文章编号:1672 - 058X(2013)05 - 0017 - 05

基于 PCA 的聚类分析在汶川地震灾情分类中的应用*

陈 丽[a],张朝元[b]
(大理学院 a. 工程学院;b. 数学与计算机学院,云南 大理 671003)

摘 要:利用主成分分析(PCA)运用 SPSS 软件对汶川地震 36 个严重受灾县市的 8 个灾情指标进行了综合分析,得到了累积贡献率为 83.403% 的 3 个主成分及其得分;然后,基于 3 个主成分的得分采用聚类分析对汶川地震 36 个严重受灾县市进行了分类;得到了全面、合理和科学的分类结果。

关键词:汶川地震;灾情分类;主成分分析;聚类分析;SPSS 软件
中图分类号:P65;O212.4 **文献标志码**:A

0 引 言

发生在 2008 年的汶川大地震给我国四川人民造成了巨大的痛苦。由于此次地震震级高、涉及面广,造成四川省多个县市均遭到了不同程度的严重破坏。正确评估各个县市的受损情况对于灾区的救援和重建有着重要的意义。根据研究,影响一个灾区灾情有诸多指标,如总人口、面积加权平均烈度、死亡和失踪人数、万人死亡和失踪率、倒塌房屋间数、万人倒塌房屋率、地质灾害危险度和万转移安率等[1]。要综合考虑以上指标的影响,必须采用多指标的综合分析方法。常见多指标的综合分析方法有模糊综合评价法、专家评分法、层次分析法和灰色聚类法等多种方法,但这些方法必须对指标进行筛选,存在筛选时的主观性和信息丢失等问题,会影响评价结果。如果采用一种可以从大量指标中挑选出几个具有代表性的主要指标的方法,则减少了信息的丢失,并可以减少分类模型的复杂性。

主成分分析和聚类分析是多元统计分析中两种重要的方法。主成分分析通过对大量指标进行综合分析得到少数的综合指标即主成分,从而大大简化了数据结构,具有较强综合信息和解释实际意义的能力,使得评价结果更具科学性、客观性和公正性[1]。目前,主成分分析在高校学生质量评价和高校教师教学质量评价等方面得到了广泛的应用。聚类分析[2]指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过程。聚类分析的目标就是在相似的基础上收集数据来分类。聚类是将数据分类到不同类或者簇的一个过程,所以同一个簇中的对象有很大的相似性,而不同簇间的对象有很大的相异性。从统计学的观点看,聚类分析是通过数据建模简化数据的一种方法。

本文结合这两种方法建立了基于主成分分析的聚类分析的汶川地震灾情分类机制。在简化指标的基础上将灾情类似的县市进行分类,有助于灾区的合理援助和重建。

收稿日期:2013 - 02 - 10;**修回日期**:2013 - 04 - 10.
* **基金项目**:云南省教育厅科学研究基金项目(2010C140).
作者简介:陈丽(1980-),女,汉族,湖北武汉人,讲师,从事物理研究和数据处理研究。

1 基于PCA的聚类分析

分别用 $x_1,x_2,x_3,\cdots,x_p$ 表示影响某次地震灾情的的 p 个指标,则某次地震 n 个县市灾情指标的数据矩阵:$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(x_{ij})_{n\times p}$,其中第 i 个县市的灾情数据为 $X_i=\begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}'(i=1,2,\cdots,n)$。

主成分分析的目的是要将 p 个影响指标 $x_1,x_2,x_3,\cdots,x_p$ 综合分析得到 $m(m<p)$ 个主成分 $z_1,z_2,\cdots,z_m$。其实质就是用尽可能少的主成分反映出尽可能多的原始信息,为下一步综合分类做好准备。

主成分的变量 $z_1,z_2,\cdots,z_m$ 与原变量影响指标 $x_1,x_2,x_3,\cdots,x_p$ 的关系通过式(1)式建立:

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \cdots\cdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases} \tag{1}$$

其中系数 l_{ij} 的确定原则:z_i 与 $z_j(i\neq j;i,j=1,2,\cdots,m)$ 相互无关;z_1 是 $x_1,x_2,x_3,\cdots,x_p$ 的一切线性组合中方差最大的,称为原变量指标 $x_1,x_2,x_3,\cdots,x_p$ 的第1主成分;z_2 是与 z_1 不相关的 $x_1,x_2,x_3,\cdots,x_p$ 所有线性组合中方差最大的,称为原变量指标 $x_1,x_2,x_3,\cdots,x_p$ 的第2主成分;$\cdots$;z_m 是与 $z_1,z_2,\cdots,z_{m-1}$ 都不相关的 $x_1,x_2,\cdots x_p$ 的所有线性组合中方差最大的,称为原变量指标 $x_1,x_2,x_3,\cdots,x_p$ 的第 m 主成分。

基于主成分分析的聚类分析的操作步骤[2,3]如下:

(1)将数据标准化。原始数据通常有不同的量纲和不同的数量级。在保持原始输入变量数据的全部信息的情况下,消除影响因素量纲的不统一和数量级的差别,必须对这些原始数据进行标准化,使每一个指标都能在分析时发挥平等的作用。具体的标准化处理公式[4]:$x_{ij}^{*}=x_{ij}/\bar{X}_j,(i=1,2,\cdots,n;\ j=1,2,\cdots,p)$,其中 $\bar{X}_j=(\sum_{i=1}^{n}x_{ij})/n\ (j=1,2,\cdots,p)$。则新的输入变量 $x_1^{*},x_2^{*},x_3^{*},\cdots,x_p^{*}$ 组成的数据矩阵为:$X^{*}=(x_{ij}^{*})_{n\times p}$。

(2)求相关系数矩阵。标准化后变量 x_i^{*} 与 x_j^{*} 的相关系数矩阵:$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$,其中 $r_{ij}=\sum_{k=1}^{n}(x_{ki}^{*}-\bar{x}_i^{*})(x_{kj}^{*}-\bar{x}_j^{*})/\sqrt{\sum_{k=1}^{n}(x_{ki}^{*}-\bar{x}_i^{*})^2\sum_{k=1}^{n}(x_{kj}^{*}-\bar{x}_j^{*})^2}(i,j=1,2,\cdots,p)$ 为相关系数。

(3)求系数矩阵 R 的特征值和特征向量。R 的特征值由特征方程 $|R-\lambda I|=0$ 求出为 $\lambda_1>\lambda_2>\cdots>\lambda_p\geqslant 0$;然后分别求出对应于特征值 λ_i 的特征向量 $T_i(i=1,2,\cdots,p)$,其中 $T_i=(t_{i1},t_{i2},\cdots,t_{ip})'$。

(4)计算主成分贡献率、累积贡献率并确定主成分个数。总方差中属于第 i 个主成分 z_i 的贡献率:$w_i=\lambda_i/\sum_{k=1}^{p}\lambda_i(i=1,2,\cdots,m)$;前 i 个主成分的贡献率之和称为 $w_1,w_2,\cdots,w_i$ 的累积贡献率:$W=\sum_{k=1}^{i}\lambda_i/\sum_{k=1}^{p}\lambda_i$ $(i=1,2,\cdots,m)$。一般取累积贡献率达80% ~ 95%的特征值 $\lambda_1,\lambda_2,\cdots,\lambda_m$ 所对应 $z_1,z_2,\cdots,z_m$ 为第1主成分,第2主成分,$\cdots$,第 $m(m\leqslant p)$ 个主成分。

(5) 计算主成分的得分。主成分 $z_1, z_2, \cdots, z_m$ 的得分矩阵:$z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix}$ 由

$$\begin{cases} z_1 = t_{11}x_1^* + t_{12}x_2^* + \cdots + t_{1p}x_p^* \\ z_2 = t_{21}x_1^* + t_{22}x_2^* + \cdots + t_{2p}x_p^* \\ \cdots\cdots \\ z_m = t_{m1}x_1^* + t_{m2}x_2^* + \cdots + t_{mp}x_p^* \end{cases}$$ 计算。

(6) 由聚类分析进行综合分类。依据累积贡献率达到 80% 及以上的 m 个主成分的得分,采用样品系统聚类分析对某次地震 n 个县市进行聚类,得到合理的聚类结果,并按照灾情指标和灾情的严重性进行综合分析。

2　实例分析

本文从参考文献[5]中收集到了 2008 年汶川地震时四川 36 个严重受灾县市的主要数据,其中影响灾情的指标包括总人口(x_1)、面积加权平均烈度(x_2)、死亡和失踪人数(x_3)、万人死亡和失踪率(x_4)、倒塌房屋间数(x_5)、万人倒塌房屋率(x_6)、地质灾害危险度(x_7)和万转移安率(x_8)。

按照以下三步对汶川地震 36 县市按照灾情指标进行分类:

第 1 步,通过 SPSS 软件利用主成分分析对收集到的灾情数据进行分析得到了表 1 中的特征根、方差贡献率和累积方差贡献率和表 2 中的特征向量矩阵即主成分系数。

表 1　特征根、方差贡献率和累积方差贡献率

成分	初始特征值			成分	初始特征值		
	合计	方差的/%	累积/%		合计	方差的/%	累积/%
z_1	4.279	53.482	53.482	z_5	0.372	4.644	94.788
z_2	1.439	17.989	71.471	z_6	0.269	3.368	98.156
z_3	0.955	11.933	83.403	z_7	0.121	1.514	99.670
z_4	0.539	6.740	90.144	z_8	0.026	0.330	100.000

表 2　特征向量矩阵(即主成分的系数矩阵)

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
x_1^*	−0.172	0.675	0.160	0.447	0.398	0.280	0.213	−0.081
x_2^*	0.407	0.303	−0.036	−0.174	−0.480	−0.006	0.690	0.068
x_3^*	0.420	−0.023	0.474	0.069	−0.128	0.071	−0.213	−0.732
x_4^*	0.370	−0.167	0.595	0.188	0.074	0.127	−0.078	0.651
x_5^*	0.314	0.538	−0.267	0.119	−0.234	−0.303	−0.589	0.186
x_6^*	0.376	−0.244	−0.237	0.440	0.402	−0.553	0.270	−0.093
x_7^*	0.365	0.170	−0.074	−0.673	0.608	0.093	−0.040	0.006
x_8^*	0.345	−0.217	−0.513	0.266	−0.026	0.704	−0.080	0.000

第 2 步,可以由表 1 确定主成分个数为 $m=3$ 时,累积贡献率为 83.403%,达到 80% 以上。由表 2 可知 3 个主成分 z_1, z_2, z_3 的表达式:

$$z_1 = -0.172x_1^* + 0.407x_2^* + 0.42x_3^* + 0.37x_4^* + 0.314x_5^* + 0.376x_6^* + 0.365x_7^* + 0.345x_8^*$$

$$z_2 = 0.675x_1^* + 0.303x_2^* - 0.023x_3^* - 0.167x_4^* + 0.538x_5^* - 0.244x_6^* + 0.17x_7^* - 0.217x_8^*$$

$$z_3 = 0.16x_1^* - 0.036x_2^* + 0.474x_3^* + 0.595x_4^* - 0.267x_5^* - 0.237x_6^* - 0.074x_7^* - 0.513x_8^*$$

通过上面3个主成分的表达式计算得到表3中36个县市灾情的3个主成分z_1, z_2, z_3得分。

表3 3个主成分的得分

序号	县(区、市)	z_1	z_2	z_3	序号	县(区、市)	z_1	z_2	z_3
1	汶川县	14.433	-2.430	11.579	19	小金县	1.447	-0.059	-1.131
2	北川县	9.849	-0.873	7.835	20	城区	0.648	0.645	-0.392
3	绵竹市	6.466	2.466	0.783	21	罗江县	1.018	0.518	-0.745
4	什邡市	5.535	2.298	-0.351	22	黑水县	1.509	0.008	-1.049
5	青川县	4.216	0.790	-0.183	23	崇州市	0.778	1.574	-0.105
6	茂县	5.242	0.029	0.320	24	剑阁县	0.980	1.671	-0.341
7	安县	3.350	1.784	-0.775	25	三台县	0.582	2.803	-0.318
8	都江堰	3.091	2.051	-0.397	26	阆中市	0.499	1.438	-0.314
9	平武县	4.177	0.095	1.388	27	盐亭县	0.813	1.348	-0.518
10	彭州市	2.628	2.560	-0.830	28	松潘县	1.036	0.048	-0.794
11	理县	4.269	-0.963	-2.733	29	苍溪县	0.525	1.498	-0.221
12	江油市	1.881	2.718	-1.047	30	芦山县	0.696	0.485	-0.225
13	利州区	1.799	1.281	-1.279	31	中江县	0.336	2.843	0.025
14	朝天区	1.523	0.522	-1.088	32	元坝区	0.729	0.698	-0.319
15	旺苍县	2.007	1.589	-0.699	33	大邑县	0.464	1.176	0.004
16	梓潼县	1.054	0.762	-0.827	34	宝兴县	0.610	0.373	-0.150
17	游仙区	0.965	1.202	-0.625	35	南江县	0.532	1.240	-0.158
18	旌阳区	1.217	1.787	-0.675	36	广汉市	0.393	1.301	-0.001

第3步,依据3个主成分z_1, z_2, z_3的得分采用以欧几里得距离为度量和Ward法为聚类方法在SPSS软件中进行聚类分析,得到如图1所示的聚类谱系图及在表4中基于3个主成分的三类、四类和五类等分类结果。分类结果下的数字表示第几类,如四类结果下的3表示此灾区属于第三类。

表4 基于3个主成分的分类结果

序号	县(区、市)	三类	四类	五类	序号	县(区、市)	三类	四类	五类
1	汶川县	1	1	1	19	小金县	3	4	5
2	北川县	1	2	2	20	城区	3	4	5
3	绵竹市	2	3	3	21	罗江县	3	4	5
4	什邡市	2	3	3	22	黑水县	3	4	5
5	青川县	2	3	3	23	崇州市	3	4	5
6	茂县	2	3	3	24	剑阁县	3	4	5
7	安县	3	4	4	25	三台县	3	4	5
8	都江堰	3	4	4	26	阆中市	3	4	5
9	平武县	2	3	3	27	盐亭县	3	4	5
10	彭州市	3	4	4	28	松潘县	3	4	5
11	理县	2	3	3	29	苍溪县	3	4	5
12	江油市	3	4	4	30	芦山县	3	4	5
13	利州区	3	4	4	31	中江县	3	4	5
14	朝天区	3	4	5	32	元坝区	3	4	5
15	旺苍县	3	4	4	33	大邑县	3	4	5
16	梓潼县	3	4	5	34	宝兴县	3	4	5
17	游仙区	3	4	5	35	南江县	3	4	5
18	旌阳区	3	4	5	36	广汉市	3	4	5

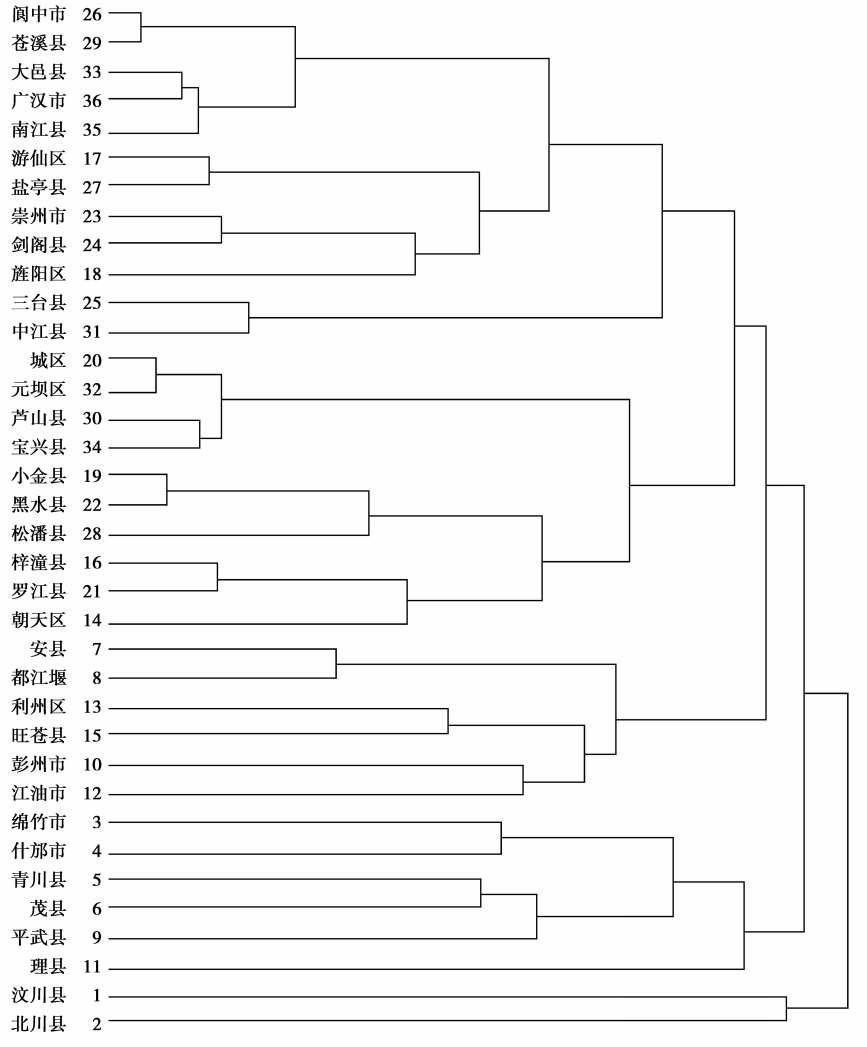

图 1　聚类谱系图

3　结　论

依据灾情指标对一个灾区进行分类属于多指标的综合分类。本文采用主成分分析法和聚类分析法对汶川地震灾情进行了分类。首先,利用主成分分析(PCA)在 SPSS 软件上对汶川地震 36 个严重受灾县市的 8 个灾情指标进行了综合分析,得到了累积贡献率为 83.403% 的 3 个主成分及其系数。然后,依据 3 个主成分表达式计算得到 36 个县市灾情的 3 个主成分得分。最后,基于 3 个主成分的得分采用系统样品聚类分析对汶川地震 36 个严重受灾县市进行了分类,得到了全面、合理和科学的分类结果。希望能为灾区的合理援助和有效重建起到积极的促进作用。

参考文献:

[1] 陈丽,张朝元. 基于主成分分析的汶川地震灾情的综合评价[J]. 洛阳师范学院学报,2012,31(2):7-10

[2] 朱建平. 应用多元统计分析[M]. 北京:科学出版社,2006

[3] 张永利,傅俊伟. 基于主成分分析方法的聚类分析方法在灾情综合分类中的应用[J]. 佳木斯大学学报:自然科学版,2011,29(2):296-299

[4] 李树清. 改进的主成分分析法在综合评估中的应用[J]. 济宁学院学报,2010,31(3):15-17

[5] 国家减灾委员会-科学技术部抗震救灾专家组. 汶川地震灾害综合分析与评价[M]. 北京:科学出版社,2008

(下转第 25 页)