应的特征向量. 则 c_1 使得 $\mu(c)$ 取极大值 λ_1 , 称 $C(x) = c_1^T x$ 为第一判别函数. 取 $c = c_2$ 是在约束条件

$$\text{Cov}(c_1^T x, c_2^T x) = 0 \quad (4-20)$$

下使 $\frac{c_2^T B c_2}{c_2^T E c_2}$ 达到最大值 λ_2 的向量. 称 $C(x) = c_2^T x$ 为第二判别

函数.一般地,除去 $c_1^T x, c_2^T x, \dots, c_{k-1}^T x$, 令 $c_k^T(x)$ 为第 k 个判别函数, 它自然要求满足

$$\text{Cov}(c_i^T x, c_k^T x) = 0 \quad (i < k, k = 2, 3, \dots, r)$$

使得

$$\frac{c_k^T B c_k}{c_k^T E c_k} = \max_{i < k} \frac{c_i^T B c_i}{c_i^T E c_i} \quad (4-21)$$

由此可以组成一个判别函数向量 $y = (y_1, y_2, \dots, y_r)'$, 其中 $y_i = c_i^T x$, $Ey = (Ey_1, \dots, Ey_r)^T$. 若 $x \in G_i$, 令

$$\mu_{iy} \triangleq E(y | x \in G_i) = (c_1^T \mu_i, c_2^T \mu_i, \dots, c_r^T \mu_i)^T \quad i = 1, 2, \dots, q \quad (4-22)$$

对于新的样本 x_0 , 由判别函数向量(4-22)知它对应一判别向量 $y_0 = (y_{01}, y_{02}, \dots, y_{0r})^T$, 它到 μ_{iy} 的 B 模距离, 由于 y_1, y_2, \dots, y_r 不相关, 故与欧氏距离等价, 即

$$\mathcal{D}^2(y_0, \mu_{iy}) = \sum_{k=1}^r (y_{0k} - \mu_{ik}^T c_k)^2 \quad i = 1, 2, \dots, q \quad (4-23)$$

我们建立费歇判别为

$$x_0 \in G_i, \text{ 若 } \mathcal{D}^2(y_0, \mu_{iy}) = \min_{1 \leq j \leq q} \mathcal{D}^2(y_0, \mu_{jy}) \quad i = 1, 2, \dots, q$$

若有不止一个总体同时使得(4-23)达到最小值, 取 G_i 右下标最小的为归入类. 至此, 我们对费歇判别给出了圆满的结果.

对于费歇判别我们还有以下几点要说明:

(i) 如果总体均值与协差阵是未知的, 则可以通过样本观察值去估计, 这与一般点估计相同. 另外, 若 q 个总体的先验概率相同即为离散均匀分布, 损失函数误判时等值, 则费歇判别与 Bayes 判别结果一致.

(ii) 如果 q 个总体的均值, 无显著性差异, 则费歇判别的效果不会很好.

二、费歇判别的综合评价问题例题分析

例 4.7^[11] 从三个不同地区采集了 56 个原油样品, 每个样品

试了五个指标: $x_1 = V$ (钒), $x_2 = \text{Fe}^{\frac{1}{2}}$ (铁 $^{\frac{1}{2}}$), $x_3 = \text{Pi}^{\frac{1}{2}}$ (铍 $^{\frac{1}{2}}$), $x_4 =$ (饱和烃) $^{-1}$, $x_5 =$ 芳香烃. 根据化学成分这 56 个样品归属于三个沙岩层(三个总体)

$$G_1: \text{wilhelm 沙岩}, \quad n_1 = 7$$

$$G_2: \text{低 mulinia 沙岩}, \quad n_2 = 11$$

$$G_3: \text{上沙岩}, \quad n_3 = 38$$

测量的原始数据从略, 算得三个总体的样本均值如下:

$$\bar{x}^{(1)} = (3.229, 6.589, 0.303, 0.150, 11.54)^T$$

$$\bar{x}^{(2)} = (4.445, 5.667, 0.344, 0.157, 5.454)^T$$

$$\bar{x}^{(3)} = (7.226, 4.634, 0.598, 0.223, 5.768)^T$$

$$\bar{x} = (6.180, 5.081, 0.511, 0.201, 6.434)^T$$

现按总的均值(4-19)式定义分别计算 B 与 E , 并利用(4-20)的要求算得 $E^{-1}B$ 的非零特征根有两个, 它们为 $\lambda_1 = 4.354$, $\lambda_2 = 0.559$. 相应的特征向量

$$c_1 = (0.312, -0.710, 2.764, 11.809, -0.235)^T$$

$$c_2 = (0.169, -0.245, -2.046, -24.453, -0.378)^T$$

故得两个判别函数:

$$y_1 = c_1^T x = 0.312x_1 - 0.710x_2 + 2.764x_3 \\ + 11.809x_4 - 0.235x_5$$

$$y_2 = c_2^T x = 0.169x_1 - 0.245x_2 - 2.046x_3 \\ - 24.453x_4 - 0.378x_5$$

以 (y_1, y_2) 形成二维判别空间. 因为 $\text{Cov}(y_1, y_2) = 0$, 故对新的样本点或被评价客体, x_0 可以利用欧氏距离建立费歇判别规则.

$$x_0 \in G_i, \text{ 若 } \mathcal{D}^2(x_0, \bar{x}^{(i)}) = \min_{j=1,2,3} \mathcal{D}^2(x_0, \bar{x}^{(j)})$$

这样我们就建立了对这个问题的费歇判别综合模型.

注意: 这个问题我们在第七章还要对比分析一次. 现在我们再举一个应用于经济领域中的综合评价问题.

例 4.8 制药厂经济效益综合指标的制定. 选用全国 20 个药

厂的四项指标来制定该厂经济效益综合指标, $x_1 = \text{总产值}/\text{消耗}$, $x_2 = \text{净产值}/\text{工资}$, $x_3 = \text{盈利}/\text{资金占用}$, $x_4 = \text{销售收入}/\text{成本}$. 现用费歇判别法建立综合评价指标. 主要步骤如下:

- (i) 按第二章标准化方法将原始数据标准化处理(略).
- (ii) 用最短距离准则, 用系统聚类法将 20 个药厂聚为两类:
 | I | 类: 1, 2, 3, 4, 5, 6, 7, 8, 12, 13, 15, 16, 18;
 | II | 类: 9, 10, 11, 14, 17, 19, 20.

表 4-9 各企业指标的数值

企业编号	企业名称	x_1	x_2	x_3	x_4
1	东北制药厂	1.611	10.59	0.69	1.67
2	北京第二制药厂	1.429	9.44	0.61	1.50
3	哈尔滨制药厂	1.447	5.97	0.24	1.25
4	江西东风制药厂	1.572	10.72	0.75	1.71
5	武汉制药厂	1.483	10.99	0.75	1.44
6	湖南制药厂	1.371	6.46	0.41	1.31
7	开封制药厂	1.665	10.51	0.53	1.52
8	西南制药厂	1.403	6.11	0.71	1.31
9	华北制药厂	2.620	21.51	1.40	2.59
10	上海第三制药厂	2.033	24.15	1.80	1.89
11	上海第四制药厂	2.015	26.86	1.93	2.02
12	山东新华制药厂	1.501	9.74	0.87	1.48
13	北京第一制药厂	1.578	14.52	1.21	1.91
14	天津制药厂	1.735	12.88	0.87	1.52
15	上海第五制药厂	1.453	17.94	0.89	1.40
16	上海第二制药厂	1.765	11.64	1.21	1.91
17	上海第六制药厂	1.532	29.42	2.52	1.80
18	杭州第一制药厂	1.488	9.23	0.81	1.45
19	福州抗生素厂	2.586	16.07	0.82	1.83
20	四川制药厂	1.992	21.63	1.01	1.89

(iii) 利用费歇判别法建立判别函数

$$U(Y_t) = Y_t^T E^{-1}(\bar{Y}^{(1)} - \bar{Y}^{(2)}) \quad (4-24)$$

其中

$$\bar{Y}^{(1)} = (1.605, 11.054, 0.716, 1.519)^T$$

$$\bar{Y}^{(2)} = (2.034, 24.714, 1.732, 2.033)^T$$

$$E^{-1}(\bar{Y}^{(1)} - \bar{Y}^{(2)}) = (-71.649, -37.933, 866.13, 46.433)^T$$

将以上数据代入(4-24)式得 $\mu(Y_t)$ 在 $t=1, 2, \dots, 20$ 的值:

$$\mu(Y_1) = 77.538, \quad \mu(Y_2) = 69.945$$

$$\mu(Y_3) = 58.037, \quad \mu(Y_4) = 79.395$$

$$\mu(Y_5) = 66.859, \quad \mu(Y_6) = 60.832$$

$$\mu(Y_7) = 70.574, \quad \mu(Y_8) = 61.288$$

$$\mu(Y_9) = 120.254, \quad \mu(Y_{10}) = 87.753$$

$$\mu(Y_{11}) = 93.789, \quad \mu(Y_{12}) = 68.716$$

$$\mu(Y_{13}) = 68.252, \quad \mu(Y_{14}) = 88.681$$

$$\mu(Y_{15}) = 70.574, \quad \mu(Y_{16}) = 65.002$$

$$\mu(Y_{17}) = 83.574, \quad \mu(Y_{18}) = 67.323$$

$$\mu(Y_{19}) = 84.967, \quad \mu(Y_{20}) = 87.753$$

利用第一判别函数的值,对这 20 个医药企业的经济效益给出综合评价.在这里已完全序化.

第七节 逐步判别

从前面章节的介绍,我们已经知道.评价中,关于综合评价的指标体系的建立是至关重要的.在第二章第二节中我们就此问题进行了讨论,指出了建立综合评价的指标体系的一些原则与条件.广义方差最小,极大不相关及聚类后选取典型指标等方法.这些方法都在不同的分析原则下,建立了较优的筛选指标的方法.本节内容是通过附加信息的检验,选择最优的评价指标体系,然后通过建立判别函数给出评价结果.

一、Wilks 统计量 A 及其性质

设有 m 个总体 G_1, G_2, \dots, G_m , 均服从 $N_p(\mu_i, V) i = 1, 2, \dots, m$. 从中抽取 n_1, n_2, \dots, n_m 个样本, 令 x_{ij} 表示从第 i 个母体中抽取的第 j 个样本, 即 $j = 1, 2, \dots, n_i; i = 1, 2, \dots, m$.

记

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \quad (\text{其中 } n = \sum_{i=1}^m n_i)$$

全部数据为

$$X = (x_{ij})_{n \times p}$$

J 为元素全为 1 的方阵. 则对总离差阵可作如下分解:

$$\begin{aligned} W &= \sum_{k=1}^m \sum_{i=1}^{n_k} (x_{ki} - \bar{x})(x_{ki} - \bar{x})^T \\ &= \sum_{k=1}^m \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^T + \sum_{k=1}^m n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T \\ &= \sum_{k=1}^m x_k^T \left(I - \frac{1}{n_k} J \right) x_k + \sum_{k=1}^m n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T \\ &= E + B \end{aligned} \quad (4-25)$$

其中

$$E = \sum_{k=1}^m x_k^T \left(I - \frac{1}{n_k} J \right) x_k$$

$$B = \sum_{k=1}^m n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T$$

当 $G_i \sim N_p(\mu_i, V) (i = 1, 2, \dots, m), \mu_i$ 都相等时

$$W \sim W_p(n-1, V)$$

$$E \sim W_p(n-m, V) \quad (4-26)$$

$$B \sim W_p(m-1, V)$$

并且 B 与 E 独立, 这里 $W_p(k, V)$ 表示 Wishart 分布. 在多元统计

分析中,是较为熟悉的内容.^[3,11]根据多元统计分析,建立 Wilks 统计量

$$\Lambda_{(p)} = \frac{|E|}{|B+E|} = \frac{|E|}{|W|} \sim \Lambda(p, n-m, m-1) \quad (4-27)$$

如果对 E, B, W 矩阵,分别做如下分块

$$E = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}_{p-r}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}_{p-r}$$

则记统计量

$$\Lambda_{(p-r)(r)} \triangleq \frac{|E_{22} - E_{21}E_{11}^{-1}E_{12}|}{|W_{22} - W_{21}W_{11}^{-1}W_{12}|}$$

时用(4-27)可得

$$\begin{aligned} \Lambda_{(p)} &= \frac{|E|}{|W|} = \frac{|E_{11}| \cdot |E_{22} - E_{21}E_{11}^{-1}E_{12}|}{|W_{11}| \cdot |W_{22} - W_{21}W_{11}^{-1}W_{12}|} \\ &= \Lambda_{(r)} \Lambda_{(p-r)(r)} \end{aligned} \quad (4-28)$$

根据 Wilks 统计量的性质知

$$\Lambda_{(p)} \sim \Lambda(p, n-m, m-1)$$

$$\Lambda_{(r)} \sim \Lambda(r, n-m, m-1)$$

$$\Lambda_{(p-r)(r)} \sim \Lambda(p-r, n-m-r, m-1)$$

因此利用(4-28)式可以推广到一般情况.

$$\Lambda_{(p)} = \Lambda_1 \Lambda_{2,1} \Lambda_{3,1,2} \cdots \Lambda_{p,1,2,\dots,(p-1)} \quad (4-29)$$

以上(4-29)式说明了 Λ 统计量的一个非常优良的特性,这正是建立逐步判别的综合评价的理论根据.

二、逐步判别的原则

(i)在 x_1, x_2, \dots, x_p 中选出一个变量,它使得 Wilks 统计量 $\Lambda_i = \frac{l_{ii}}{W_{ii}}, i=1, 2, \dots, p$ 达到最小.即从统计的观点看也就是附加其上的信息量最大,因而对总体判别具有较大贡献,其中 l_{ii} 的定义从 $E = (l_{ij})_{p \times p}$ 中求得,以下同此.不妨设 $\Lambda_1 = \min_{1 \leq i \leq p} \{\Lambda_i\}$,并对

Λ_1 进行显著性检验, 如果 Λ_1 不显著, 则无法用判别分析进行综合评价; 否则, 进入下一步.

(ii) 计算未选变量与已选变量(无妨设为 x_1) 配合的 Λ 值, 并选择使其达到最小的变量作为第二个选入变量, 今

$$\Lambda_{1,2} = \min_{2 \leq i \leq m} \begin{vmatrix} l_{11} & l_{1i} \\ l_{i1} & l_{ii} \\ \hline W_{11} & W_{1i} \\ W_{i1} & W_{ii} \end{vmatrix}$$

如此继续到选入第 r 个变量, 无妨设为 x_1, x_2, \dots, x_r , 则未入选的 $p-r$ 个变量中, 选一个与已选变量配合计算.

$$\Lambda_{1,2,\dots,r,l} = \min_{r+1 \leq l \leq p} \begin{vmatrix} l_{11} & \dots & l_{1r} & l_{1l} \\ \vdots & & \vdots & \vdots \\ l_{r1} & \dots & l_{rr} & l_{rl} \\ \hline l_{l1} & \dots & l_{lr} & l_{ll} \\ \hline W_{11} & \dots & W_{1r} & W_{1l} \\ \vdots & & \vdots & \vdots \\ W_{r1} & \dots & W_{rr} & W_{rl} \\ W_{l1} & \dots & W_{lr} & W_{ll} \end{vmatrix}$$

由于 $\Lambda_{1,2,\dots,r,l} = \Lambda_{(r)} \cdot \Lambda_{l \cdot 1,2,\dots,r}$, 所以上式极小值就等价于 $\Lambda_{l \cdot 1,2,\dots,r}$ 极小. 计算 $\Lambda_{l \cdot 1,2,\dots,r}$, 并检验第 $r+1$ 个变量 x_l 能否提供附加信息, 否则进入第(iv)步.

(iii) 考查已选入的 r 个变量中, 是否有其重要性的变化, 及时剔除不能提供附加信息的变量. 如果对于第 l 个引入变量 x_l , 计算 $\Lambda_{l \cdot 1,2,\dots,l-1,l+1,\dots,r}$ 选择达到极小的 l , 看是否显著, 若不显著剔除该变量继续进行该步骤; 如果显著, 则进入第(ii)步, 入选变量.

(iv) 当既没有人选变量, 又没有剔除变量时, 对未选的 $p-r$ 个变量, 进行附加信息检验, 利用统计量

$$\Lambda_{(p-r) \cdot (r)} = \frac{|E_{22} - E_{21}E_{11}^{-1}E_{12}|}{|W_{22} - W_{21}W_{11}^{-1}W_{12}|}$$

$$\sim \Lambda(p-r, n-m-r, m-1)$$

对于给定显著性水平, 如果 Λ 不显著, 说明余下变量不能提供更多的信息, 选变量过程结束, 转入建立判别函数第(v)步. 否则说明余下 $p-r$ 个变量尚有可供附加信息, 按第(iii)步, 选第 $r+1$ 个变量.

(v) 利用选出的变量建立判别函数. 可以利用距离法、贝叶斯判别法和费歇判别法, 有关方法见本章第三、四、六节内容.

三、应用举例

例 4.9^[3] 在研究砂基液化的问题中, 选用 9 个有关因素, 震级 x_1 , 震中 x_2 (km), 土的类型 (砾土为 0, 否则为 1) x_3 , 砂的类型 (粉砂为 0, 否则为 1) x_4 , 水深 (m) x_5 , 土深 (m) x_6 , 贯入值 x_7 , 最大地面加速度 (g) x_8 , 地震持续时间 (s) x_9 , 今从已液化与未液化的地层中分别抽 12 与 23 个样本. 数据见表 4-10.

现对表述 9 个变量, 采用 Wilks 统计量, 经计算, x_7, x_8, x_9 选入, 其余变换全部剔除. 经计算 (为了方便使协方差阵不退化, 我们不计算 x_3 与 x_4) 有

$$\bar{y}^{(1)} = \begin{pmatrix} 7.358 \\ 73.667 \\ 1.458 \\ 6.000 \\ 15.250 \\ 0.172 \\ 49.500 \end{pmatrix}, \quad \bar{y}^{(2)} = \begin{pmatrix} 7.687 \\ 69.609 \\ 2.043 \\ 5.239 \\ 6.348 \\ 0.216 \\ 70.348 \end{pmatrix}$$

表 4-10 砂基液化数据

编号	组别	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	I	6.6	39	0	0	1.0	6.0	6	0.12	20
2	I	6.6	39	0	0	1.0	6.0	12	0.12	20
3	I	6.1	47	0	0	1.0	6.0	6	0.08	12

续表

编号	组别	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
4	I	6.1	47	0	0	1.0	6.0	12	0.08	12
5	I	8.4	32	1	0	2.0	7.5	19	0.35	75
6	I	7.2	6	0	0	1.0	7.0	28	0.30	30
7	I	8.4	113	0	0	3.5	6.0	18	0.15	75
8	I	7.5	52	0	0	1.0	6.0	12	0.16	40
9	I	7.5	52	0	0	3.5	7.5	6	0.16	40
10	I	8.3	113	1	0	0.0	7.5	35	0.12	180
11	I	7.8	172	0	0	1.0	3.5	14	0.21	45
12	I	7.8	172	0	0	1.5	3.0	15	0.21	45
13	II	8.4	32	0	0	1.0	5.0	4	0.35	75
14	II	8.4	32	0	0	2.0	9.0	10	0.35	75
15	II	8.4	32	0	0	2.5	4.0	10	0.35	75
16	II	6.3	11	0	0	4.5	7.5	3	0.20	15
17	II	7.0	8	0	0	4.5	4.5	9	0.25	30
18	II	7.0	8	0	0	6.0	7.5	4	0.25	30
19	II	7.0	8	0	0	1.5	6.0	1	0.25	30
20	II	8.3	161	0	0	1.5	4.0	4	0.08	70
21	II	8.3	161	0	1	0.5	2.5	1	0.08	70
22	II	7.2	6	0	0	3.5	4.0	12	0.30	30
23	II	7.2	6	0	0	1.0	3.0	3	0.30	30
24	II	7.2	6	0	1	1.0	6.0	5	0.30	30
25	II	5.5	6	0	0	2.5	3.0	7	0.18	18
26	II	8.4	113	0	0	3.5	4.5	6	0.15	75
27	II	8.4	113	0	0	3.5	4.5	8	0.15	75
28	II	7.5	52	0	0	1.0	6.0	6	0.16	40
29	II	7.5	52	0	0	1.0	7.5	8	0.16	40
30	II	8.3	97	0	0	0.0	6.0	5	0.15	180
31	II	8.3	97	0	0	2.5	6.0	5	0.15	180
32	II	8.3	89	0	0	0.0	6.0	10	0.16	180
33	II	8.3	56	0	0	1.5	6.0	13	0.25	180
34	II	7.8	172	0	0	1.0	3.5	6	0.21	45
35	II	7.8	283	0	0	1.0	4.5	6	0.10	45

现利用前面讲过的费歇判别法,建立判别函数为

$$u(x) = 0.0158x_7 - 0.6321x_8 - 0.0012x_9$$

取阈值

$$\bar{\mu} = \frac{1}{2}[\mu(\bar{y}^{(1)}) + \mu(\bar{y}^{(2)})] = -0.03156$$

将数值代入判别函数,与原训练样品比较得到回报效果,如表 4-11.与距离判别法^[3]的结果比较(见表 4-12),我们发现结果差不多.可见筛选变量的逐步判别法还是有意义的.

表 4-11

	I	II
I	12	0
II	2	21

表 4-12

	I	II
I	11	1
II	0	23

第五章 其他综合评价方法

实际中遇到的综合评价问题是多种多样的,本章在前面几章的基础上介绍其他的几种综合评价方法,它们各具特色,分别适合于处理不同情况下的综合评价问题.

第一节 距离综合评价方法

综合评价是通过描述被评价事物的多个指标来进行的,如果将指标看成变量,则在几何上将形成一个高维空间,而每个被评价事物由反映它的多个指标值在该空间中决定一个点.因此,从几何角度来看,综合评价的对象就是高维空间中的一些点,综合评价问题就变为对这些点作出总体评价或排序.受到聚类 and 判别分析的启示,一个直观而自然的想法就是在空间中确定出参考点,比如最优样本点、最劣样本点等,然后计算各样本点到参考点的距离,距最优样本点越近越好,距最劣样本点越远越好,这就是距离综合评价方法的基本思想.关于具体的排序方法,一些文献中已提出了好几种,比如文献[23,26]介绍的 TOPSIS 法,文献[24]提出的投影法,等等.下面先介绍再讨论.

不妨设用 p 个指标对 n 个事物进行综合评价,原始数据构成如下矩阵:

$$X' = (x'_{ij})_{n \times p}, i = 1, 2, \dots, n; j = 1, 2, \dots, p.$$

距离综合评价需要经过以下几个步骤:

1. 指标同向化

如果 p 个指标中有逆指标或适度指标,则利用第二章所介绍的方法将其转化为正指标.转化后数据矩阵记为

$$X = (x_{ij})_{n \times p}, i = 1, 2, \dots, n; j = 1, 2, \dots, p$$

2. 无量纲化

选用合适的方法对数据进行无量纲化,变换后数据阵记为

$$Y' = (y'_{ij}), i = 1, 2, \dots, n; j = 1, 2, \dots, p$$

TOPSIS 法一般采用 (2-18) 式变换数据.

3. 构造加权数据矩阵

设已确定出各指标的权重为 w_1, w_2, \dots, w_p , 以它们为主对角线元素构造对角矩阵 W , 即

$$\begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_p \end{pmatrix}_{p \times p} \triangleq W$$

则加权数据矩阵为

$$Y'W = Y = (y_{ij})_{n \times p}$$

或

$$y_{ij} = w_j y'_{ij}, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, p$$

4. 确定参考样本(虚拟的样本)

一般以最优样本(也称为理想样本)和最劣样本(也称为负理想样本)为参考样本. 由于指标已正向化, 所以可用所有参评样本中各指标的最大值构成最优样本, 用各指标的最小值构成最劣样本. 分别用 Y^+ 和 Y^- 表示如下:

$$Y^+ = (y_1^+, y_2^+, \dots, y_p^+)^T$$

$$Y^- = (y_1^-, y_2^-, \dots, y_p^-)^T$$

其中

$$y_j^+ = \max_{1 \leq i \leq n} \{y_{ij}\}$$

$$y_j^- = \min_{1 \leq i \leq n} \{y_{ij}\}, j = 1, 2, \dots, p$$

5. 计算距离

有关文献采用了以下几种形式:

(1) 样本点到最优样本点的距离^[25]

$$D_i^+ = \sqrt{\sum_{j=1}^p (y_{ij} - y_j^+)^2} \quad (5-1)$$

$$i = 1, 2, \dots, n$$

D_i^+ 越小,第 i 个样本点距最优样本点越近,表明第 i 个被评事物总体表现越好.当然为了排序方便,也可以将 D_i^+ 再做一些变换,比如

$$Z_i = \frac{100}{D_i^+} \quad (5-2)$$

用 Z_i 作为评价分值更符合习惯.

(2)样本点到最优样本点的相对接近度^[26]

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \quad (5-3)$$

其中 D_i^- 为样本点到最劣样本点的距离

$$D_i^- = \sqrt{\sum_{j=1}^p (y_{ij} - y_j^-)^2} \quad (5-4)$$

$$i = 1, 2, \dots, n$$

C_i 越大,表明样本点与最优样本点的相对距离越近.

(3)样本点在两参考点连线上的射影到最劣样本点的距离^[23]

设最优样本和最劣样本对应点分别为 A^+ 和 A^- ,第 i 个样本点为 A_i ,则该三点在 p 维空间中构成一个平面三角形,如图 5-1 所示, H 为 A_i 在线 A^-A^+ 上的投影, A^- 到 H 的距离 d_i 就是向量 $\overrightarrow{A^-A_i}$ 在向量 $\overrightarrow{A^-A^+}$ 上的射影 $\overrightarrow{A^-H}$ 的长度.记加权数据阵 Y 中第 i 行即第 i 个样本数据为

$$Y_i = (y_{i1}, y_{i2}, \dots, y_{ip})^T$$

易得

$$d_i = \frac{(Y_i - Y^-)^T \cdot (Y^+ - Y^-)}{\|Y^+ - Y^-\|}$$

$$= \frac{\sum_{j=1}^p (y_{ij} - y_j^-)(y_j^+ - y_j^-)}{\sqrt{\sum_{j=1}^p (y_j^+ - y_j^-)^2}} \quad (5-5)$$

上式中“·”为向量内积,“||”为欧氏范数.最优样本 Y^+ 和最劣样本 Y^- 的取法可以保证 $d_i \geq 0$.

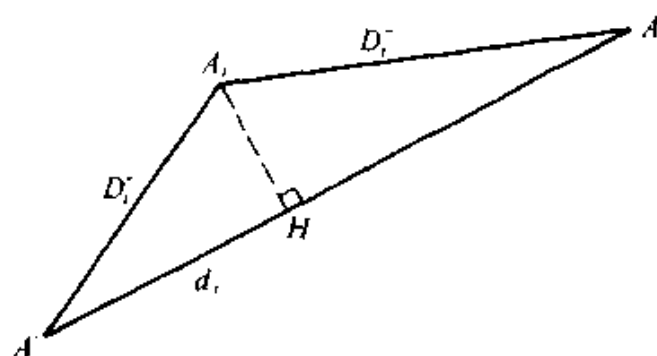


图 5-1 距离评价方法图示

6. 由上步计算结果对 n 个样本比较排序

下面通过一个例子熟悉一下上述过程.

例 5-1 某医院医疗工作质量的综合评价^[26]. 该医院 1980 年至 1986 年几项工作质量指标值(见表 5-1), 试用距离方法对其进行综合评价.

表 5-1 某医院 1980~1986 年几项工作质量指标情况

年度	好转率	床位周转次数	平均病床工作日	平均费用值
1980	95.3	29.4	331.1	47.2
1981	96.3	26.9	332.9	41.8
1982	95.9	24.7	329.4	50.5
1983	97.3	28.4	356.6	67.7
1984	98.1	28.8	378.7	40.8
1985	97.0	25.4	332.4	41.2
1986	97.2	25.5	333.5	41.9

在四个评价指标中,好转率、病床周转次数、平均病床工作日三个指标均为正指标,平均费用值为逆指标,首先将它正向化.对平均费用值取倒数后,发现数值太小(与其他三个指标相比),为了减少计算误差,给倒数值再乘以 100,可得数据矩阵为

$$X = \begin{bmatrix} 95.3 & 29.4 & 331.1 & 2.119 \\ 96.3 & 26.9 & 332.9 & 2.392 \\ 95.9 & 24.7 & 329.4 & 1.980 \\ 97.3 & 28.4 & 356.6 & 1.477 \\ 98.1 & 28.8 & 378.7 & 2.451 \\ 97.0 & 25.4 & 332.4 & 2.427 \\ 97.2 & 25.5 & 333.5 & 2.387 \end{bmatrix}_{7 \times 4}$$

采用公式(2-18)进行无量纲化,即 X 中每一元素除以其所在列所有元素平方和的算术平方根,可得数据矩阵

$$Y' = \begin{bmatrix} 0.3724 & 0.4105 & 0.3654 & 0.3639 \\ 0.3763 & 0.3756 & 0.3673 & 0.4108 \\ 0.3747 & 0.3449 & 0.3635 & 0.3400 \\ 0.3802 & 0.3965 & 0.3935 & 0.2536 \\ 0.3833 & 0.4021 & 0.4179 & 0.4209 \\ 0.3790 & 0.3546 & 0.3668 & 0.4168 \\ 0.3798 & 0.3560 & 0.3680 & 0.4099 \end{bmatrix}_{7 \times 4}$$

本例中各指标权数均为 1,因而加权数据阵 $Y = Y'$.

根据数据阵 Y' 确定出两个参考样本分别为

$$Y^+ = (0.3833, 0.4105, 0.4179, 0.4209)^T$$

$$Y^- = (0.3724, 0.3449, 0.3635, 0.2536)^T$$

然后计算各样本点与参考点的距离并排序.分别采用(5-1)、(5-3)、(5-5)式计算,所得结果见表 5-2,除 1985 年和 1986 年的排序结果稍有差异外,其余的排序结果都是一致的.

还有两个问题需要说明一下.第一,分别采用(5-1)、(5-3)、(5-5)式会有不同的排序结果,那么就存在一个选择使用的问题.采用(5-1)式时,在以 A^+ 为中心的同一球上的样本点无法区分

表 5-2 评价结果

年 度	D_i^+		$C_i = \frac{D_i^-}{D_i^+ + D_i^-}$		d_i	
	D_i^+ 值	排序	C_i 值	排序	d_i 值	排序
1980	0.07826	5	0.6211	5	0.6463	5
1981	0.06268	2	0.7189	2	0.8075	2
1982	0.11780	6	0.4232	6	0.4094	6
1983	0.16970	7	0.2618	7	0.1442	7
1984	0.00840	1	0.9566	1	0.9751	1
1985	0.07597	4	0.6830	3	0.7804	3
1986	0.07479	3	0.6772	4	0.7691	4

优劣,采用(5-5)式时,在与 A^+A^- 垂直的同一平面上的样本点无法区分优劣,这两种形式有一个共同特点:没有考虑样本点距两参考点连线的远近(即图 5-1 中 A_i 到点 H 的距离),而正是这种远近反映了事物在各方面发展的均衡性.对(5-3)式可分解如下:

$$C_i = D_i^- \cdot \frac{1}{D_i^+ + D_i^-}$$

第二个因式 $\frac{1}{D_i^+ + D_i^-}$ 反映了事物发展的均衡性,如果事物在各方面发展不均衡,则样本点距离两参考点连线就远些, $\frac{1}{D_i^+ + D_i^-}$ 较小;反之,若事物发展比较均衡,则 $\frac{1}{D_i^+ + D_i^-}$ 较大.采用(5-3)式,只有当样本点距最劣样本点远(一定程度上讲离最优样本点就近些)且发展较均衡时才能取得较大的评价值.因此,从这个角度讲可优先考虑采用(5-3)式.第二,选择不同的距离也影响到评价结果.一般来讲,在各评价指标之间相关性不大时,多采用欧氏距离.如果各指标间相关性较大时则宜选用马氏距离,一定程度上可以

消除指标间的相关性对综合评价结果的影响.习惯上多采用欧氏距离,而在选择评价指标时要尽量避免强相关.

第二节 灰色关联度评价法

1982年,华中理工大学邓聚龙教授首先提出了灰色系统的概念,并建立了灰色系统理论.之后,灰色系统理论得到了较深入的研究,并在许多方面获得了成功的应用.灰色系统理论认为,人们对客观事物的认识具有广泛的灰色性,即信息的不完全性和不确定性,因而由客观事物所形成的是一种灰色系统,即部分信息已知、部分信息未知的系统,比如社会系统、经济系统、生态系统等都可以看作是灰色系统.人们对综合评价的对象——被评价事物的认识也具有灰色性,因而可以借助于灰色系统的相关理论来研究综合评价问题.下面首先介绍灰色关联分析方法,然后探讨其在综合评价中应用的一些问题.

一、灰色关联分析(GRA)方法

灰色关联分析是一种多因素统计分析方法,它是以各因素的样本数据为依据用灰色关联度来描述因素间关系的强弱、大小和次序的.如果样本数据列反映出两因素变化的态势(方向、大小、速度等)基本一致,则它们之间的关联度较大;反之,关联度较小.与传统的多因素分析方法(相关、回归等)相比,灰色关联分析对数据要求较低且计算量小,便于广泛应用.

GRA分析的核心是计算关联度,下面通过一个例子来说明计算关联度的思路和方法.表5-3是某地区1990~1995年国内生产总值的统计资料.现在提出这样的问题:该地区三次产业中,哪一产业产值的变化与该地区国内生产总值(GDP)的变化态势更一致呢?也就是哪一产业与GDP的关联度最大呢?这样的问题显然是很有实际意义的.一个很自然的想法就是分别将三次产业产值的时间序列与GDP的时间序列进行比较,为了能够比较,先对

表 5-3 某地区国内生产总值统计资料(百万元)

年份	国内生产总值	第一产业	第二产业	第三产业
1990	1988	386	839	763
1991	2061	408	846	808
1992	2335	422	960	953
1993	2750	482	1258	1010
1994	3356	511	1577	1268
1995	3806	561	1893	1352

各序列进行无量纲化,这里采用均值化法.各序列的均值分别为:2716,461.5,1228.83,1025.67,表 5-3 中每列数据除以其均值可得均值化序列(如表 5-4 所示).粗略地想一下,两序列变化的态势是表现在其对应点的间距上.如果各对应点间距均较小,则两序列变化态势的一致性较强,否则,一致性弱.分别计算各产业产值与 GDP 在对应期的间距(绝对差值),结果见表 5-5.接下来似乎应

表 5-4

年份 t	GDP $x_0(t)$	第一产业 $x_1(t)$	第二产业 $x_2(t)$	第三产业 $x_3(t)$
1990	0.7320	0.8364	0.6828	0.7440
1991	0.7588	0.8819	0.6885	0.7878
1992	0.8597	0.9144	0.7812	0.9291
1993	1.0125	1.0444	1.0237	0.9847
1994	1.2356	1.1073	1.2833	1.2363
1995	1.4013	1.2156	1.5405	1.3182

表 5-5

年份 t	$\Delta_{01}(t) =$ $ x_0(t) - x_1(t) $	$\Delta_{02}(t) =$ $ x_0(t) - x_2(t) $	$\Delta_{03}(t) =$ $ x_0(t) - x_3(t) $
1990	0.1044	0.0492	0.0119
1991	0.1231	0.0704	0.0289
1992	0.0547	0.0785	0.0694
1993	0.0319	0.0112	0.0278
1994	0.1284	0.0477	0.0006
1995	0.1857	0.1392	0.0832

该是对三个绝对差值序列分别求平均再进行比较,就可以解决问题了.但如果仔细观察表 5-5 中数据就会发现绝对差值数据序列的数据间存在着较大的数量级差异(最大为 0.1857,最小的为 0.0006,相差 300 多倍),不能直接进行综合,还需要对其进行一次规范化.设 $\Delta(\max)$ 和 $\Delta(\min)$ 分别表示表 5-5 中绝对差值 $\Delta_{0i}(t)$ 的最大数和最小数,则

$$0 \leq \Delta(\min) \leq \Delta_{0i}(t) \leq \Delta(\max)$$

因而

$$0 \leq \frac{\Delta(\min)}{\Delta(\max)} \leq \frac{\Delta_{0i}(t)}{\Delta(\max)} \leq 1$$

显然 $\frac{\Delta_{0i}(t)}{\Delta(\max)}$ 越大,说明两序列 $(x_i$ 和 $x_0)$ 变化态势一致性弱,反之,一致性强,因此可考虑将 $\frac{\Delta_{0i}(t)}{\Delta(\max)}$ 取倒反向.为了使规范化后数据在 $[0,1]$ 内,可考虑

$$\frac{\Delta(\min)/\Delta(\max)}{\Delta_{0i}(t)/\Delta(\max)}$$

由于在一般情况下 $\Delta(\min)$ 可能为零(即某个 $\Delta_{0i}(t)$ 为零),故将上式改进为

$$\frac{\Delta(\min)/\Delta(\max) + \rho}{\Delta_{0i}(t)/\Delta(\max) + \rho} \triangleq \xi_{0i}(t)$$

ρ 在 0 和 1 之间取值,上式可变形为

$$\xi_{0i}(t) = \frac{\Delta(\min) + \rho \Delta(\max)}{\Delta_{0i}(t) + \rho \Delta(\max)} \quad (5-6)$$

$$i = 1, 2, 3, \quad t = 1990, \dots, 1995$$

$\xi_{0i}(t)$ 称为序列 x_i 和序列 x_0 在第 t 期的灰色关联系数(常简称为关联系数).由(5-6)式可以看出, ρ 取值的大小可以控制 $\Delta(\max)$ 对数据转化的影响, ρ 取较小的值,可以提高关联系数间差异的显著性,因而称 ρ 为分辨系数.利用(5-6)式对表 5-5 中绝对差值 $\Delta_{0i}(t)$ 进行规范化,取 $\rho = 0.4$,结果见表 5-6.以 $\xi_{01}(1990)$ 计算

为例:

$$\Delta(\min) = 0.0006, \quad \Delta(\max) = 0.1857$$

$$\begin{aligned}\xi_{01}(1990) &= \frac{0.0006 + 0.4 \times 0.1857}{0.1044 + 0.4 \times 0.1857} \\ &= 0.4191\end{aligned}$$

同样可计算出表 5-6 中其余关联系数.

表 5-6

年份 t	$\xi_{01}(t)$	$\xi_{02}(t)$	$\xi_{03}(t)$
1990	0.4191	0.6067	0.8687
1991	0.3796	0.5178	0.7257
1992	0.5808	0.4903	0.5213
1993	0.7055	0.8761	0.7338
1994	0.3696	0.6141	1.000
1995	0.2881	0.3510	0.4758

最后分别对各产业与 GDP 的关联系数序列求算术平均可得

$$\begin{aligned}r_{01} &= \frac{1}{6}(0.4191 + 0.3796 + 0.5808 + 0.7055 \\ &\quad + 0.3696 + 0.2881) \\ &= 0.4571\end{aligned}$$

$$\begin{aligned}r_{02} &= \frac{1}{6}(0.6067 + 0.5178 + 0.4903 + 0.8761 \\ &\quad + 0.6141 + 0.3510) \\ &= 0.5760\end{aligned}$$

$$\begin{aligned}r_{03} &= \frac{1}{6}(0.8687 + 0.7257 + 0.5213 + 0.7338 \\ &\quad + 1.000 + 0.4758) \\ &= 0.7209\end{aligned}$$

r_{0i} 称为序列 x_0 和 $x_i (i=1,2,3)$ 的灰色关联度. 由于 $r_{03} > r_{02} >$

r_{01} ,因而第三产业产值与 GDP 的关联度最大,其次是第二产业、第一产业。

从上例可以看出,灰色关联分析需要经过以下几个步骤:

1. 确定分析序列

在对所研究问题定性分析的基础上,确定一个因变量因素和多个自变量因素.设因变量数据构成参考序列 X_0' ,各自变量数据构成比较序列 X_i' ($i = 1, 2, \dots, n$), $n + 1$ 个数据序列形成如下矩阵:

$$(X_0', X_1', \dots, X_n') = \begin{bmatrix} x_0'(1) & x_1'(1) & \cdots & x_n'(1) \\ x_0'(2) & x_1'(2) & \cdots & x_n'(2) \\ \vdots & \vdots & & \vdots \\ x_0'(N) & x_1'(N) & \cdots & x_n'(N) \end{bmatrix}_{N \times (n+1)} \quad (5-7)$$

其中

$$X_i' = (x_i'(1), x_i'(2), \dots, x_i'(N))^T, i = 0, 1, 2, \dots, n$$

N 为变量序列的长度。

2. 对变量序列进行无量纲化

一般情况下,原始变量序列具有不同的量纲或数量级,为了保证分析结果的可靠性,需要对变量序列进行无量纲化.无量纲化后各因素序列形成如下矩阵:

$$(X_0, X_1, \dots, X_n) = \begin{bmatrix} x_0(1) & x_1(1) & \cdots & x_n(1) \\ x_0(2) & x_1(2) & \cdots & x_n(2) \\ \vdots & \vdots & & \vdots \\ x_0(N) & x_1(N) & \cdots & x_n(N) \end{bmatrix}_{N \times (n+1)} \quad (5-8)$$

常用的无量纲化方法有均值化法(见(5-9)式)、初值化法(见(5-10)式)等。

$$x_i(k) = \frac{x_i'(k)}{\frac{1}{N} \sum_{k=1}^N x_i'(k)} \quad (5-9)$$

$$x_i(k) = \frac{x_i'(k)}{x_i'(1)} \quad (5-10)$$

$$i = 0, 1, \dots, n; k = 1, 2, \dots, N$$

3. 求差序列、最大差和最小差

计算(5-8)中第一列(参考序列)与其余各列(比较序列)对应的绝对差值,形成如下绝对差值矩阵:

$$\begin{bmatrix} \Delta_{01}(1) & \Delta_{02}(1) & \cdots & \Delta_{0n}(1) \\ \Delta_{01}(2) & \Delta_{02}(2) & \cdots & \Delta_{0n}(2) \\ \vdots & \vdots & & \vdots \\ \Delta_{01}(N) & \Delta_{02}(N) & \cdots & \Delta_{0n}(N) \end{bmatrix}_{N \times n}$$

其中

$$\Delta_{0i}(k) = |x_0(k) - x_i(k)| \quad (5-11)$$

$$i = 1, 2, \dots, n; k = 1, 2, \dots, N$$

绝对差值阵中最大数和最小数即为最大差和最小差:

$$\max_{\substack{1 \leq i \leq n \\ 1 \leq k \leq N}} \{\Delta_{0i}(k)\} \triangleq \Delta(\max) \quad (5-12)$$

$$\min_{\substack{1 \leq i \leq n \\ 1 \leq k \leq N}} \{\Delta_{0i}(k)\} \triangleq \Delta(\min) \quad (5-13)$$

4. 计算关联系数

对绝对差值阵中数据作如下变换:

$$\xi_{0i}(k) = \frac{\Delta(\min) + \rho \Delta(\max)}{\Delta_{0i}(k) + \rho \Delta(\max)} \quad (5-14)$$

得到关联系数矩阵:

$$\begin{bmatrix} \xi_{01}(1) & \xi_{02}(1) & \cdots & \xi_{0n}(1) \\ \xi_{01}(2) & \xi_{02}(2) & \cdots & \xi_{0n}(2) \\ \vdots & \vdots & & \vdots \\ \xi_{01}(N) & \xi_{02}(N) & \cdots & \xi_{0n}(N) \end{bmatrix}_{N \times n} \quad (5-15)$$

式中分辨系数 ρ 在(0,1)内取值,一般情况下依据(5-15)中数据情况多在0.1至0.5取值, ρ 越小越能提高关联系数间的差异.关联系数 $\xi_{0i}(k)$ 是不超过1的正数, $\Delta_{0i}(k)$ 越小, $\xi_{0i}(k)$ 越大,它反

映第 i 个比较序列 X_i 与参考序列 X_0 在第 k 期的关联程度.

5. 计算关联度

比较序列 X_i 与参考序列 X_0 的关联程度是通过 N 个关联系数(即(5-15)中第 i 列)来反映的,求平均就可得到 X_i 与 X_0 的关联度

$$r_{0i} = \frac{1}{N} \sum_{k=1}^N \xi_{0i}(k) \quad (5-16)$$

6. 依关联度排序

对各比较序列与参考序列的关联度从大到小排序,关联度越大,说明比较序列与参考序列变化的态势越一致.

从上边也可以看出,关联度的几何含义为比较序列与参考序列曲线的相似与一致程度.如果两序列的曲线形状接近,则两者关联度就较大,反之,两者关联度就较小.

二、用灰色关联分析进行综合评价

灰色关联分析的目的是揭示因素间关系的强弱,其操作对象是因素的时间序列,最终的结果表现为通过关联度对各比较序列做出排序.综合评价的对象也可以看作是时间序列(每个被评事物对应的各项指标值),并且往往需要对这些时间序列做出排序,因而可以借助于灰色关联分析来进行.比较序列自然是由被评事物的各项指标值构成的序列,那么参考序列是什么呢?考虑到要用比较序列与参考序列的关联度来对各比较序列排序,参考序列应该是一个理想的比较标准.受到距离评价方法的启示,可选最优样本数据作为参考序列,与其关联度越大则越好.

设用 p 个指标 x_1, x_2, \dots, x_p (不失一般性,设其均为正向指标),对 n 个样本进行评价,无量纲化后形成如下数据矩阵:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n \times p}$$

其中第 i 个样本数据为 $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T, i = 1, 2, \dots, n$. 构造最优样本

$$X_0 = (x_{01}, x_{02}, \dots, x_{0p})^T$$

其中

$$x_{0j} = \max_{1 \leq i \leq n} \{x_{ij}\}, j = 1, 2, \dots, p$$

由以下公式可计算出样本 $X_i (i = 1, 2, \dots, n)$ 与最优样本 X_0 的关联度 r_{0i} .

$$\Delta_{0i}(j) = |x_{ij} - x_{0j}| \quad (5-17)$$

$$\xi_{0i}(j) = \frac{\min_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} \Delta_{0i}(j) + \rho \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} \Delta_{0i}(j)}{\Delta_{0i}(j) + \rho \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}} \Delta_{0i}(j)} \quad (5-18)$$

$$i = 1, 2, \dots, n, j = 1, 2, \dots, p$$

$$r_{0i} = \sum_{j=1}^p w_j \xi_{0i}(j), i = 1, 2, \dots, n \quad (5-19)$$

上式中, $w_j (j = 1, 2, \dots, p)$ 是指标 $x_j (j = 1, 2, \dots, p)$ 的归一化权重. 最后由 $r_{0i} (i = 1, 2, \dots, n)$ 即可对 n 个样本排出优劣顺序. 下面看一个完整的例子.

例 5.2 麦棉两熟小麦配套品种(系)的灰色关联度评价.

评价对象是“鲁西北棉区麦棉两熟小麦配套品种筛选”课题中的 10 个小麦品种(系), 依据是 1989~1993 年度在山东省陵县进行的试验测试结果, 数据见表 5-7^[28]. 评价指标即为小麦品种的一些表现性状, 共 11 个. 有些指标是逆指标, 但表 5-7 中的数据均已作过正向化处理(见表下注).

首先构造最优样本——理想品种, 理想品种在各性状(指标)上要符合麦棉两熟小麦配套品种的要求, 其各性状值要优于或同于参试品种相应性状的最优值. X_0 如表 5-7 中第一行所示. 借助于理想品种的性状值对数据进行无量纲化, 即每个指标数值除以“理想品种”相应的指标数值可得无量纲化数据(见表 5-8).

表 5-7 各参试品种与理想品种的主要性状平均值

品 种	1 成熟 期	2 亩穗数 (万)	3 穗粒 数(个)	4 千粒重 (g)	5 产量 (公斤 /亩)	6 抗锈 病	7 抗白 粉病	8 抗冻 性	9 抗倒 性	10 株高	11 株型	其 他
X_0' (理 想品种)	4	42	31	48	425	4	4	4	5	18	3	
X_1' (鲁麦 20)	4	39.7	25.3	39.1	333.8	1	2	3	5	12.5	3	
X_2' (泰 山 10 号)	1	38.3	27.1	40.8	360.0	1	2	3	4	3.7	2	
X_3' (86 中 15)	0	40.6	27.7	43	411.0	4	3	4	4	8.6	2	
X_4' (85 中 33)	1	37.7	26.9	47.1	406.0	4	3	3	5	17.2	2	
X_5' (德 农 10 号)	2	39.7	30.1	41.2	418.5	1	1	3	2	2.5	3	
X_6' (鲁麦 15)	1	41.3	27.8	43.5	424.5	4	4	4	4	14.9	3	
X_7' (轮早 3 号)	3	38.6	26.1	43.2	369.9	4	3	3	4	11.4	2	
X_8' 泰早 2 号	3	39.1	24.2	44.9	361.1	3	3	3	4	4.9	2	
X_9' (邯 87—1)	2	38.3	24.0	45.1	352.4	3	2	3	4	8.9	2	
X_{10}' (淮 60169)	1	32.9	29.4	42.1	346.1	3	2	3	2	0.8	3	

注:成熟期以供试品种最晚熟日期记为零,每早熟 1 天记为 1,以此类推;抗锈病、抗白粉病、抗冻性、抗倒性均以 5 减去其相应抗性级别;株高是以 80 减去实际株高(cm);株型分松散、中间和紧凑型分别记作 1,2,3。

表 5-8 无量纲化数据

指标 品种	1	2	3	4	5	6	7	8	9	10	11
X_0	1	1	1	1	1	1	1	1	1	1	1
X_1	1	0.95	0.82	0.81	0.79	0.25	0.5	0.75	1	0.69	1
X_2	0.25	0.91	0.87	0.85	0.85	0.25	0.5	0.75	0.8	0.21	0.67
X_3	0	0.97	0.89	0.90	0.97	1	0.75	1	0.8	0.48	0.67
X_4	0.25	0.90	0.87	0.98	0.96	1	0.75	0.75	1	0.96	0.67
X_5	0.5	0.95	0.97	0.86	0.98	0.25	0.25	0.75	0.4	0.14	1

续表

指标 品种	1	2	3	4	5	6	7	8	9	10	11
X_6	0.25	0.98	0.90	0.91	1	1	1	1	0.8	0.83	1
X_7	0.75	0.92	0.84	0.90	0.87	1	0.75	0.75	0.8	0.63	0.67
X_8	0.75	0.93	0.78	0.94	0.85	0.75	0.75	0.75	0.8	0.27	0.67
X_9	0.5	0.91	0.77	0.94	0.83	0.75	0.5	0.75	0.8	0.49	0.67
X_{10}	0.25	0.78	0.95	0.88	0.81	0.75	0.5	0.75	0.4	0.04	1

表 5-9 计算关联系数

指标		1	2	3	4	5	6	7	8	9	10	11
差 序 列	Δ_{01}	0	0.05	0.18	0.19	0.21	0.75	0.5	0.25	0	0.31	0
	Δ_{02}	0.75	0.09	0.13	0.15	0.15	0.75	0.5	0.25	0.2	0.79	0.33
	Δ_{03}	1	0.03	0.11	0.10	0.03	0	0.25	0	0.2	0.52	0.33
	Δ_{04}	0.75	0.10	0.13	0.02	0.04	0	0.25	0.25	0	0.04	0.33
	Δ_{05}	0.5	0.05	0.03	0.14	0.02	0.75	0.75	0.25	0.6	0.86	0
	Δ_{06}	0.75	0.02	0.10	0.09	0	0	0	0	0.2	0.17	0
	Δ_{07}	0.25	0.08	0.16	0.10	0.13	0	0.25	0.25	0.2	0.37	0.33
	Δ_{08}	0.25	0.07	0.22	0.06	0.15	0.25	0.25	0.25	0.2	0.73	0.33
	Δ_{09}	0.5	0.09	0.23	0.06	0.17	0.25	0.5	0.25	0.2	0.51	0.33
	Δ_{10}	0.75	0.22	0.05	0.12	0.19	0.25	0.5	0.25	0.6	0.96	0
关 联 系 数	ξ_{01}	1	0.91	0.74	0.72	0.70	0.40	0.50	0.67	1	0.62	1
	ξ_{02}	0.4	0.85	0.79	0.77	0.77	0.4	0.5	0.67	0.71	0.39	0.6
	ξ_{03}	0.33	0.94	0.82	0.83	0.94	1	0.67	1	0.71	0.49	0.6
	ξ_{04}	0.4	0.83	0.79	0.96	0.93	1	0.67	0.67	1	0.93	0.6
	ξ_{05}	0.5	0.91	0.94	0.78	0.96	0.4	0.4	0.67	0.45	0.37	1
	ξ_{06}	0.4	0.96	0.83	0.85	1	1	1	1	0.71	0.75	1
	ξ_{07}	0.67	0.86	0.76	0.83	0.79	1	0.67	0.67	0.71	0.57	0.60
	ξ_{08}	0.67	0.88	0.69	0.89	0.77	0.67	0.67	0.67	0.71	0.41	0.60
	ξ_{09}	0.5	0.85	0.68	0.89	0.75	0.67	0.5	0.67	0.71	0.50	0.60
	ξ_{10}	0.4	0.69	0.91	0.81	0.72	0.67	0.5	0.67	0.45	0.34	1

由(5-17)式可得绝对差序列 Δ_{0i} , ($i=1, 2, \dots, 10$), 其中 $\Delta(\max)=1, \Delta(\min)=0$. 取 $\rho=0.5$, 由(5-18)式可得关联系数 $\xi_{0i}(j)$, $i=1, 2, \dots, 10, j=1, 2, \dots, 11$, 结果见表 5-9. 最后由

(5-19)式可得各参试品种与理想品种的关联度见表5-10.

表 5-10 最终结果

参试品种	$\rho=0.5$				$\rho=0.4$			
	简单平均		加权平均		简单平均		加权平均	
	关联度	位次	关联度	位次	关联度	位次	关联度	位次
X_1	0.751	4	0.751	4	0.717	4	0.716	4
X_2	0.623	10	0.639	10	0.576	10	0.593	10
X_3	0.759	3	0.782	3	0.727	3	0.754	2
X_4	0.798	2	0.783	2	0.768	2	0.752	3
X_5	0.671	7	0.725	6	0.635	7	0.693	6
X_6	0.864	1	0.876	1	0.844	1	0.860	1
X_7	0.740	5	0.750	5	0.698	5	0.709	5
X_8	0.693	6	0.709	7	0.648	6	0.664	7
X_9	0.665	8	0.674	8	0.618	8	0.627	8
X_{10}	0.651	9	0.668	9	0.609	9	0.625	9

注:各指标权数分别为 0.12,0.11,0.08,0.07,0.20,0.08,0.07,0.10,0.03,0.06,0.08.

按照灰色关联分析的原则,关联度大的品种与理想品种最为接近,是最适宜的品种.从表5-10可以看出,不论是简单平均还是加权平均,品种 X_6 (鲁麦15号)的关联度最大,这说明鲁麦15号与理想品种最为接近,是鲁西北棉区麦棉两熟条件下最适宜的小麦配套品种;品种 X_4 和 X_3 次之,也是该棉区麦棉两熟较适宜的小麦配套品种;其他品种与前边三个品种比较起来与理想品种的关联度都较低,不适宜在该棉区麦棉两熟田推广.

通过对表5-9中的关联系数的进一步分析还可以了解到各品种的特点和存在的问题,如品种 X_6 之所以综合性状好、关联度高,主要是由于该品种在亩穗数、产量及抗逆性等方面与理想品种关联系数较高;品种 X_5 虽然在穗粒数、产量等方面与理想品种关联系数高,但由于其在抗逆性及株高等方面与理想品种关联系数较低,因此关联度低,综合性状稍差.

为了说明权数在计算关联度中的作用,表5-10列出了关联

度的两种计算结果,即关联系数的简单算术平均和加权算术平均.可以看出,加权后各品种的位次稍有改变,比如品种 X_5 超过了品种 X_8 .这主要是由于 X_5 在两个大权重指标(X_5 和 X_2)上与“理想品种”的关联系数较大的缘故.另外,为了对分辨系数的作用有一个直观的认识,表 5-10 还列出了 $\rho = 0.4$ 时的计算结果,可以看出, ρ 取较小的值能够提高评价结果(即关联度)的区分能力,这是灰色关联度评价法的一个显著特点.

第三节 DEA 方法

DEA 为 Data Envelopment Analysis 的简称,即数据包络分析.它是以相对效率概念为基础,根据多指标投入和多指标产出对相同类型的单位(部门或企业)进行相对有效性或效益评价的一种新方法.自从 1978 年,由著名运筹学家查恩斯、库伯以及罗兹首先提出 C^2R 模型并用于评价部门间的相对有效性以来,DEA 方法不断得到完善并在实际中被广泛运用,特别是在对非单纯盈利的公共服务部门,如学校、医院,某些文化设施等的评价方面被认为是一个有效的方法.

下面先从一个简单的例子理解一下 DEA 方法的基本思想.

假设有 5 个生产任务相同的工厂(DEA 方法称它们为决策单元,即 Decision Making Unit,简记为 DMU),比如 5 家水泥厂或 5 家纺织厂等,每个工厂都有两种投入和一种产出,其具体数据见表 5-11.

表 5-11 各厂情况

工厂(DMU)	A	B	C	D	E
投入 1	10	5	1	3	1
投入 2	17	1	1	2	2
产出	120	20	6	24	10

如何对 5 个工厂生产情况的“好坏”进行评价呢？一个很自然的想法是：那些相对投入少而产出多的工厂应是“好”的，反之，则“不好”。为了便于比较，现把 5 个 DMU 的各项投入和产出按比例变化，使其产出相同，见表 5-12。这样就可以只比较其投入了。

表 5-12 各厂调整后的情况

DMU	A	B	C	D	E
投入 1	10	30	20	15	12
投入 2	17	6	20	10	24
产出	120	120	120	120	120

现以两项投入为坐标建立投入平面，如图 5-2 所示。5 个 DMU 分别对应于平面上 5 个等产出点。显然，靠近左下方的 DMU 应该是“好”的，而远离左下方的 DMU 则“不好”。将位于左下方的 A, D, B 依次连接起来，再由 A 向上做纵轴的平行线，由 B 向右做横轴的平行线，可以得到一个“凸包”，这是使所有 5 个 DMU 均位于其右上方的“最小凸包”，A, D, B 位于凸包的边界上，C, E 位于其内部。将 C, E 与原点相连，分别交凸包边界于

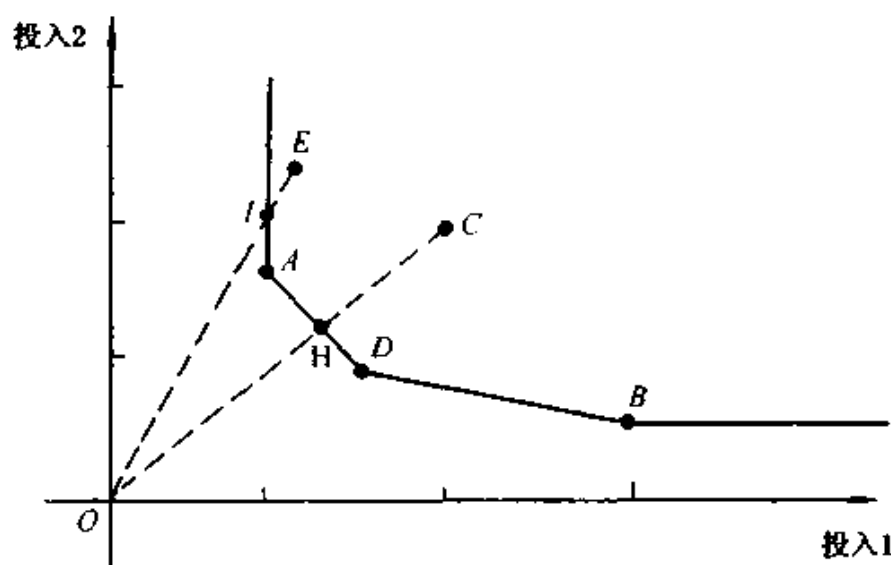


图 5-2

H, I 点. H 由 A 和 D 组合而成, 可以看作是一个虚构的 DMU. 经过简单的计算可知, H 是由 A 的 $\frac{5}{12}$ 和 D 的 $\frac{7}{12}$ 组合而成, 其投入和产出分别为

$$\text{投入 1: } \frac{5}{12} \times 10 + \frac{7}{12} \times 15 = 12.92$$

$$\text{投入 2: } \frac{5}{12} \times 17 + \frac{7}{12} \times 10 = 12.92$$

$$\text{产出: } \frac{5}{12} \times 120 + \frac{7}{12} \times 120 = 120$$

由于 H 是 A 与 D 的组合, 因而在实际生产中是可以实现的, 其两项投入均是 C 两项投入的

$$\frac{12.92}{20} \times 100\% = 64.6\%$$

显然 C 不是相对有效的. 同样, I 也可以看作是一个虚构的 DMU, 其投入分别为 10 和 20, 产出为 120, 其投入是 E 的 83.3%, 因而 E 也不是相对有效的. 很显然, 越偏向右上方, 有效性越差. 相对而言, 位于凸包边界上的 A, D, B 是有效的.

在上面的例子中, DEA 方法评价的对象 DMU 是同类型的工厂, 事实上, DEA 方法评价的对象并不局限于这种真正的生产活动, 它们可以是广义的“生产”活动. 比如: 多所大学、多家医院、多个空军基地、多家银行等都可以作为 DMU, 其基本的特点就是有相同种类的投入和产出, 这些投入产出数据就是评价相对有效性的依据. 判断某个 DMU 是否为相对有效的, 就是看是否有一个虚构的 DMU (它是实际观察到的 DMU 的某种组合) 比它更“好” (相同产出条件下投入更少或相同投入情况下产出更多), 若有这样的 DMU, 则原 DMU 不是相对有效的, 否则, 是相对有效的, 这是 DEA 方法评价的基本思路. 对于多投入多产出情况下的评价则必须借助于线性规划模型.

1978 年, 著名运筹学家查恩斯 (A. Charnes)、库伯 (W. W. Cooper) 及罗兹 (E. Rhodes) 提出了第一个模型, 该模型被命名为 C^2R 模型. 从生产函数的角度讲, C^2R 模型是用来研究多

个 DMU (即多项“生产”活动)的技术有效性和规模有效性的。1985 年查恩斯、库伯、格拉尼 (B. Golany)、赛福德 (L. Seiford) 和斯图茨 (J. Stutz) 提出了另一个模型, 称为 C^2GS^2 模型, 该模型是用来研究多个 DMU 的技术有效性的。下面着重从应用的角度介绍这两个最基本的模型。

一、评价规模有效性和技术有效性的 C^2R 模型

1. C^2R 模型

假设有 n 个 DMU, 每个 DMU 都有 m 种投入和 s 种产出, 如表 5-13 所示。其中

x_{ij} : 第 j 个 DMU 的第 i 种投入总量

y_{rj} : 第 j 个 DMU 的第 r 种产出总量

$i = 1, 2, \dots, m; j = 1, 2, \dots, n; r = 1, 2, \dots, s$ 。

表 5-13

DMU 投入与产出	1	2	...	j	...	n
投入 1	x_{11}	x_{12}	...	x_{1j}	...	x_{1n}
投入 2	x_{21}	x_{22}	...	x_{2j}	...	x_{2n}
\vdots	\vdots	\vdots		\vdots		\vdots
投入 m	x_{m1}	x_{m2}	...	x_{mj}	...	x_{mn}
产出 1	y_{11}	y_{12}	...	y_{1j}	...	y_{1n}
产出 2	y_{21}	y_{22}	...	y_{2j}	...	y_{2n}
\vdots	\vdots	\vdots		\vdots		\vdots
产出 s	y_{s1}	y_{s2}	...	y_{sj}	...	y_{sn}

各 DMU 的投入与产出可用向量表示为

$$X_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T$$

$$Y_j = (y_{1j}, y_{2j}, \dots, y_{sj})^T$$

依据 DEA 方法的评价思想可构造如下的线性规划模型

$$(D) \begin{cases} \min \theta = V_D \\ \text{s. t.} \quad \sum_{j=1}^n \lambda_j X_j \leq \theta X_{j_0} \\ \sum_{j=1}^n \lambda_j Y_j \geq Y_{j_0} \\ \lambda_j \geq 0, j = 1, 2, \dots, n \end{cases} \quad (5-20)$$

其中 $\lambda_j (j=1, 2, \dots, n)$ 为 n 个 DMU 的某种组合权重, $\sum_{j=1}^n \lambda_j X_j$ 和 $\sum_{j=1}^n \lambda_j Y_j$ 分别为按这种权重组合的虚构 DMU 的投入和产出向量, X_{j_0} 和 Y_{j_0} 为所评价的第 j_0 个 DMU 的投入和产出向量. 模型(5-20)的含义非常明显, 即找 n 个 DMU 的某种组合, 使得其产出在不低于第 j_0 个 DMU 的产出的条件下投入尽可能地减小. 该模型是从产出不变, 投入减小的角度构造的, 用于研究投入的有效性. 同理, 从投入不变、产出增加的角度可构造模型来研究产出的有效性:

$$(D') \begin{cases} \max \alpha = V_{D'} \\ \text{s. t.} \quad \sum_{j=1}^n \lambda_j X_j \leq X_{j_0} \\ \sum_{j=1}^n \lambda_j Y_j \geq \alpha Y_{j_0} \\ \lambda_j \geq 0, j = 1, 2, \dots, n \end{cases} \quad (5-21)$$

即找 n 个 DMU 的某种组合, 使得其投入在不高于第 j_0 个 DMU 投入的条件下产出尽可能地增大. 引入松弛变量, 以上两个模型可写为

$$(D) \begin{cases} \min \theta = V_D \\ \text{s. t.} \quad \sum_{j=1}^n \lambda_j X_j + S^- = \theta X_{j_0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = Y_{j_0} \\ \lambda_j \geq 0, j = 1, 2, \dots, n \\ S^+ \geq 0, S^- \geq 0 \end{cases} \quad (5-22)$$

$$(D') \begin{cases} \max \alpha = V_D' \\ \text{s. t. } \sum_{j=1}^n \lambda_j X_j + S^- = X_{j_0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = \alpha Y_{j_0} \\ \lambda_j \geq 0, j = 1, 2, \dots, n \\ S^- \geq 0, S^+ \geq 0 \end{cases} \quad (5-23)$$

其中

$$S^- = (s_1^-, s_2^-, \dots, s_m^-)^T$$

$$S^+ = (s_1^+, s_2^+, \dots, s_s^+)^T$$

s_i^- 和 s_r^+ ($i = 1, 2, \dots, m; r = 1, 2, \dots, s$) 为松弛变量。

以上四个模型统称为 C^2R 模型。以下主要讨论模型 (D) ，研究投入的有效性。

这里有一个问题需要说明一下，有关资料在介绍 DEA 模型时是先引入效率评价指标，利用效率评价指标建立分式规划模型 (\bar{P}) ，再通过 Charnes - Cooper 变换将 (\bar{P}) 化为一个等价的线性规划模型 (P) ，最后得到 (P) 的对偶规划模型 (D) 。由于实际中直接应用的都是模型 (D) ，本书在不影响对 DEA 方法的理解和应用的前提下，直接给出模型 (D) 。对以上转化过程有兴趣的读者可参阅文献 [6] 或 [7]。

2. DEA 有效性 (C^2R)

关于弱 DEA 有效性和 DEA 有效性最初是根据上边提到的模型 (P) 定义的。这里从应用的角度不加证明地直接给出由模型 (D) 判断 DEA 有效性的定理，它本质上就是在模型 (D) 下弱 DEA 有效和 DEA 有效的定义。这样处理更易于对 DEA 方法的理解。

定理 5.1 对于线性规划 (D) ，有

(i) 若 (D) 的最优值 $V_D = 1$ ，则第 j_0 个 DMU 为弱 DEA 有效，反之亦然。

(ii) 若 (D) 的最优值 $V_D = 1$ ，并且它的每个最优解 λ^* ，

S^{*-}, S^{*+}, θ^* 都有

$$S^{*-} = 0, \quad S^{*+} = 0 \text{ (每个分量都为零)}$$

则第 j_0 个 DMU 为 DEA 有效, 反之亦然.

下面对模型 (D) 解的情况作更详细的分析, 进一步弄清弱 DEA 有效和 DEA 有效的直观含义.

如果模型 (D) 的最优值 $V_D < 1$, 这说明存在一个虚构的 DMU (它是 n 个 DMU 的某种组合), 其产出不低于第 j_0 个 DMU 的产出, 而其各项投入均比第 j_0 个 DMU 的投入小 (不超过第 j_0 个 DMU 投入的 V_D 倍). 由定理 5.1, 第 j_0 个 DMU 不是弱 DEA 有效的, 当然更不是 DEA 有效的, 也就是说, 它是非 DEA 有效的, 并且 V_D 越小, 其有效性越差.

如果模型 (D) 的最优值 $V_D = 1$, 设其最优解为

$$\lambda^*, S^{*-}, S^{*+}, \theta^*$$

关于松弛变量可能出现以下几种情况:

(1) $S^{*-} \neq 0, S^{*+} = 0$, 即 S^{*-} 的 m 个分量中有部分分量大于零, 但不可能全部分量大于零, 否则与 $V_D = \theta^* = 1$ 矛盾.

这说明可以通过对实际的 n 个 DMU 按 λ^* 各分量进行组合得到一个虚构的 DMU, 其部分投入 (S^{*-} 中大于零的部分分量所对应的投入) 小于第 j_0 个 DMU 的相应投入, 而其各项产出等于 (由于 $S^{*+} = 0$) 第 j_0 个 DMU 的各项产出.

(2) $S^{*-} = 0, S^{*+} \neq 0$, 即 S^{*+} 的 s 个分量中有部分大于零, 但不可能全部分量大于零, 否则与 $V_D = \theta^* = 1$ 矛盾.

这说明可以通过对实际的 n 个 DMU 按 λ^* 各分量进行组合得到一个虚构的 DMU, 其各项投入等于 (因为 $S^{*-} = 0$) 第 j_0 个 DMU 的投入, 而其部分产出 (与 S^{*+} 中大于零的分量对应的产出) 高于第 j_0 个 DMU 的相应产出.

(3) 对每个解都有 $S^{*-} = 0, S^{*+} = 0$, 即 S^{*-} 中 m 个分量和 S^{*+} 中 s 个分量全部为零.

这说明不存在虚构的 DMU 比第 j_0 个 DMU 更好 (即产出不

低而投入更少).也就是说,要保持第 j_0 个 DMU 的各项产出,不仅各项投入不能整体按比例减少,而且连部分投入也不能再减小.

以上(1)和(2)就是弱 DEA 有效的情形,(3)就是 DEA 有效的情形,可见 DEA 有效比弱 DEA 有效“更有效”.

在图 5-2 中, A, D, B 就属于情形(3),为 DEA 有效, E, C 为非 DEA 有效,虚构的 I 属于情形(1),为弱 DEA 有效.

在弄清了 C^2R 下 DEA 有效性含义后,实际中如何对多个 DMU 进行相对有效性评价呢?由定理 5.1 可知,若用模型(D)判定某个 DMU 为 DEA 有效的,需要检查其所有的解 $\lambda^*, S^{*-}, S^{*+}, \theta^*$ 都满足

$$\theta^* = V_D = 1, S^{*-} = 0, S^{*+} = 0$$

如果只有 $\theta^* = 1$,但并非所有的 $S^{*-} = 0, S^{*+} = 0$,还不能保证第 j_0 个 DMU 的 DEA 有效性.但是对于(D)要判断所有的 $S^{*-} = 0, S^{*+} = 0$ 并不是一件易事,因此在实际中经常直接使用的并非(D),而是一个稍加变化了的模型——具有非阿基米德无穷小 ϵ 的 C^2R 模型.

非阿基米德无穷小 ϵ 是一个小于任何正数而大于零的数(在实际使用中常取为一个足够小的正数,比如 10^{-6}),具有非阿基米德无穷小 ϵ 的 C^2R 模型也有三种形式(\bar{P}_ϵ), (P_ϵ), (D_ϵ).下面直接给出模型(D_ϵ)和判断定理:

$$(D_\epsilon) \begin{cases} \min [\theta - \epsilon(\hat{E}^T S^- + E^T S^+)] = V_D \\ \text{s. t. } \sum_{j=1}^n \lambda_j X_j + S^- = \theta X_{j_0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = Y_{j_0} \\ \lambda_j \geq 0 \quad j = 1, 2, \dots, n \\ S^- \geq 0, S^+ \geq 0 \end{cases} \quad (5-24)$$

其中 $\hat{E} = (1, \dots, 1)_{1 \times m}^T$ 和 $E = (1, \dots, 1)_{1 \times s}^T$ 分别为元素全取 1 的 m

维和 s 维列向量.

可以看出模型 (D_ϵ) 和 (D) 的区别在于目标函数不同, (D_ϵ) 将松弛变量也放入目标函数中, 由 (D_ϵ) 判定 DEA 有效性可依据如下定理:

定理 5.2 设 ϵ 为非阿基米德无穷小, 模型 (D_ϵ) 的最优解为

$$\lambda^*, S^{*-}, S^{*+}, \theta^*$$

(i) 若 $\theta^* = 1$, 则第 j_0 个 DMU 为弱 DEA 有效.

(ii) 若 $\theta^* = 1$, 且 $S^{*-} = 0, S^{*+} = 0$, 则第 j_0 个 DMU 为 DEA 有效.

由定理 5.2 可知, 用模型 (D_ϵ) 判定某个 DMU 为 DEA 有效, 只需检查一个解满足 $\theta^* = 1$ 且 $S^{*-} = 0, S^{*+} = 0$ 即可, 并不需要检查所有的解. 因而模型 (D_ϵ) 在实际中经常使用. 关于模型 (D_ϵ) 的求解, 需要用到单纯形解法, 本书不予介绍, 读者可参阅线性规划方面的文献. 下面看一个应用的例子.

例 5.3 设有四个决策单元, 其投入和产出数据见表 5-14. 试用模型 (D_ϵ) 判定各 DMU 的 DEA 有效性.

表 5-14

DMU	1	2	3	4
投入 1	1	3	3	4
投入 2	3	1	3	2
产出	1	1	2	1

对于 DMU_1 , 模型 (D_ϵ) 为

$$\left\{ \begin{array}{l} \min [\theta - \epsilon(s_1^- + s_2^- + s_1^+)] \\ \text{s. t. } \lambda_1 + 3\lambda_2 + 3\lambda_3 + 4\lambda_4 + s_1^- = \theta \\ \quad 3\lambda_1 + \lambda_2 + 3\lambda_3 + 2\lambda_4 + s_2^- = 3\theta \\ \quad \lambda_1 + \lambda_2 + 2\lambda_3 + \lambda_4 - s_1^+ = 1 \\ \quad \lambda_j \geq 0 \quad j = 1, 2, 3, 4 \\ \quad s_1^- \geq 0, s_2^- \geq 0, s_1^+ \geq 0 \end{array} \right.$$

利用单纯形解法可得如下最优解

$$\lambda^* = (1, 0, 0, 0)^T, s_1^{*-} = s_2^{*-} = s_1^{*+} = 0, \theta^* = 1$$

因而 DMU₁ 为 DEA 有效. 同样可以判定 DMU₂ 和 DMU₃ 为 DEA 有效.

对于 DMU₄, 模型(D_ε)为

$$\begin{cases} \min [\theta - \epsilon(s_1^- + s_2^- + s_1^+)] \\ \text{s. t. } \lambda_1 + 3\lambda_2 + 3\lambda_3 + 4\lambda_4 + s_1^- = 4\theta \\ \quad 3\lambda_1 + \lambda_2 + 3\lambda_3 + 2\lambda_4 + s_2^- = 2\theta \\ \quad \lambda_1 + \lambda_2 + 2\lambda_3 + \lambda_4 - s_1^+ = 1 \\ \quad \lambda_j \geq 0, \quad j = 1, 2, 3, 4 \\ \quad s_1^- \geq 0, s_2^- \geq 0, s_1^+ \geq 0 \end{cases}$$

由单纯形方法可得其最优解为

$$\lambda^* = \left(0, \frac{3}{5}, \frac{1}{5}, 0\right)^T, s_1^{*-} = s_2^{*-} = s_1^{*+} = 0, \theta^* = \frac{3}{5},$$

由于 $\theta^* < 1$, 故 DMU₄ 为非 DEA 有效.

3. DEA 有效性(C²R)的经济含义

如果把相同类型的 DMU 看成是某种“生产”活动, 则 DEA 有效性具有一定的经济含义: 在 C²R 模型下为 DEA 有效的 DMU, 从生产函数的角度讲, 既是“技术有效”的, 也是“规模有效”的. 要理解这一点需要首先弄清以下两个问题, 即生产可能集和生产函数.

(1) 生产可能集

生产可能集是指一些可能的生产活动的集合:

$$T = \{(X, Y) \mid \text{产出向量 } Y \text{ 可以由投入向量 } X \text{ 生产出来}\}$$

由于 n 个 DMU 是实际的某种“生产”活动, 因而 $(X_j, Y_j) \in T, j = 1, 2, \dots, n$. 以这 n 个生产活动为基础, 可以构造其他可能的生产活动. 下面是几条构造规则, 即生产可能集几条公理.

(i) 凸性. 对 $\forall (X, Y) \in T$ 和 $(\hat{X}, \hat{Y}) \in T$ 以及任意的 $\lambda \in [0, 1]$, 均有

$$\begin{aligned} & \lambda(X, Y) + (1 - \lambda)(\hat{X}, \hat{Y}) \\ &= (\lambda X + (1 - \lambda)\hat{X}, \lambda Y + (1 - \lambda)\hat{Y}) \in T \quad (5-25) \end{aligned}$$

即若以 X 的 λ 倍与 \hat{X} 的 $1 - \lambda$ 倍之和投入, 就可得到 Y 的 λ 倍与 \hat{Y} 的 $1 - \lambda$ 倍之和的产出.

(ii) 锥性. 对任意的 $(X, Y) \in T$ 及数 $k \geq 0$, 均有

$$k(X, Y) = (kX, kY) \in T \quad (5-26)$$

即以 X 的 k 倍投入, 获得 Y 的 k 倍的产出是可能的.

(iii) 无效性. 对任意的 $(X, Y) \in T$, 当 $\hat{X} \geq X$ 时均有 $(\hat{X}, Y) \in T$; 当 $\hat{Y} \leq Y$ 时均有 $(X, \hat{Y}) \in T$. 即以较多的投入进行生产总是可能的, 获得较少的产出也总是可能的.

(iv) 最小性. 生产可能集是满足 (i) ~ (iii) 的所有集合的交集.

由 $(X_j, Y_j), (j = 1, 2, \dots, n)$ 决定的满足以上 (i) ~ (iv) 的生产可能集是唯一的, 有

$$\begin{aligned} T = \{ (X, Y) \mid \sum_{j=1}^n X_j \lambda_j \leq X, \sum_{j=1}^n Y_j \lambda_j \geq Y, \\ \lambda_j \geq 0, j = 1, 2, \dots, n \} \triangleq T_1 \end{aligned} \quad (5-27)$$

事实上, 由于 $(X_j, Y_j) \in T_1, j = 1, 2, \dots, n$, 凸性和锥性可以保证其任意组合 $(\sum_{j=1}^n \lambda_j X_j, \sum_{j=1}^n \lambda_j Y_j) \in T_1, \lambda_j \geq 0, j = 1, 2, \dots, n$. 无效性又可使满足

$$X \geq \sum_{j=1}^n \lambda_j X_j, Y \leq \sum_{j=1}^n \lambda_j Y_j \quad (5-28)$$

的生产活动 $(X, Y) \in T_1$, 显然满足 (5-28) 的 (X, Y) 的集合是满足 (i)、(ii)、(iii) 的最小集合.

考虑到模型 (D), 由于 $(X_{j_0}, Y_{j_0}) \in T_1$, 因而

$$\sum_{j=1}^n \lambda_j X_j \leq X_{j_0}$$

$$\sum_{j=1}^n \lambda_j Y_j \geq Y_{j_0} \quad (5-29)$$

其中 $\lambda_j \geq 0, j = 1, 2, \dots, n$. 可以看出模型(D)表示在生产可能集 T_1 内, 当产出 Y_{j_0} 保持不减的条件下尽量将投入量 X_{j_0} 按同一比例 θ 减小.

(2) 生产函数

为了方便, 这里以单投入单产出的情况来说明. 生产函数 $y = f(x)$ 表示投入为 x 时所能获得的最大产出. 一般来说, 随着投入 x (即生产规模) 的增大, 最大产出 y 也随之增大, 生产函数为增函数; 当投入 x 较小时, 最大产出 y 也较小, 随着 x 的逐渐增大, y 也不断增大, 并且增加速度逐渐加快; 当 x 增加到一定程度时, y 的增加速度达到最大; 随着 x 的继续增大, y 的增加速度逐渐减小, 生产函数表现为“S”形的递增函数, 如图 5-3 所示. 用数学语言来描述: 设 $A(x_1, y_1)$ 为生产函数曲线的拐点, 当 $x \in (0, x_1)$ 时, 边际函数 $y'(x)$ 递增, 生产函数的二阶微商 $y''(x) > 0$, 规模收益递增; 当 $x = x_1$ 时, 边际函数 $y'(x)$ 达到最大, $y''(x) = 0$; 当 $x \in (x_1, +\infty)$ 时, 边际函数 $y'(x)$ 递减, $y''(x) < 0$, 规模收益递减. 可见, A 点对应的 DMU 既是技术有效的, 也是规模有效的, 而位于生产函数曲线上的 B 点对应的 DMU 由于规模过大, 仅是技术有效而非规模有效 (由生产函数含义可知, 位于生产函数曲线上的点所对应的 DMU 均是技术有效的), C 点对应的 DMU 则既不是技术有效的, 也不是规模有效的.

(3) DEA 有效 (C^2R) 的经济含义

首先用模型(D)对图 5-3 中 A, B, C 三个 DMU 进行 DEA 相对有效性评价, 这里用图形来考虑. 由 A, B, C 所决定的生产可能集 T_1 为图 5-3 中 AOx 角状区域. 在此区域内, 对 DMU_A 来说, 要保证其产出 y_1 不减, 其投入 x_1 显然不能再减少, 因而 DMU_A 为 DEA 有效, 它既是技术有效的, 也是规模有效的; 对于 DMU_B , 要保持其产出 y_2 不减, 其投入还可降至 x_2' , 因而 DMU_B 为非 DEA 有效, 事实上, 它是技术有效而非规模有效; 对于

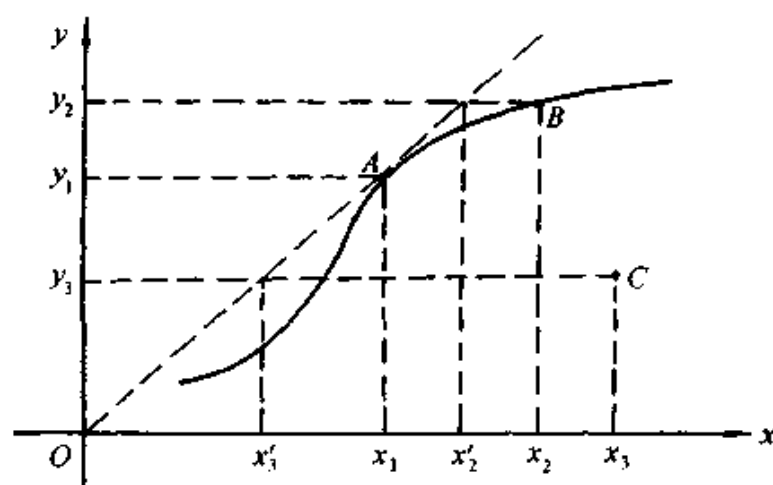


图 5-3

DMU_C, 要保持其产出 y_3 不减, 其投入可降至 x_3' , 因而也是非 DEA 有效的, 事实上, 其既非技术有效也非规模有效。

由上可以看出, 如果某个 DMU 是 DEA 有效的 (C^2R), 则从生产理论来讲, 它既是技术有效的, 也是规模有效的。否则, 它或不为技术有效的, 或不为规模有效的。

二、评价技术有效性的 C^2GS^2 模型

由上边的讨论知道, C^2R 模型是以 T_1 为生产可能集的, T_1 满足生产可能集公理的 (i) ~ (iv)。下面考虑生产可能集 T_2 。 T_2 仅满足生产可能集公理的 (i)、(iii)、(iv), 而不满足锥性。

$$T_2 = \left\{ (X, Y) \left| X \geq \sum_{j=1}^n \lambda_j X_j, Y \leq \sum_{j=1}^n \lambda_j Y_j \right. \right. \\ \left. \left. \sum_{j=1}^n \lambda_j = 1, \lambda_j \geq 0, j = 1, 2, \dots, n \right\} \quad (5-30)$$

在空间中, T_1 是一个凸锥, T_2 是一个凸多面体。 C^2GS^2 是以 T_2 为生产可能集的, 对于 n 个 DMU, 模型 (D) 为

$$(D) \begin{cases} \min \theta = V_D \\ \text{s. t. } \sum_{j=1}^n \lambda_j X_j \leq \theta X_{j_0} \\ \sum_{j=1}^n \lambda_j Y_j \geq Y_{j_0} \\ \sum_{j=1}^n \lambda_j = 1 \\ \lambda_j \geq 0, \quad j = 1, 2, \dots, n \end{cases} \quad (5-31)$$

或

$$(D) \begin{cases} \min \theta = V_D \\ \text{s. t. } \sum_{j=1}^n \lambda_j X_j + S^- = \theta X_{j_0} \\ \sum_{j=1}^n \lambda_j Y_j - S^+ = Y_{j_0} \\ \sum_{j=1}^n \lambda_j = 1 \\ \lambda_j \geq 0, \quad j = 1, 2, \dots, n \\ S^+ \geq 0, \quad S^- \geq 0 (\text{每一分量非负}) \end{cases} \quad (5-32)$$

模型中各种符号含义同前. 可以看出, C^2GS^2 模型(D)与 C^2R 模型(D)相比, 只需在约束条件中加上

$$\sum_{j=1}^n \lambda_j = 1 \quad (5-33)$$

而这也正是生产可能集 T_1 与 T_2 的区别. 关于 C^2GS^2 模型有一些与 C^2R 类似的结果, 下面首先给出在 C^2GS^2 模型(D)下 DEA 有效性的含义.

定义 5.1 若模型(D)(5-32)的最优解 $\lambda^*, S^{*-}, S^{*+}, \theta^*$ 满足 $V_D^* = \theta^* = 1$, 则称第 j_0 个 DMU 为弱 DEA 有效(C^2GS^2).

定义 5.2 若模型(D)(5-32)的任一最优解 $\lambda^*, S^{*-},$

S^{*-+}, θ^* 不仅满足 $V_D^* = \theta^* = 1$, 而且 $S^{*-+} = 0, S^{*-} = 0$, 则称第 j_0 个 DMU 为 DEA 有效(C^2GS^2).

基于与 C^2R 模型相同的原因, 引入非阿基米德无穷小 ϵ , 可得 (D_ϵ) 模型:

$$(D_\epsilon) \begin{cases} \min[\theta - \epsilon(\hat{E}^T S^- + E^T S^+)] \\ \text{s.t.} \quad \sum_{j=1}^n \lambda_j X_j + S^- = \theta X_{j_0} \\ \quad \quad \sum_{j=1}^n \lambda_j Y_j - S^+ = Y_{j_0} \\ \quad \quad \sum_{j=1}^n \lambda_j = 1 \\ \quad \quad \lambda_j \geq 0, j = 1, 2, \dots, n \\ \quad \quad S^- \geq 0, S^+ \geq 0 \end{cases} \quad (5-34)$$

其中符号含义同前.

定理 5.3 设 (D_ϵ) (5-34) 的最优解为

$$\lambda^*, S^{*-}, S^{*-+}, \theta^*$$

(i) 若 $\theta^* = 1$, 则 DMU_{j_0} 为弱 DEA 有效(C^2GS^2).

(ii) 若 $\theta^* = 1$, 且 $S^{*-} = 0, S^{*-+} = 0$, 则 DMU_{j_0} 为 DEA 有效(C^2GS^2).

例 5.4 设有三个 DMU 数据如表 5-15. 试评价其 DEA 有效性(C^2GS^2).

表 5-15 投入产出表

DMU	1	2	3
投入	1	3	4
产出	2	3	1

对于 $DMU_1, (D_\epsilon)$ 为

$$\left\{ \begin{array}{l} \min [\theta - \epsilon(s_1^- + s_1^+)] \\ \text{s. t. } \lambda_1 + 3\lambda_2 + 4\lambda_3 + s_1^- = \theta \\ \quad 2\lambda_1 + 3\lambda_2 + \lambda_3 - s_1^+ = 2 \\ \quad \lambda_1 + \lambda_2 + \lambda_3 = 1 \\ \quad \lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0 \\ \quad s_1^- \geq 0, s_1^+ \geq 0 \end{array} \right.$$

其最优解为

$$\lambda^* = (1, 0, 0)^T, s_1^{*-} = s_1^{*+} = 0, \theta^* = 1$$

由定理 5.3 知, DMU₁ 为 DEA 有效(C²GS²). 同样, DMU₂ 也为 DEA 有效(C²GS²).

对 DMU₃, (D_e) 为

$$\left\{ \begin{array}{l} \min [\theta - \epsilon(s_1^- + s_1^+)] \\ \text{s. t. } \lambda_1 + 3\lambda_2 + 4\lambda_3 + s_1^- = 4\theta \\ \quad 2\lambda_1 + 3\lambda_2 + \lambda_3 - s_1^+ = 1 \\ \quad \lambda_1 + \lambda_2 + \lambda_3 = 1 \\ \quad \lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0 \\ \quad s_1^- \geq 0, s_1^+ \geq 0 \end{array} \right.$$

其最优解为

$$\lambda^* = (1, 0, 0)^T, s_1^{*-} = s_1^{*+} = 0, \theta^* = \frac{1}{4}$$

由定理 5.3 知, DMU₃ 为非 DEA 有效(C²GS²).

在 C²R 模型下 DEA 有效的 DMU, 从生产函数角度讲, 既是技术有效的, 也是规模有效的, 那么在 C²GS² 模型下 DEA 有效有何经济含义呢? 为此, 再来考虑图 5-3 中的三个 DMU, 为了方便将其移至图 5-4 中. 生产可能集 T₂ 是 NBAM_x 围成的右方区域, 在此区域内, 对 A, B 两个 DMU 来说, 要保持其产出不减, 投入已无法再减小, 因而这两个 DMU 均为 DEA 有效(C²GS²). 对于 DMU_C 来说, 要保持其产出不减, 其投入还可再降低, 因而 DMU_C

为非 DEA 有效(C^2GS^2). 可见, 在 C^2GS^2 模型下的 DEA 有效仅是技术有效的, 而不一定是规模有效的. 由此也可以看出, 不同的模型具有不同的评价作用. 若对同一组 DMU, 将两个模型结合起来使用, 就可以进一步弄清每个 DMU 的规模有效性和技术有效性问题. 比如图 5-4 中三个 DMU, 由 C^2R 模型可知 DMU_A 既是技术有效的也是规模有效的, 对于 DMU_B 和 DMU_C 则只知其或不为技术有效, 或不为规模有效, 而通过 C^2GS^2 模型则可知 DMU_B 是技术有效而非规模有效, DMU_C 不是技术有效的.

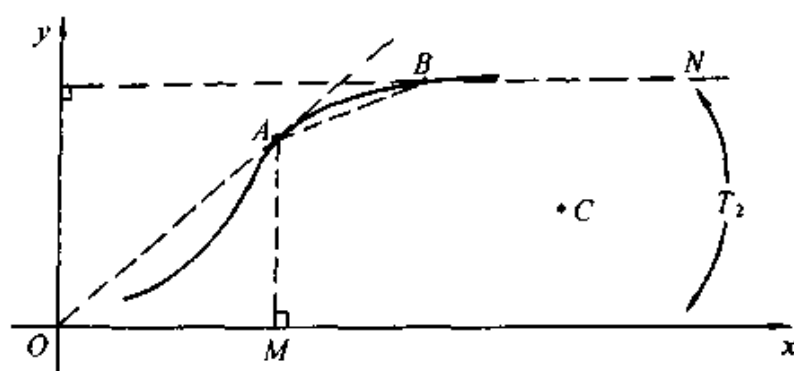


图 5-4

上边主要讨论了从“产出不变, 投入减小”角度构造的 C^2R 和 C^2GS^2 两种模型, 这类模型称为基于投入的模型, 直接用于研究投入的有效性; 从“投入不变, 产出增加”角度构造的模型称为基于产出的模型(比如(5-23)), 其直接用于研究产出的有效性. 这两类模型是从不同的方面分析“生产”活动(DMU)的相对有效性, 它们对应模型的最优解之间存在着一定的关系. 以 C^2R 模型(D)为例, 设(D)(见(5-22))的最优解为

$$\lambda^*, S^{*-}, S^{*+}, \theta^*$$

(D')(见(5-23))的最优解为

$$\lambda^0, S^{0-}, S^{0+}, \alpha^0$$

则有

$$\lambda^0 = \frac{1}{\theta^*} \lambda^*, S^{0-} = \frac{1}{\theta^*} S^{*-}$$

$$S^{0+} = \frac{1}{\theta^*} S^{*+}, \alpha^0 = \frac{1}{\theta^*} \quad (5-35)$$

也就是说,求出基于投入模型的最优解,就可以知道对应的基于产出模型的最优解.实际中常用基于投入的模型进行分析.

以上介绍了 DEA 方法中的两个基本模型,由于实际生产过程和经济活动的多样性,在这两个模型基础上又派生出了一些新的 DEA 模型,比如: C^2W 模型可以处理无穷多个决策单元;锥比率模型 C^2WH 可以处理具有过多投入和产出的情况,并且锥的选取可以体现决策者或评价者的偏好;还有 $C-D$ 型 DEA 模型,含有偏好信息的 DEA 模型等.有关这些模型的介绍可参阅文献[6].

在进入实际应用前还需要强调以下几个问题:

(1)评价对象是同种类型的 DMU,既可以横向对比,比如以多个棉纺企业作为不同的 DMU,也可以纵向对比,比如以不同年份的情况作为不同的 DMU.

(2)这里的指标和前面几章接触到的评价指标有所不同.前面接触到的指标多为投入和产出的综合指标,比如劳动生产率、资金利税率等.而这里是将投入指标与产出指标分离开来,比如以资金、职工人数等为投入指标,而以总产值,利税总额等为产出指标.

(3)由于 DEA 方法并不直接对指标数据进行综合,因而建立模型前无需对数据进行无量纲化处理.可以证明,某个 DMU 的相对有效性评价结果与各投入产出指标的量纲选取无关.

(4)通过对 DEA 模型求解可以将参评的多个 DMU 分成三类:第一类是 DEA 有效的 DMU,第二类是仅为弱 DEA 有效的 DMU,第三类是非 DEA 有效的 DMU.这三类显然已经序化,依次由“好”到“不好”,对于第一类 DMU,DEA 方法并不作出排序,对于第二类 DMU,也不作出排序,对于第三类 DMU 可按各 DMU 的相对有效性值(即模型(D)的最优值 θ^*)来排序, θ^* 越小其相对有效性越差.另外,更为重要的是,对第二类和第三类 DMU 可以找出其“生产”过程中问题所在,为管理提供更为丰富的信息.

最后,看两个应用实例.

例 5.5 中国城市宏观经济状况评价^[6].

本例是 1989 年 Charnes 等人首次利用 DEA 方法对中国 28 个主要城市的宏观经济活动评价的一些结果. 所选投入和产出共 6 个指标:

$$\begin{array}{l} \text{投入} \left\{ \begin{array}{l} \text{劳动力人数(Labor)} \\ \text{年流动资金(WF)} \\ \text{用于扩大再生产的投资额(INV)} \end{array} \right. \\ \text{产出} \left\{ \begin{array}{l} \text{年工业生产总值(GIOV)} \\ \text{年利税总额(PT)} \\ \text{商品零售总额(RS)} \end{array} \right. \end{array}$$

依据表 5-16 和表 5-17 中数据用 C^2R 模型(D_e)进行各城市宏观经济活动的相对有效性评价. 以 DMU_6 (广州)为例, 模型(D_e)简写如下(1983 年):

表 5-16 1983 年各城市投入产出指标数据

DMU 号	城市	产出(万元)			投入		
		GIOV	PT	RS	Labor (万人)	WF (万元)	INV (万元)
1	上海	6,785,798	1,594,957	1,088,699	483.01	1,397,736	616,961
2	北京	2,505,984	545,140	835,745	371.95	855,509	385,453
3	天津	2,292,025	406,947	473,600	268.23	685,584	341,941
4	沈阳	1,158,016	135,939	336,165	202.02	452,713	117,424
5	武汉	1,244,124	204,909	317,709	197.93	471,650	112,634
6	广州	1,187,130	190,178	605,037	178.96	423,124	189,743
7	哈尔滨	658,910	86,514	239,760	148.04	367,012	97,004
8	重庆	993,238	1,411,954	353,896	184.93	408,311	111,904
9	南京	854,188	135,327	239,360	123.33	251,542	91,861
10	西安	606,743	78,357	208,188	116.91	305,316	91,710
11	成都	736,545	114,365	298,112	129.62	295,812	92,409
12	长春	454,684	67,154	233,733	106.26	198,703	53,499
13	太原	494,196	78,992	118,553	89.70	210,891	95,642
14	大连	842,854	149,186	243,361	109.26	282,209	84,202
15	青岛	776,285	116,974	234,875	85.50	184,992	49,357

续表

DMU 号	城市	产出(万元)			投入		
		GIOV	PT	RS	Labor (万人)	WF (万元)	INV (万元)
16	兰州	490,998	117,854	118,924	72.17	222,327	73,907
17	济南	482,448	67,857	158,250	76.18	161,159	47,977
18	抚顺	515,237	114,883	101,231	73.21	144,163	43,312
19	鞍山	625,517	173,099	130,423	86.72	190,043	55,326
20	昆明	382,880	74,126	123,968	69.09	158,436	66,640
21	苏州	867,467	65,229	262,876	77.69	135,046	46,198
22	杭州	830,142	128,279	242,773	97.42	206,926	66,120
23	宁波	521,684	37,245	184,055	54.96	79,563	43,192
24	无锡	869,973	86,859	194,416	67.00	144,092	43,350
25	常州	604,715	55,989	127,586	46.30	100,431	31,428
26	南通	601,299	37,088	224,855	65.12	96,873	28,112
27	宜昌	145,792	11,816	24,442	20.09	50,717	54,650
28	长沙	319,218	31,726	169,051	69.81	117,790	30,976

表 5-17 1984 年各城市投入产出指标数据

DMU 号	城市	产出(万元)			投入		
		GIOV	PT	RS	Labor (万人)	WF (万元)	INV (万元)
1	上海	7,443,700	1,692,100	1323,800	487.44	1,594,040	718,953
2	北京	2,817,200	589,600	1,016,600	375.42	955,124	522,032
3	天津	2,514,900	443,600	566,200	273.69	782,585	371,314
4	沈阳	1,337,100	162,700	410,300	208.82	489,676	138,017
5	武汉	1,377,600	232,900	382,900	199.99	519,686	142,688
6	广州	1,333,600	209,700	614,500	181.90	480,392	259,092
7	哈尔滨	758,947	102,893	298,563	150.93	410,727	95,775
8	重庆	1,157,572	166,978	412,991	188.23	470,104	134,031
9	南京	973,951	166,056	294,927	126.48	296,534	130,325
10	西安	668,107	83,735	248,680	122.70	335,329	105,474
11	成都	834,600	128,455	856,868	133.13	335,605	103,431
12	长春	540,428	86,288	298,182	109.31	277,571	65,906

续表

DMU 号	城市	产出(万元)			投入		
		GIOV	PT	RS	Labor (万人)	WF (万元)	INV (万元)
13	太原	541,923	87,296	145,379	94.45	204,998	130,349
14	大连	918,508	169,458	290,519	113.16	309,779	115,381
15	青岛	849,956	128,665	274,982	87.54	209,197	64,903
16	兰州	540,857	130,638	141,709	73.70	250,558	86,045
17	济南	546,065	82,058	192,074	77.64	181,067	52,858
18	抚顺	547,122	140,122	120,739	74.19	158,850	52,715
19	鞍山	686,383	201,297	152,881	89.69	203,720	76,580
20	昆明	450,743	89,031	171,123	75.52	184,676	78,686
21	苏州	922,572	61,174	255,685	70.73	136,273	13,424
22	杭州	1,008,736	137,108	298,717	68.10	235,282	12,365
23	宁波	664,434	62,610	217,554	58.58	95,712	7,454
24	无锡	1,093,882	97,857	214,078	69.27	174,731	13,906
25	常州	709,278	69,343	150,142	47.97	111,573	10,502
26	南通	694,295	38,004	256,040	67.77	105,075	10,317
27	宜昌	162,454	12,841	29,041	20.07	55,384	1,847
28	长沙	359,956	38,869	201,795	72.37	133,826	4,322

$$\begin{aligned}
 & \min [\theta - \epsilon(s_1^- + s_2^- + s_3^- + s_1^+ + s_2^+ + s_3^+)] \\
 & \text{s. t. } 483.01\lambda_1 + 371.95\lambda_2 + \cdots + 69.81\lambda_{28} + s_1^- = 178.96\theta \\
 & \quad 1397736\lambda_1 + 855509\lambda_2 + \cdots + 117790\lambda_{28} + s_2^- = 423124\theta \\
 & \quad 616961\lambda_1 + 385453\lambda_2 + \cdots + 30976\lambda_{28} + s_3^- = 189743\theta \\
 & \quad 6785798\lambda_1 + 2505984\lambda_2 + \cdots + 319218\lambda_{28} - s_1^+ = 1187130 \\
 & \quad 1594957\lambda_1 + 545140\lambda_2 + \cdots + 31726\lambda_{28} - s_2^+ = 190178 \\
 & \quad 1088699\lambda_1 + 835745\lambda_2 + \cdots + 169051\lambda_{28} - s_3^+ = 605037 \\
 & \quad \lambda_j \geq 0, j = 1, 2, \cdots, 28 \\
 & \quad s_1^- \geq 0, s_2^- \geq 0, s_3^- \geq 0 \\
 & \quad s_1^+ \geq 0, s_2^+ \geq 0, s_3^+ \geq 0
 \end{aligned}
 \tag{D_c}$$

(5-36)

其最优解为

$$\lambda^* = (0, 0, 0, 0, 0, 1, 0, \dots, 0)_{1 \times 28}^T, s_1^{*-} = s_2^{*-} = s_3^{*-} = 0$$

$$s_1^{*+} = s_2^{*+} = s_3^{*+} = 0, \theta^* = 1$$

由定理 5-2 可知,广州市在 1983 年是 DEA 有效的(C^2R).同理,可得广州市 1984 年 C^2R 模型(D_e)的最优解为

$$\lambda^* = (0, \dots, 0, 1.250, 0.084, 0, 0, 0.871, 0, 0)_{1 \times 28}^T$$

$\uparrow \qquad \qquad \uparrow \qquad \qquad \uparrow$
 第 22 个 第 23 个 第 26 个

$$s_1^{*-} = s_2^{*-} = 0, s_3^{*-} = 187206.3$$

$$s_1^{*+} = 587328.1, s_2^{*+} = s_3^{*+} = 0$$

$$\theta^* = 0.8192 < 1$$

故广州市在 1984 年宏观经济活动是非 DEA 有效的(C^2R).广州市 DEA 有效性值的变化率为

$$\frac{0.8192 - 1}{1} \times 100\% = -18.10\%$$

即广州市 1984 年 DEA 有效性值比 1983 年下降了 18.10%. 对其他 27 个 DMU 进行同样的评价,结果见表 5-18.

表 5-18 28 个城市相对有效性值

DMU 号	城市	1983 年有效性	1984 年有效性	变化率(%)
1	上海	1.0000	1.0000	0.00
2	北京	0.7732	0.7674	-0.75
3	天津	0.6578	0.6385	-2.93
4	沈阳	0.5602	0.5071	-9.49
5	武汉	0.7199	0.5477	-23.91
6	广州	1.0000	0.8192	-18.10
7	哈尔滨	0.5041	0.4764	-5.50
8	重庆	0.6255	0.5555	-11.19
9	南京	0.6707	0.6752	0.67
10	西安	0.5522	0.4877	-11.08
11	成都	0.7190	0.6661	-7.30

续表

DMU 号	城市	1983 年有效性	1984 年有效性	变化率(%)
12	长春	0.6916	0.6663	-3.66
13	太原	0.4576	0.4831	5.57
14	大连	0.7837	0.7072	-9.76
15	青岛	1.0000	0.8468	-15.32
16	兰州	0.6565	0.5907	-10.02
17	济南	0.6847	0.6542	-4.50
18	抚顺	0.9212	0.9397	2.01
19	鞍山	1.0000	1.0000	0.00
20	昆明	0.6040	0.6324	4.70
21	苏州	1.0000	1.0000	0.00
22	杭州	0.8683	1.0000	20.48
23	宁波	1.0000	1.0000	0.00
24	无锡	1.0000	1.0000	0.00
25	常州	1.0000	0.9927	-0.70
26	南通	1.0000	1.0000	0.00
27	宜昌	0.5357	0.6324	18.05
28	长沙	0.7141	1.0000	40.04

表 5-19

		实际值(1)	松弛变量(2)	目标改进值(3)
投入	Labor	181.9	0.0	149.0(=181.9×0.8192)
	WF	480392.0	0.0	393557.3(=480392.0×0.8192)
	INV	259092.0	187206.3	25052.7(=259092.0×0.8192-187206.3)
产出	GIOV	1333600.0	587328.1	1920928.1(=1333600.0+587328.1)
	PT	209700.0	0.0	209700.0
	RS	614500.0	0.0	614500.0

针对 1984 年广州市宏观经济活动的非 DEA 有效性,由其模型最优解可得各项指标的目标改进值(见表 5-19)。

下面对总体情况作一分析,从表 5-18 中可以看出,上海(DMU₁)、鞍山(DMU₁₉)、苏州(DMU₂₁)、宁波(DMU₂₃)、无锡

(DMU₂₄)和南通(DMU₂₆)连续二年均为 DEA 有效城市,而中国东北地区的最大工业基地沈阳(DMU₄)不仅连续二年为非相对有效城市,而且有效性值较低.东北地区的另一个重要工业城市哈尔滨(DMU₇)以及长江上游重要城市重庆(DMU₈)情况也与沈阳类似.

另外,一些城市的有效性值从 1983 年到 1984 年发生了较大的变化,例如南方重镇广州(DMU₆)从 1983 年的有效变成 1984 年的非有效,这说明即使广州本身 1984 年的经济状况比 1983 年有很大进步,但就 1984 年的投入/产出来看,广州相对于上海、鞍山、长沙、杭州、苏州、无锡、南通等城市(均为相对有效城市),工业综合效益还不够好.同时,由于长沙从非相对有效城市变为相对有效城市,亦可看出它在这 28 个城市中间,应该说宏观经济运行状况有了较大改善.

从地域上看,地处我国西北和西南地区的西安(DMU₁₀)、太原(DMU₁₃)、兰州(DMU₁₆)、昆明(DMU₂₀),由于地理、交通、人才等方面的原因,有效性值均较低,这是我国经济发展中的一个十分值得注意的问题.

另外,从表 5-18 的最后一列看出,一半以上(15 个)的城市的有效性值具有负增长率.我们认为,这一般并不说明这些城市 1984 年的经济状况不如 1983 年,因为各城市的有效性值是各年份内 28 个城市之间比较后的相对结果,即是“横”向评估的结果,因此,各 DMU 本身不同年份的有效性值缺乏有说服力的可比性.为了说明这一点,我们再把 1983 年与 1984 年的数据放在一起组成一个含有 56 个样本的参考集(1983 年的上海为 DMU₁,1983 年的北京为 DMU₂,……,1984 年的上海为 DMU₂₉,1984 年的北京为 DMU₃₀,……),并用同样的 DEA 模型进行评估,结果见表 5-20.

从表 5-20 中可以看出,56 个 DMU 中只有广州和鞍山这两个城市具有负变化率,而总的平均变化率为 19.490,这说明我国主要城市 1984 年宏观经济状况要比 1983 年好,这个结论与我国

表 5-20

DMU 号 (1983)	有效性值	城市	DMU 号 (1984)	有效性值	变化率(%)
1	1.0000	上海	29	1.0000	0.00
2	0.6987	北京	30	0.7674	9.83
3	0.6320	天津	31	0.6385	1.03
4	0.4455	沈阳	32	0.5071	13.83
5	0.5043	武汉	33	0.5477	8.61
6	0.8407	广州	34	0.8192	-2.56
7	0.4012	哈尔滨	35	0.4764	18.74
8	0.5071	重庆	36	0.5555	9.54
9	0.6080	南京	37	0.6752	11.05
10	0.4361	西安	38	0.4877	11.83
11	0.5242	成都	39	0.6610	26.10
12	0.5758	长春	40	0.6663	15.72
13	0.4171	太原	41	0.4760	14.12
14	0.6426	大连	42	0.7072	10.05
15	0.7852	青岛	43	0.8468	7.85
16	0.5550	兰州	44	0.5907	6.43
17	0.5705	济南	45	0.6542	14.67
18	0.8368	抚顺	46	0.8871	6.00
19	0.9692	鞍山	47	0.9380	-3.22
20	0.5360	昆明	48	0.6324	17.99
21	0.9346	苏州	49	1.0000	7.00
22	0.7343	杭州	50	1.0000	36.18
23	0.9747	宁波	51	1.0000	2.60
24	0.9454	无锡	52	1.0000	5.78
25	0.9329	常州	53	0.9927	6.41
26	0.9607	南通	54	1.0000	4.09
27	0.4595	宜昌	55	0.9876	114.93
28	0.6353	长沙	56	1.0000	57.41

经济发展的实际情况是一致的。

例 5.6 棉纺织企业经济效益评价^[29]. 包头纺织总厂某年的生产活动状况与同行业相比不是非常理想, 有关投入和产出各项

指标及数据见表 5-21,为了找出生产经营活动中问题所在,需要与同行业相比较,试用 C^2GS^2 模型解决这一问题。

表 5-21

评价 指标 DMU (企业)	产出指标		投入指标		
	利税总额 (十万元)	工业总产值 (百万元)	流动资金 (百万元)	固定资产净值 (百万元)	职工人数 (百人)
1. 上海第十七棉纺厂	589.7	299.0	33.5	38.3	95.0
2. 包头纺织总厂	84.3	87.4	53.6	30.5	80.7
3. 牡丹江纺织厂	142.6	136.7	99.7	55.4	102.2
4. 天津第二棉纺厂	324.6	200.2	29.4	51.7	101.6
5. 上海第七棉纺厂	270.6	142.3	13.9	21.8	48.6
6. 天津第一棉纺厂	358.9	205.3	28.1	17.0	89.4
7. 石家庄第四棉纺厂	343.0	214.7	29.2	30.3	87.5
8. 武汉第二棉纺厂	204.4	144.9	47.0	40.5	91.4
9. 上海第十二棉纺厂	415.6	213.8	27.5	27.3	74.4
10. 郑州第四棉纺厂	206.9	137.1	38.2	23.8	94.3

对于 DMU_2 ——包头纺织总厂, C^2GS^2 模型(D)为

$$\begin{aligned}
 (D) \quad & \begin{cases} \min \theta \\ \text{s.t.} \quad 33.5\lambda_1 + 53.6\lambda_2 + \cdots + 38.2\lambda_{10} + s_1^- = 53.6\theta \\ \quad \quad 38.3\lambda_1 + 30.5\lambda_2 + \cdots + 23.8\lambda_{10} + s_2^- = 30.5\theta \\ \quad \quad 90.5\lambda_1 + 80.7\lambda_2 + \cdots + 94.3\lambda_{10} + s_3^- = 80.7\theta \\ \quad \quad 589.7\lambda_1 + 84.3\lambda_2 + \cdots + 206.9\lambda_{10} - s_1^+ = 84.3 \\ \quad \quad 299.0\lambda_1 + 87.4\lambda_2 + \cdots + 137.1\lambda_{10} - s_2^+ = 87.4 \\ \quad \quad \sum_{j=1}^{10} \lambda_j = 1 \\ \quad \quad \lambda_j \geq 0 \quad j=1, 2, \dots, 10 \\ \quad \quad s_1^- \geq 0, s_2^- \geq 0, s_3^- \geq 0, s_1^+ \geq 0, s_2^+ \geq 0 \end{cases} \quad (5-37)
 \end{aligned}$$

其最优解为

$$\lambda^* = (0, 0, 0, 0, 0.8303, 0.1697, 0, 0, 0, 0)$$

$$s_1^{*-} = 20.5689,$$

$$s_2^{*-} = s_3^{*-} = 0, s_1^{*+} = 201.2872$$

$$s_2^{*+} = 65.5930, \theta^* = 0.6880 < 1$$

显然为非 DEA 有效(C^2GS^2). 由以上最优解可得包头棉纺总厂各项指标的目标改进值, 以流动资金为例, 有

$$53.6 \times 0.688 - 20.5689 = 16.3079$$

同样可得其他指标的目标改进值, 结果见表 5-22, 最后一列给出了从实际值到目标值变化的百分比. 从表 5-22 可以看出, 该厂与其他厂相比各项投入均偏高, 而产出偏低, 特别是在实现利税总额方面差距更大.

表 5-22

指 标	实际值	目标值	差距占实际值的百分比(%)
流动资金	53.6	16.3079	69.57
固定资产净值	30.5	20.9840	31.2
职工人数	80.7	55.5216	31.2
利税总额	84.3	285.5872	238.77
工业总产值	87.4	152.9930	75.01

第六章 模糊综合评价

模糊综合评价是借助模糊数学的一些概念,对实际的综合评价问题提供一些评价的方法,它与概率、统计的方法是不同的.因为客观事物的不确定性有两大类:一类是事物对象是明确的,但出现的规律有不确定性,例如晴天、下雨、下雪,这是明确的,但出现规律不确定;另一类是事物对象本身不明确,例如年轻、年老、严重、不严重等这一类程度上的差别没有截然的分界线.后一类对象的不确定性是与分类的不确定有关,也即一个对象是否属于某一类,可以是也可以不是,所以首先要对集合的概念给以拓广,引入模糊集合的概念,一个元素 x 可以属于 A 集合,也可以不属于 A 集合,引入隶属度——隶属函数这一概念,这就导出了模糊数学的构架.这一工作是由美国控制论专家查德(L. A. Zadeh)奠定基础的.自 1965 年首次发表这一工作以来,模糊数学有了很大的发展,得到了广泛的应用.

模糊综合评价就是以模糊数学为基础,应用模糊关系合成的原理,将一些边界不清,不易定量的因素定量化、进行综合评价的一种方法.本章第一节首先从实例出发熟悉模糊综合评价方法,后边几节分别对几个重要的环节予以讨论.

第一节 模糊综合评价的基本程序

本节先从一个例子谈起.假设一个服装厂的设计部门设计出 10 种老年春秋装,那么,哪些服装最适合于市场的需求?工厂将如何安排生产?这是决策者必须解决的问题.很显然,对 10 种服装的综合评价结果将有助于决策.

这里评价的对象为 10 种服装,评价内容为服装适合市场需求

或者说受消费者欢迎的程度.不妨用以下三个指标(或者从以下三个方面)来评价:

花色式样 —— x_1

价格贵贱 —— x_2

耐穿程度 —— x_3

与前边所接触到的指标相比,这三个指标有一个特点,即它们是主观或定性的指标,具有模糊性.因此,可以考虑用模糊方法来评价.

设 10 种服装构成一个普通集合:

$$F = \{F_1, F_2, \dots, F_{10}\}$$

这里 $F_i (i=1, 2, \dots, 10)$ 表示工厂设计出的第 i 种服装.首先构造 F 的几个等级模糊子集:

$H = \{\text{很受消费者欢迎的服装}\}$

$I = \{\text{比较受消费者欢迎的服装}\}$

$J = \{\text{消费者认为一般的服装}\}$

$K = \{\text{不受消费者欢迎的服装}\}$

很显然,某种服装对以上四个模糊集合的属性并不是“非此即彼”,而往往是“亦此亦彼”.如果能得到每种服装总体上对以上四个模糊子集的隶属度,那么问题就基本解决了.为此,先从单方面(指标)考虑,即单因素评价,然后再综合得到总的结果.

先考虑第一种服装(F_1),为了得到 F_1 对 H, I, J, K 的隶属度,可以聘请一批顾客或专门人员对 F_1 分别从 x_1, x_2, x_3 三个方面给出自己的看法.比如,这批人员中有 20% 的人认为 F_1 的花色式样很好,分别有 70% 和 10% 的人认为较好和一般,没有人认为不好.很自然的,可以用 0.2, 0.7, 0.1, 0 分别作为 F_1 从花色式样来看对 H, I, J, K 的隶属度,这样,就得到一个模糊向量

$$(0.2, 0.7, 0.1, 0) \triangleq R|x_1$$

同样的方法,可以得到 F_1 从另外两个指标来看对 H, I, J, K 的隶属度,设已知为

$$(0, 0.4, 0.5, 0.1) \triangleq R|x_2$$

$$(0.2, 0.3, 0.4, 0.1) \triangleq R|x_3$$

将对 F_1 的这些单因素评价结果放到一起就得到一个模糊关系矩阵(也称隶属关系矩阵)

$$R = \begin{Bmatrix} R|x_1 \\ R|x_2 \\ R|x_3 \end{Bmatrix} = \begin{bmatrix} 0.2 & 0.7 & 0.1 & 0 \\ 0 & 0.4 & 0.5 & 0.1 \\ 0.2 & 0.3 & 0.4 & 0.1 \end{bmatrix}$$

它反映了 F_1 从各因素来看对各等级模糊子集的隶属程度.

下面考虑各因素的权重问题, 设因素集合为

$$X = \{x_1, x_2, x_3\}$$

构造 X 的模糊子集

$$M = \{\text{评价服装的重要因素}\}$$

假设 x_1, x_2, x_3 对 M 的隶属度分别为 0.2, 0.5, 0.3, 其构成模糊权向量

$$A = (0.2, 0.5, 0.3)$$

显然, A 的三个分量大小就反映了三个因素的相对重要性. 有了权向量, 就可以对单因素评价结果进行综合, 这需要用到模糊变换.

设 A 为模糊向量, R 为模糊关系矩阵:

$$A = (a_1, a_2, \dots, a_p)$$

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pm} \end{bmatrix}$$

模糊变换是指

$$A \circ R = (b_1, b_2, \dots, b_m) \triangleq B$$

其中

$$b_j = (a_1 \wedge r_{1j}) \vee (a_2 \wedge r_{2j}) \vee \cdots \vee (a_p \wedge r_{pj}) \quad j = 1, 2, \dots, m$$

“ \wedge ”表示取小运算, “ \vee ”表示取大运算. 可以看出, 模糊变换与普通向量和矩阵间的乘法运算过程是一样的, 只是将乘法运算改为

取小运算,将加法运算改为取大运算.

由模糊变换可得

$$\begin{aligned} A \circ R &= (0.2, 0.5, 0.3) \begin{pmatrix} 0.2 & 0.7 & 0.1 & 0 \\ 0 & 0.4 & 0.5 & 0.1 \\ 0.2 & 0.3 & 0.4 & 0.1 \end{pmatrix} \\ &= (0.2, 0.4, 0.5, 0.1) \\ &\triangleq B \end{aligned}$$

B 中第一个分量计算如下:

$$\begin{aligned} &(0.2 \wedge 0.2) \vee (0.5 \wedge 0) \vee (0.3 \wedge 0.1) \\ &= 0.2 \vee 0 \vee 0.2 \\ &= 0.2 \end{aligned}$$

同理可算出其余三个分量.

B 为模糊综合评价结果向量,它反映了 F_1 总体上对 H, I, J, K 的隶属程度.

同样的方法,可以得到 F_2, F_3, \dots, F_{10} 的模糊综合评价结果向量,从中就可以看出各种服装适合市场需求的程度,有助于决策.

从这个例子可以看出,模糊综合评价是通过构造等级模糊子集把反映被评事物的模糊指标进行量化(即确定隶属度),然后利用模糊变换原理对各指标综合,一般需要按以下程序进行:

1. 确定评价对象的因素论域

$$U = \{u_1, u_2, \dots, u_p\}$$

也就是 P 个评价指标.

2. 确定评语等级论域

$$V = \{v_1, v_2, \dots, v_m\}$$

即等级集合,每一个等级可对应一个模糊子集.上例中 $m=4$,

$$V = \{v_1, v_2, v_3, v_4\}$$

其中

v_1 ——很受欢迎——→ H

v_2 ——比较受欢迎——→ I

$$\begin{aligned} v_3 &\text{—— 一般 ——} \rightarrow J \\ v_4 &\text{—— 不受欢迎 ——} \rightarrow K \end{aligned}$$

一般情况下,评语等级数 m 取 $[3, 7]$ 中的整数,如果 m 过大,那么语言难以描述且不易判断等级归属.如果 m 太小又不符合模糊综合评价的质量要求. m 取奇数的情况较多,因为这样可以有一个中间等级,便于判断被评事物的等级归属.具体等级可以依据评价内容用适当的语言描述,比如评价产品的竞争力可取 $V = \{\text{强, 中, 弱}\}$,评价地区的社会经济发展水平可取 $V = \{\text{高, 较高, 一般, 较低, 低}\}$,评价经济效益可取 $V = \{\text{好, 较好, 一般, 较差, 差}\}$,等等.

3. 进行单因素评价,建立模糊关系矩阵 R

在构造了等级模糊子集后,就要逐个对被评事物从每个因素 $u_i (i = 1, 2, \dots, p)$ 上进行量化,也就是确定从单因素来看被评事物对各等级模糊子集的隶属度 $(R | u_i)$,进而得到模糊关系矩阵

$$R = \begin{bmatrix} R | u_1 \\ R | u_2 \\ \dots \\ R | u_p \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ r_{21} & r_{22} & \dots & r_{2m} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & r_{pm} \end{bmatrix}_{p \times m}$$

矩阵 R 中第 i 行第 j 列元素 r_{ij} 表示某个被评事物从因素 u_i 来看对 v_j 等级模糊子集的隶属度.一个被评事物在某个因素 u_i 方面的表现是通过模糊向量 $(R | u_i) = (r_{i1}, r_{i2}, \dots, r_{im})$ 来刻画的,而在其他评价方法中多是由一个指标实际值来刻画的,因此,从这个角度讲模糊综合评价要求更多的信息.

4. 确定评价因素的模糊权向量 $A = (a_1, a_2, \dots, a_p)$

一般情况下, p 个评价因素对被评事物并非是同等重要的,各单方面因素的表现对总体表现的影响也是不同的,因此在合成之前要确定模糊权向量.在模糊综合评价中,权向量 A 中的元素 a_i 本质上是因素 u_i 对模糊子集 $\{\text{对被评事物重要的因素}\}$ 的隶属度,因而一般用模糊方法来确定,并且在合成之前要归一化.

5. 利用合适的合成算子将 A 与各被评事物的 R 合成得到各被评事物的模糊综合评价结果向量 B .

R 中不同的行反映了某个被评价事物从不同的单因素来看对各等级模糊子集的隶属程度. 用模糊权向量 A 将不同的行进行综合就可得到该被评事物从总体上来看对各等级模糊子集的隶属程度, 即模糊综合评价结果向量 B . 模糊综合评价的模型为

$$\begin{aligned} A \circ R &= (a_1, a_2, \dots, a_p) \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pm} \end{pmatrix} \\ &= (b_1, b_2, \dots, b_m) \\ &\triangleq B \end{aligned}$$

其中 b_j 是由 A 与 R 的第 j 列运算得到的, 它表示被评事物从整体上看对 v_j 等级模糊子集的隶属程度.

6. 对模糊综合评价结果向量进行分析

每一个被评事物的模糊综合评价结果都表现为一个模糊向量, 这与其他方法中每一个被评事物得到一个综合评价值是不同的, 它包含了更丰富的信息. 对不同的一维综合评价值可以方便地进行比较并排序, 而对不同的多维模糊向量进行比较排序就不那么方便了, 具体分析方法我们将在第四节讨论.

以上为模糊综合评价的六个基本步骤, 其中第 3 步和第 5 步为比较核心的两步. 第 3 步为模糊单因素评价, 本质上是求隶属度, 在实际应用中往往要凭经验来选取合适的方法, 并且工作量相当大. 第 5 步的合成本质上是对模糊单因素评价结果的综合, 真正体现了综合评价.

第二节 模糊单因素评价

模糊单因素评价方法与因素(指标)的属性有关, 不同属性的

因素可以采用不同的方法.

一、主观或定性指标的模糊评价方法

一般而言,主观或定性的指标都具有一定程度的模糊性.比如上节例中的 x_1 (花色式样)就是一个主观的定性指标,因为对某种服装的花色式样,不同的人会有不同的看法(主观性)并且这种看法是一种定性的描述.对其评价为“好看”、“比较好看”等都具有模糊性.另外,再比如,人们对社会保障的信任程度、对环境质量的满意程度、产品质量高低等都属于这类指标.对这类指标的模糊评价方法主要是以模糊统计试验为依据的等级比重法.

下面结合上节的例子来说明.

设 W 是一个论域,比如 F_i 表示第 i 种服装,

$$W = \{F_1, F_2, \dots, F_{10}\}$$

F_1 为 W 中一个元素.再考虑 W 上的一个子集合

$$H^* = \{\text{花色式样好看的服装}\}$$

显然,对于某一个人来说, H^* 是一个普通的集合,但是对于不同的人来说, H^* 可能是不相同的.因为不同的人判断 F_1 是否属于 H^* ,有时会得到 $F_1 \in H^*$,有时会得到 $F_1 \notin H^*$.因此,对一批评价者来说 H^* 是 W 上一个运动的、边界可变的普通子集合,它就对应着一个模糊集合

$$H = \{\text{花色式样好看的服装}\}$$

F_1 对 H 的隶属度可表示为

$$\mu_H(F_1) = \lim_{n \rightarrow \infty} \frac{F_1 \in H^* \text{ 的次数}}{n} \quad (6-1)$$

其中 n 为试验次数(这里为评价人员数),随着 n 的增大,比重

$$\frac{F_1 \in H^* \text{ 的次数}}{n} \quad (6-2)$$

趋于 $[0,1]$ 中的一个稳定数,它就是 F_1 对 H 的隶属度.

一般地,设等级论域为

$$V = \{v_1, v_2, \dots, v_r, \dots, v_m\}$$

分别对应模糊子集 E_1, E_2, \dots, E_m , 如何确定某评价对象 F 对 $E_j (j=1, 2, \dots, m)$ 的隶属度呢? 可让一批评价者 (共 n 人) 分别给出其对该问题的看法并统计结果, 见表 6-1.

表 6-1

等 级	v_1	v_2	\dots	v_j	\dots	v_m
认为 F 属于某等级的人数	n_1	n_2	\dots	n_j	\dots	n_m
等级对应的模糊子集	E_1	E_2	\dots	E_j	\dots	E_m
F 对 E_j 的隶属度 $\mu_{E_j}(F)$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	\dots	$\frac{n_j}{n}$	\dots	$\frac{n_m}{n}$

其中 $n_1 + n_2 + \dots + n_m = n$.

用等级比重法确定隶属度时, 为了保证可靠性, 一般要注意两个问题: 第一, 评价者人数不能太少, 因为根据模糊统计试验, 只有当试验次数 n 充分大时, 等级比重才趋于隶属度. 第二, 评价者必须对被评事物有相当的了解, 特别是一些涉及专业方面的评价, 比如对酒的综合评价, 如果请一些不会喝酒的人, 也许所有的酒都会是一个样.

例 6.1 广东省五个地区生存质量的模糊综合评价^[30].

本次评价采用了世界卫生组织生存质量表. 因素论域为

$$U = \{u_1, u_2, u_3, u_4, u_5\}$$

其中

- u_1 : 生理领域;
- u_2 : 心理领域;
- u_3 : 社会关系领域;
- u_4 : 环境领域;
- u_5 : 独立性领域.

每个领域分别包含若干个条目. 评语等级论域为

$$V = \{\text{极差}(1), \text{差}(2), \text{一般}(3), \text{好}(4), \text{极好}(5)\}$$

共进行了 564 例的现场问卷调查, 每个被调查者依据其自身生存

质量状况对各因素的具体条目按 V 中 5 个等级进行判断,最后分地区进行统计.表 6-2 为中山地区调查统计结果,表 6-3 为湛江地区调查统计结果,其余地区的统计结果从略.

表 6-2 中山地区生存质量情况

评价因素	所包含 条目 个数	评 价 等 级				
		1	2	3	4	5
生理领域	16	70(70)	213(213)	574(574)	645(645)	258(258)
心理领域	24	57(38)	371(247.33)	999(666)	1038(692)	383(255.33)
社会关系领域	8	25(50)	72(144)	235(470)	439(878)	178(356)
环境领域	32	131(65.5)	598(299)	1600(800)	1127(563.5)	367(183.5)
独立性领域	16	77(77)	117(117)	342(342)	658(658)	723(723)

* 圆括号中数字为以 16 条为标准,将条目数目标标准化后得到,以下均同.

表 6-3 湛江地区生存质量情况

评价因素	所包含 条目 个数	评 价 等 级				
		1	2	3	4	5
生理领域	16	129(129)	230(230)	572(572)	530(530)	196(196)
心理领域	24	107(71.33)	406(270.67)	996(664)	851(576.33)	228(152)
社会关系领域	8	22(44)	95(190)	265(530)	362(724)	141(282)
环境领域	32	211(105.5)	546(273)	1480(740)	988(494)	283(141.5)
独立性领域	16	111(111)	215(215)	391(391)	566(566)	473(473)

利用等级比重法可得单因素评价结果,以中山地区为例,从生理领域来看对各等级模糊子集的隶属度为

$$\begin{aligned}
 R_1|u_1 &= \left(\frac{70}{1760}, \frac{213}{1760}, \frac{574}{1760}, \frac{645}{1760}, \frac{258}{1760} \right) \\
 &= (0.04, 0.12, 0.33, 0.37, 0.15)
 \end{aligned}$$

其中

$$1760 = 70 + 213 + 574 + 645 + 258$$

同样可得其他单因素评价结果.最后可得两地区的模糊关系矩阵为

$$R_1 = \begin{bmatrix} 0.04 & 0.12 & 0.33 & 0.37 & 0.15 \\ 0.02 & 0.13 & 0.35 & 0.36 & 0.13 \\ 0.03 & 0.08 & 0.25 & 0.46 & 0.19 \\ 0.03 & 0.16 & 0.42 & 0.29 & 0.10 \\ 0.04 & 0.06 & 0.18 & 0.34 & 0.38 \end{bmatrix}_{5 \times 5}$$

$$R_2 = \begin{bmatrix} 0.08 & 0.14 & 0.35 & 0.32 & 0.12 \\ 0.04 & 0.16 & 0.38 & 0.33 & 0.09 \\ 0.02 & 0.11 & 0.30 & 0.41 & 0.16 \\ 0.06 & 0.16 & 0.42 & 0.28 & 0.08 \\ 0.06 & 0.12 & 0.22 & 0.32 & 0.27 \end{bmatrix}_{5 \times 5}$$

其余地区的模糊关系矩阵从略,后面几节将继续讨论.

二、客观或定量指标的模糊评价方法

1. 隶属函数法

首先以人的年龄—— x 为例说明隶属函数. 设模糊集合

$$L = \{\text{年轻人}\}$$

显然,对 L 的隶属度随年龄 x 而变化,因而对 L 的隶属度可以看作 x 的函数,记为 $\mu_L(x)$,比如模糊集合创始人扎德曾给出

$$\mu_L(x) = \begin{cases} 1 & , 0 \leq x \leq 25 \\ \frac{1}{1 + \left(\frac{x-25}{5}\right)^2} & , 25 < x \leq 200 \end{cases} \quad (6-3)$$

一般地,设 X 为一论域, L 为一模糊集合,若对 $\forall x \in X$ 都对应一个对 L 的隶属度,即存在一个映射

$$\begin{aligned} \mu_L: X &\longrightarrow [0,1] \\ x &\longrightarrow \mu_L(x) \end{aligned}$$

则称 $\mu_L(x)$ 为 L 关于 X 的隶属函数.

客观或定量指标一般都有数值表现结果,如果能得到它们对等级模糊子集的隶属函数,求隶属度就非常方便了.从模糊学可以知道,隶属函数的分布主要有戒上型、戒下型和中间型三种,并且

有一些具体的函数形式^[8],因此实际评价中只需依据指标的特性和等级模糊集合选择一种合适的形式并确定其中参数即可。

例 6.2 大学毕业生专业学习水平的模糊综合评价。

以往常用的方法是将各门专业课的学习成绩求加权平均,然后排序。如果要给出一个总的鉴定意见,一般是通过划分数段来确定,比如不低于 90 分为优秀,75 分至 90 分为良好,60 分至 75 分为合格,60 分以下为不合格。若甲、乙、丙三名学生的专业课加权平均成绩分别为 90.5, 89, 76, 则他们总的鉴定意见分别为优秀、良好、良好。这样的结论显然有一定的不合理性:甲与乙仅 1.5 分之差就分属于两个等级,而乙与丙相差 13 分仍属一个等级。如果断定某学生的所属等级时附之以隶属度,则能提供更为丰富的评价信息,一定程度上可以克服以上的不合理性。

设因素论域

$$U = \{u_1, u_2, \dots, u_p\}$$

其中

u_i :第 i 门专业课学习成绩, $i = 1, 2, \dots, p$

评语等级论域为

$$V = \{\text{优秀, 良好, 合格, 不合格}\}$$

对应的等级模糊子集分别为 H, I, J, K 。

设 x 为某门专业课的分数, $x \in [0, 100] = X$, 下面分别构造隶属函数 $\mu_H(x), \mu_I(x), \mu_J(x), \mu_K(x)$ 。

(1) $\mu_H(x)$

显然, x 越大则对 H 的隶属度越大, 因此应选择戒下型的隶属函数。考虑升半哥西分布

$$\mu(x) = \begin{cases} 0, & x \leq a \\ \{1 + [\alpha(x - a)]^{-\beta}\}^{-1}, & x > a \end{cases} \quad (6-4)$$

其中 $\alpha > 0, \beta > 0$ 。如图 6-1 所示, 可取 $a = 80$, 即若成绩不高于 80, 则对 H 的隶属度为零, 若成绩高于 80 才对 H 有非零的隶属度, 并且成绩越高, 对 H 的隶属度越大。考虑到 $x = 100$ 时应有 $\mu_H(x) = 1$, 故可考虑将 (6-4) 修改为

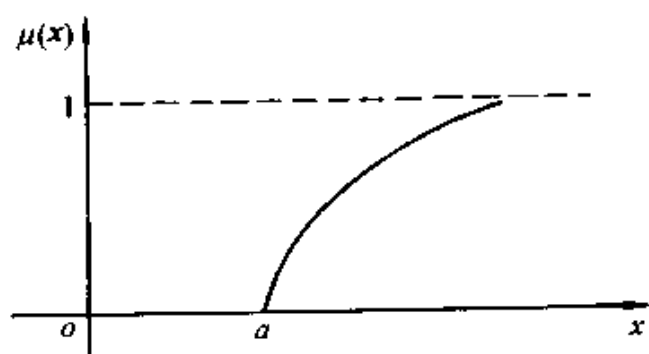


图 6-1

$$\mu(x) = \begin{cases} 0, & 0 \leq x \leq 80 \\ \{1 + [\alpha(x - 80)]^{-\beta}\}^{-1}, & 80 < x < 90 \\ \frac{x}{100}, & 90 \leq x \leq 100 \end{cases} \quad (6-5)$$

实际中常取 $\beta=2$. 另外, 为了保证 $\mu(x)$ 的连续性, 应有

$$\lim_{x \rightarrow 90^-} \{1 + [\alpha(x - 80)]^{-2}\}^{-1} = 0.9 \quad (6-6)$$

不难求得满足(6-6)的 $\alpha = \frac{3}{10}$, 因此

$$\mu_H(x) = \begin{cases} 0, & 0 \leq x \leq 80 \\ \left\{1 + \left[\frac{3}{10}(x - 80)\right]^{-2}\right\}^{-1}, & 80 < x < 90 \\ \frac{x}{100}, & 90 \leq x \leq 100 \end{cases} \quad (6-7)$$

(2) $\mu_K(x)$

考虑降半哥西分布

$$\mu(x) = \begin{cases} 1, & x \leq a \\ \{1 + [\alpha(x - a)]^\beta\}^{-1}, & x > a \end{cases} \quad (6-8)$$

其中 $\alpha > 0, \beta > 0$, 如图 6-2 所示. 取 $a = 55, \beta = 2$, 再令 $\mu(60) = \frac{1}{2}$, 可得 $\alpha = \frac{1}{5}$, 故有

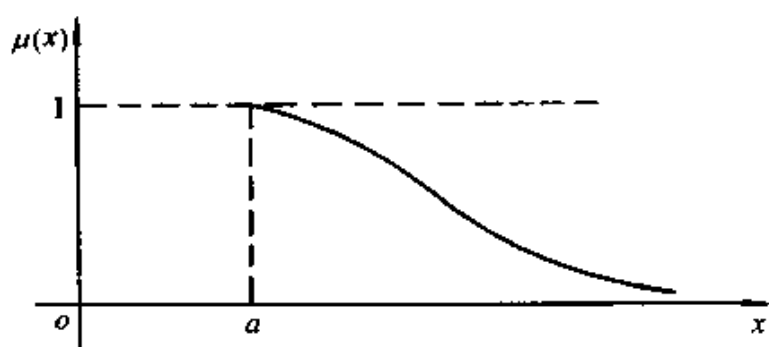


图 6-2

$$\mu_K(x) = \begin{cases} 1, & x \leq 55 \\ \left[1 + \left(\frac{x-55}{5}\right)^2\right]^{-1}, & 55 \leq x \leq 100 \end{cases} \quad (6-9)$$

(3) $\mu_I(x), \mu_J(x)$

直接考虑哥西分布, 分别取 $a = 80$ 和 $a = 67.5$, 取 $\alpha = 1/7.5$, 有

$$\mu_I(x) = \left[1 + \left(\frac{x-80}{7.5}\right)^2\right]^{-1} \quad (6-10)$$

$$\mu_J(x) = \left[1 + \left(\frac{x-67.5}{7.5}\right)^2\right]^{-1} \quad (6-11)$$

四个隶属函数分别如图 6-3(a)~(d)所示.

假设某学生的四门专业课成绩分别为 79, 88, 94, 70, 将其分别代入(6-7), (6-10), (6-11)和(6-9)可得如下隶属关系矩阵 R .

因 素	模 糊 子 集				
		H	I	J	K
u_1		$\begin{bmatrix} 0 & 0.98 & 0.30 & 0.042 \\ 0.85 & 0.47 & 0.118 & 0.022 \\ 0.94 & 0.22 & 0.074 & 0.016 \\ 0 & 0.36 & 0.9 & 0.1 \end{bmatrix} = R$			
u_2					
u_3					
u_4					

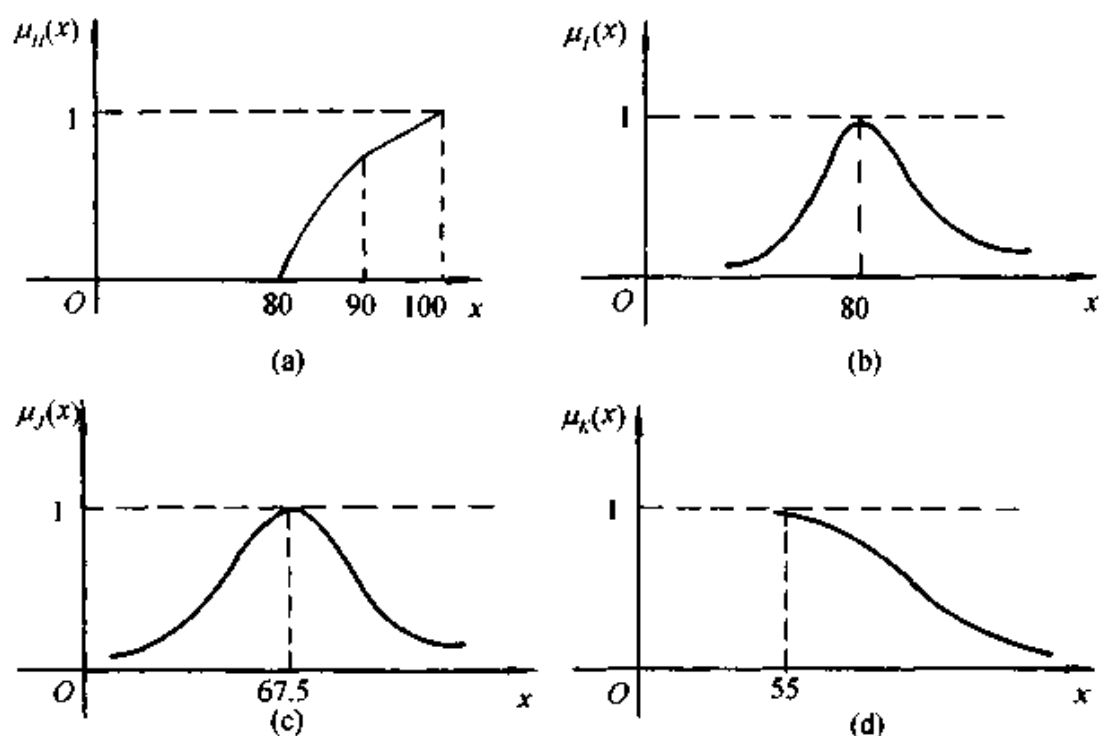


图 6-3

本例中的因素为各门专业课成绩,论域 $X = [0, 100]$ 相同,因而只确定了 4 个隶属函数.一般而言,若有 p 个因素(指标), m 个等级模糊子集,则需要确定出 $p \times m$ 个隶属函数.因此该方法工作量较大,实际应用不很普遍.另外,本例采用的是曲线型的隶属函数,当然也可以选用较为简单的折线型隶属函数,这在后边的例子中将会看到.

2. 频率法

该方法是先划分指标值在不同等级的变化区间,然后以指标值的历史资料数据在各等级变化区间内出现的频率作为对各等级模糊子集的隶属度.

例 6.3 工业企业经济效益的模糊综合评价.^[32]

首先对每一指标确定等级变化范围,具体结果见表 6-4.

某企业在过去一年(12 个月)中,各项月度指标数值属于表 6-4 中对应区间的次数(见表 6-5).

表 6-4 指标等级范围

等级 数值区域 评价指标	优	良	一般	较差	差	甚差
产品销售率 (%)	98 以上	98~96	96~94	94~92	92~91	91 以下
资金利税率 (%)	17 以上	17~16	16~15	15~14	14~13	13 以下
工业增加值率 (%)	9 以上	9~8	8~7	7~6	6~5	5 以下
成本费用利润率 (%)	30 以上	30~29	29~28	28~27	27~26	26 以下
全员劳动生产率 (元/人)	1050 以上	1050~1040	1040~1030	1030~1020	1020~1010	1010 以下
营运资金周转率 (次/月)	0.29 以上	0.29~0.28	0.28~0.27	0.27~0.26	0.26~0.25	0.25 以下

注:表中数值区域为左开右闭区间,如 98~96 为(96,98]。

表 6-5

次数(隶属度) 等级 评价指标	优	良	一般	较差	差	甚差	合计
产品销售率	0(0.000)	3(0.250)	6(0.500)	2(0.167)	0(0.000)	1(0.083)	12(1)
资金利税率	2(0.167)	3(0.250)	6(0.500)	1(0.083)	0(0.000)	0(0.000)	12(1)
工业增加值率	2(0.167)	3(0.250)	4(0.333)	2(0.167)	1(0.083)	0(0.000)	12(1)
成本费用利润率	0(0.000)	4(0.333)	5(0.417)	2(0.167)	1(0.083)	0(0.000)	12(1)
全员劳动生产率	1(0.083)	5(0.417)	5(0.417)	1(0.083)	0(0.000)	0(0.000)	12(1)
营运资金周转率	2(0.167)	3(0.250)	3(0.250)	2(0.167)	1(0.083)	1(0.083)	12(1)

以第一行为例,该企业在过去的 12 个月中,产品销售率有 3 个月在 96% ~ 98% 之间,6 个月在 94% ~ 96% 之间,2 个月在 92% ~ 94% 之间,有 1 个月在 91% 以下.即在 12 次中有 0 次优,3 次良,6 次一般,2 次较差,1 次差,这些次数除以 12 可得各等级出现的频率,以此作为对各等级模糊子集的隶属度,见表 6-5 中括号内数据.

这种方法操作方便,工作量小,但是比较粗糙,指标值的等级区间划分会影响到评价结果.另外,以频率作为隶属度需要较多的历史数据,这样可能会导致时间跨度较大,因而评价的意义就值得怀疑.

第三节 单因素模糊评价的综合

通过模糊单因素评价可以得到隶属关系矩阵 R , R 反映了一个被评事物在各因素方面对各等级模糊子集的隶属情况,那么总体情况如何呢?这就需要对单因素评价结果进行综合.

一、模糊权向量的确定

设因素论域

$$U = \{u_1, u_2, \dots, u_p\}$$

U 上的模糊子集

$$M = \{\text{对评价内容重要的因素}\}$$

因素 u_i 对 M 的隶属度为 a_i' , 则模糊权向量为

$$A = \{a_1, a_2, \dots, a_p\}$$

其中

$$a_i = \frac{a_i'}{\sum_{i=1}^p a_i'}, \quad i = 1, 2, \dots, p \quad (6-12)$$

对模糊权向量的确定多采用专家估计法,即请几位专家分别估计出 $u_i (i = 1, 2, \dots, p)$ 对 M 的隶属度,然后对不同专家的估计

结果求平均并归一化就可以得到 A 。

例 6.4 导弹性能的综合评价。

因素论域为

$$U = \{\text{威力}(u_1), \text{有效性}(u_2), \text{机动能力}(u_3)\}$$

U 上模糊子集

$$M = \{\text{评价导弹性能的重要因素}\}$$

现有三位专家分别估计出 $u_i (i=1, 2, 3)$ 对 M 的隶属度见表 6-6。因而, $A = (0.4, 0.3, 0.3)$ 。

表 6-6

	u_1	u_2	u_3
专家 1	0.75	0.56	0.65
专家 2	0.8	0.61	0.57
专家 3	0.85	0.63	0.58
平均	0.8	0.6	0.6
归一化	0.4	0.3	0.3

二、选择模糊合成算子

模糊综合评价的原理是模糊变换, 模型为

$$\begin{aligned}
 A \circ R &= (a_1, a_2, \dots, a_p) \circ \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pm} \end{pmatrix} \\
 &= (b_1, b_2, \dots, b_m) \quad (6-13)
 \end{aligned}$$

其中“ \circ ”为模糊合成算子 $M(\frac{*}{*}, *)$, “ $\frac{*}{*}$ ”和“ $*$ ”是模糊变换的两种运算, 具体的

$$b_j = (a_1 \frac{*}{*} r_{1j}) * (a_2 \frac{*}{*} r_{2j}) * \cdots * (a_p \frac{*}{*} r_{pj}) \quad j = 1, 2, \dots, m \quad (6-14)$$

常用的模糊合成算子有以下几种:

1. $M(\wedge, \vee)$ 算子

$$b_j = \bigvee_{i=1}^p (a_i \wedge r_{ij}) \\ = \max_{1 \leq i \leq p} \{ \min(a_i, r_{ij}) \}, \quad j = 1, 2, \dots, m \quad (6-15)$$

2. $M(\cdot, \vee)$ 算子

$$b_j = \bigvee_{i=1}^p (a_i \cdot r_{ij}) = \max_{1 \leq i \leq p} \{ a_i r_{ij} \} \quad j = 1, 2, \dots, m \quad (6-16)$$

3. $M(\wedge, \oplus)$ 算子

“ \oplus ”是有界和运算,即在有界限制下的普通加法运算.对 t 个实数 x_1, x_2, \dots, x_t 有

$$x_1 \oplus x_2 \oplus \dots \oplus x_t = \min \left\{ 1, \sum_{k=1}^t x_k \right\}$$

“ \sum ”是连续有界和符号,上式可写为

$$\sum_{k=1}^t x_k = \min \left\{ 1, \sum_{k=1}^t x_k \right\}$$

利用 $M(\wedge, \oplus)$ 算子,有

$$b_j = \sum_{i=1}^p (a_i \wedge r_{ij}) \\ = \min \left\{ 1, \sum_{i=1}^p \min(a_i, r_{ij}) \right\} \quad j = 1, 2, \dots, m \quad (6-17)$$

4. $M(\cdot, \oplus)$

$$b_j = \sum_{i=1}^p (a_i \cdot r_{ij}) \\ = \min \left(1, \sum_{i=1}^p a_i r_{ij} \right), \quad j = 1, 2, \dots, m \quad (6-18)$$

模糊合成算子 $M(\cdot, \circledast)$ 由两步运算组成,第一步运算“ \cdot ”用于 a_i 对 r_{ij} 的修正,第二步运算“ \circledast ”用于对修正后的 r_{ij} ($i = 1, 2, \dots, p$) 的综合.上边四个算子在第一步分别用了取小和乘法运算,第二步分别用了取大和有界和运算.从体现权数作用来讲,第一步用乘法运算较为合适,从综合的角度来讲,第二步用有界和运

算较为合适,它可以保证充分利用 R 阵提供的各方面信息.表 6-7 对四个算子作了一些比较.对综合评价而言, $M(\cdot, \oplus)$ 较为合适并在实际中常被采用.

表 6-7

算子	$M(\wedge, V)$	$M(\cdot, V)$	$M(\wedge, \oplus)$	$M(\cdot, \oplus)$
比较内容				
体现权数作用	不明显	明显	不明显	明显
综合程度	弱	弱	强	强
利用 R 的信息	不充分	不充分	比较充分	充分
类型	主因素决定型	主因素突出型	不均衡平均型	加权平均型

例 6.5(续例 6.2) 由有关教师、学生和教学管理人员结合四门专业课程的学时数和课程内容等共同确定的模糊权向量为

$$A = (0.35, 0.20, 0.25, 0.20)$$

采用 $M(\cdot, \oplus)$ 算子可得该生专业课学习的模糊综合评价结果向量

$$A \circ R = (0.35, 0.20, 0.25, 0.20) \begin{pmatrix} 0 & 0.98 & 0.30 & 0.042 \\ 0.85 & 0.47 & 0.118 & 0.022 \\ 0.94 & 0.22 & 0.074 & 0.016 \\ 0 & 0.36 & 0.9 & 0.1 \end{pmatrix}$$

$$= (0.405, 0.564, 0.327, 0.043)$$

$$\triangleq B$$

B 说明了该生专业学习的总体情况对各等级模糊子集的隶属程度.

第四节 模糊综合评价结果向量的分析

模糊综合评价的结果是被评事物对各等级模糊子集的隶属度,它构成一个模糊向量,而不是一个点值,因而它能提供的信息比其他方法更丰富.若对多个事物比较并排序,就需要进一步处理

模糊综合评价结果向量. 本节介绍几种常用的方法.

一、最大隶属度原则

设模糊综合评价结果向量为

$$B = (b_1, b_2, \dots, b_m) \quad (6-19)$$

若 $b_r = \max_{1 \leq j \leq m} \{b_j\}$, 则被评事物总体上来讲隶属于第 r 等级, 这就是最大隶属原则. 在例 6.5 中

$$B = (0.405, 0.564, 0.327, 0.043)$$

因此该生专业学习总体评价为“良好”. 稍加留心就会发现, 这种实际中最常用的方法在某些情况下使用会显得很勉强, 损失信息较多, 甚至得出不合理的评价结果. 比如 $B = (0.2, 0.7, 0.68)$ 就属于这样的情况. 由此可见, 最大隶属原则的使用是有条件的. 文献[1]探讨了最大隶属原则的有效性问题, 并提出了改进的方法.

对(6-19)定义

$$\beta = \max_{1 \leq j \leq m} \{b_j\} / \sum_{j=1}^m b_j \quad (6-20)$$

$$\gamma = \sec_{1 \leq j \leq m} \{b_j\} / \sum_{j=1}^m b_j \quad (6-21)$$

其中 $\sec_{1 \leq j \leq m} \{b_j\}$ 表示 B 中的第二大分量, β 和 γ 分别是 B 中最大分量与次大分量占各分量总和的比例, 易得 $\beta \in \left[\frac{1}{m}, 1\right]$, $\gamma \in \left[0, \frac{1}{2}\right]$. 再由 β 和 γ 定义

$$\beta' = \frac{\beta - \frac{1}{m}}{1 - \frac{1}{m}} = \frac{m\beta - 1}{m - 1} \quad (6-22)$$

$$\gamma' = \frac{\gamma - 0}{\frac{1}{2} - \gamma} = 2\gamma \quad (6-23)$$

则 $\beta' \in [0, 1]$, $\gamma' \in [0, 1]$, 再定义

$$\alpha = \frac{\beta'}{\gamma'} = \frac{m\beta - 1}{2\gamma(m-1)} \quad (6-24)$$

可以证明, $\alpha \in [0, +\infty)$, 由 α 的定义可以看出, α 越大, 最大隶属度原则有效度越高, 因此, 可以用 α 指标对最大隶属度原则的有效性进行度量. 表 6-8 给出了一般情况下 α 的不同范围对应的最大隶属度原则的有效性情况.

表 6-8

α	B	最大隶属度原则的有效性
$+\infty$	$(0, \dots, 0, 1, 0, \dots, 0)$	完全有效
$[1, +\infty)$		非常有效
$[0.5, 1)$		比较有效
$(0, 0.5)$		低效
0	$(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m})$	完全失效

对于例 6.5 中的 B, 有

$$\alpha = \frac{4 \times \frac{0.564}{0.405 + 0.564 + 0.327 + 0.043} - 1}{2 \times \frac{0.405}{0.405 + 0.564 + 0.327 + 0.043} \times (4-1)} = 0.38$$

一般情况下, 当 $\alpha < 0.5$ 时, 最大隶属度原则是低效的, 不宜采用. 文献[1]提出了此时应使用加权平均求隶属等级的方法.

二、加权平均原则

加权平均原则是基于这样的思想: 将等级看作一种相对位置, 使其连续化. 为了能定量处理, 不妨用“1, 2, 3, \dots , m ”依次表示各等级, 并称其为各等级的秩. 然后用 B 中对应分量将各等级的秩加权求和, 得到被评事物的相对位置. 这就是加权平均原则, 可表示为

$$A = \frac{\sum_{j=1}^m b_j^k \cdot j}{\sum_{j=1}^m b_j^k} \quad (6-25)$$

其中 k 为待定系数($k=1$ 或 $k=2$), 目的是控制较大的 b_j 所起的作用. 可以证明, 当 $k \rightarrow \infty$ 时, 加权平均原则就是最大隶属原则.

对例 6.5 中的 B , 有

$$A_{k=2} = \frac{1 \times 0.405^2 + 2 \times 0.564^2 + 3 \times 0.327^2 + 4 \times 0.043^2}{0.405^2 + 0.564^2 + 0.327^2 + 0.043^2} = 1.9$$

即该生专业学习水平为良好稍偏优一点.

对于多个被评事物可以依据其等级位置进行排序.

三、模糊向量单值化

如果给各等级赋以分值, 然后用 B 中对应的隶属度将分值加权求平均就可以得到一个点值, 便于比较排序.

设给 m 个等级依次赋以分值 c_1, c_2, \dots, c_m , 一般情况下(等级是由高到低或由好到差), $c_1 > c_2 > \dots > c_m$, 且间距相等, 则模糊向量可单值化为

$$c = \frac{\sum_{j=1}^m b_j^k c_j}{\sum_{j=1}^m b_j^k} \quad (6-26)$$

其中 k 的含义与作用同(6-25)中的 k . 多个被评事物可以依据(6-26)式由大到小排出次序.

以上几种处理方法可依据评价的目的来选用, 如果只需给出某事物一个总体评价结论, 则用第一种方法, 如果需要序化, 可选用后两种方法.

例 6.6 (续例 6.1) 广东省五个地区生存质量的模糊综合评价.

例 6.1 中已得到了五个地区的隶属关系矩阵. 各因素的权重采用对比排序法确定, 每位被调查者在作自身生存质量评定后, 接着给五个因素按在他心目中的重要性排序, 最不重要的排在第 1 位, 其次排在第 2 位, …… , 最重要的排在第 5 位, $1, 2, \dots, 5$ 为各

因素的秩次,各因素的权重由下式计算

$$w'_j = \frac{\sum_{k_j=1}^5 f_{k_j} \cdot \log m^{k_j}}{\sum_{k_j=1}^5 f_{k_j}} \quad (6-27)$$

$$w_j = \frac{w'_j}{\sum w'_j} \quad (6-28)$$

其中

m ——评价因素的个数

k_j ——第 j 个因素的秩次

f_{k_j} —— k_j 出现的频数

将调查统计结果代入上式,即可得各因素的权重,组成如下模糊权向量

$$A = (0.30, 0.20, 0.15, 0.12, 0.24)$$

下面采用算子 $M(\cdot, \oplus)$ 进行合成,对中山地区有

$$A \circ R = (0.30, 0.20, 0.15, 0.12, 0.24) \begin{bmatrix} 0.04 & 0.12 & 0.33 & 0.37 & 0.15 \\ 0.02 & 0.13 & 0.35 & 0.36 & 0.13 \\ 0.03 & 0.08 & 0.25 & 0.46 & 0.19 \\ 0.03 & 0.16 & 0.42 & 0.29 & 0.10 \\ 0.04 & 0.06 & 0.18 & 0.34 & 0.38 \end{bmatrix}$$

$$= (0.03, 0.11, 0.30, 0.37, 0.20)$$

$$\triangleq B_{\text{中山}}$$

同样可得其他四个地区的模糊综合评价结果向量为

$$B_{\text{湛江}} = (0.06, 0.13, 0.33, 0.33, 0.15)$$

$$B_{\text{梅州}} = (0.03, 0.10, 0.24, 0.43, 0.20)$$

$$B_{\text{韶关}} = (0.03, 0.10, 0.32, 0.40, 0.16)$$

$$B_{\text{广州}} = (0.04, 0.12, 0.33, 0.33, 0.18)$$

给五个等级(极差、差、一般、好、极好)分别赋以秩次 5, 4, 3, 2, 1,

由(6-25)可得

$$A_{\text{中山}} = \frac{0.03 \times 5 + 0.11 \times 4 + 0.30 \times 3 + 0.37 \times 2 + 0.20 \times 1}{0.03 + 0.11 + 0.30 + 0.37 + 0.20} = 2.41$$

$$A_{\text{湛江}} = 2.62, \quad A_{\text{梅州}} = 2.33$$

$$A_{\text{韶关}} = 2.45, \quad A_{\text{广州}} = 2.51$$

因而五个地区生存质量由好到差的排序依次为

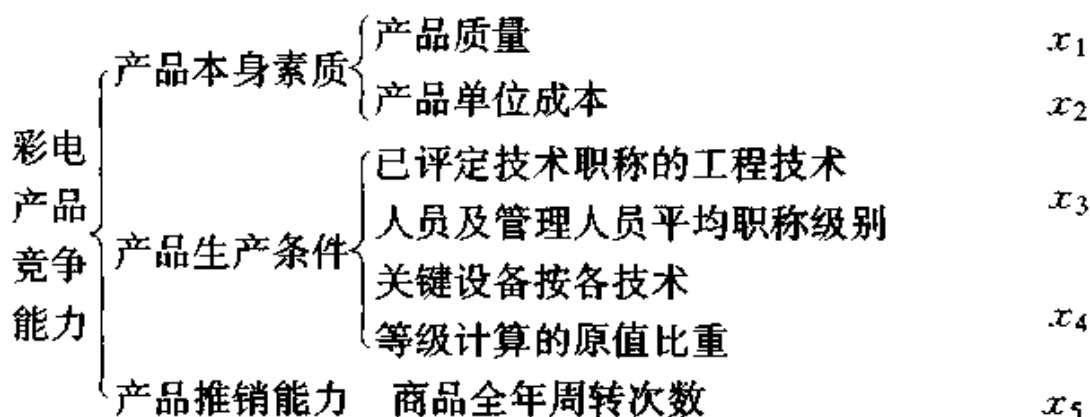
梅州 中山 韶关 广州 湛江

利用模糊向量单值化也可得到同样的排序结果,读者可以自己来完成.

另外,对于本例,由于参评对象数目较少,仅五个地区,因此也可以考虑用隶属度直接对比的方法来排序.首先,按最大隶属原则,中山、梅州、韶关三地区生存质量隶属于等级“好”,而湛江和广州两地区隶属于“好”和“一般”之间,因而相对于前三个地区稍差一些.对前三个地区再考虑它们对等级“极好”和“好”的隶属度大小不难排出顺序:梅州地区生存质量相对最好,以下依次是中山、韶关.对于后两个地区同样可以得出结论:广州稍好.综上,通过隶属度直接对比的方法也得出了与前边一致的排序结果.

例 6.7 彩电产品竞争能力的模糊综合评价^[34].

产品竞争能力的强弱直接关系到企业的存亡,因而评价产品的竞争能力具有重要的意义.彩电产品的竞争能力可以从以下三个方面用 5 个指标来反映:



其中

商品全年周转次数 = 商品全年销售额 / 年平均库存额
因而因素论域为

$$U = \{x_1, x_2, x_3, x_4, x_5\}$$

等级论域取为

$$V = \{v_1 \text{——强}, v_2 \text{——中}, v_3 \text{——弱}\}$$

本次评价以 1985 年全国工业普查资料为基础数据来源,评价对象为全国 16 个地区.本例仅对天津市彩电产品竞争力作一评价.

1. 针对不同的指标,结合数据资料选用合适的单因素评价方法.

(1) x_1 的单因素评价

产品质量是一个定性指标.由于该次普查中对主要工业产品质量实行三级分等,即优等品、一等品、合格品,因而可以将这三个等级分别与 v_1, v_2, v_3 对应,直接以等级品率作为产品质量的单因素评价结果.天津市彩电产品的等级品率为:一等品率 41%,合格品率 59%,因而

$$R|x_1 = (0, 0.41, 0.59)$$

(2) x_4 的单因素评价

该次普查中,把关键设备的技术水平划分为国际水平、国内先进水平、国内一般水平、国内落后水平四个等级.不妨把最后两个等级合并变为三个等级,这样就可以把三个等级的彩电生产线按原值计算的比重作为 x_4 的单因素评价结果.天津市当时只拥有一条彩电生产线,属国内先进水平,因而

$$R|x_4 = (0, 1, 0)$$

(3) x_2 的单因素评价

单位产品成本是一个定量的逆指标,这里采用折线型隶属函数法进行单因素评价.首先确定 x_2 对应的强、中、弱三个等级模糊子集的代表值,16 个地区 x_2 的平均值为 1182 元/台, x_2 高于 1182 元/台的有 8 个地区,其 x_2 平均值为 1241 元/台, x_2 低于 1182 元/台的有 8 个地区,其 x_2 平均值为 1123 元/台.因而可用

以上三个均值作为三个代表值,构造的隶属函数如图 6-4 所示.

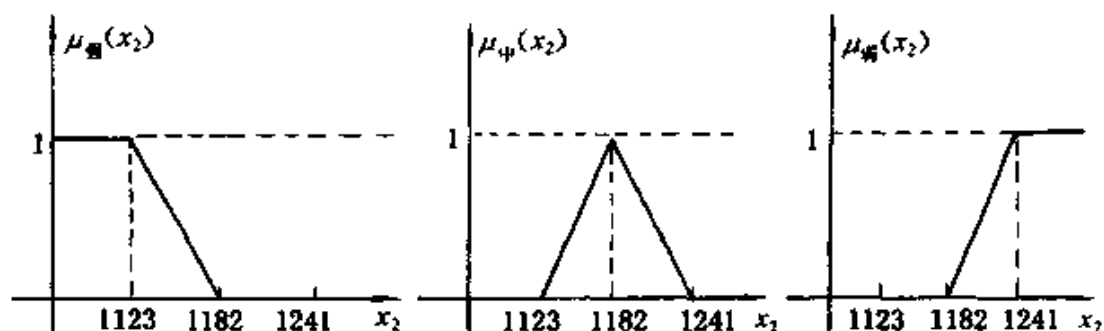


图 6-4

$$\mu_{\text{强}}(x_2) = \begin{cases} 1, & x_2 \leq 1123 \\ \frac{1182 - x_2}{1182 - 1123}, & x_2 \in (1123, 1182) \\ 0, & x_2 \geq 1182 \end{cases} \quad (6-29)$$

$$\mu_{\text{中}}(x_2) = \begin{cases} 0, & x_2 \leq 1123 \text{ 或 } x_2 \geq 1241 \\ \frac{x_2 - 1123}{1182 - 1123}, & x_2 \in (1123, 1182] \\ \frac{1241 - x_2}{1241 - 1182}, & x_2 \in (1182, 1241) \end{cases} \quad (6-30)$$

$$\mu_{\text{弱}}(x_2) = \begin{cases} 0, & x_2 \leq 1182 \\ \frac{x_2 - 1182}{1241 - 1182}, & x_2 \in (1182, 1241) \\ 1, & x_2 \geq 1241 \end{cases} \quad (6-31)$$

对于天津市, $x_2 = 1130$ 元/台. 由以上三式可得

$$R|x_2 = (0.88, 0.12, 0)$$

(4) x_5 的单因素评价

类似于 x_2 , 先确定几个代表点. 但由于数据原因不能用同 x_2 一样的方法, 这里作一点权宜处理. 全国彩电产品周转次数是 20 次/年, 它可以作为“中”的代表值. 就 x_5 而言, 其极端下限值为 0,

取 0~20 的三等分点 6.67 为等级“弱”的代表值,以 20 为中心取 6.67 的对称点 $20 + (20 - 6.67) = 33.33$ 为等级“强”的代表值,可构造 x_5 的隶属函数(如图 6-5 所示):

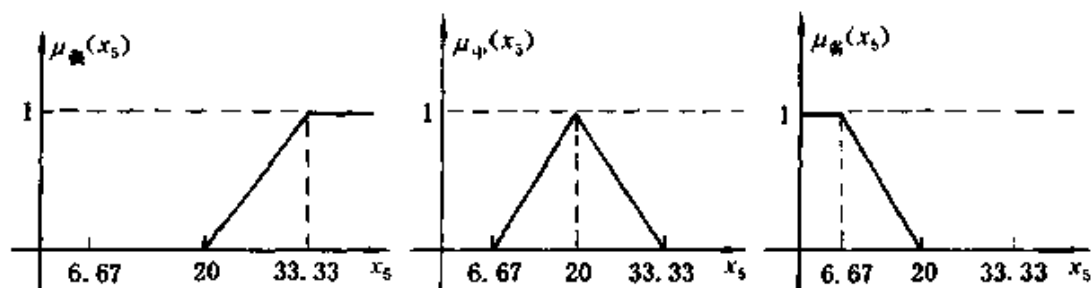


图 6-5

$$\mu_{\text{强}}(x_5) = \begin{cases} 0, & x_5 \leq 20 \\ \frac{x_5 - 20}{33.33 - 20}, & x_5 \in (20, 33.33) \\ 1, & x_5 \geq 33.33 \end{cases} \quad (6-32)$$

$$\mu_{\text{中}}(x_5) = \begin{cases} 0, & x_5 \leq 6.67 \text{ 或 } x_5 \geq 33.33 \\ \frac{x_5 - 6.67}{20 - 6.67}, & x_5 \in (6.67, 20] \\ \frac{33.33 - x_5}{33.33 - 20}, & x_5 \in (20, 33.33) \end{cases} \quad (6-33)$$

$$\mu_{\text{弱}}(x_5) = \begin{cases} 1, & x_5 \leq 6.67 \\ \frac{20 - x_5}{20 - 6.67}, & x_5 \in (6.67, 20) \\ 0, & x_5 \geq 20 \end{cases} \quad (6-34)$$

天津市彩电产品全年周转次数为 46.04 次/年,故有

$$R|x_5 = (1, 0, 0)$$

(5) x_3 的单因素评价

x_3 是个比例指标,先得考虑将其转化为数值指标.不妨给各级技术职称分别赋值,高级、中级、初级分别赋值 3,2,1,然后利用下式来转化:

$$x = \frac{3 \times s + t \times 2 + r \times 1}{s + t + r} \quad (6-35)$$

其中 $s:t:r$ 为各级职称比例. 高级职称太多了, x 值偏大; 反之, x 值偏小. 显然这是一适度指标. 若合理的职称比例为

$$\text{高级} : \text{中级} : \text{初级} = 1 : 3 : 6$$

则由(6-35)可得该指标适度值为

$$\frac{3 \times 1 + 2 \times 3 + 1 \times 6}{1 + 3 + 6} = 1.5$$

很显然, 该指标上下限分别为 3 和 1. 如图 6-6 所示, 将 $[1, 1.5]$ 和 $[1.5, 3]$ 分别 5 等分, 取两侧分点 1.1 和 2.7 作为等级“弱”的代表值, 取 1.3 和 2.1 为等级“中”的代表值, 取 1.5 为等级“强”的代表值, 构造隶属函数如图 6-7 所示.

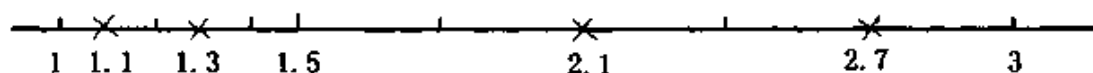


图 6-6

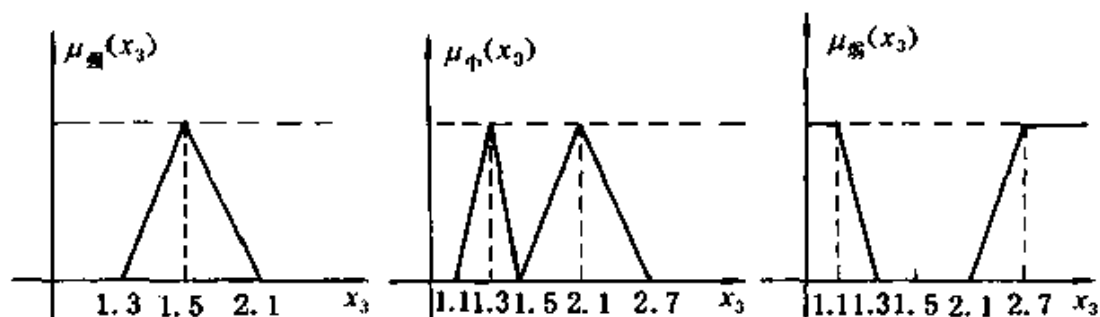


图 6-7

$$\mu_{\text{强}}(x_3) = \begin{cases} 0, & x_3 \leq 1.3 \\ \frac{x_3 - 1.3}{1.5 - 1.3}, & x_3 \in (1.3, 1.5] \\ \frac{2.1 - x_3}{2.1 - 1.5}, & x_3 \in (1.5, 2.1) \\ 0, & x_3 \geq 2.1 \end{cases} \quad (6-36)$$

$$\mu_{\text{弱}}(x_3) = \begin{cases} 1, & x_3 \leq 1.1 \\ \frac{1.3 - x_3}{1.3 - 1.1}, & x_3 \in (1.1, 1.3) \\ 0, & x_3 \in [1.3, 2.1] \\ \frac{x_3 - 2.1}{2.7 - 2.1}, & x_3 \in (2.1, 2.7) \\ 1, & x_3 \geq 2.7 \end{cases} \quad (6-37)$$

$$\mu_{\text{中}}(x_3) = \begin{cases} 0, & x_3 \leq 1.1 \\ \frac{x_3 - 1.1}{1.3 - 1.1}, & x_3 \in (1.1, 1.3] \\ \frac{1.5 - x_3}{1.5 - 1.3}, & x_3 \in (1.3, 1.5] \\ \frac{x_3 - 1.5}{2.1 - 1.5}, & x_3 \in (1.5, 2.1] \\ \frac{2.7 - x_3}{2.7 - 2.1}, & x_3 \in (2.1, 2.7] \\ 0, & x_3 \geq 2.7 \end{cases} \quad (6-38)$$

天津市彩电行业的职称比例为 0:31:69, 利用(6-35)可转化为

$$\frac{0 \times 3 + 31 \times 2 + 69 \times 1}{0 + 31 + 69} = 1.31$$

代入(6-36), (6-37), (6-38)可得

$$R|x_3 = (0.05, 0.95, 0)$$

综上可得天津市彩电产品的单因素评价结果为

$$R = \begin{bmatrix} 0 & 0.41 & 0.59 \\ 0.88 & 0.12 & 0 \\ 0.05 & 0.95 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}_{5 \times 3}$$

2. 由专家咨询的方法可得 5 个因素的模糊权向量为

$$A = (0.30, 0.30, 0.15, 0.15, 0.10)$$

3. 采用 $M(\cdot, \oplus)$ 可得天津市彩电产品的模糊综合评价结果为

$$B = A \circ R = (0.37, 0.45, 0.18)$$

4. 模糊向量单值化

分别给 v_1, v_2, v_3 赋以分值 30, 20, 10. 由 (6-26) 式可得天津地区的评价值为

$$\frac{30 \times 0.37^2 + 20 \times 0.45^2 + 10 \times 0.18^2}{0.37^2 + 0.45^2 + 0.18^2} = 22.81$$

同样的过程可以得到其他地区的模糊评价结果, 进而比较排序.

第五节 多级模糊综合评价

前边主要讨论了单级模糊综合评价, 而对较复杂的事物往往需要进行多级模糊综合评价. 以例 6-2 为例, 可用同样的方法再进行专业基础课和公共基础课学习水平的模糊综合评价, 下面直接给出该生在这两方面的模糊综合评价结果向量:

$$B_2 = (0.301, 0.548, 0.265, 0.05)$$

$$B_3 = (0.102, 0.404, 0.578, 0.04)$$

经有关教师和教学管理人员共同确定的三类课程: 专业课、专业基础课、公共基础课的模糊权向量为

$$A = (0.35, 0.35, 0.30)$$

用三方面的模糊综合评价结果向量组成隶属关系矩阵可再进行更高一层的模糊综合评价, 得到该生总体学习水平的模糊综合评价结果向量:

$$\begin{aligned} A \circ R &= (0.35, 0.35, 0.30) \begin{pmatrix} 0.405 & 0.564 & 0.327 & 0.043 \\ 0.301 & 0.548 & 0.265 & 0.05 \\ 0.102 & 0.404 & 0.578 & 0.04 \end{pmatrix} \\ &= (0.278, 0.51, 0.381, 0.045) \end{aligned}$$

这就反映了该生的总体学习情况.

一般地,多级模糊综合评价需要经过以下步骤:

1. 把因素论域按某种属性分成 s 个子集.

$$U = \bigcup_{i=1}^s u_i \quad (6-39)$$

其中

$$u_i = \{u_{i1}, u_{i2}, \dots, u_{ip_i}\}, \quad i = 1, 2, \dots, s$$

2. 对每一个 u_i 进行单级模糊综合评价.

设评语等级论域为

$$V = \{v_1, v_2, \dots, v_m\}$$

u_i 中各因素的模糊权向量为

$$A_i = (a_{i1}, a_{i2}, \dots, a_{ip_i})$$

$$\sum_{r=1}^{r=p_i} a_{ir} = 1 \quad (6-40)$$

u_i 的单因素评价结果为 R_i (p_i 行, m 列), 单级评价模型为

$$A_i \circ R_i = (b_{i1}, b_{i2}, \dots, b_{im})$$

$$\triangleq B_i \quad i = 1, 2, \dots, s \quad (6-41)$$

3. 将 u_i 看作一个综合因素, 用 B_i 作为它的单因素评价结果, 可得隶属关系矩阵

$$R = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_s \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{s1} & b_{s2} & \cdots & b_{sm} \end{bmatrix}$$

设综合因素 u_i ($i = 1, 2, \dots, s$) 的模糊权向量为

$$A = (a_1, a_2, \dots, a_s)$$

则二级模糊综合评价模型为

$$A \circ R = (b_1, b_2, \dots, b_m) \triangleq B \quad (6-42)$$

如果第一步划分中 u_i ($i = 1, 2, \dots, s$) 仍较多, 则可继续划分得到三级或更高级的模型.

二级模型见图 6-8.

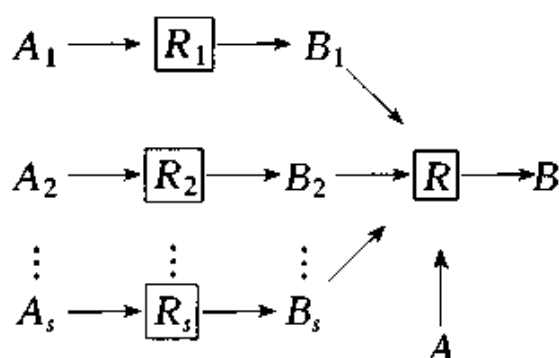


图 6-8 二级模型示意图

例 6.8 战略导弹效能的多级模糊综合评价^[35].

为了分析和评价战略导弹系统的效能,可建立指标体系,见图 6-9.

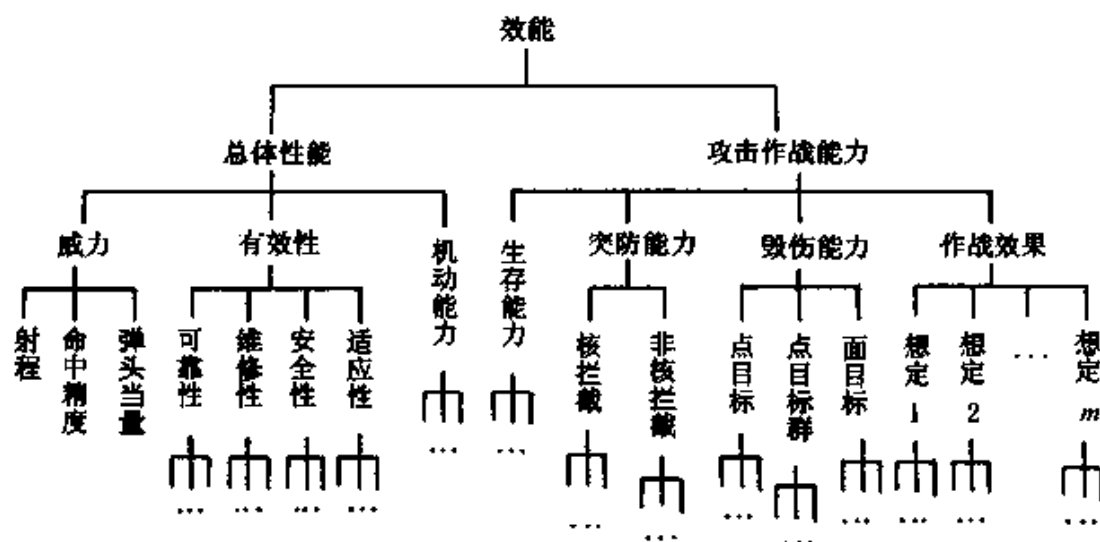


图 6-9

评语等级分为 5 级,即:好、较好、一般、较差、差.在此仅考虑某型号战略导弹的有效性评价,并直接给出有关初步结果.

根据指标分析,分别给出可靠性和维修性的模糊评价结果向量为

$$B_1 = (0.3, 0.5, 0.2, 0, 0)$$

$$B_2 = (0, 0.2, 0.5, 0.2, 0.1)$$

安全性有三项指标,设其模糊权向量和单因素评价结果分别为

$$A_3 = (0.4, 0.3, 0.3)$$

$$B_3 = \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.1 & 0 \\ 0.3 & 0.3 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.3 & 0.2 & 0 \end{bmatrix}$$

因而

$$B_3 = A_3 \circ R_3$$

$$= (0.4, 0.3, 0.3) \circ \begin{bmatrix} 0.3 & 0.4 & 0.2 & 0.1 & 0 \\ 0.3 & 0.3 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.3 & 0.2 & 0 \end{bmatrix}$$

$$= (0.24, 0.37, 0.23, 0.13, 0.03)$$

适应性有四项指标, B_4 可计算如下(A_4 与 R_4 设为已知):

$$B_4 = A_4 \circ R_4$$

$$= (0.3, 0.4, 0.1, 0.2) \circ \begin{bmatrix} 0.1 & 0.3 & 0.4 & 0.2 & 0 \\ 0.2 & 0.4 & 0.3 & 0.1 & 0 \\ 0 & 0.5 & 0.3 & 0.1 & 0.1 \\ 0.4 & 0.3 & 0.1 & 0.1 & 0.1 \end{bmatrix}$$

$$= (0.19, 0.36, 0.29, 0.13, 0.03)$$

有效性的四个方面的模糊权向量为

$$A = (0.4, 0.2, 0.2, 0.2)$$

二级模糊综合评价如下:

$$A \circ R = A \circ \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{bmatrix}$$

$$= (0.4, 0.2, 0.2, 0.2) \circ \begin{bmatrix} 0.3 & 0.5 & 0.2 & 0 & 0 \\ 0 & 0.2 & 0.5 & 0.2 & 0.1 \\ 0.24 & 0.37 & 0.23 & 0.13 & 0.03 \\ 0.19 & 0.36 & 0.29 & 0.13 & 0.03 \end{bmatrix}$$

$$= (0.206, 0.386, 0.284, 0.092, 0.003)$$

这就得到了该型号导弹有效性的模糊综合评价结果向量。

第六节 舍弃等级论域的模糊综合评价

在模糊综合评价中需要确定评语等级论域 $V = \{v_1, v_2, \dots, v_m\}$, 一般情况下, $3 \leq m \leq 7$. 正是由于有了各等级的模糊子集, 才可以对被评事物进行模糊综合评价, 得到综合评价结果向量, 反映被评事物对各等级模糊子集的隶属程度. 实际中有这样一种作法: 取 $m = 1$, $V = \{v_1\}$, 即只保留第一个等级而将其他等级舍弃掉, 从评语上来看, v_1 一般对应于“优秀”、“强”、“很好”, 等. 这样, 对每一个被评事物而言, 其隶属关系矩阵变成一个列向量, 综合评价结果变成一个一维模糊向量, 即一个点值. 事实上, 正因为如此, 常常是对多个被评事物同时进行评价, 隶属关系矩阵 R , 如表 6-9 所示.

表 6-9

因 素 \ 事 物				
	1	2	...	n
u_1	r_{11}	r_{12}	...	r_{1n}
u_2	r_{21}	r_{22}	...	r_{2n}
\vdots	\vdots	\vdots		\vdots
u_p	r_{p1}	r_{p2}	...	r_{pn}

其中第 i 列 ($i = 1, 2, \dots, n$) 为第 i 个被评事物从各因素看对这一个等级模糊子集的隶属度, 设模糊权向量为

$$A = (a_1, a_2, \dots, a_p),$$

则

$$\begin{aligned}
 A \circ R &= (a_1, a_2, \dots, a_p) \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pn} \end{pmatrix} \\
 &= (b_1, b_2, \dots, b_n) \quad (6-43) \\
 &\triangleq B
 \end{aligned}$$

显然 B 中第 i 个分量 b_i 表示第 i 个被评事物从总体上看对这一个等级模糊子集的隶属度, b_i 越大, 表示第 i 个被评事物表现越好, 因此可以直接用 b_i 对 n 个被评事物进行排序。

例 6.9 用舍弃等级论域的模糊综合评价方法选择主导产业^[2]。

主导产业的选择正确与否, 不仅直接决定着资源的配置、产业政策的确立以及产业结构的演进方向, 而且事关经济增长的快慢。正因为如此, 一些国家或地区在制定经济发展规划时, 都十分重视主导产业的选择。

明确主导产业的特征是正确选择主导产业的前提。主导产业的主要特征是:

- (1) 国民经济的增长对该产业有较大的需求;
- (2) 该产业能够快速有效地吸收技术进步成果;
- (3) 该产业对其他产业的发展有较大的带动作用, 其他产业的发展反过来又将诱导该产业的进一步发展;
- (4) 在国民经济中占有较大的份额。

上述四个方面的特征为具体选择主导产业提供了理论依据, 但这些定性的原则难以综合起来考虑, 要作出明确的选择, 还应进行定量的综合分析。下面以某省为例, 选择了五项指标对国民经济各产业进行综合定量分析。这五项指标是:

1. 需求收入弹性

就是当人均国民收入增长 1% 时该产业的需求增长率。计算公式是

$$e_i = \frac{\Delta X_i}{X_i} / \frac{\Delta y}{y} \quad (6-44)$$

式中 e_i 为 i 产业的需求收入弹性, $\frac{\Delta X_i}{X_i}$ 是 i 产业的需求增长率, $\frac{\Delta y}{y}$ 为人均国民收入增长率。

产业需求收入弹性越大,说明该产业随经济增长而有更大的市场需求,对经济增长有较大的贡献。因此,应优先发展需求收入弹性大的产业。

2. 技术进步率

该指标反映了主导产业能够迅速有效地吸收技术进步成果的特征。其计算公式为

$$a_i = x_i - \alpha_i k_i - \beta_i l_i \quad (6-45)$$

式中 a_i 为 i 产业的技术进步率, x_i, k_i, l_i 分别为 i 产业的产出增长率、资金增长率和劳动力增长率, α_i, β_i 分别为资金产出弹性和劳动力产出弹性。一般地, $\alpha_i = 0.3$ 或 0.4 , $\beta_i = 0.7$ 或 0.6 。

3. 产业关联度

在现代经济中,产业之间相互联系、相互依存、相互制约地推动整个国民经济的发展。为了体现产业间的联系程度,设置了制约度和依存度两个指标。

(1) 制约度

$$f_i = n \sum_j q_{ij} / \sum_i \sum_j q_{ij} \quad (6-46)$$

(2) 依存度

$$b_j = n \sum_i q_{ij} / \sum_i \sum_j q_{ij} \quad (6-47)$$

(6-46)和(6-47)式中的 q_{ij} 为 Leontief 逆矩阵 $(I - A)^{-1}$ 的 i 行 j 列元素 (A 为直接消耗系数矩阵), n 为产业个数, f_i 为 i 产业制约度, b_j 为 j 产业依存度。

制约度 f_i 是当 i 产业增加一个单位的最终产品时所有产业总产值增加量之和,因此 f_i 度量了 i 产业对其他产业发展的制约

程度;而依存度 b_j 是当所有产业的最终产品都增加一个单位产值时 j 产业总产值的增加量. 因此 b_j 度量了 j 产业的发展在多大程度上依赖于其他产业的发展. 国民经济发展中的主导产业应该是制约度和依存度比较大的产业.

4. 产业份额

产业份额就是 i 产业在整个国民经济中所占比重, 计算公式为

$$g_i = X_i / \sum_{j=1}^n X_j \quad (6-48)$$

设置这个指标是要说明主导产业应当是已经成为国民经济的主要产业, 也就是说, 不能期待那些尚为幼稚的产业迅速成为主导产业. 主导产业不仅应当有发展潜力, 而且应当有发展基础.

主导产业的选择其实是一个多指标综合评价问题. 而且对一个产业能否成为主导产业的判断往往具有鲜明的模糊性. 因此, 可以尝试用模糊综合评价的方法来确定主导产业. 这里, 我们采用舍去评语等级论域的方法来选择主导产业.

按舍去评语等级论域方法的步骤, 首先要确定各评价指标. 我们选择产业构成 (g_i)、制约度 (f_i)、依存度 (b_i)、技术进步率 (a_i) 和需求收入弹性 (e_i) 五个指标. 某省 25 个产业的这五项指标实际值如表 6-9 所示.

其次, 要根据各指标的性质和数值分布特点确定隶属函数. 由对主导产业的特征分析可知, 主导产业应满足如下条件:

- (1) 产业的需求收入弹性大于 1;
- (2) 产业的技术进步率应大于总平均技术进步率;
- (3) 产业的构成份额应大于平均比重;
- (4) 产业的制约度大于平均制约度;
- (5) 产业的依存度应大于平均依存度.

根据以上条件和表 6-9 中的数据特点, 可确定出各指标的隶属函数如下:

- (1) 需求收入弹性的隶属函数

$$\mu(e_i) = \begin{cases} 0, & e_i \leq 1 \\ \frac{1}{1 + [5(e_i - 1)]^{-2}}, & e_i > 1 \end{cases} \quad (6-49)$$

(2) 技术进步率的隶属函数

$$\mu(a_i) = \begin{cases} 0, & a_i \leq 7.2376 \\ \frac{a_i - 7.2376}{12.7624}, & 7.2376 < a_i < 20 \\ 1, & a_i \geq 20 \end{cases} \quad (6-50)$$

(3) 产业构成的隶属函数

$$\mu(g_i) = \begin{cases} 0, & g_i \leq 4 \\ \frac{1}{1 + (g_i - 4)^{-2}}, & g_i > 4 \end{cases} \quad (6-51)$$

(4) 制约度的隶属函数

$$\mu(f_i) = \begin{cases} 0, & f_i \leq 0.4458 \\ \frac{1}{1 + [5(f_i - 0.4458)]^{-2}}, & f_i > 0.4458 \end{cases} \quad (6-52)$$

(5) 依存度的隶属函数

$$\mu(b_i) = \begin{cases} 0, & b_i \leq 0.3955 \\ \frac{1}{1 + [5(b_i - 0.3955)]^{-2}}, & b_i > 0.3955 \end{cases} \quad (6-53)$$

表 6-10

行 业	产业构成	制约度	依存度	技术进步率	需求收入弹性
农业	22.3582	0.3131	0.1021	0	0.38
林业	1.9359	0.0483	0.1867	0	0.68
牧业	7.3323	0.1375	0.6536	0	0.89
副业、渔业	4.1346	0.4445	0.0829	0	1.57
黑色冶金工业	2.7784	0.8115	0.2728	25.76	1.99
有色金属工业	2.2878	0.3041	0.383	19.54	1.27
电力工业	2.9198	0.7903	0.2294	13.62	1.15

续表

行 业	产业构成	制约度	依存度	技术进步率	需求收入弹性
煤炭及炼焦工业	2.8573	0.6006	0.314	2.24	1.13
石油工业	0.0218	79.7626	0.478	0	1.22
化学工业	4.5672	0.5397	0.3464	9.14	1.75
机械工业	11.1832	0.1983	0.377	27.1	1.98
建筑材料工业	2.3313	1.0368	0.2822	13.1	1.4
森林工业	1.1743	0.8049	0.2089	4.36	0.9
食品工业	8.6194	0.0648	0.472	4.72	1.018
纺织工业	1.4539	0.5747	0.3518	4.18	1.105
缝纫工业	1.2098	0.1311	0.7062	4.18	0.981
皮革工业	0.2875	0.2307	0.4815	0	0.893
造纸工业	0.5504	0.6315	0.6523	10.1	2.035
文教及艺术用品工业	0.5969	0.2197	0.6506	10.1	0.918
其他工业	1.1523	0.6887	0.5513	11.8	1.669
建筑业	9.9992	0	0.6756	21	1.069
铁路运输业	1.2084	0.748	0.2503	0	1.35
公路及其他交通运输业	2.6935	0.6049	0.4675	0	1.25
邮电业	0.2099	0.4572	0.5037	0	1.14
商业、饮食、物资供销业	6.1367	0.3178	0.2086	0	1.008

将表 6-10 中的数据代入上述各隶属函数就可计算出模糊关系矩阵 R , 计算结果如表 6-11 所示。

表 6-11

行 业	产业构成	制约度	依存度	技术进步率	需求收入弹性
农业	0.9970	0.0000	0.0000	0.0000	0.0000
林业	0.0000	0.0000	0.0000	0.0000	0.0000
牧业	0.9174	0.0000	0.6248	0.0000	0.0000
副业、渔业	0.0178	0.0000	0.0000	0.0000	0.8904
黑色冶金工业	0.0000	0.7698	0.0000	1.0000	0.9608
有色金属工业	0.0000	0.0000	0.0000	0.9640	0.6457
电力工业	0.0000	0.7479	0.0000	0.5001	0.3600
煤炭及炼焦工业	0.0000	0.3746	0.0000	0.0000	0.2970
石油工业	0.0000	1.0000	0.1454	0.0000	0.5475
化学工业	0.2434	0.1806	0.0000	0.1491	0.9336
机械工业	0.9810	0.0000	0.0000	1.0000	0.9600
建筑材料工业	0.0000	0.8972	0.0000	0.4592	0.8000
森林工业	0.0000	0.7632	0.0000	0.0000	0.0000

续表

行 业	产业构成	制约度	依存度	技术进步率	需求收入弹性
食品工业	0.9552	0.0000	0.1276	0.0000	0.0080
纺织工业	0.0000	0.2935	0.0000	0.0000	0.2161
缝纫工业	0.0000	0.0000	0.7070	0.0000	0.0000
皮革工业	0.0000	0.0000	0.1560	0.0000	0.0000
造纸工业	0.0000	0.4630	0.6225	0.2243	0.9640
文教及艺术用品工业	0.0000	0.0000	0.6139	0.2243	0.0000
其他工业	0.0000	0.5960	0.3777	0.3575	0.9180
建筑业	0.9730	0.0000	0.6623	1.0000	0.1064
铁路运输业	0.0000	0.6954	0.0000	0.0000	0.7538
公路及其他交通运输业	0.0000	0.3876	0.1147	0.0000	0.6098
邮电业	0.0000	0.0032	0.2264	0.0000	0.3289
商业、饮食、物资供销业	0.8203	0.0000	0.0000	0.0000	0.0016

第三,确定评价指标的权数向量 A . 经过定性、定量分析,本例确定技术进步率和需求收入弹性两项指标的权数均为 0.25,制约度和依存度两项指标赋权均为 0.20,产业构成的权数为 0.10.

第四,选取乘与有界和算子,将 A 与 R 合成,得到评价结果向量 B

$$B = A \circ R$$

B 中第 i 个分量表明了第 i 个产业属于主导产业的程度,计算结果见表 6-12.

表 6-12

行 业	隶属度	位次	行 业	隶属度	位次
农业	0.0997	22	食品工业	0.1231	20
林业	0.0000	25	纺织工业	0.1127	21
牧业	0.2167	14	缝纫工业	0.1414	18
副业、渔业	0.2244	13	皮革工业	0.0312	24
黑色冶金工业	0.6442	1	造纸工业	0.5142	3
有色金属工业	0.4024	7	其他工业	0.5136	4
电力工业	0.3646	9	文教及艺术用品工业	0.1799	15
煤炭及炼焦工业	0.1492	17	建筑业	0.5064	5
石油工业	0.3660	8	铁路运输业	0.3275	11
化学工业	0.3311	10	公路及其他交通运输业	0.2529	12
机械工业	0.5881	2	邮电业	0.1281	19
建筑材料工业	0.4943	6	商业、饮食、物资供销业	0.0824	23
森林工业	0.1526	16			

根据表 6-12 中的隶属度,结合定性分析,就可选出近期该省的主导产业是:

(1) 以铝、钢铁、铁合金为重点的黑色冶金工业和有色金属工业;

(2) 以汽车、电子产品、仪器仪表为重点的机械工业;

(3) 以水泥和墙体材料为重点的建筑材料工业及与此相关的建筑业;

(4) 水火互济的电力工业;

(5) 以磷化工、煤化工、植物精细化工为重点的化学工业.

第七章 多维标度法

综合评价问题就所用的指标来看,类型是很多的,既有定量指标,又有定性指标的类型,这一章的重点是介绍用统计中的多维标度(multidimensional scaling)法来进行综合评价,它与主成分方法是不同的,但是又有主成分方法的一些思路.为此我们要对定性的变量如何度量它们之间的相关性作一些讨论,有关的专门著作,可以参阅文献[4].

第一节 相似性度量

在定量的指标分析中,相关系数是一个重要的量,它能相当好地反映变量之间的线性相关的程度.对于定性的指标,是否有类似的量来反映呢?这就是 Pearson 的 χ^2 、熵等这一类概念.我们先作一些介绍,然后来谈它们的应用.有些结论我们不去详细证明(因为涉及更多的准备知识),而只是谈如何用于解决问题,怎样理解这些结论.

一、Pearson 的 χ^2

设 x, y 都是离散的随机变量, x 可以是 r 个状态 x_1, x_2, \dots, x_r 之一, x_i 不一定是数,但我们可以用指定的数来代表这一状态,因此就可以让 $x_i = i$, 当 $x = i$ 时,表示处于第 i 种状态,所以这样的离散变量就是定性资料的一种概率论的描述方法.对于 y , 类似地,它可以处于 y_1, y_2, \dots, y_c 这 c 个状态之一.于是它们的联合分布为

$$\begin{cases} p(x = x_i, y = y_j) = \pi_{ij} \\ i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c \end{cases} \quad (7-1)$$

它们各自的边缘分布为

$$\begin{cases} p(x = x_i) = \sum_{j=1}^c \pi_{ij} \triangleq \pi_{i\cdot} \\ p(y = y_j) = \sum_{i=1}^r \pi_{ij} \triangleq \pi_{\cdot j} \\ i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c \end{cases} \quad (7-2)$$

相应的条件分布为

$$\begin{cases} p(y = y_j | x = x_i) = \frac{\pi_{ij}}{\pi_{i\cdot}} \triangleq \pi_{(i)j} \\ p(x = x_i | y = y_j) = \frac{\pi_{ij}}{\pi_{\cdot j}} \triangleq \pi_{i(j)} \\ i = 1, 2, \dots, r; \quad j = 1, 2, \dots, c \end{cases} \quad (7-3)$$

如果观察了 (x, y) 的 n 个样本, 就可以计算出 $x = x_i, y = y_j$ 的频数 n_{ij} , 因此

$$n_{ij}/n, i = 1, 2, \dots, r; j = 1, 2, \dots, c$$

就是 (x, y) 的 n 个样本中处于 $x = x_i, y = y_j$ 的频率. 大数定律告诉我们频率会趋于概率, 只要 n 相当大. 所以如果一个量是用 π_{ij} 来定义的, 那么用频率 n_{ij}/n 代替 π_{ij} 后, 就得到一个相应的统计量.

Pearson 的 χ^2 是用 π_{ij} 及相应的 $\pi_{i\cdot}$ 和 $\pi_{\cdot j}$ 来导出的. 如果 x 与 y 是相互独立的, 则自然有

$$\begin{aligned} p(x = x_i, y = y_j) &= p(x = x_i)p(y = y_j) \quad (7-4) \\ &= \pi_{i\cdot} \pi_{\cdot j} \end{aligned}$$

对 $i = 1, 2, \dots, r; j = 1, 2, \dots, c$ 都成立, 因此

$$\pi_{ij} - \pi_{i\cdot} \pi_{\cdot j}$$

就反映了实际的 (x, y) 与独立的 x, y 之间的差距, 这个差距越大 (不必考虑正负号), 就表示 x 与 y 的相关性越强, 所以

$\sum_{i,j} (\pi_{ij} - \pi_{i\cdot} \pi_{\cdot j})^2$ 就是总的差距. 考虑到不同的 π_{ij} 应有不同的权

重,出现概率越小的,权重就应大些,所以用 $\frac{1}{\pi_{i\cdot}\pi_{\cdot j}}$ 作为权重, $\pi_{i\cdot}$, $\pi_{\cdot j}$ 是独立时出现的概率.这就导出

$$\varphi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(\pi_{ij} - \pi_{i\cdot}\pi_{\cdot j})^2}{\pi_{i\cdot}\pi_{\cdot j}} \quad (7-5)$$

这个量, Pearson 最初是用 φ^2 表示这个量,所以现在有些书上还是用 φ^2 来表示,然而可以证明用频率 n_{ij}/n 代替上式中的 π_{ij} 之后,用

$$n_{i\cdot} = \sum_{j=1}^c n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^r n_{ij}$$

相应的 $n_{i\cdot}/n, n_{\cdot j}/n$ 代 $\pi_{i\cdot}, \pi_{\cdot j}$ 之后,在 x 与 y 的确是独立的前提下,可以证明它的极限分布是 χ^2 分布,因此后来常用 χ^2 表示下述统计量:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(nn_{ij} - n_{i\cdot}n_{\cdot j})^2}{nn_{i\cdot}n_{\cdot j}} \quad (7-6)$$

它的极限分布是自由度为 $(r-1)(c-1)$ 的 χ^2 分布.

现在来看 φ^2 的性质:

(i) $0 \leq \varphi^2 \leq \min(r, c) - 1$;

(ii) $\varphi^2 = 0$ 的充分必要条件是 x 与 y 独立;

(iii) $\varphi^2 = 1$ 的充分必要条件是概率为 1 的 y 可以由 x 确定或 x 由 y 确定.

这(i),(ii),(iii)三条中,(ii)是明显的,因为 $\varphi^2 = 0$ 的充要条件是 $\pi_{ij} - \pi_{i\cdot}\pi_{\cdot j} = 0$ 对 i, j 成立;而(i),(iii)这两条并不明显,但我们不去证明了.

由于(i),自然想到引入

$$\theta^2 = \varphi^2 / (\min(r, c) - 1) \quad (7-7)$$

后, θ^2 就在 $[0, 1]$ 之内,它与相关系数 ρ 的平方在 $[0, 1]$ 之内相仿, θ^2 就可以作为一种 x 与 y 关联性的度量值.有关的一些理论上的讨论,我们放在附录中给以介绍.

二、最优预测系数^[5]

仍然和一中的情况相同,把一个变量 x 对 y 的预测能力作为它们之间关联性的度量,当然也可以用 y 对 x 的预测能力来度量.

当没有 x 的信息时, y 自身的边缘分布相应的概率为 $\pi_{\cdot 1}, \pi_{\cdot 2}, \dots, \pi_{\cdot c}$, 记

$$\pi_{\cdot m} = \max_{1 \leq j \leq c} \pi_{\cdot j} \quad (7-8)$$

这里“ m ”作为足标,表示最大值相应的标号. 于是用 m 来预测 y 会出现的状态后,犯错误的概率就是 $1 - \pi_{\cdot m}$, 记为 p_1 , 即 $p_1 = 1 - \pi_{\cdot m}$.

如果已知 x 的状态为 x_i , 此时考虑 y 对 $x = x_i$ 的条件概率, 哪一个状态出现可能性最大, 就以它为“ m ”, 因为与 $x = x_i$ 有关, 于是记为 $m(i)$, 这就导出

$$\pi_{(i)m(i)} = \max_{1 \leq j \leq c} \pi_{(i)j}$$

注意到

$$\pi_{(i)j} = \pi_{ij} / \pi_{\cdot j}$$

因此由

$$\pi_{im(i)} = \max_{1 \leq j \leq c} \pi_{ij}$$

得相应的出错概率为

$$p_{2(i)} = 1 - \pi_{im(i)} = 1 - \frac{\pi_{im(i)}}{\pi_{\cdot i}} \quad (7-9)$$

对每种状态出错的概率求平均, 得

$$\begin{aligned} p_2 &= \sum_{i=1}^r \pi_{\cdot i} \cdot p_{2(i)} \\ &= \sum_{i=1}^r \pi_{\cdot i} \cdot \left(1 - \frac{\pi_{im(i)}}{\pi_{\cdot i}} \right) \\ &= 1 - \sum_{i=1}^r \pi_{im(i)} \end{aligned} \quad (7-10)$$

今 $\pi_{\cdot m} = \sum_{i=1}^r \pi_{im} \leq \sum_{i=1}^r \pi_{i m(i)}$, 所以 $p_2 \leq p_1$, 也即 x 的状态已知这一条件会改善出错的规律, 这样就可以定义

$$\lambda_y = \frac{p_1 - p_2}{p_1}$$

类似地, 有 λ_x , 记

$$\pi_{g\cdot} = \max_{1 \leq i \leq r} \pi_{i\cdot}$$

$$\pi_{g(j)j} = \max_{1 \leq i \leq r} \pi_{ij}$$

则 λ_y 与 λ_x 可以表示为

$$\begin{cases} \lambda_y = (\sum_{i=1}^r \pi_{i m(i)} - \pi_{\cdot m}) / (1 - \pi_{\cdot m}) \\ \lambda_x = (\sum_{j=1}^c \pi_{g(j)j} - \pi_{g\cdot}) / (1 - \pi_{g\cdot}) \end{cases} \quad (7-11)$$

再引入平均的 $\overline{p_1}$ 和 $\overline{p_2}$,

$$\overline{p_1} = \frac{1}{2}((1 - \pi_{\cdot m}) + (1 - \pi_{g\cdot}))$$

$$\overline{p_2} = \frac{1}{2}[(1 - \sum_{i=1}^r \pi_{i m(i)}) + (1 - \sum_{j=1}^c \pi_{g(j)j})]$$

再定义 x, y 间最优预测系数为

$$\lambda = \frac{\overline{p_1} - \overline{p_2}}{\overline{p_1}} = \frac{\sum_i \pi_{i m(i)} + \sum_j \pi_{g(j)j} - \pi_{\cdot m} - \pi_{g\cdot}}{2 - \pi_{\cdot m} - \pi_{g\cdot}} \quad (7-12)$$

在(7-12)中用样本相应的频率来代替概率, 就得

$$\hat{\lambda} = \frac{\sum_i n_{i m(i)} + \sum_j n_{g(j)j} - n_{\cdot m} - n_{g\cdot}}{2n - n_{\cdot m} - n_{g\cdot}} \quad (7-13)$$

三、熵及关联信息量

熵是反映随机变量不确定性的一个重要的量, 对上面讨论过

的随机变量 x, y 而言, 每一种分布概率就相应有一个熵, x 的概率分布为

$$\begin{array}{cccc} x_1 & x_2 & \cdots & x_r \\ \pi_{1.} & \pi_{2.} & \cdots & \pi_{r.} \\ \\ y_1 & y_2 & \cdots & y_c \\ \pi_{.1} & \pi_{.2} & \cdots & \pi_{.c} \\ \\ (x_1, y_1) & \cdots & (x_i, y_i) & \cdots & (x_r, y_c) \\ \pi_{11} & \cdots & \pi_{ij} & \cdots & \pi_{rc} \end{array}$$

x 相应的熵

$$H(x) = - \sum_{i=1}^r \pi_{i.} \ln \pi_{i.}$$

y 相应的熵

$$H(y) = - \sum_{j=1}^c \pi_{.j} \ln \pi_{.j}$$

(x, y) 相应的熵

$$H(x, y) = - \sum_{i,j=1}^{r,c} \pi_{ij} \ln \pi_{ij}$$

它们之间的条件概率分布, 也有相应的熵, 用 $H_x(y)$ 表 y 对 x 的平均条件熵, 即

$$H_x(y) = - \sum_{i=1}^r \sum_{j=1}^c \pi_{ij} \ln \pi_{ij} / \pi_{i.}$$

同样地有

$$H_y(x) = - \sum_{i=1}^r \sum_{j=1}^c \pi_{ij} \ln \pi_{ij} / \pi_{.j}$$

什么是 x 对 y 的信息呢? 就是已知 x 之后, y 的不确定性就减少了, 即熵变小了, 少掉的熵正是 X 提供给 Y 的信息量, 用 $I(x, y)$ 表示它, 于是有

$$I(x, y) = H(y) - H_x(y)$$

同理 y 给 x 的信息量是

$$I(y, x) = H(x) - H_y(x)$$

注意到 (x, y) 的熵

$$\begin{aligned} H(x, y) &= H(x) + H_x(y) \\ &= H(y) + H_y(x) \end{aligned}$$

因此

$$I(x, y) = I(y, x)$$

这样我们可以引出关联信息量 $RI(x, y)$ 的定义

$$RI(x, y) = \frac{I(x, y)}{\sqrt{H(x)H(y)}} \quad (7-14)$$

很明显, $RI(x, y)$ 有如下的性质:

$$(i) RI(x, y) = RI(y, x) \text{ (对称性);} \quad (7-15)$$

$$(ii) RI(x, y) = 0 \text{ 的充要条件是 } x \text{ 与 } y \text{ 独立;} \quad (7-16)$$

(iii) $RI(x, y) = \sqrt{H(y)/H(x)}$ 的充要条件是 y 概率为 1 地可写成 x 的函数.

第二节 托格森(Torgerson)方法

设有 n 个样本, 每个样本有 r 个指标, 因此可以用 r 维欧氏空间中 n 个向量 $x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}$ 来表示. 用

$$d_{ij}^2 = \|x_{(i)} - x_{(j)}\|^2 = (x_{(i)} - x_{(j)})^T (x_{(i)} - x_{(j)})$$

表示 $x_{(i)}$ 与 $x_{(j)}$ 之间的距离的平方. 而对每一个 i , 可以给出一个矩阵 $B_i = (b_{ml}^{(i)})$, 它是一个 $n \times m$ 的矩阵, 其中 $b_{ml}^{(i)}$ 的定义是

$$b_{ml}^{(i)} = \frac{1}{2} (d_{il}^2 + d_{im}^2 - d_{lm}^2) \quad (7-17)$$

从(7-17)式知道

$$\begin{aligned} b_{ml}^{(i)} &= \frac{1}{2} [(x_{(i)} - x_{(l)})^T (x_{(i)} - x_{(l)}) \\ &\quad + (x_{(i)} - x_{(m)})^T (x_{(i)} - x_{(m)}) \\ &\quad - (x_{(l)} - x_{(m)})^T (x_{(l)} - x_{(m)})] \end{aligned}$$

注意到向量内积的公式

$$(a - b)^T (a - b) = a^T a + b^T b - 2a^T b$$

因此就有

$$b_{ml}^{(i)} = (x_{(m)} - x_{(i)})^T (x_{(l)} - x_{(i)}) \quad (7-18)$$

对一切 $i, m, l = 1, 2, \dots, n$ 成立. 这就知道每一个 B_i 都是向量的内积矩阵, 因而是非负定的.

于是可以考虑一个很一般的问题, 如果对 n 个对象, 只知道它们之间的相对距离, 即两两之间的距离 d_{ij} , 但不知道它们是在什么空间之内 (因为距离可以不是欧氏的, 也可以是维数很高的), 这时, 利用 (7-17) 可以定义矩阵 B_i , 但 B_i 不一定是非负定 (因为距离不一定是由内积导出的), 但只要 $\text{rank} B_i = r$, 就可以求它的正特征根所相应的标准正交的特征向量, 由它们来重现这 n 个对象的相对位置, 这样就可以将高维空间 (或不知维数的空间) 对象用低维的点来表示, 而尽可能保持原有的相对结构 (彼此之间的距离). 例如将天空中恒星的散布情况用平面上的点来描述, 而使相对距离不变; 又如将地球 (球面) 上的城市散布情况用平面上的点来表示, 等等, 所以这一方法称为多维标度法, 它就用多维欧氏空间的尺度 (标尺、标度) 来重现对象的散布情况 (因为原始的距离可以不是欧氏的, 维数很高的). 但如果由 (7-17) 定义所得 B_i 是非负定的, 那就有更强的结论, 这就是下面的定理.

定理 7.1 (Young-Householder) 若事先给定的距离 d_{ij} 导出的 B_i 是非负定的, B_i 中元素 $b_{ml}^{(i)}$ 是由 (7-17) 定义的, 且 $\text{rank} B_i = r$, 则在 r 维欧氏空间中可以找到 n 个向量 $x_{(1)}, \dots, x_{(n)}$, 使得以 $x_{(i)}$ 为起点的向量 $x_{(j)} - x_{(i)}, j = 1, 2, \dots, n$ 所形成的内积矩阵正好是 B_i , 因而 d_{ij} 就是这些 $x_{(i)}$ 所相应的欧氏距离, 并且维数 r 不可能再降低.

这个定理的直观意义是很明显的, 它给出了重现内部结构的一种构造性的方法. 下面我们先论述一般情况下如何达到这一目的, 然后再举例说明.

现在给出具体的计算方法, 如何从距离矩阵 (d_{ij}) 出发来找出表示各点的向量 $x_{(1)}, \dots, x_{(n)}$.

设给了距离矩阵 $D = (d_{ij})$, 它是一个 $n \times n$ 的矩阵, 若存在一个 i 使相应的

$$b_{lm}^{(i)} = \frac{1}{2}(d_{il}^2 + d_{im}^2 - d_{lm}^2)$$

形成的矩阵 B_i 是非负定的, 而且由 $B_i^T = B_i, \text{rank} B_i = r$, 得

$$B_i = (b_{lm}^{(i)}) = (u_1, u_2, \dots, u_r) \begin{bmatrix} \lambda_1 & 0 \\ & \ddots \\ 0 & \lambda_r \end{bmatrix} \begin{bmatrix} u'_1 \\ \vdots \\ u'_r \end{bmatrix}$$

u_j 是矩阵 B_i 相应的非 0 特征根 λ_j 所对应的标准特征向量. 令

$$\begin{aligned} X_{n \times r} &= (u_1 u_2 \cdots u_r) \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ & \ddots \\ 0 & \sqrt{\lambda_r} \end{bmatrix} \\ &= (\sqrt{\lambda_1} u_1 \quad \sqrt{\lambda_2} u_2 \cdots \sqrt{\lambda_r} u_r) \end{aligned}$$

于是 X 中 n 个行向量就构成所要空间的点. 这是因为上述 X 的 n 行就是 n 个 r 维的向量, 即

$$X = \begin{bmatrix} x_{(1)}^T \\ \vdots \\ x_{(n)}^T \end{bmatrix}$$

而

$$\begin{aligned} XX^T &= \begin{bmatrix} x_{(1)}^T \\ \vdots \\ x_{(n)}^T \end{bmatrix} (x_{(1)} \cdots x_{(n)}) \\ &= (x_{(i)}^T x_{(j)}) \\ &= B_i \end{aligned} \quad (7-19)$$

所以这些 $x_{(1)}, \dots, x_{(n)}$ 正是以 $x_{(i)}$ 为原点相应的距离阵, 这是因为

$$x_{(j)}^T x_{(k)} = b_{jk}^{(i)} = \frac{1}{2}(d_{ij}^2 + d_{ik}^2 - 2d_{jk}^2)$$

当 $j = k$ 时

$$x_{(j)}^T x_{(j)} = d_{ij}^2 \quad (7-20)$$

当 $j = i$ 时

$$x_{(i)}^T x_{(i)} = d_{ii}^2 = 0$$

这样就找到了解. 这一段内容实际上也就是定理 7.1 的证明, 这个证明是构造性的.

对一般的距离来说, 只就距离满足的几条公理是得不到由 (7-17) 给出定义相应 B_i 的非负定性, 因此上述方法似乎是不适用的. 但 B_i 一定是对称的. 利用对称阵的正交变换化成对角阵, 就可以知道它与非负定阵的区别, 只在于有一部分特征根会小于 0, 于是选用 B_i 的正特征根与相应的特征向量, 就可以仿照上面的方法进行, 这时主要的差别就是拟合的点并不能完全与原来的距离一样, 现在来说明这一点.

设 B_i 的正特征根为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$, 共有 r 个, 与前面一样, 相应的特征向量记为 u_1, u_2, \cdots, u_r , 令

$$X_{n \times r} = (u_1 u_2 \cdots u_r) \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ & \ddots \\ 0 & \sqrt{\lambda_r} \end{bmatrix}$$

X 的每一行代表一个点, 它是 r 维的点. 它的内积矩阵并不是 B_i , 它只是反映了它的一部分, 我们用

$$a_{1r} = \sum_{i=1}^r \lambda_i / \sum_{i=1}^n |\lambda_i| \text{ 或 } a_{2r} = \sum_{i=1}^r \lambda_i^2 / \sum_{i=1}^n \lambda_i^2$$

来反映代表的程度, 称为拟合的优度, 这类似于分布的拟合优度.

托格森的方法称为 MDS 的经典解法, 下面给出有关的求解步骤和例子.

托格森方法的求解步骤为

(i) 从给定的距离矩阵 $D = (d_{ij})$ 依 (7-17) 的公式求出 B_i , 选一个 B_i 作为 $B = (b_{lm})$.

(ii) 求 B 的正特征根及相应的特征向量, 记为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ 及 u_1, u_2, \cdots, u_r .

$$(iii) \text{ 构造 } X = (u_1 u_2 \cdots u_r) \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ & \ddots \\ 0 & \sqrt{\lambda_r} \end{bmatrix}.$$

(iv) 计算拟合优度 a_{1r} 或 a_{2r} .

实际上, 我们可以对每个 B_i 先求出相应的特征根, 按拟合优度最好的 (也就是最大的) B_i 来求 X , 给出表达彼此距离的点.

例 7.1 莫尔斯(Morse)电码混同率的综合评价分析.

表 7-1 Morse 电码表

A	· —	J	· — — —	S	· · ·	2	· · — — —
B	— · · ·	K	— · —	T	—	3	· · · — —
C	— · — ·	L	· — · ·	U	· · —	4	· · · · —
D	— · ·	M	— —	V	· · · —	5	· · · · ·
E	·	N	— ·	W	· — —	6	— · · · ·
F	· · — ·	O	— — —	X	— · · —	7	— — · · ·
G	— — ·	P	· — — ·	Y	— · — —	8	— — — · ·
H	· · · ·	Q	— — · —	Z	— — · ·	9	— — — — ·
I	· ·	R	· — ·	1	· — — —	0	— — — — —

(i) 相似性度量的建立. 为了研究电码的相似性, 找一些不懂电码的人进行实验, 以一定的时间间隔成对快速发送 Morse 电码; 以判断两种电码为一种电码信号的人数比例作为该两电码的混合率, 并建立相似性度量; 现以 10 个阿拉伯数字由罗斯科普弗 (Rothkopf) 于 1957 年建立的数据阵为基础.

表 7-2 莫尔斯电码混合率的罗斯科普弗数据 (%)

	1	2	3	4	5	6	7	8	9	10
1	84	63	13	8	10	8	19	32	57	55
2	62	89	54	20	5	14	20	21	16	11
3	18	64	86	31	23	41	16	17	8	10
4	5	26	44	89	42	44	32	10	3	3
5	14	10	30	69	90	42	24	10	6	5
6	15	14	26	24	17	86	69	14	5	14
7	22	29	18	15	12	61	85	70	20	13
8	42	29	16	16	9	30	60	89	61	26
9	57	39	9	12	4	11	42	56	91	78
10	50	26	09	11	5	22	17	52	81	94

为了建立相似性度量阵, 我们将其中关于主对角线对称的每一对元素用它们的算术平均数代替就得一个 10×10 的对称阵, 记为 $S = (s_{ij})$. 此时可以定义距离

$$d_{ij} = (s_{ii} + s_{jj} - 2s_{ij})^{\frac{1}{2}} \quad (7-21)$$

$$i, j = 1, 2, \dots, 10$$

为了避免逐个计算 B_i , 再行比较, 这里采用中心化的方法, 由 d_{ij} 直接给出一个 $B = (b_{ij})$, 其中 b_{ij} 由下式确定: 为了有一般性, 设 S 是 $n \times n$ 的阵,

$$b_{ij} = s_{ij} - \frac{1}{n} \sum_{t=1}^n s_{it} - \frac{1}{n} \sum_{t=1}^n s_{jt} + \frac{1}{n^2} \sum_{\alpha=1}^n \sum_{\beta=1}^n s_{\alpha\beta} \quad (7-22)$$

于是求得 b_{ij} 为表 7-3.

表 7-3 B 的元素表

	1	2	3	4	5	6	7	8	9	10
1	45.66	52.36	-17.89	-25.59	-16.29	-23.09	-18.39	-2.89	17.21	16.26
2	25.01	52.36	26.46	-8.24	-19.94	-17.94	-13.54	-14.04	-11.44	-10.89
3	-17.89	26.46	57.56	10.36	3.16	3.86	-16.94	-18.44	-16.34	-21.79
4	~	~	~	63.16	33.46	5.66	-9.14	-20.64	-26.4	-22.99
5	~	~	~	~	71.76	4.96	-10.84	-20.34	-24.74	-21.19
6	~	~	~	~	~	55.16	29.86	-14.14	-28.04	-14.49
7	~	~	~	~	~	~	45.56	24.56	-9.34	-21.79
8	~	~	~	~	~	~	~	47.56	17.16	1.21
9									49.76	41.81
10										59.86

(ii) 求出 B 的特征值为 187.3, 121.2, 96.0, 55.8, 47.0, 32.1, 9.1, 3.8, 0.0, -4.0 相应的特征向量: 取前二个得

$$y_1 = \sqrt{\lambda_1} u_1 = (4.23, 0.28, -3.74, -5.61, -5.35, -3.80, -0.92, 3.02, 6.19, 5.69)^T$$

$$y_2 = \sqrt{\lambda_2} u_2 = (3.20, 5.79, 4.20, 6.45, 0.00, -3.93, -5.47, -3.68, -0.64, -0.12)^T$$

该二向量形成平面标度的两个坐标, 如图 7-1.

此例中, 虽然没有利用全部的正特征根所对应的特征向量, 而

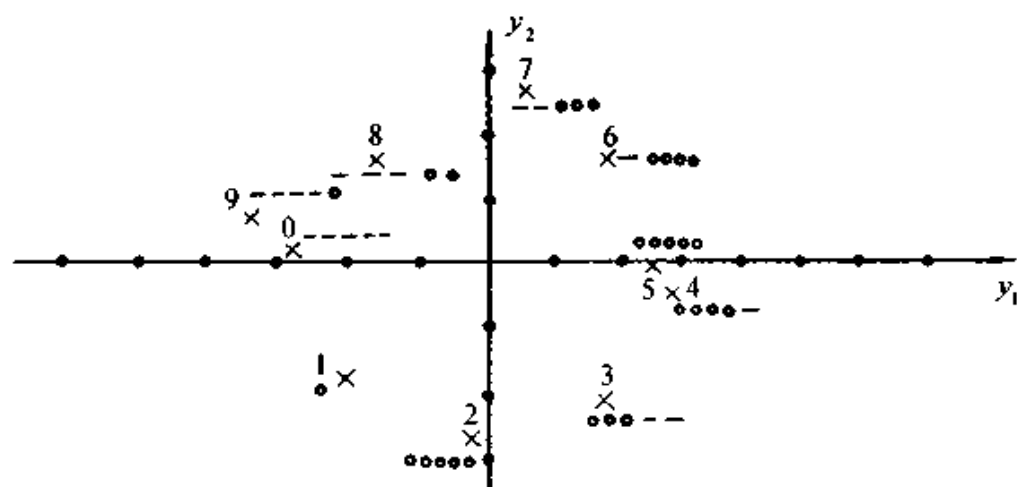


图 7-1

只用了最前面的二个, 但 $a_1 r = \sum_{i=1}^2 \lambda_i / \sum_{i=1}^{10} (\lambda_i) = 56\%$, 从图形上看已可以明显反映出相互的联系. 从 y_1 的轴来看, 从左到右, 是“·”数增加, “-”数减少; 从 y_2 轴来看, 远离原点时, “·”与“-”组合较多, 靠近原点时, 比较单一. 从“0”与“9”, “4”与“5”的位置很近, 就可以看出, 这两对中, 每一对的两个成员是易于混淆的.

这个例子说明了托格森方法的特点, 在图形上显示出相对位置后, 易于比较分析. 另一方面它也表明, 由它本身并不能直接给出一种序的结构, 排出对象的“优”、“劣”次序, 但为我们按某种标准来排序作好了准备.

与托格森方法相比, 要求前提条件更少的是日本林知己夫的准计量性 MDS 方法, 它是从对象间的相似性出发来分析的.

设 ρ_{ij} 是 n 个对象第 i 个与第 j 个相似性的值, 但不要求有对称性, 即 ρ_{ij} 与 ρ_{ji} 可以不同, ρ_{ii} 一般认为应该是 1, 但也可以无定义. 我们从简单的情形开始, 逐步复杂化.

给定了相似矩阵 $P = (\rho_{ij})_{n \times n}$ 后, 考虑在实轴上如何寻找 n 个点 x_1, \dots, x_n , 使得 $(x_i - x_j)^2$ 能反映由 P 给出的相似性? ρ_{ij} 是相似性的值, 那么 $-\rho_{ij}$ 就可以认为是不相似的值. 于是林知己夫用下述模型来求 x_1, \dots, x_n 的值.

给定 $P = (\rho_{ij})_{n \times n}$ 阵之后, 求 x_1, \dots, x_n 满足:

$$\begin{cases} \sum_{i=1}^n x_i = 0 & (\text{即 } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0) \\ \sum_{i=1}^n x_i^2 = 1 \\ \text{使 } Q = - \sum_{i \neq j} \rho_{ij} (x_i - x_j)^2 \text{ 达到最大} \end{cases} \quad (7-23)$$

注意到上式中 Q 是 x_1, \dots, x_n 的二次型, 所以求解问题是一个二次规划问题, 但对 (7-23) 这一特殊类型, 可以清晰地求出它的解.

将 Q 整理为 x_1, \dots, x_n 的二次型的规范形式, 因为记 $\rho_{ii} = 1$ 之后,

$$\begin{aligned} Q &= - \sum_{i=1}^n \sum_{j \neq i} \rho_{ij} (x_i - x_j)^2 \\ &= - \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} (x_i - x_j)^2 \\ &= - \sum_{i,j=1}^n \rho_{ij} (x_i^2 + x_j^2 - 2x_i x_j) \\ &= \sum_{i=1}^n \left(- \left(\sum_{j=1}^n (\rho_{ij} + \rho_{ji}) \right) \right) x_i^2 + 2 \sum_{i,j=1}^n \rho_{ij} x_i x_j \end{aligned}$$

$$\text{令} \quad \begin{cases} a_{ii} = - \sum_{j \neq i} a_{ij}, & i = 1, 2, \dots, n \\ a_{ij} = \rho_{ij} + \rho_{ji}, & i \neq j, i, j = 1, 2, \dots, n \end{cases} \quad (7-24)$$

之后, 就有

$$Q = x^T A x \quad (7-25)$$

其中

$$A = (a_{ij}), \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

很明显, (7-25) 中的 A 是对称的, 因此利用对称阵的特性:

$$\max_{x^T x = 1} x^T A x = \lambda_1 \quad (7-26)$$

λ_1 是 A 的最大特征根, 这就求出 (7-23) 的解是 A 矩阵 λ_1 所对应的特征向量, $x_* = (x_1^*, \dots, x_n^*)^T$ 就是这个特征向量的 n 个坐标, 这就在直线上排出了顺序.

注意到 x_* 与 $-x_*$ 都满足 $Ax = \lambda_1 x$, 因此从小到大的序也可以是从大到小的序. 现在来进一步看 (7-26) 给出的解有什么特点.

细心的读者会问, λ_1 对应的特征向量 x_* 是否满足

$$\sum_{i=1}^n x_i^* = 0 \quad (7-27)$$

这个条件? 很显然, 这个条件不一定会满足, 但这个条件只是把 n 个对象的“中心”规定在原点而已, 所以不在原点也是可以的. 条件 (7-27) 可以写成

$$\mathbf{1}^T x_* (= x_*^T \mathbf{1}) = 0$$

$\mathbf{1}$ 就是坐标全为 1 的向量.

若给定的 P 是对称的, 即 $\rho_{ij} = \rho_{ji}$, 此时可选 $\rho_{ii} = 0$, 相应的

$$A = 2P$$

因此问题求解, 直接从 P 出发就可求得, 只需用 P 的最大特征根所对应的特征向量作为解.

若 ρ_{ij} 均为非负, 或均为非正, 此时

$$Q = - \sum_{i=1}^n \sum_{j=1}^n \rho_{ij} (x_i - x_j)^2$$

(取 $\rho_{ii} = 0$)

或恒为非正, 或恒为非负, 因此相应的 (7-25) 中的 A 或是非正定, 或是非负定, 同样可求解.

如果相似矩阵 P 是由距离或其他的不相似性度量导出的, 则当 D 是距离时, 令

$$P = -D = (-d_{ij})$$

就可按上述方法处理, 这样就导出了利用距离矩阵来将对象排序的方法, 它就弥补了托格森方法不能排序的缺点.

容易想到, 这里对相似性的度量 ρ_{ij} 没作什么要求, 所以总可

以作一平移,令

$$r_{ij} = \rho_{ij} + c$$

使得所有的 $r_{ij} \geq 0$, 这样的平移对结论会有什么影响呢?

这时 r_{ij} 相应的 A 矩阵记为 $\tilde{A} = (\tilde{a}_{ij})$, ρ_{ij} 相应的 $A = (a_{ij})$, 易见

$$\begin{aligned}\tilde{a}_{ij} &= a_{ij} + 2c \\ \tilde{a}_{ii} &= a_{ii} + 2(n-1)c\end{aligned}$$

因此

$$\tilde{A} = A + 2c(11^T + (n-2)I_n)$$

所以当 $Ax = \lambda x$ 时

$$\tilde{A}x = Ax + 2c(11^T + (n-2)I_n)x$$

若 $x^T 1 = (1^T x) = 0$, 则上式就是

$$\tilde{A}x = \lambda x + 2c(n-2)x = (\lambda + 2c(n-2))x$$

它的特征向量是不改变的.

例 7.2^[4] 这里用 21 个已知金矿为例, 按其储量大小, 由大到小排列编号, 每个矿用 67 个地质、物化探标志作为变量, 如地质年代、岩性、磁场特征、重力场特征等. 所有定量标志自然保留原始数据, 定性标志用 0, 1 值的量表示, 该矿有某一特性为 1, 没有为 0, 这样就得原始数据矩阵

$$\bar{Y} = (y_{ij})$$

它是 21×67 的矩阵.

取

$$\bar{\rho}_{ij} = \sum_{a=1}^{67} y_{ia} y_{ja}$$

也即第 i 矿与第 j 矿配合状况为相似性. 令

$$\text{当 } i \neq j \text{ 时, } \rho_{ij} = \bar{\rho}_{ij} - 1, \rho_{ii} = - \sum_{j \neq i} \rho_{ij}$$

$$i, j = 1, 2, \dots, n$$

然后求 A 的特征值, 用前二个特征值对应的特征向量, 得

$\lambda_1 = 126.78, \lambda_2 = 99.91$ 相应的特征向量:

$y_1 = (-0.0099, -0.0014, -0.0445, -0.0194, -0.0206,$
 $-0.0516, -0.0376, -0.0539, -0.0461, 0.0387,$
 $-0.0981, -0.0477, -0.0674, -0.3282, -0.0427,$
 $0.9205, -0.0076, -0.0522, 0.0030, 0.0540, -0.0872)$

$y_2 = (-0.0408, -0.0364, -0.0767, -0.0235, -0.0309,$
 $-0.0715, -0.0465, 0.0585, -0.0531, -0.0528,$
 $-0.0572, -0.0836, -0.1097, -0.0670, -0.0328,$
 $-0.0492, -0.0507, 0.9310, 0.2241, -0.1240)$

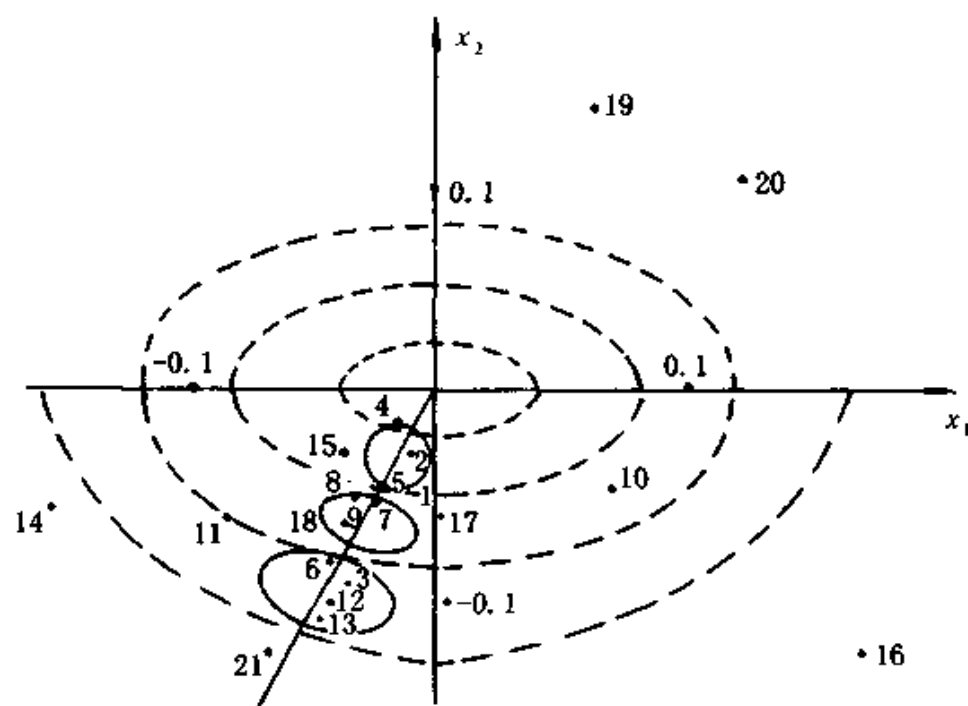


图 7-2

构建二维形如图 7-2. 可见用 1, 2, 4, 5 号矿, 7, 8, 9, 10 号矿, 11, 12, 13 号矿, 16, 19, 20, 21 号矿作为大、中、中小、小型矿的训练样本是建立预测模型可取的标准单元. 从标度图上看, 第 15, 18, 17 号矿可能还有未探明的潜在未知储量; 第 19, 16, 14, 20 号矿, 因结构异常, 有进一步研究的必要, 这正是该方法应用于综合评价所要解决的问题.

第三节 K-L方法

考虑先给定了不相似的度量值,用 s_{ij} 表示 n 个对象中第 i 个与第 j 个不相似的值, K-L 方法是在 r 维空间中, 求出向量 $x_{(1)}, \dots, x_{(n)}$, 使它们之间的距离平方

$$\begin{aligned}d_{ij}^2 &= \|x_{(i)} - x_{(j)}\|^2 \\&= \sum_{\alpha=1}^r (x_{i\alpha} - x_{j\alpha})^2\end{aligned}$$

与 s_{ij} 在总体上相当接近, 即存在 L 使

$$W^2 = \sum_{i \neq j} (s_{ij} - d_{ij}^2 - L)^2$$

达到最小. 当 $r=1$ 时, 就要求

$$W^2 = \sum_{i \neq j} (s_{ij} - (x_i - x_j)^2 - L)^2$$

达到最小.

实际求解时, 将 $-s_{ij}$ 作为相似性度量值, 利用上一节林知己夫的方法求出 x_1, \dots, x_n , 但是并不能使 s_{ij} 与 $(x_i - x_j)^2$ 相当接近. 此时 s_{ij} 与 $d_{ij}^2 = (x_i - x_j)^2$ 都是已知的, 因此可以求 K, L 两个值, 使

$$\sum_{i \neq j} (s_{ij} - K(x_i - x_j)^2 - L)^2$$

达到最小. 把 s_{ij} 当作因变量 y_{ij} , $d_{ij}^2 = (x_i - x_j)^2$ 当作自变量 u_{ij} , 上述问题是一个典型的回归问题, K 相当于回归系数, L 是回归常数, 所以 K, L 的公式立即可以导出. 注意到此时 y_{ij}, u_{ij} 是双足标, 各有 $n(n-1)$ 个值, 只须将一元线性回归的公式直接套用, 就可求出 K, L 的表达式.

记 $m = n(n-1)$, $\sum_{i \neq j}$ 表示对 $i \neq j$ 的 $n(n-1)$ 对 (i, j) 足标求和, 则

$$\begin{cases} K = \frac{m \sum_{i \neq j} s_{ij} \cdot d_{ij}^2 - (\sum_{i \neq j} s_{ij}) (\sum_{i \neq j} d_{ij}^2)}{m \sum_{i \neq j} d_{ij}^4 - (\sum_{i \neq j} d_{ij}^2)^2} \end{cases} \quad (7-28)$$

$$\begin{cases} L = \frac{1}{m} \sum_{i \neq j} s_{ij} - K \frac{1}{m} \sum_{i \neq j} d_{ij}^2 \end{cases} \quad (7-29)$$

有了(7-28)和(7-29),就可以将(7-29)的 L 作为 L_0 ,取 $x_i^{(0)} = \sqrt{K} x_i$, $K > 0$ 在一般情况下是能满足的(因为 s_{ij} 大时 d_{ij}^2 也应较大),于是求 y_i 使

$$W^2 = \sum_{i \neq j} (s_{ij} - L_0 - (x_i^0 - x_j^0 + y_i - y_j)^2)^2$$

达到最小,求得改进的 y_{i*} 之后,令

$$x_i^1 = x_i^0 + y_i, i = 1, 2, \dots, n$$

又可求 K_1, L_1 使

$$\sum_{i \neq j} (s_{ij} - K_1 (x_i^{(1)} - x_j^{(1)})^2 - L_1)^2$$

达到最小,如此不断迭代,就可以求解.这一迭代的公式可以很方便求出,迭代到改进很小时就可以停止.这就是 K - L 方法.

上面谈到的是用一维向量来表示,也即用实轴上的点表示.如果用平面上的点、 r 维空间的点来表示,又该怎样进行呢?

如果用平面上的点来表示,该方法是将它逐步一维化,即将上面这样求出的一维的点作为各点的第一个坐标,即 $x_{11}, x_{21}, \dots, x_{n1}$,这 n 个坐标已经有了, s_{ij} 也是有的,现在令

$$\begin{aligned} \tilde{s}_{ij} &= s_{ij} - (x_{i1} - x_{j1})^2, i \neq j \\ i, j &= 1, 2, \dots, n \end{aligned} \quad (7-30)$$

把 \tilde{s}_{ij} 作为 s_{ij} ,重复前面的方法,可以求 x_{i2} 使

$$W^2 = \sum_{i \neq j} (\tilde{s}_{ij} - (x_{i2} - x_{j2})^2 - L)^2$$

达到最小.虽然 \tilde{s}_{ij} 不一定全是非负的,但可以平移,使它们全为非负,上一节已讨论过这一做法不会影响求解,所以整个过程就不难重复进行.

容易看出,用这种方法,维数可以不断地增加,达到想要的维

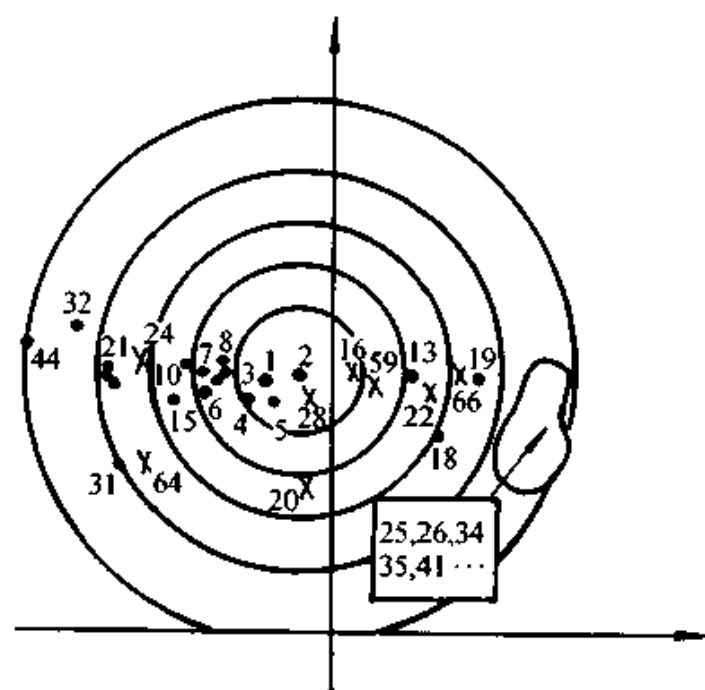


图 7-3

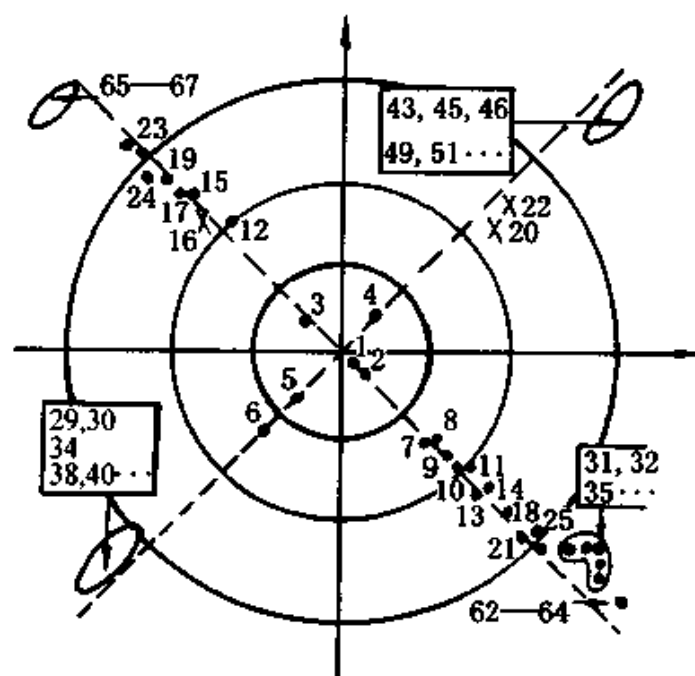


图 7-4

数.

怎样来衡量所得的表示点好坏的程度呢? 很自然, 应用所得的解 x_1, \dots, x_n 与 L 的值, 它们相应的与 s_{ij} 的平均差距来度量. 也即用

$$\frac{1}{n(n-1)} \sum_{i \neq j} (s_{ij} - (x_i - x_j)^2 - L)^2$$

来衡量.

例 7.3^[5] 莹石矿床的分析.

取我国 67 个已知的莹石矿床作为对象, 每个矿床取 33 个有关的地质标志值作为变量, 得一个 67×33 的矩阵, 然后导出不相似性的矩阵 $S = (s_{ij})$. 这些原始资料我们都略去, 着重于将求得的点表示清楚, 选用平面上的点来表示, 其结果是图 7-3. 采用林知己夫的方法, 所得结果是图 7-4, 我们一并给出, 以便比较.

第四节 谢帕尔德方法

谢帕尔德(Shepard)方法是从相似性矩阵出发求得 r 维空间的点 $x_{(1)}, \dots, x_{(n)}$, 使得 $\{x_{(i)}\}$ 之间的距离能与相似性尽量相符合. 记

$$x_{(i)} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ir} \end{bmatrix}, i = 1, 2, \dots, n$$

于是

$$d_{ij}^2 = \sum_{\alpha=1}^r (x_{i\alpha} - x_{j\alpha})^2, i, j = 1, 2, \dots, n$$

如果对于已给定的 n 个对象的相似性矩阵为

$$P = P^T = (\rho_{ij}), i, j = 1, 2, \dots, n, i \neq j$$

(ρ_{ii} 不赋值, 用 * 号标记),

很自然会提出如下的要求, 希望找出的 $X(i)$ 相应的矩阵 (d_{ij}) 与 (ρ_{ij}) 之间满足:

$$\begin{cases} \text{若 } \rho_{ij} = \rho_{ik}, & \text{则 } d_{ij} = d_{ik} \\ \text{若 } \rho_{ij} > \rho_{ik}, & \text{则 } d_{ij} < d_{ik} \end{cases} \quad (7-31)$$

因此对 P 中的 ρ_{ij} 可以按大小顺序排出(共 $\frac{n(n-1)}{2}$ 个)

$$\rho_{i_1 j_1} \geq \rho_{i_2 j_2} \geq \cdots \geq \rho_{i_m j_m}, \quad m = \frac{1}{2} n(n-1)$$

则相应地应该有

$$d_{i_1 j_1} \leq d_{i_2 j_2} \leq \cdots \leq d_{i_m j_m}$$

也就是对表示点的要求降低了,只要求按距离的近、远来排序和按相似性的大小排序相一致,而不要在量上有什么接近.衡量这个表示好坏如何,就看它与上述理想的情形(7-31)有多大差距.我们先看一个简单的例子,如何来计算这一差距,然后引出一般的定义,最后再给出求解的方法.

若给定四个点的相似性矩阵 P 和找出的表示点距离矩阵 D 为

$$P = \begin{bmatrix} * & 8 & 2 & 4 \\ & * & 6 & 9 \\ & & * & 3 \\ & & & * \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0.4 & 0.8 & 0.5 \\ & 0 & 1.2 & 0.7 \\ & & 0 & 1.5 \\ & & & 0 \end{bmatrix}$$

很明显 $\rho_{24}=9$ 是 P 中最大的,但 D 中 $d_{24}=0.7$ 是排在第3位(距离由小到大),因为 0.4 和 0.5 都比它小,它相当于 P 中由大到小第三位的 6,因此错位的程度是 $\rho_{ij} - \rho(d_{ij}) = 9 - 6 = 3$. 如果我们用 $S(\rho_{ij})$ 表示 ρ_{ij} 由大到小排的名次, S^{-1} 是它的反函数, $S(d_{ij})$ 表示 d_{ij} 由小到大的名次,于是上述 $\rho_{ij} - \rho(d_{ij}) = 9 - 6$ 就是

$$\rho_{24} - S^{-1}(S(d_{24}))$$

因此就可以导出一个优化的目标函数

$$\delta = \frac{\alpha}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n (\rho_{ij} - S^{-1}(S(d_{ij})))^2 \quad (7-32)$$

它的值越小越好.实际寻找时,就不断改进上述 δ 使它越来越小,直到无法改进为止.

为了计算上的方便,我们总假定

$$0 \leq \rho_{ij} \leq 1$$

若不然,令

$$\rho_{ij} = \frac{\rho_{ij} - \min \rho_{ij}}{\max \rho_{ij} - \min \rho_{ij}}, i \neq j; i, j = 1, 2, \dots, n$$

就可以达到要求. 因此以下总假定 $0 \leq \rho_{ij} \leq 1$.

下面给出求解的迭代算法.

首先给出一组初始点, 这可以利用 $-\rho_{ij}$ 为不相似性, 导出一组指定维数的表示点, 这些点用 $x_{(i)}^{(0)}, i = 1, 2, \dots, n$ 表示, 于是

$$\delta^{(0)} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} (\rho_{ij} - S^{-1}(S(d_{ij}^{(0)})))^2$$

$$d_{ij}^{(0)} = \|x_{(i)}^{(0)} - x_{(j)}^{(0)}\|, i \neq j; i, j = 1, 2, \dots, n$$

现在来进行第一次迭代, 降低 $\delta^{(0)}$ 的值, 令

$$x_{(i)}^{(1)} = x_{(i)}^{(0)} + \Delta x_{(i)}^{(0)}, i = 1, 2, \dots, n$$

为了方便, 我们将复合函数 $S^{-1}(S(d_{ij}))$ 记为 $\rho(d_{ij})$, 于是令

$$\begin{cases} a_{ij}^{(0)} = \alpha(\rho_{ij} - \rho(d_{ij}^{(0)})) \frac{x_{(i)}^{(0)} - x_{(j)}^{(0)}}{d_{ij}^{(0)}} \\ \text{其中 } \alpha > 0 \text{ 是给定的常数} \\ b_{ij}^{(0)} = \beta(\rho_{ij} - \bar{\rho}) \frac{x_{(i)}^{(0)} - x_{(j)}^{(0)}}{d_{ij}^{(0)}} \\ \text{其中 } \beta > 0 \text{ 是常数, } \bar{\rho} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \rho_{ij} \end{cases} \quad (7-33)$$

于是修正量

$$\Delta x_{(i)}^{(0)} = \sum_{i \neq j} (a_{ij}^{(0)} + b_{ij}^{(0)})$$

这样就可得一次迭代后的 $x_{(i)}^{(1)}, i = 1, 2, \dots, n$, 以 $x_{(i)}^{(1)}$ 作为上述步骤的 $x_{(i)}^{(0)}$, 再进行迭代, 如此下去, 直到迭代后与迭代前差别在指定范围之内就可停止. α, β 是两个选择的常数, 目的是使迭代收敛的速度尽可能快.

下面用一个例子来说明这一方法的用途.

例 7.4 Morse 电码混同率的评价.

前面曾对 Morse 电码混同率作过一些讨论,现在列出 1957 年 Rothkopf 进行的实验,所得的相似性度量表(见表 7-4).取前两个正特征根,得平面图 7-5.从图 7-5 可以看出:

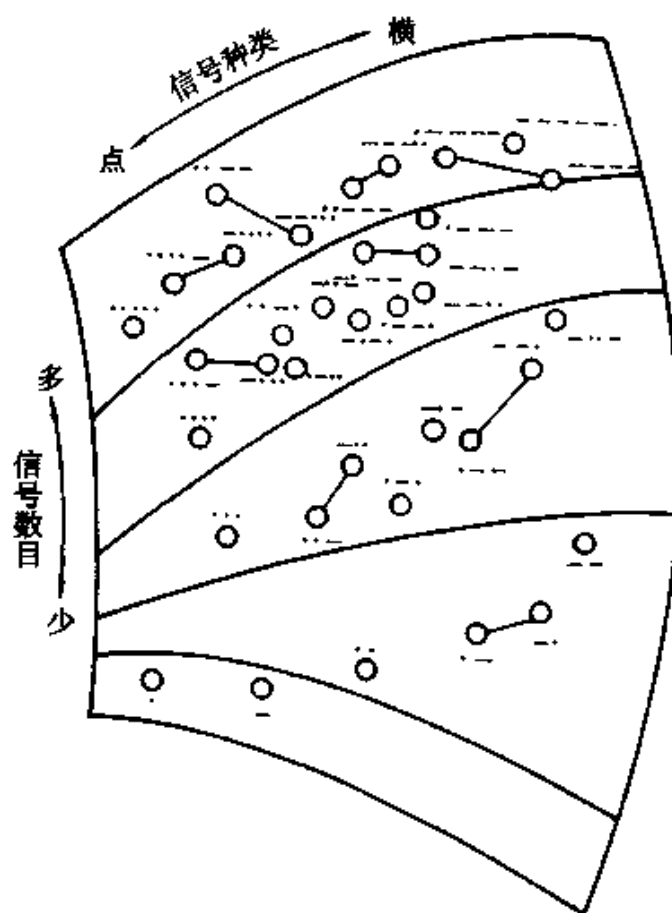


图 7-5 谢帕尔德所求得莫尔斯电码的空间配置及其解释

(i)在平面上相距越近的混同率越高;

(ii)“·”多的偏左,“-”多的偏右;

(iii)码多的居上,码少的居下.

这些特点明显地反映了这一方法的识别能力.

表 7-4 关于莫尔斯电码混同率的罗思科普弗数据(单位:%)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	1	2	3	4	5	6	7	8	9	0
A	92	04	08	13	03	14	10	13	46	05	22	03	25	34	06	06	09	35	23	06	37	13	17	12	07	03	02	07	05	05	08	06	05	06	02	03
B	06	84	37	31	05	28	17	21	05	19	34	40	06	10	12	22	25	16	18	02	18	34	08	84	30	42	12	17	14	40	32	74	43	17	04	04
C	04	38	87	17	04	29	13	67	11	19	24	35	14	03	09	51	34	24	14	06	06	11	14	32	82	38	13	15	31	14	10	30	28	24	18	12
D	08	62	17	86	07	23	40	36	09	13	81	56	08	07	09	27	09	45	29	06	17	20	27	40	15	33	03	09	06	11	09	19	08	10	05	06
E	06	13	14	06	97	02	04	04	17	01	05	06	04	04	05	01	05	10	07	67	03	03	02	05	06	05	04	03	05	03	05	02	04	02	03	03
F	04	51	33	19	02	90	10	29	05	33	16	50	07	06	10	42	12	35	14	02	21	27	25	19	27	13	08	16	47	25	26	24	21	05	05	05
G	09	18	27	38	01	14	90	06	05	22	33	16	14	13	82	52	23	21	05	03	15	14	32	21	23	39	15	14	05	10	04	10	17	23	20	11
H	08	45	23	25	09	32	08	87	10	10	09	29	05	08	08	14	08	17	37	04	36	59	09	33	14	11	03	09	15	43	70	35	17	04	03	03
I	64	07	07	13	10	08	06	12	93	03	05	10	13	30	07	03	05	19	35	16	10	05	08	02	05	07	02	05	08	09	06	08	05	02	04	05
J	07	09	38	09	02	24	08	05	04	85	22	31	08	03	21	63	47	11	02	07	09	09	09	22	32	28	67	66	33	15	07	11	28	29	26	23
K	05	24	38	73	01	17	25	11	05	27	91	33	10	12	31	14	31	22	02	02	23	17	33	63	16	18	05	09	17	08	08	18	14	13	05	06
L	02	69	43	45	10	24	12	26	09	30	27	86	06	02	09	37	36	28	12	05	16	19	20	31	25	59	12	13	17	15	26	29	56	16	07	03
M	24	12	05	14	07	17	29	08	08	11	23	08	96	62	11	10	15	20	07	09	13	04	21	09	18	08	05	07	06	06	05	07	11	07	10	04
N	31	04	13	30	08	12	10	16	13	03	16	08	59	93	05	09	05	23	12	10	16	04	12	04	06	11	05	02	03	04	04	06	02	02	10	02
O	07	07	20	06	05	09	76	07	02	39	25	10	04	08	86	37	35	10	03	04	11	14	25	35	27	27	19	17	07	07	06	18	14	11	20	12
P	05	22	33	12	05	36	22	12	03	78	14	46	05	06	21	83	43	22	09	04	12	19	19	19	41	30	34	44	24	11	15	17	24	23	25	13
Q	08	20	38	11	04	15	10	05	02	27	23	36	07	06	22	51	91	11	02	03	06	14	12	37	50	63	34	32	17	12	09	27	40	58	37	24
R	13	14	16	23	05	34	26	15	07	12	21	37	14	12	12	29	08	87	16	02	25	23	62	14	12	13	07	10	13	04	07	12	07	09	06	02
S	17	24	05	30	11	26	05	59	16	03	13	10	06	17	06	06	03	18	96	09	56	24	12	10	06	07	08	02	02	15	28	09	05	05	02	02
T	13	10	01	05	46	03	06	06	14	06	14	07	06	05	06	11	04	04	07	96	08	05	04	02	02	06	05	05	03	03	03	08	07	06	14	06

U	14	29	12	32	04	32	11	34	21	07	44	32	11	13	06	20	12	40	51	06	93	57	34	17	09	11	06	06	16	34	10	09	09	07	04	03	U
V	05	17	24	16	09	29	06	39	05	11	26	43	04	01	09	17	10	17	11	06	32	92	17	57	35	10	10	14	28	79	44	36	25	10	01	05	V
W	09	21	30	22	09	36	25	15	04	25	29	18	15	06	26	30	25	61	12	04	19	20	86	22	25	22	10	22	19	16	05	09	11	06	03	07	W
X	07	64	45	10	03	28	11	06	01	35	50	42	10	08	24	32	61	10	12	03	12	17	21	91	48	26	12	20	24	27	16	57	29	16	17	06	X
Y	09	23	62	15	04	26	22	09	01	30	12	14	05	06	14	30	52	05	07	04	06	13	21	44	86	23	26	44	40	15	11	26	22	33	23	16	Y
Z	03	46	45	18	02	22	17	10	07	23	21	51	11	02	15	59	72	14	04	03	09	11	12	36	42	87	16	21	27	09	10	25	66	47	15	15	Z
1	02	05	10	03	03	05	13	04	02	29	05	14	09	07	14	30	28	09	04	02	03	12	14	17	19	22	84	63	13	08	01	08	19	32	57	55	1
2	07	14	22	05	04	20	13	03	25	26	09	14	02	03	17	37	28	06	05	03	06	10	11	17	30	13	62	89	54	20	05	14	20	21	16	11	2
3	03	08	21	05	04	32	06	12	02	23	06	13	05	02	05	37	19	09	07	06	04	16	06	22	25	12	18	64	86	31	23	41	16	17	08	10	3
4	06	19	19	12	06	25	14	16	07	21	13	19	03	03	02	17	29	11	09	03	17	55	08	37	24	03	05	26	44	89	42	44	32	10	03	03	4
5	08	45	15	04	02	45	04	67	07	14	04	41	02	00	04	13	07	09	27	02	14	45	07	45	10	10	14	10	30	69	90	42	24	10	00	05	5
6	07	80	30	17	04	23	04	14	02	11	11	27	06	02	07	16	30	11	14	03	12	30	09	58	38	39	15	14	23	24	17	86	69	14	05	14	6
7	06	33	22	14	05	25	06	04	06	24	13	32	07	06	07	36	39	12	06	02	03	13	09	30	30	50	22	29	18	15	12	61	85	70	20	13	7
8	03	23	40	06	03	15	15	06	02	33	10	14	03	06	14	12	45	02	06	04	06	07	05	24	35	50	42	29	16	16	16	30	60	89	61	26	8
9	03	14	23	03	01	06	14	05	02	30	06	07	16	11	10	31	32	05	06	07	06	03	08	11	21	24	57	39	09	12	04	11	42	56	91	78	9
0	09	03	11	02	05	07	14	09	05	30	08	03	02	03	25	21	29	02	03	04	05	03	02	12	15	20	50	26	09	11	05	22	17	52	81	94	0

第五节 克拉斯卡尔方法

克拉斯卡尔(Kruskal)方法的基本想法与谢帕尔德方法是一样的,只是在距离的选择上可以不是欧氏的,而用闵可夫斯基的距离,它比欧氏距离更广泛些.另一个不同之处是上节的方法以 $P = (\rho_{ij})$ 为主,而对 $D = (d_{ij})$ 只用它的顺序关系,这节的方法正好相反,以距离之间的差距大小为衡量标准,而只用到 P 中 ρ_{ij} 的顺序关系.

以下求和号 $\sum_{i,j}$ 表示这两种求和:

- (i) 当 $P^T = P$ 时,它表示 $\sum_{1 \leq i < j \leq n}$;
- (ii) 当 P 不是对称阵时,它表示 $\sum_{i \neq j}$.

同样地

$$m = \begin{cases} \frac{1}{2}n(n-1), & \text{当 } P^T = P \\ n(n-1), & \text{当 } P^T \neq P \end{cases}$$

现在我们来看如何去求解,优化的目标是什么.假定已给好相似性度量矩阵 $P = (\rho_{ij})$,在指定 r 维空间中用 n 个点表示它们,相互之间的距离为 d_{ij} ,于是按 ρ_{ij} 的大小顺序,将 d_{ij} 也可以依次排出.若将 ρ_{ij} 依从大到小的顺序放在横轴上,不考虑 ρ_{ij} 的大小,于是可用 $1, 2, \dots, m$ 这 m 个点表示,对应的 d_{ij} 就是纵轴 y 的值,因此 P 与 D 的点图为图 7-6.

如果 d_{ij} 与 ρ_{ij} 是一致的,则图形应该是一条单调上升的直线,也即有

$$\rho_{ij} > \rho_{kl} \Rightarrow d_{ij} \leq d_{kl}$$

$$\rho_{ij} = \rho_{kl} \Rightarrow d_{ij} = d_{kl}$$

单调性的要求一般是不满足的,因此要求新的点 $x_{(1)}, \dots, x_{(n)}$,使它们的距离 \hat{d}_{ij} 能合于单调性的要求,而与初始表示的距离尽可能接近,也即要求 \hat{d}_{ij} 满足

$$Q = \sum_{i,j} (d_{ij} - \bar{d}_{ij})^2 \quad (7-34)$$

达到最小.

从这个要求可以看出, 初始的表示 $x_{(1)}, \dots, x_{(n)}$ 是很重要的, 只是在原有的表示上作一些调整, 初始表示可以用前面讨论过的任何一种方法去求得.

用 $\rho_1 < \rho_2 < \dots < \rho_m$ 表示 P 中的 ρ_{ij} 按由小到大排出的全体 m

个值, 每个 ρ_l 实际上是某一个 ρ_{ij} , 于是将对应的 d_{ij} 写在 ρ_l 下面, 即有

$$\begin{array}{ccccccc} \rho_1 & < & \rho_2 & < & \rho_3 & < & \dots < \rho_m \\ d_1 & & d_2 & & d_3 & & \dots & d_m \end{array}$$

$\rho_l = \rho_{ij}$ 时, $d_l = d_{ij}$.

现在用分段平均法求 \bar{d}_l (即 \bar{d}_{ij}). 将上述按 $\rho_1 < \rho_2 < \dots < \rho_m$ 排的 d_1, d_2, \dots, d_m 分成 β 段, 每一段有若干个 d_l , 其中第 b 段有 m_b 个 d_l , 记为 $d_{b1}, d_{b2}, \dots, d_{bm_b}$, 求它的均值, 即

$$\bar{d}_b = \frac{1}{m_b} \sum_{a=1}^{m_b} d_{ba}, \quad b = 1, 2, \dots, \beta \quad (7-35)$$

只要段分得合适, 就会出现

$$\bar{d}_1 < \bar{d}_2 < \dots < \bar{d}_\beta$$

于是每一段中的 d_{ba} 均由 \bar{d}_b 来代替, 而能保持单调性.

具体步骤为

1. 先令 $\bar{d}_1 = d_1$;

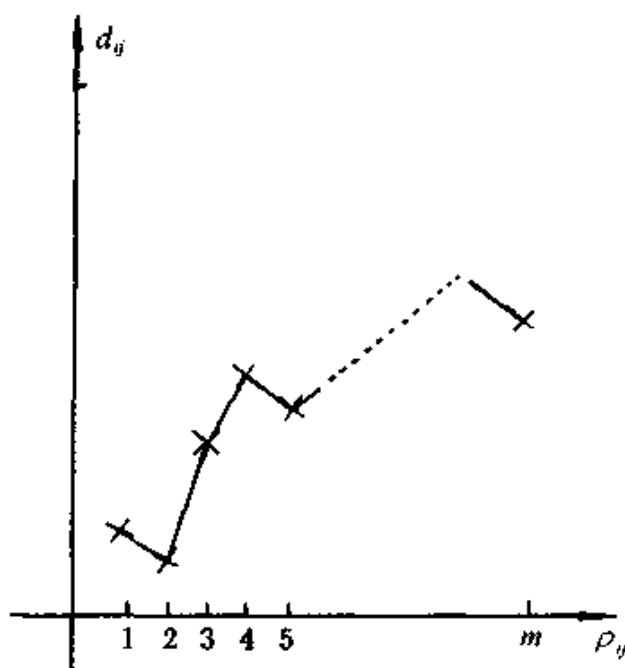


图 7-6 示意图

2. 若已有 $\hat{d}_1 < \hat{d}_2 < \cdots < \hat{d}_{k-1}$, 就取 $\hat{d}_k = d_k$. 当 $\hat{d}_k > \hat{d}_{k-1}$ 时, 这个 \hat{d}_k 不必动. 当 $\hat{d}_k < \hat{d}_{k-1}$ 时, 就要调整如下:

对 $i = 1, 2, \cdots, k-2$, 寻求合于条件

$$\frac{1}{1+i}(\hat{d}_k + \sum_{i=1}^i \hat{d}_{k-i}) \geq \hat{d}_{k-i-1} \quad (7-36)$$

的最小的 i , 找到了这个 i 就令

$$\hat{d}_k = \hat{d}_{k-1} = \cdots = \hat{d}_{k-i} = \hat{d}_*$$

\hat{d}_* 就是合于(7-36)要求最小 i 所对应的左端的值. 于是就有

$$\hat{d}_1 \leq \hat{d}_2 \leq \cdots \leq \hat{d}_{k-i-1} = \hat{d}_{k-i} = \cdots = \hat{d}_k$$

若找不到(7-36)这样的 i , 就令

$$\hat{d}_* = \frac{1}{k} \sum_{i=1}^k \hat{d}_i$$

$\hat{d}_1, \cdots, \hat{d}_k$ 均调整为 \hat{d}_* .

3. 这一步骤一直进行到 \hat{d}_m 为止.

上述算法, 最后一定可以将 \hat{d}_i 调整好且具有单调性. 用什么来衡量调整到 \hat{d}_i 后的好坏呢? 这里用

$$\eta = \left(\frac{\sum_{i,j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} (d_{ij} - \bar{d})^2} \right)^{\frac{1}{2}} \quad (7-37)$$

来衡量, 其中 $\bar{d} = \frac{1}{m} \sum_{i,j} d_{ij}$.

从(7-37)的表达式, 就可以看出, 只要调整各表示点的坐标, 就可以调整 \hat{d}_i 的值, (7-37)是坐标的二次函数的比值, 因此可以用多元函数的求极值方法, 逐次调整, 使 η 值不断下降, 这些在原则上都是不困难的. 要列出有关的算法和公式是很繁的, 这里就略去了.

下面举一个例子说明这一应用.

例 7.5^[5] 林知己夫于 1976 年对广岛市 1120 名居民进行就城市问题的统计调查, 回收已填好表 802 份, 就有关问题按居民认为的重要程度——非常重要、重要、比较重要、不重要四级评价. 将调查项目看作有序变量, 算出它们两两间相似性度量, 以此求各项目的标度, 给出综合结果. 关于交通问题, 设置调查项目如下:

A_1 :走路时

- ①人行道与车行道分开;
- ②应严格遵守交通规则;
- ③设置步道桥以减轻交通负担;
- ④为身体不便的人走路安全想方设法;
- ⑤将人行道铺装得更好以利于行走.

A_2 :利用电车和公共汽车时

- ⑥从家到车站要近;
- ⑦增加运行时间使清晨和深夜均有车可乘;
- ⑧增加运行车次以减少等车时间;
- ⑨车内不混乱;
- ⑩道路不混杂,易于通行;
- ⑪调整线路,到处均有车通行;
- ⑫身体不便的人和老年人上下车方便.

A_3 :利用自行车与出租车时

- ⑬应很好遵守交通规则;
- ⑭道路不混杂,易于通行;
- ⑮停车场地好;
- ⑯车道铺装良好;
- ⑰利用出租车,不需等待即可乘坐.

B. 作为快适交通的条件

- ⑱有安全行走的人行道;
- ⑲利用电车和公共汽车安全便利;
- ⑳利用自用车和出租车安全便利.

C. 总之

- ㉑有安全快适的交通.

解 该题给出了一个 802×21 的原始数据阵,按照 Kruskal 方法,建立相似性数据阵,并求出 $r=5$ 时的适合度 $\eta=0.054$.

表 7-5 交通问题中各项目按克拉斯卡尔方法计算的结果

	1	2	3	4	5
1	1.196	-0.303	0.758	0.161	0.141
2	0.283	0.342	0.754	0.704	0.311
3	-0.182	-0.305	0.769	-0.776	-0.570
4	0.033	0.710	0.678	-0.193	-0.341
5	-0.379	-0.378	0.492	-0.179	0.440
6	-0.802	-0.589	-0.012	-0.109	-0.224
7	-0.290	-0.821	-0.246	-0.254	-0.969
8	-0.227	-0.680	-0.434	-0.079	-0.632
9	-0.575	-0.285	-0.423	-0.319	-0.264
10	-0.528	0.074	-0.668	-0.445	-0.130
11	-0.288	-0.013	-0.420	0.350	-0.389
12	-0.155	0.607	0.103	-0.062	-0.307
13	0.289	0.586	0.266	0.662	0.333
14	-0.264	0.081	-0.444	0.241	0.491
15	-0.146	0.220	-0.388	0.254	0.998
16	-0.155	-0.039	-0.016	0.335	0.604
17	-0.162	-0.588	-0.340	-0.327	0.268
18	0.847	0.392	0.078	0.533	-0.199
19	0.472	0.227	-0.201	0.196	-0.173
20	0.319	0.442	-0.109	-0.017	0.339
21	0.707	0.320	-0.197	-0.677	0.271

因为 $r=5$, 用 5 个坐标中第三与第五个坐标形成平面来表示, 得二维空间配置: 虽然我们从一侧面来反映样本结果, 但至少关于调查项目的分类在这里得到了清晰的反映。

多维标度标准化后标度值的重心在原点, 那在五维空间下, 同心五维球面上的点, 对于城市交通问题来说具有同等重要的程度。这种空间结构为城市交通规划提供了较为科学的依据。

本例也可利用同样的相似性数据用因子分析方法进行。取前 17 个基本调查项目的数据, 得 $P=8$ 时收敛性较好, 方差贡献率见表 7-7。

有兴趣的读者可以将前两个因子作轴看一下调查项目的空间配置, 这里从略。

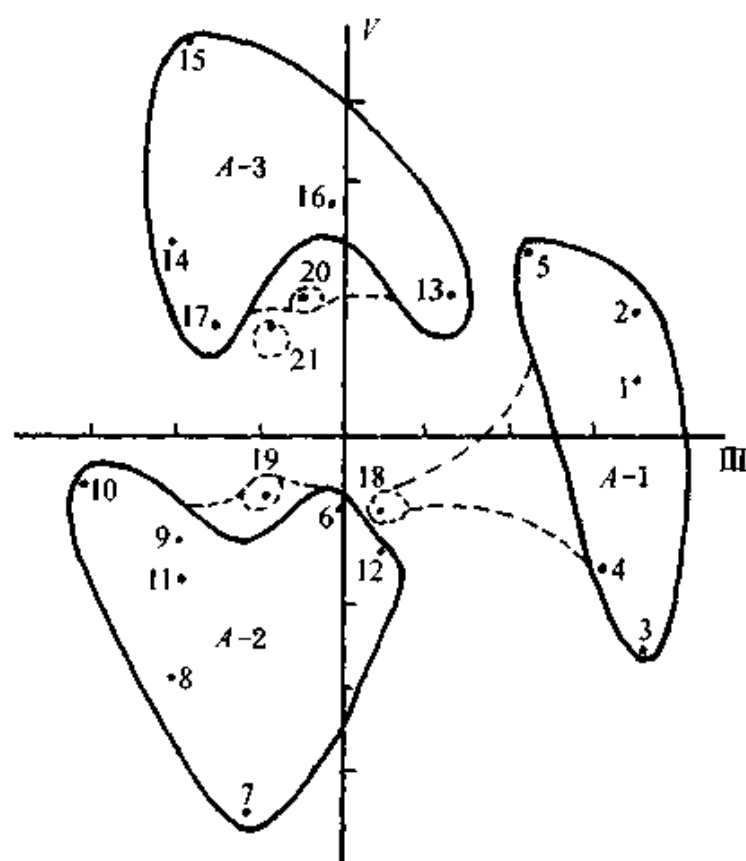


图 7-7 交通问题中各项目按克拉斯卡尔方法所求得分类

表 7-6

方差	5.2917	1.8698	1.3329	1.0061	0.9480	0.8321	0.7306	0.6731
贡献率	0.4172	0.1474	0.1051	0.0793	0.0747	0.0656	0.0576	0.0531
累计 贡献率	41.7%	56.4%	67%	75%	82.4%	88.9%	94.6%	100%

第六节 最小维数分析法(MDA 方法)

当不相似性无法度量时,两个对象之间不相似的程度只能分成等级,此时不相似矩阵 P 中的元素都是正整数,无妨设为 $1, 2, \dots, G$,而每个数只代表一种等级,当 $i = j$ 时, ρ_{ij} 无法定义, P 矩阵只有 $n(n-1)$ 个元素.

此时若 n 个对象用 r 维的 n 个点表示, $x_{(1)}, \dots, x_{(n)}$ 是 n 个

点,

$$x_{(i)}^T = (x_{i1}, \dots, x_{ir})$$

$$d_{ij}^2 = \sum_{a=1}^r (x_{ia} - x_{ja})^2 \stackrel{\text{记}}{=} m_{ij}$$

$$i \neq j, i, j = 1, 2, \dots, n$$

于是将 m_{ij} 也分为 G 类, 使类内的方差尽可能小, 类间的方差尽可能大. 这只要将 m_{ij} 按由小到大顺序排列, 用聚类分析中的最优分割法分成 G 类就可完成. 这样的分割就与给定的 P 没有任何关系, 因此寻找这种分割时, 应与 P 中 ρ_{ij} 的取值的分类有关, 所以优化的目标函数中就必须能体现出 P 中不相似性的影响. 为此, 我们引入

$$\delta_{ij}(g) = \begin{cases} 1 & \rho_{ij} = g, \\ 0 & \rho_{ij} \neq g, \end{cases} \quad g = 1, 2, \dots, G$$

和前面林知己夫的方法类似, 我们也是先定出一维的表示, 然后再考虑如何增加维数, 求出更好的表示.

所以先是求 n 个数 x_1, \dots, x_n , 使目标函数达到最大, 为了给出目标函数的表达式, 先引入一些记号, 记

$$m_{ij} = d_{ij}^2 = (x_i - x_j)^2$$

$$\bar{m} = \frac{1}{n(n-1)} \sum_{i \neq j} m_{ij} \quad (7-38)$$

$$f_g = \sum_{i \neq j} \delta_{ij}(g) (= P \text{ 中取值为 } g \text{ 的 } \rho_{ij} \text{ 的个数})$$

$$\sigma^2 = \frac{1}{n(n-1)} \sum_{i \neq j} (m_{ij} - \bar{m})^2 \quad (7-39)$$

$$\sigma_b^2 = \frac{1}{n(n-1)} \sum_{g=1}^G f_g (\bar{m}_g - \bar{m})^2 \quad (7-40)$$

其中

$$\bar{m}_g = \frac{1}{f_g} \sum_{i \neq j} m_{ij} \delta_{ij}(g) \quad (7-41)$$

上式中 σ^2 是 m_{ij} 全部数据的变异, σ_b^2 是按 P 中的值分类后, 类之间离异程度. 熟悉统计分析的都了解, 总的变异可以分解为两部

分:组内变异 σ_w^2 和组间变异 σ_b^2 , 即有

$$\sigma^2 = \sigma_b^2 + \sigma_w^2 \quad (\text{或 } \sigma_w^2 = \sigma^2 - \sigma_b^2)$$

而表示的好坏, 就是表示的 $|x_i|$ 应使

$$\eta^2 = \sigma_b^2 / \sigma^2 \quad (7-42)$$

达到最大, 也即要

$$\eta^2 = 1 - \sigma_w^2 / \sigma^2$$

达到最大.

怎样用迭代方法来使 η^2 的值变大呢? 设第 0 步(初始)的值选定为 $x_1^{(0)}, \dots, x_n^{(0)}$, 于是迭代一次后的值为 $x_j^{(1)}, j=1, 2, \dots, n$, 且

$$x_j^{(1)} = x_j^{(0)} + y_j^{(1)}, j = 1, 2, \dots, n \quad (7-43)$$

利用(7-43), 可以求出相应组间差 $\sigma_b^2(1), \eta^2(1), \sigma^2(1)$, 此时

$$\begin{aligned} m_{ij}^{(1)} &= (x_i^{(0)} + y_i^{(1)} - x_j^{(0)} - y_j^{(1)})^2 \\ &= m_{ij}^{(0)} - 2m_{ij}^{(0)}(y_i^{(1)} - y_j^{(1)}) + (y_i^{(1)} - y_j^{(1)})^2 \end{aligned}$$

注意 $\delta_{ij}(g), f_g$ 都是确定的常数, 因此 η^2 只是 m_{ij} 的函数, 将 η^2 用 m_{ij} 的线性函数(它的一次泰勒展式)来近似, 就可以得到一个增量 y_j 的函数、再加上约束条件就是一个典型的线性规划问题, 现在来推导这个规划问题. 我们用微分的形式来推导比较方便. 首先要注意已有的约束条件是什么, 它们如何限制了 y_i 的取值, 然后目标函数 η^2 如何变成线性的. 为了方便, 我们将 $m_{ij}^{(0)}$ 都用 m_{ij} 表示, $y_i^{(1)}$ 都用 y_i 表示, $x_i^{(0)}$ 都用 x_i 表示, 这样推导公式时, 符号简单.

约束条件共有两种四条:

(i) 是问题本身可以附加的.

因为我们只是要求一种相对表示, 所以无妨要求所选的点始终满足:

$$\sum_{i=1}^n x_i = 0, \sum_{i=1}^n x_i^2 = n(n-1)$$

写成对 y_j 的条件, 自然有

$$\sum_{j=1}^n (x_j + y_j) = \sum_{j=1}^n x_j + \sum_{j=1}^n y_j = 0$$

因此必须有

$$\sum_{j=1}^n y_j = 0 \quad (7-44)$$

从 $1 = \frac{1}{n(n-1)} \sum_{i=1}^n x_i^2 = \frac{1}{n(n-1)} \sum_{i=1}^n (x_i + y_i)^2$, 就得到

$$\sum_{i=1}^n (x_i)^2 = \sum_{i=1}^n x_i^2 + 2 \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2$$

即有

$$0 = 2 \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2$$

就得

$$\sum_{i=1}^n y_i (y_i + 2x_i) = 0 \quad (7-45)$$

(ii)是与优化目标及算法有关的。

我们希望改进后解仍然要与 P 中 ρ_{ij} 所取的值(归类的值)有更好的一致性,这就是(7-41)中的 \bar{m}_g 与 g 值的大小有同样的单调序(因为在通常情况下, ρ_{ij} 的值越大表示越不相似),这就是

$$\bar{m}_g \leq \bar{m}_{g+1}, g = 1, 2, \dots, G-1 \quad (7-46)$$

从计算角度考虑,每次步长 α 不能太长, α 必须是 $(0, 1)$ 中的一个数,也即

$$|y_i| \leq \alpha |x_i|, i = 1, 2, \dots, n \quad (7-47)$$

这(7-44)~(7-47)四个条件就是约束条件

现在来看目标函数如何在迭代中改进,此时利用 $\sigma^2 \equiv 1$ 就知道

$$\eta^2 = \sigma_b^2 = \frac{1}{n(n-1)} \sum_{g=1}^G f_g \bar{m}_g^2 - \bar{m}^2$$

因此, $n(n-1)\eta^2 = \sum_{g=1}^G f_g \bar{m}_g^2 - n(n-1)\bar{m}^2$, 于是

$$n(n-1)d\eta^2 = 2 \sum_{g=1}^G f_g \bar{m}_g d\bar{m}_g - n(n-1)2\bar{m}d\bar{m}$$

$$\begin{aligned}
&= 2 \sum_{g=1}^G f_g \bar{m}_g \frac{1}{f_g} \sum_{i \neq j} \delta_{ij}(g) dm_{ij} \\
&\quad - 2n(n-1) \bar{m} \frac{1}{n(n-1)} \sum_{i \neq j} dm_{ij} \\
&= 2 \sum_{i \neq j} \left[\left(\sum_{g=1}^G \delta_{ij}(g) \bar{m}_g \right) - \bar{m} \right] dm_{ij}
\end{aligned}$$

因此用 $y_i - y_j$ 代入上述 dm_{ij} , 就得线性规划的目标函数(它是 y_i 的线性函数):

$$F = \sum_{i \neq j} \left(\sum_{g=1}^G \delta_{ij}(g) \bar{m}_g - \bar{m} \right) (y_i - y_j) \quad (7-48)$$

这样我们就将每一步要解的线性规划问题推导出来了.

细看一下, 似乎约束条件(7-45)不是线性的, 但很容易将它化成线性的.

因为从 $x_i \rightarrow x_i + y_i$ 后, x_i 相应的 σ^2 记为 $\sigma_{(0)}^2$, $x_i + y_i$ 相应的记为 $\sigma_{(1)}^2$, 于是有近似式

$$\sigma_{(1)}^2 = \sigma_{(0)}^2 \left(1 + \sum_{i=1}^n b_i y_i \right) \quad (7-49)$$

可见要保持 $\sigma_{(1)}^2 = \sigma_{(0)}^2 = 1$, 其必要而充分的条件是

$$\sum_{i=1}^n b_i y_i = 0$$

其中 b_i 可用 x_i 表示为

$$\begin{aligned}
b_i = & \frac{8}{n(n-1)} \sum_{j \neq i} (x_i - x_j) \left[(x_i - x_j)^2 \right. \\
& \left. - \frac{1}{n(n-1)} \sum_{l \neq j} (x_l - x_j)^2 \right]
\end{aligned}$$

这一推导过程我们就略去了, 读者可以参阅文献[5]的有关内容. 下面用例子来说明这一方法的应用.

例 7.5^[5] 关于水稻品种的综合评价.

取 17 个品种的水稻, 各品种来源情况如下:

1, 2, 3, 4, 5 日本栽培种

6, 7, 8, 9 印度野生种

10 泰国野生种

11,12,13 热带美洲野生种

14 非洲栽培种

15,16,17 非洲野生种

(i) 现用两品种 i, j 杂交, 测定其杂交后的结实度, 作为它们之间的相似性度量, 令

$$s_{ij} = \begin{cases} 1, i, j \text{ 杂交后结实非常好} \\ 2, i, j \text{ 杂交后结实好} \\ 3, i, j \text{ 杂交后结实不好} \\ 4, i, j \text{ 杂交后结实未成熟.} \end{cases}$$

则 s_{ij} 构成对象 i 与 j 之间的一种非相似性度量数据, 见表 7-7.

表 7-7 17 个水稻品种之间的非相似性等级

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1		1	3	3	1	1	1	1	2	1	4	4	4	4	4	4	4
2			2	2	1	1	2	2	1	1	4	4	4	4	4	4	4
3				1	1	1	3	1	3	1	4	4	4	4	4	4	4
4					1	2	3	1	2	3	4	4	4	4	4	4	4
5						1	2	1	1	1	4	4	4	4	4	4	4
6							1	1	1	1	3	4	3	3	4	3	4
7								1	1	3	4	3	3	4	3	4	4
8									2	1	3	4	4	4	4	4	4
9										2	2	4	4	4	4	4	4
10											3	4	4	4	4	4	4
11												1	1	4	4	4	4
12													1	4	4	4	4
13														4	4	4	4
14															1	1	1
15																1	1
16																	1
17																	

(ii)取 $e_{ij} = s_{ij}$, 用 e_{ij} 型数量化方法计算初值并且依

$$x'_u = x_u / 4 \sqrt{\sigma^2}$$

来正则化数据. 然后用林知己夫数量化方法(见本章第二节)求出一维表示的点坐标, 列于表 7-8 中第一列数值; 然后考虑线性规划问题求解, 每次迭代求解后 η^2 的值列于表 7-9 中, 共迭代了 20 次. 可以看出, 在第 15 次 $\eta^2 = 0.46089$ 为最大, 于是就用这一步的解的第一个坐标. 然后求第二个坐标, 重复前面的步骤, 得第二个坐标同样是 20 次规划求解, 从表 7-9 第二列的数值看出, 第 19 次 η^2 的值最大, 用它的解作为第二个坐标. 将第一、第二坐标的结果列在表 7-10 中.

表 7-8 各个坐标的迭代初值

对象的 序号 u	第一坐标的初值 x_{u1}^0	第二坐标的初值 x_{u2}^0	第三坐标的初值 x_{u3}^0
1	0.31131	-0.60563	0.19289
2	0.31349	-0.62114	0.10285
3	0.31559	-0.71286	-0.14635
4	0.31944	-0.72020	-0.07448
5	0.31399	-0.60998	0.11746
6	0.23396	-0.03626	-0.09746
7	0.26569	0.00324	1.62313
8	0.31581	-0.35837	-0.07739
9	0.32019	-0.28116	-0.16619
10	0.31943	-0.34576	-0.41602
11	0.42152	0.77843	-0.74737
12	0.46927	0.91428	0.07359
13	0.44568	1.40566	0.09235
14	-1.06179	0.31493	-0.25026
15	-1.05972	0.27956	0.19367
16	-1.06366	0.30963	-0.20766
17	-1.18018	0.23562	-0.21275

表 7-9 迭代次数及各个坐标相应的 η^2 值

迭代次数 l	第一坐标相应的 $\eta^{2,l}$	第二坐标相应的 $\eta^{2,l}$
0	0.44787	0.75980
1	0.44200	0.79212
2	0.45103	0.80786
3	0.44388	0.82213
4	0.45763	0.83203
5	0.45212	0.83653
6	0.45875	0.84018
7	0.45302	0.84450
8	0.45634	0.84999
9	0.45187	0.84671
10	0.46046	0.85359
11	0.45400	0.84987
12	0.45923	0.85132
13	0.45908	0.85844
14	0.45474	0.86004
15	0.46089	0.86051
16	0.45193	0.86176
17	0.45105	0.86618
18	0.45128	0.86428
19	0.46061	0.86748
20	0.45279	0.86653

表 7-10 二维标度结果

u	x_{u1}	x_{u2}
1	0.16562	-0.50696
2	0.16524	-0.59295
3	0.14638	-0.55411
4	0.13893	-0.55064
5	0.17882	-0.54065
6	0.05573	-0.05146
7	0.16180	0.00051
8	0.23718	-0.41171
9	0.25425	-0.37865
10	0.26045	-0.42583
11	0.59865	0.64906
12	0.72740	0.87182
13	0.59898	0.82329
14	-0.89070	0.45258
15	-0.94234	0.40362
16	-0.89696	0.49436
17	-0.95944	0.41770

将二维表示的点用平面上的点标出,就是图 7-8.

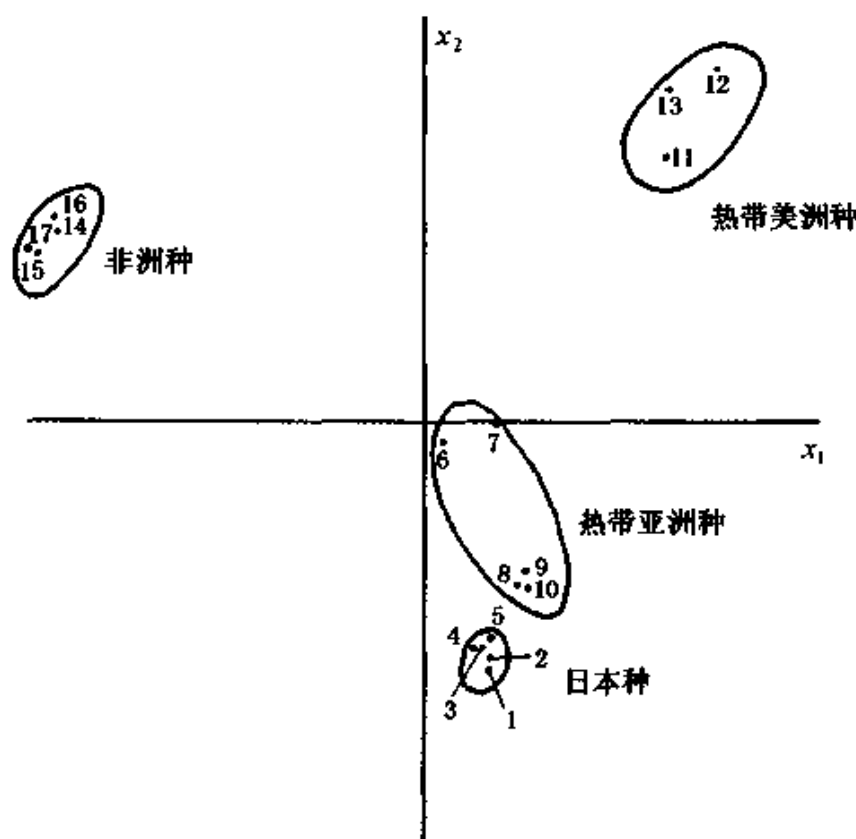


图 7-8 17 个水稻品种的二维标度图

从图 7-8 可以看出,这种表示分类很清楚.

第七节 综合评价方法优选探讨

面对一个实际的综合评价问题,如何从众多的评价方法中选出一个合适的方法以得到客观合理的评价结果?这仍是综合评价理论中未完全解决的问题.这里提出“三位一体”的思想与其他的選擇方法相结合会对这一问题提供一个较为满意的结果.

一、“三位一体”的基本思想

“三位”即指:评价目的,被评价事物,评价方法,其归宿为合适的评价方法.

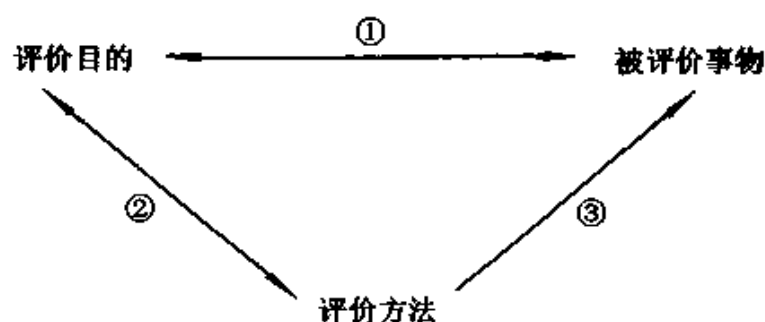


图 7-9

如图 7-9 所示,有三层含义:

1. 评价目的和被评价事物都是评价者本身(或管理者)所确定的,二者之间具有一致性.比如,某主管部门要对其下属企业进行经济效益综合评价,评价目的是了解下属企业的经营情况,为管理提供依据,同时奖励先进,鞭策落后,使以后的工作再上一个新台阶,被评价事物为下属企业的经济效益.

2. 合适的评价方法一方面由评价者的目的所决定,或者说,评价方法的选取要与评价者的目的相一致,能充分体现评价者本身的目的和意愿.比如,在上面的下属企业经济效益评价中,如果管理者认为先进的企业应该是在各方面均衡发展,即各方面都位居前列,不均衡发展的企业不应该是先进的企业,要体现管理者这样的思想就应在综合评价的最后一个环节上选取乘法来进行综合,因为乘法综合才能体现出事物发展的非均衡性,而加法综合则正好掩盖了这种非均衡性.采用乘法综合所得的综合评价价值高的企业才是体现管理者意愿的先进企业.又比如高等学校入学考试的目的是从众多的考生中为高等学校选拔优秀的人才,为了能更好地满足这一目的,可以采用一定的方法拉大分数的间距以便于录取.

3. 合适的评价方法另一方面由被评价物本身的特点所决定,或者说,评价方法的特点要与被评事物的特点相吻合.比如,上面所提的企业经济效益评价,一般而言反映经济效益的各项指标均为客观的定量指标,因而可以选择适合于确定性指标的一些方法,

比如常规方法,主成分方法等;如果所选指标较多或者指标间有较强的相关性,则可以考虑用主成分的评价方法或因子分析的方法,如果被评企业仅一个,则我们可以选一些有训练的样本作为参考系运用判别分析的方法,或者选一些无训练样本作为参考系采用聚类分析的方法.又比如,我们对同类产品作综合评价,描述产品的各项指标,诸如产品质量、产品性能、产品外观、产品价格等多为主观或定性的指标,因而可以选用模糊综合评价方法或 AHP 方法.再比如,我们要对多个投资方案作出评价,而反映投资方案的各项指标诸如净现值、收益率、投资回收期等均是随机指标,如果我们能够得到这些随机指标(变量)的估计值及其方差,则可以考虑用随机指标的综合评价方法,同时考虑指标值和方差,否则我们可借用确定性指标评价方法仅对指标估计值进行评价.

综上所述,“三位一体”的基本思想是“三位”之间相互一致,评价方法的选取主要取决于评价者本身的目的(意愿)和被评价事物的特点.

二、方法的优选

针对某一评价的实际问题,符合“三位一体”思想的方法可能还不止一种,这种情况下我们可以考虑如下的两种思路:

1. 组合评价法

组合评价法即是对同样的被评事物由不同方法所得的不同评价结果再次进行综合(组合),得到一个最终的综合评价结果,作为管理和决策的依据,具体方法可参阅文献[41].

2. 等级相关系数法

对于不同评价方法所得评价排序结果计算两两之间的等级相关系数,如果某方法的结果与其他方法结果之间的等级相关系数都较大,则认为这一方法最优,也就是以该方法评价结果作为最终的评价结果.具体可参阅文献[40].

显然“三位一体”思想与优选法的结合会给我们在实际评价中提供一个清晰的思路.

参 考 文 献

- [1] 邱东.多指标综合评价的系统分析.中国统计出版社,1991
- [2] 张崇甫等.统计分析方法及其应用.重庆大学出版社,1995
- [3] 张尧庭、方开泰.多元统计分析引论.科学出版社,1982
- [4] 张尧庭等.定性资料的统计分析.广西师范大学出版社,1991
- [5] 周光亚、夏立显.非定量数据分析及其应用.科学出版社,1993
- [6] 盛昭瀚等.DEA理论、方法与应用.科学出版社,1996
- [7] 魏权龄.评价相对有效性的DEA方法.中国人民大学出版社,1987
- [8] 萧筱南.实用模糊学.亚洲出版社,1993
- [9] 刘新平等.标准分数及其应用.西北工业大学出版社,1997
- [10] 王学仁等.实用多元统计分析.上海科技出版社,1984
- [11] 孙文爽等.多元统计分析.高等教育出版社,1994
- [12] 王国梁等.多变量经济数据统计分析.陕西科学技术出版社,1993
- [13] 刘星.外国直接投资的系统评价方法及应用.数量经济技术经济研究,7,1994
- [14] 刘贤龙.我国普通高等教育发展水平的统计分析.数理统计与管理,4,1998
- [15] 范金等.上证30指数二届成分股的聚类分析.数理统计与管理,9,1998
- [16] 陈述云.综合评价中指标的客观赋权方法.上海统计,6,1995
- [17] 王硕平.用数学方法选取社会经济指标.统计研究,6,1986
- [18] 张尧庭等.几种选取部分代表性指标的统计方法.统计研究,1,1990
- [19] 陈述云.多指标综合评价方法新探.统计与预测,3,1993
- [20] 孟生旺.多指标综合评价中权数的选择.统计研究,2,1993
- [21] 余芳东.国外综合国力研究方法的评价.统计研究,6,1993
- [22] 国涓等.经济效益综合评价的相对比较算法.财经问题研究,7,1997
- [23] 何晓亚.企业经济效益综合评价方法及应用.数理统计与管理,4,1997
- [24] 王应明.多指标决策与评价的新方法——投影法.统计与决策,4,1998
- [25] 陈述云.多指标综合评价的多元分析方法.青海统计,1,1992
- [26] 刘玉秀等.TOPSIS法用于医院工作质量的多指标综合评价.中国卫生统计,1993
- [27] 谭学瑞等.灰色关联分析:多因素统计分析新方法.统计研究,3,1995
- [28] 王明友等.关联序分析在小麦品种(系)综合评价中的应用.数理统计与管理,3,1997
- [29] 卢刚.利用DEA方法综合评价棉纺织企业经济效益.统计与管理,3,1991

- [30] 史明丽等. 模糊综合评价法在社区人群生存质量评价中的应用. 中国卫生统计, 1997
- [31] 彭天好. 机械产品设计方案的模糊综合评价. 模糊系统与数学, 2, 1997
- [32] 王云亮等. 工业企业经济效益综合评价的一个有效方法: 模糊综合评判法. 上海统计, 8, 1996
- [33] 韩正忠等. 运用隶属函数对毕业生水平的模糊综合评价. 运筹与管理, 3, 1997
- [34] 杨曾武等. 工业产品竞争能力的定量描述. 统计与管理, 2, 1991
- [35] 徐培德. 导弹效能分析的模糊综合评判模型. 模糊系统与数学, 1, 1997
- [36] 于恒兰. 综合评价的多元分析方法. 安徽大学学报(哲社版), 3, 1993
- [37] 吴国富等. 综合动态指数及复杂系统的综合评价. 数量统计与管理, 3, 1996
- [38] 陶凤梅等. e_{ij} 型数量化方法在小生素质综合评价中的应用. 数理统计与管理, 3, 1998
- [39] 张道宏等. 我国城市第三产业发展水平的综合评判. 当代经济科学, 1, 1998
- [40] 陈述云等. 多指标综合评价方法及其优化选择研究. 数理统计与管理, 5, 1994
- [41] 郭显光. 一种新的综合评价方法——组合评价方法. 统计研究, 5, 1999
- [42] 胡永宏等. 企业社会效益的统计分析与综合评价. 当代经济科学, 4, 1998



C0504356

名 词 索 引

三画

- 广义方差(2.2)
- 广义条件方差极小(2.2)

四画

- 方差(2.2)
- 无量纲化方法(2.3)
- 中间距离法(4.1)
- 贝叶斯(Bayes)判别(4.5)

五画

- 功效系数法(2.1)
- 主成分分析(3.1)
- 可变类平均法(4.1)

六画

- 权(2.4)
- 协方差(2.2)
- 因子分析(3.4)
- 因子贡献率(3.4)
- 灰色关联度(5.2)
- 多因素模糊综合评价(6.4)
- 多级模糊综合评价(6.5)
- 关联信息量(7.1)
- 托格森(Torgerson)方法(7.2)

七画

- 极大方差旋转(3.4)
- 极小极大距离法(4.2)
- 极大不相关(2.2)
- 均值(2.2)
- 克拉斯卡尔(Kruskal)方法(7.5)

八画

- 隶属度(6.1)
- 经验选点(4.2)

林知己夫方法(7.2,7.6)

九画

- 相邻指标比较法(2.4)
- 重心法(4.1)
- 界值确定法(4.2)
- 费歇(Fisher)判别法(4.6)
- 类平均法(4.2)

十一画

- 综合国力(2.1)
- 综合经济效益指数(2.1)
- 斜旋转(3.4)
- 随机选点(4.2)
- 距离公理(4.1)
- 距离综合评价(5.1)

十二画

- 最短距离法(4.1)
- 最长距离法(4.1)
- 最优预测系数(7.1)
- 谢帕尔德(Shepard)方法(7.4)

十四画

- 模糊单因素评价(6.2)

十五画

- 德尔菲(Delphi)方法(2.4)
- 熵(7.1)
- C^2R 模型(5.3)
- C^2GS^2 模型(5.3)
- DEA有效性(5.3)
- DEA弱有效性(5.3)
- K-L方法(7.3)
- Pearson的 χ^2 (7.1)

