基于 R 型聚类-因子分析的指标体系简化方法①

陆　可[1], 邹启鸣[1], 李　鸣[1], 吴金南[2]

[1](安徽工业大学 管理科学与工程学院, 马鞍山 243032)
[2](安徽工业大学 商学院, 马鞍山 243032)

摘　要: 评价指标体系过于复杂, 容易产生冗余信息, 增加计算与分析的难度. 本文针对评价指标体系的简化方法展开研究, 提出了基于 R 型聚类-因子分析的代表元提取方法. 该方法具有定量控制代表元的信息丢失率、代表元实际含义易于解释等优点, 同时避免了因子分析难以处理独立指标的问题. 实验结果表明本文所提出的R型聚类–因子分析指标体系简化方法可以在没有先验知识可用的情况下, 有效提高指标体系的简明性与有效性, 同时保持指标信息的完备性.

关键词: 指标体系简化; 代表元; 因子分析; R 型聚类

Simplification Method of Index System Based on R Cluster Analysis and Factor Analysis

LU Ke[1], ZOU Qi-Ming[1], LI Ming[1], WU Jin-Nan[2]

[1](School of Management Science and Engineering, Anhui University of Technology, Maanshan 243032, China)
[2](Business College, Anhui University of Technology, Maanshan 243032, China)

Abstract: An excessively complicated index system may increase the complexity of computation. This work proposes a simplification method of index system based on R cluster analysis and factor analysis to extract the representative elements. This method can control the information loss rate of the representation element and explain the actual meaning of the representation element easily. It also avoids the weakness of factor analysis which is difficult to handle the problem with independent factor. The experiment results show that this method can improve the simplicity and effectiveness of index system significantly, and maintain the completeness of the index information without a prior knowledge is available.

Key words: simplification method of index system; representation element; factor analysis; R cluster analysis

1　引言

评价指标体系, 是指由表征评价对象各方面特性及其相互联系的多个指标所构成的、具有内在结构的有机整体[1]. 评价指标体系要求具备完备性、简明性、有效性[2]. 但是当我们在做探索性研究时, 由于不了解研究对象的专业背景知识, 为求完备性, 所建立的指标体系往往存在无关指标过多, 以及指标之间关联性过高等问题, 导致指标体系缺乏简明性与有效性.

R 型聚类作为一种无监督的学习方法, 可以在没有先验知识可用的情况下, 通过对若干个指标进行数量化相似程度计算, 把相关指标聚集成类, 然后通过提取每一个类中的代表元来反映类中原始指标的信息, 进而基于代表元建立更加简洁有效的新指标体系.

近年来有许多学者利用 R 型聚类方法简化指标体系, 但是提取代表元的具体做法各有不同. 王奎实通过调查, 专家咨询确定了 17 个人才特征指标, 进一步利用R型聚类, 将17个特征指标聚为10类, 并从每个类中随机选取一个指标作为代表元, 形成新的人才评价指标体系[3]. Tao 利用变异系数刻画指标分辨信息的能力, 指出变异系数与信息分辨能力呈正相关. 对原始指标进行 R 型聚类后, 保留每一个聚类中变异系数最大的指标, 作为对应聚类的代表元, 由代表元集合

① 基金项目:国家自然基金项目(71371013)
收稿时间:2015-08-27;收到修改稿时间:2015-10-26

构成新的指标体系[4]. 何国民利用复相关系数衡量代表元与类内原始指标的相关程度. 选择每一个聚类中复相关系数最大的指标作为代表元, 形成新的指标体系[5]. 显然, 随机选择法受主观影响较大, 不能成为一种通用的方法. 而变异系数最大法与复相关系数最大法的共同点是在聚类中选择一个原始指标作为代表元. 但是, 这种方法的鲁棒性较差, 它虽然可以在类内指标相关性很大的情况下达到较好的效果, 但是如果类内指标相关性不是很大, 就会造成信息的大量丢失.

针对传统方法信息丢失严重的问题, 本文提出一种基于因子分析的 R 型聚类代表元提取方法. 因子分析法可以使用少数因子代表原始指标的大部分信息, 利用累积方差贡献率定量控制信息丢失, 这些因子具有线性相关性不显著以及具备命名解释性的特点[6]. 本文将在第二部分简要介绍 R 型聚类的算法流程与因子分析的基本原理. 以此为基础, 在第三部分进一步叙述 R 型聚类-因子分析法的具体内容. 本文的第四部分, 在不同的指标体系下将 k-means 聚类算法应用于等多个数据集进行实验比较, 并以聚类纯净度与平均误差平方和作为聚类效果及指标体系简化方法有效性的评价指标. 实验结果表明, 本文提出的 R 型聚类-因子分析法较传统方法有更好的鲁棒性与有效性.

2 相关概念

2.1 R 型聚类

R 型聚类分析是对评价指标进行量化相关程度划分的层次聚类方法. 其原理是将每个指标初始化为独立的子类, 然后找到相关系数最小的 2 个指标聚成一类, 再依次增加聚类对象, 直到将所有的指标归到一类为止, 通过选择合适的置信水平, 将指标划分成确定数目的聚类[7].

假设原始指标集 $S=\{s_1,s_2,...s_n\}$, 一共有 n 个指标. 样本集合为 $X=\{x_1,x_2,...,x_m\}$, 一共有个 m 样本, 每个样本 n 维. 不妨设

$$x_i=\{x_{i1},x_{i2},...,x_{in}\}$$
$$x_j=\{x_{j1},x_{j2},...,x_{jn}\}$$
$$i,j=1,2,3,...,m$$

R 型聚类的主要步骤包括:

step 1. 将每个指标初始化为独立的子类, 记为 $C_i\ (i=1,2,...,m)$. 聚类之间的相关程度由类平均相关系数定量描述, 即对于两个类 C_a, C_b 的类平均相关系数 $R(C_a,C_b)$ 为

$$R_{a,b}=\frac{1}{|C_a||C_b|}\sum_{s_i\in C_1}\sum_{s_j\in C_2}p_{ij} \tag{1}$$

其中 p_{ij} 为 Pearson 相关系数.

$$p_{ij}=\frac{N\sum_{r=1}^{N}x_{ri}'x_{rj}'-\sum_{r=1}^{N}x_{ri}'\sum_{r=1}^{N}x_{rj}'}{\sqrt{N\sum_{r=1}^{N}x_{ri}'^2-(\sum_{r=1}^{N}x_{ri}')^2}\sqrt{N\sum_{r=1}^{N}x_{rj}'^2-(\sum_{r=1}^{N}x_{rj}')^2}}$$

$$s_i\in C_a,s_j\in C_b \tag{2}$$

将类平均相关系数矩阵 M_0 初始化为:

$$M_0=(R_{ij})_{n\times n}$$

step 2. 找到 M_0 中的数值最大的一项, 设其为 R_{ab}. 将 C_a,C_b 合并为一类, 更新类平均相关系数矩阵 $M_1=(R_{ij})_{(n-1)\times(n-1)}$

step 3. 重复执行步骤 2, 直到全部指标被聚集到单一聚类中

step 4. 选择合适的置信水平, 输出聚类结果

$$C'=\{C_1',C_2',...,C_l'\}$$

R 型聚类分析能客观地反映指标之间的内在关系. 进行 R 型聚类后, 指标体系的结构更加直观, 降低了我们处理数据的难度.

2.2 因子分析

因子分析是一种从指标集合中提取共性因子的数学模型[8]. 它的基本思想是通过研究指标集合的相关系数矩阵, 将原指标的数据结构分解成公共因子和特殊因子两部分, 利用少数几个公共因子去描述所有指标的信息, 从而实现数据降维. 其数学模型为:

$$s_i=a_{i1}F_1+a_{i2}F_2+...+a_{it}F_t+\varepsilon_i,i=1,2,...,n \tag{3}$$

称 $F_j(j=1,2,...,t)$ 为 $s_i(i=1,2,...,n)$ 的公共因子, ε_i 为 s_i 的特殊因子. $A=(a_{ij})_{n\times t}$ 为因子载荷矩阵, a_{ij} 反映了第 i 个指标 s_i 在第 j 个公共因子 F_j 上的相对重要性.

设有 m 个样本, 每个样本共有 n 个维度, 构成 $m\times n$ 阶的样本数据矩阵

$$X=\begin{bmatrix}x_{11} & x_{12} & ... & x_{1n}\\ x_{21} & x_{22} & ... & x_{2n}\\ ... & & & \\ x_{m1} & x_{m2} & ... & x_{mn}\end{bmatrix}$$

记 $\sum$ 是 X 的协方差矩阵, $\sum$ 的特征值及相应的

正交化特征向量分别为 $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_n \ge 0$ 及 $e_1, e_2, ..., e_n$，X 的第 i 个列向量为 s_i，对数据矩阵 X 的 n 个列向量 $s_1, s_2, ..., s_n$ 作线性组合为[9]:

$$F_i = e_{1i}s_1 + e_{2i}s_2 + ... + e_{ni}s_n, \quad i = 1, 2, ..., n \tag{4}$$

其中 $e_i = [e_{1i}\ e_{2i}\ ...\ e_{ni}]^T$ 由下列原则确定

1) $F_i, F_j (i \ne j, i, j = 1, 2, ..., n)$ 不相关，即 $Cov(Y_i, Y_j) = 0, i \ne j$.

2) 线性组合 F 的方差尽量大.

每一个公因子所提取的信息量可以用方差定量描述，且 $Var(F_i) = e_i^T \sum e_i = \lambda_i, \quad i = 1, 2..., m$，称 $\lambda_k / \sum_{i=1}^{m} \lambda_i$ 为第 k 个公因子 F_k 的方差贡献率，$\sum_{i=1}^{t} \lambda_i / \sum_{i=1}^{m} \lambda_i$ 为前 t 个公因子 $F_1, ..., F_t$ 的累积方差贡献率，累积方差贡献率越大表明相应的若干个公因子所反映的综合信息量越大[9].

因子分析法可以使用少数几个公共因子集中原始指标的大部分信息，从而实现数据降维. 它通过计算累积方差贡献率，对公因子的信息贡献率做出定量的描述. 因子载荷反映了原始指标体系与公共因子的相关性，有助于理解公共因子的实际含义. 但是使用因子分析法的前提条件是原始指标之间存在较强的相关性，存在相互独立指标的指标体系容易导致因子分析失效.

3 R型聚类–因子分析法

本文所提出的 R 型聚类–因子分析法是在 R 型聚类的基础上使用因子分析法提取聚类代表元，进而由代表元构成新指标体系. 利用相关系数对指标进行 R 聚类后，每一个聚类中的指标之间呈现较高的相关性，在这个基础上使用因子分析法可以达到更好的效果. 因此，该方法可以在无监督的情况下利用 R 型聚类初步简化指标体系结构并发挥因子分析法信息丢失可控，公共因子命名清晰度高的优势，同时避免独立指标导致因子分析法失效的问题.

R 型聚类–因子分析法的具体步骤如下:

Step1. 对指标数据进行标准化处理，并剔除变异系数小的指标

Step2. 对指标进行 R 型聚类，输出指标聚类集合

Step3. 对每个聚类中的指标进行因子分析，通过将公共因子的累积信息贡献率与预先设定的截断系数比较，获得指标聚类的代表元

Step4. 由代表元集合构成新指标体系

下面几个小节将具体介绍以上各个步骤.

3.1 数据预处理

3.1.1 指标数据标准化

为了消除指标的量纲与自身水平对定量分析结果的影响，通过式(4)、式(5)、式(6)对原始数据进行标准化处理:

$$x_{ij}^{'} = \frac{x_{ij} - \overline{x_j}}{s_j} \tag{5}$$

其中，$\overline{x_j}$ 是 x_j 的均值，

$$x_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^{'}$$

s_j 是 x_j 的标准差，

$$s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}^{'} - \overline{x_j}\right)^2}$$

3.1.2 剔除变异系数小的指标

变异系数是客观反映数据离散程度的指标，可以用来刻画指标的信息分辨能力，一般认为指标的变异系数越大，信息分辨能力就越强，反之，则越弱[10]. 变异系数较小的指标对聚类结果的影响程度也较小，通过剔除这些指标可以初步实现指标体系简化. 第 j 个指标的变异系数为:

$$v_j = \frac{s_j}{\overline{x_j}} \tag{6}$$

3.2 R 型聚类过程

本文采用平均相关系数法对经过预处理的指标集 $S = \{s_1, s_2, ...s_n\}$ 进行R型聚类，聚类结束后通过选择合适的置信水平确定聚类数目 l，输出聚类结果 $C^{'} = \{C_1^{'}, C_2^{'}, ..., C_l^{'}\}$.

目前，对于聚类数目 l 的确定还没有一种普遍适用的方法. 为避免聚类数目的确定受主观随意性影响，赵宇哲等人提出对聚类后的指标进行非参数 K-W 检验来判断聚类数目 l 的合理性[11]. 非参数 K-W 检验的原假设是不同的指标无显著差异. 若所有指标的显著性水平 Sig > 0.05，则接受原始假设，即同类指标之间无显著差异，聚类数目合理；反之，则应重新确定聚类数目.

3.3 利用因子分析提取代表元

对 R 型聚类后的每一个聚类使用因子分析获得一系列公共因子, 通过将累积信息贡献率与截断系数比较, 选择累积信息贡献率最大的一个或多个公共因子作为每一个聚类的代表元, 并由代表元构成新指标体系.

3.3.1 求类内原始指标的代表元

设类$C_g^{'}$中有d个指标, $\sum$是其协方差矩阵, $\sum$的特征值及相应的正交化特征向量分别为$\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_d \ge 0$及$e_1, e_2, ..., e_d$则类$C_g^{'}$的第$i$个公共因子为

$$F_i = e_i^T C_g^{'} = e_{1i}s_1 + e_{2i}s_2 + ... + e_{di}s_d, \quad s_i \in C_g^{'}, i = 1, 2, ..., d \tag{7}$$

3.3.2 选取前t个公共因子作为代表元

t为使式(8)成立的最小正整数

$$\sum_{i=1}^{t}\lambda_i / \sum_{i=1}^{d}\lambda_i \ge \alpha \tag{8}$$

式中α($0 < \alpha < 1$)为预先设定的截断系数. 只有当公共因子的累积信息贡献率达到α时, 这些公共因子才能够成为代表元. α的设定保证了信息丢失被控制在了一个可以接受的范围内. 由于指标经 R 型聚类后, 类内指标具有较强的相关性, 所以一般t的取值为 1.

确定了公因子数目之后, 我们可以得到类内各个指标的因子模型:

$$s_i = a_{i1}F_1 + a_{i2}F_2 + ... + a_{it}F_t + \varepsilon_i, \quad i = 1, 2, ..., d \tag{9}$$

因子模型的因子载荷有助于我们解释公因子的实际意义, 有利于研究的深入进行.

使用该方法可以提取出l个聚类的w个代表元$F = \{F_1, F_2, ..., F_w\}$, 其中$w \ge l$, 可将其定义为新指标集合.

4 算法复杂度分析与算法表现

设聚类C_g中含有d个指标, 并且数据集含有N个样本.

复相关系数最大法主要涉及的操作包括计算聚类C_g中的指标对于聚类中其他指标的复相关系数, 以及对聚类内所有指标的复相关系数进行排序. 该算法需要构造$X_1, X_2, .., X_{d-1}$的线性组合, 并通过计算该线性组合与y之间的简单相关系数作为变量y与$X_1, X_2, .., X_{d-1}$之间的复相关系数. 矩阵求逆是构造线性组合的主要步骤, 其复杂度为$O(Nd^3)$. 由于需要计算聚类中每一个指标的复相关系数, 并且计算简单相关系数与对指标排序的复杂度均低于矩阵求逆, 所以复相关系数最大法的算法复杂度为$O(Nd^4)$

变异系数最大法需要计算聚类C_g中所有指标的变异系数并进行排序. 计算所有指标变异系数的复杂度为$O(Nd)$. 假设使用冒泡排序算法对指标的变异系数进行降序排序, 其算法复杂度为$O(d^2)$. 由于一般情况下$N \bullet d$, 所以变异系数最大法的算法复杂度可以认为是$O(Nd)$.

因子分析法的主要操作是对协方差矩阵的求解, 所以其算法复杂度为$O(Nd^2)$.

根据以上分析, 可得到各算法的复杂度如表 1 所列. 由表 1 可以发现因子分析法的算法复杂度介于复相关系数最大法与变异系数最大法之间.

表 1 算法复杂度情况表

算法	复杂度
复相关系数最大法	$O(Nd^4)$
变异系数最大法	$O(Nd)$
因子分析法	$O(Nd^2)$

为了进一步分析所提出算法的性能, 将因子分析法与变异系数最大法, 复相关系数最大法分别进行实验比较. 实验中用到的数据集包括: *Iris* 数据集, *Wine* 数据集, *Ionoshere* 数据集, *segmentation* 数据集, *Musk* 数据集. 以上数据集的基本情况如表 2.

表 2 数据集基本情况表

数据集	实例数	属性数	类别数
Iris	150	4	3
Wine	178	13	3
Ionoshere	351	32	2
Segmentation	2100	19	7
Musk	6598	168	2

通过利用不同的简化方法得到原始指标体系的新指标体系. 在各指标体系上将$k-means$聚类算法应用于不同数据集. 通过比较聚类效果的好坏评价新指标体系的有效性. 聚类效果的评价标准可以分为外部度量标准与内部度量标准[12].

外部度量标准针对的是数据集聚类个数与样本正确分类为已知的情况. 本文通过聚类纯净度(pur)进行衡量. 对于聚类C_k, 其纯净度为根据已知的正确类

标识, 计算得到的正确分类样本数占整个聚类的比例, 即

$$pur(C_k)=\frac{NT_k}{N_k} \tag{10}$$

其中 N_k 为聚类 C_k 的样本数, NT_k 为聚类 C_k 正确聚类的样本数.

整个聚类的纯净度 $pur(C)$ 是所有聚类纯净度的均值, 纯净度越高则聚类效果越好.

内部度量标准处理的数据集结构未知, 利用数据的特征与量值评价聚类效果, 本文利用误差平方和(V)进行度量. 误差平方和是反映聚类结果紧密度的指标, 其值越小则聚类越紧密, 聚类效果越好. 由于各指标体系的指标数目不同, 因此求误差平方和的平均值, 其定义为:

$$V=(\sum_{C_k\in C}\sum_{i\in C_k}f(i,u_k))/d \tag{11}$$

其中 u_k 是聚类 C_k 的聚类中心, $f(i,u_k)$ 是样本 i 与其所在聚类的聚类中心之间的距离函数, d 为对应指标体系的指标总数.

将截断系数 α 取作 0.75, 对每一个指标体系在数据集上重复实验 500 次取评价指标的平均值, 实验结果如表 3, 表 4 所示.

表 3　聚类纯净度比较

数据集	原始指标体系	变异系数最大法	复相关系数最大法	因子分析法
Iris	0.8792	0.8187	0.6607	0.8016
Wine	0.9415	0.9209	0.9267	0.9548
Ionoshere	0.6096	0.5686	0.5629	0.5976
Segmentation	0.5407	0.6303	0.6198	0.6489
Musk	0.5378	0.5189	0.5455	0.5693

表 4　平均误差平方和比较

数据集	原始指标体系	变异系数最大法	复相关系数最大法	因子分析法
Iris	1.7121	2.065	2.1781	2.084
Wine	3.4850	3.5124	3.4418	3.493
Ionoshere	16.013	18.446	18.971	16.996
Segmentation	23.267	22.066	21.367	19.764
Musk	217.759	219.546	215.545	212.375

本文方法相比于传统方法所得到的指标体系在不同的数据集上均能达到较好的聚类效果, 如在 *Musk* 数据集上, 本文所采用方法的聚类纯度比变异系数最大法与复相关系数最大法分别高出 5.04%, 2.38%, 其平均误差平方和比后两者分别减少了 7.171, 3.195. 尤其在某些数据集上, 因子分析法的聚类效果甚至高于利用原始指标体系得到的聚类结果, 如 *segmentation* 数据集. 证明了本文方法相比于传统方法具备更好的鲁棒性, 所得到的新指标体系具有更好的代表性与有效性. 这是因为变异系数最大法与复相关系数最大法的基本思想是通过选取指标聚类中某个具有代表性的指标作为聚类中心, 这种方法只能够提取出部分指标的信息, 容易造成信息的大量丢失. 而利用因子分析法提取指标聚类的聚类中心既能够综合利用各指标的信息, 又不过于依赖某个指标, 有效地突出了数据的主要特征, 去除了次要特征的干扰, 提高了指标体系的科学性与可靠性.

5　结束语

本文针对利用 R 型聚类对指标体系简化的方法进行了研究, 指出传统的 R 型聚类代表元提取方法存在信息丢失严重的问题, 并提出基于因子分析的 R 型聚类代表元提取方法. 本文方法不仅充分发挥了因子分析法信息丢失程度可控, 公共因子实际含义易于解释的优势, 而且避免了它难以处理独立指标的缺点. 最后通过实验结果比较发现改进的方法能够在提高指标体系的简明性的同时, 减少信息丢失. 但是, 这种方法相比于变异系数最大法加大了算法的复杂度, 对于规模较大的数据集存在局限性. 因而如何减少该方法的算法复杂度有待进一步研究.

参考文献

1 姜春林,孙军卫,田文霞.人文社会科学成果评价若干指标内涵及其关系.情报杂志,2013,32(11):43−50.

2 Cole DC, Eyles J, Gibson BL. Indicators of human health in ecosystems: What do we measure? The Science of the Total Environment, 1998, (224): 201−213.

3 王奎实,马兰芝,胡文科.进入人才市场的人才特征评估——R 型聚类法应用试验.科学学究,1997,15(3):60−64.

4 Tao LY, Li ZD, Zhang M. The establishment of production capacity evaluation indicator system based on R-cluster and coefficient of variation. Advanced Design and Manufacturing Technology III. Switzerland. 2013: 2565−2569.

5 何国民,马良宏.一种新的 R 型聚类分析方法——复相关 R 型聚类分析法.武汉体育学院学报,2002,36(3):144−146.

6 韩宝燕.因子分析的数学模型及其发展评价.科技信息,2013,(11):47.

7 王寿超,李杰,王菊,徐志璐,房春生.R 型聚类与模糊聚类分析在源解析中的应用.安徽农业科学,2011,39(29):17757−17759,17761.

8 Sun P. The application of factor analysis in the study on cultural industry competitiveness evaluation index system. Materials Science, Computer and Information, 2014: 5132−5135.

9 孙刘平,钱吴永.基于主成分分析法的综合评价方法的改进.数学的实践与认识,2009,39(18):15−20.

10 Wang ZB, Qiu BZ. Fuzzy c-means clustering algorithm based on coefficient of variation. Advances in Applied Sciences, Engineering and Technology II, Switzerland. 2014, 998-999: 873−877.

11 赵宇哲,刘芳.生态港口评价指标体系的构建—基于 R 聚类、变异系数与专家经验的分析.科研管理,2015,36(2):124−132.

12 Liu YC. Understanding and enhancement of internal clustering validation measures. Institute of Electrical and Electronics Engineers Inc, 2013, 43(3): 982−994.