技术方法

主成分分析在遥感影像数据中的实例应用

* 密长林[1],马爱功[1],张晓东[2],孙景广[1],杨雪莲[1]

(1.临沂市国土资源局,山东 临沂　276000;2.宁夏回族自治区地质调查院,宁夏 银川　750000)

摘要:主成分分析(Principal Component Analysis)是根据变量之间的相互关系,尽可能不丢失信息地用几个综合性指标表示多个变量的方法。在多(高)光谱图像中,由于各波段的数据间具有相关性,因此包含许多冗余信息。通过主成分分析法可以把遥感图像中所含的大部分信息用少数波段表示出来,这样就可以几乎不丢失数据但可以减少数据量,消除冗余信息。在遥感数据处理时用主成分分析法作数据分析前的预处理,以达到数据压缩和图像增强的效果,更加有利于影像信息提取。文章对主成分分析在遥感图像处理中的实际应用进行了实例示范应用研究。

关键词:主成分分析;特征值;遥感

中图分类号:TP751　　**文献标识码**:B

1　主成分分析方法原理及意义

1.1　原理

主成分分析是统计学中一种可以简化数据量的方法。在处理多元样本数据时,常遇到多元变量之间存在相关关系的情形,使得数据的分析复杂化,使用主成分分析就可以把多元变量化为少数几个独立的综合变量。其统计意义方法为:把 n 个随机变量的总方差分为 k 个不相关的随机变量的方差和,即 $\lambda_1,\lambda_2\cdots,\lambda_k$,使第Ⅰ主成分的方差最大,用$(\lambda_1 \mid \Sigma)\times 100\%$来反映第Ⅰ主成分的贡献率,它表明该主成分综合变量信息量的强调。分析过程中,特征根反映各主成分方差之大小;特征根累积百分率代表所含主成分对总方差的贡献率;特征向量反映各原指标对主成分的贡献大小;其符号表示原指标改变对主成分的增减效果。从原始变量到新变量是一个正交变换(坐标变换)。设有 $X=(X_1,\cdots,X_p)'$ 是一个 P 维随机变量,有二阶矩,记 $\mu=E(X),\Sigma=Var(X)$。考虑它的线性变换:

$$Y_1 = I_1'X = l_{11}X_1 + \cdots + l_{p1}X_p$$

$$Y_p = I_p'X = l_{1p}X_1 + \cdots + l_{pp}X_p$$

$$Var(Y_i) = I_i' \sum I_i$$

$$Cov(Y_i,Y_j) = I_i' \sum I_j \qquad i,j = 1,\cdots,p$$

如果要用 Y_1 尽可能多地保留原始的 X 的信息,经典的办法是使 Y_1 的方差尽可能大,这需要对线性变换的系数 I_1 加以限制,一般要求它是单位向量 $I_1' I_1=1$。其他的各 Y_i 也希望尽可能多地保留 X 的信息,但前面的 $Y_1,\cdots,Y_{i-1}$ 已保留的信息就不再保留,即要求 $Cov(Y_i,Y_j)=0,j=1,\cdots i-1$,同时对 I_i 也有 $I_i' I_i=1$ 的要求,在这样的条件下使 $Var(Y_i)$ 最大。设协方差阵 Σ 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$,相应的单位特征向量分别为 $a_1,a_2,\cdots,a_p$(当特征根有重根时单位特征向量不唯一)。这时 X 的第 i 个主成分为 $Y_i=a_i' X(i=1,\cdots,p)$ 且 $Var(Y_i)=\lambda_i$[1]。

$A=(a_1\cdots a_p)$,$Y(Y_1\cdots Y_p)$

则 A 为正交阵,$Y=A'X$,$Var(Y)=\wedge$,且 $\sum_1^p\lambda_i=\sum_1^p\sigma_{ii}$,其中 σ_{ii} 为 Σ 的主对角线元素。

$$\wedge=\begin{pmatrix}\lambda_1 & & \\ & \cdots & \\ & & \lambda_p\end{pmatrix}$$

* **收稿日期**:2013-03-26;**修订日期**:2013-05-29;**编辑**:陶卫卫

作者简介:密长林(1973—),男,山东临沂人,高级工程师,主要从事国土资源管理信息化及评价研究,E-mail:76369@126.com。

$\lambda_k / \sum_{1}^{p} \lambda_i$→主分量 Y_k 的贡献率

$\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{p} \lambda_i$→主分量 $Y_1, \cdots, Y_m$ 的累计贡献率

1.2　在遥感应用中的意义

在遥感图像处理中主成分分析又称作 K－L 变换或主分量分析。遥感多光谱影像波段多，一些波段的遥感数据之间有不同程度的相关性，造成了数据冗余。PCA 的作用就是保留主要信息，降低数据量，从而达到增强或提取某些有用信息的目的。对某一 n 个波段的多光谱图像实行一个线性变换，即对该多光谱图像组成的光谱空间乘以各个线性变换矩阵，产生一个新的各维量相互正交的光谱空间，从而形成一幅新的包含 n 个波段的多光谱图像。特征值 λ 的大小表征了信息量的多少以及每个分量的相对重要性。第一主分量包含最大的信息量，第二主分量表示了没有被第一主分量表示的数据的最大变化量。第一主分量、第二主分量、…、一直到第 n 主分量所含的信息量依次减少，到最后几乎为零。与此同时，噪声信号逐渐增加，最后一个分量几乎全为噪声。从几何意义来看，变换后的主分量空间坐标系与变换前的多光谱空间坐标系相比，旋转了一个角度，而且新的坐标系的坐标轴一定指向数据信息量较大的方向。以二维空间为例，假定某图像像元的分布为椭圆状，那么经过旋转后新坐标系的坐标轴一定分别沿椭圆的长半轴和短半轴方向——主分量，而且长半轴这一方向信息量最大[2]。在遥感图像数据处理时常常运用 K－L 变换作数据分析前的预处理，主要包括以下几个方面：

(1)数据压缩：在处理多光谱或高光谱数据时，数据波段多，处理起来数据量很大。以 TM 影像为例，进行 K－L 变换后，7 维的多光谱空间变换成 7 维的主分量空间，这时亮度不再与地物光谱信息直接关联，但第一或前二或前三个主分量，已包含了绝大多数的地物信息，足够分析使用，同时数据量却大大地减少了。应用中常常只取前三个主分量做假彩色合成，数据量可以减少到 43%，既实现了数据压缩，也可以作为分类的特征选择。

(2)图像增强：K－L 变换后的几个主分量，信噪比大，噪声相对小，因此突出了主要信息，达到了增强图像的目的。此外，将其他增强手段与之结合使用，会收到更好的效果[3]。

(3)信息提取：主成分分析在信息提取方面应用广泛，如地质岩体蚀变信息的提取、分类纯净像元的提取、森林生物量的提取等。

2　应用

PCA 在多(高)光谱遥感数据处理过程中有着广泛的应用，是一种基于原始波段的线性特征变换方法，它具有用于简化数据空间维数，寻找综合因子表达，进行样本特征排序“利于分类特征选取等作用[4]。下面就以 TM 遥感影像数据进行主成分分析。实验区选择 2007 年 7 月 4 日的 130034 景子区，通过调用遥感数据，在 ERDAS 中建立模型(图 1)，对各波段按协方差计算公式调用算法实施统计计算，得到 6 个波段的协方差矩阵(Covariance Matrix)，如表 1 所列。

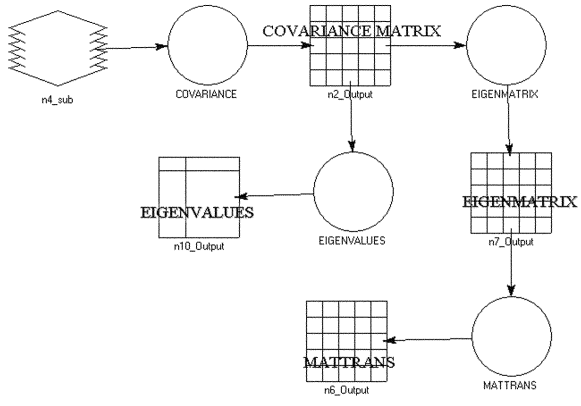

图 1　计算模型

表 1　实验区 TM 影像协方差矩阵

波段	波段 1	波段 2	波段 3	波段 4	波段 5	波段 7
波段 1	112.848	141.278	214.190	69.545	183.396	207.741
波段 2	141.278	181.703	275.971	102.155	241.182	269.939
波段 3	214.190	275.971	427.432	144.180	371.228	418.572
波段 4	69.545	102.155	144.180	202.842	163.723	148.349
波段 5	183.396	241.182	371.228	163.723	366.242	391.893
波段 7	207.741	269.939	418.572	148.349	391.893	436.606

亦即经正交变换 $Y = A^T x$ 后，将多波段遥感数据 x 的信息影射到 Y 主成分坐标空间中，相应地 K—L 变换的具体坐标变换方程表达为：

$$y_1 = 0.259x_1 + 0.337x_2 + 0.518x_3 + 0.212x_4 + 0.478x_5 + 0.526x_7$$

$$y_2 = -0.121x_1 - 0.070x_2 - 0.196x_3 + 0.955x_4 + 0.076x_5 - 0.157x_7$$

$$y_3 = -0.379x_1 - 0.379x_2 - 0.441x_3 - 0.148x_4 + 0.584x_5 + 0.394x_7$$

$$y_4 = -0.644x_1 - 0.269x_2 + 0.514x_3 + 0.085x_4 - 0.388x_5 + 0.303x_7$$

$$y_5 = -0.278x_1 - 0.068x_2 + 0.425x_3 - 0.105x_4 + 0.524x_5 - 0.672x_7$$

$$y_7 = 0.531x_1 - 0.813x_2 + 0.232x_3 + 0.056x_4 + 0.005x_5 + 0.004x_7$$

在实验区中，计算变换处理后各分量特征值，并计算各分量特征值对总体方差的贡献率及累计贡献率。由表 2 可知，仅第一主分量的贡献率就高达总体方差的 89.19%，第一、第二主分量累计贡献率达 97.62%，前三主分量累计贡献率达 99.5%，换言之，在实验区遥感研究中，用第一、第二、第三主分量就几乎具有对原有六波段数据信息的较全面表达，因此可知，主成分变换处理具有较好的对原始波段数据压缩与综合的作用，可达到减少原始波段数据冗余特征和冗余数据量、剔除或消除噪音特征的效果。另外，通过上面所列正交坐标变换方程还可看出，原始 6 个波段在数据变换处理过程中所具有的重要程度是有区别的，这主要可通过线性变换方程中的系数来确定。具体在第一主分量线性变换中，单位特征向量所对应的原始 4 波段的权重具有最大值 0.644，其次分别为 7 波段的 0.531 和 3 波段的 0.379，在第二主分量线性变换中，7 波段具有权重最大值 0.813，其次是 3 渡段 0.379，1 波段 0.337 等。这些信息对于有关参与分类技术处理的特征波段选取是非常有用的[5]。

表 2　实验区 TM 影像协方差特征向量矩阵

主成分	Pc1	Pc2	Pc3	Pc4	Pc5	Pc7
波段 1	0.259	-0.121	-0.379	-0.644	-0.278	0.531
波段 2	0.337	-0.070	-0.379	-0.269	-0.068	-0.813
波段 3	0.518	-0.196	-0.441	0.514	0.425	0.232
波段 4	0.212	0.955	-0.148	0.085	-0.105	0.056
波段 5	0.478	0.076	0.584	-0.388	0.524	0.005
波段 7	0.526	-0.157	0.394	0.303	-0.672	0.004
特征值	1540.976	145.713	32.548	4.629	3.009	0.797
贡献率	0.8919	0.0843	0.0188	0.0027	0.0017	0.0005
累计贡献率	0.8919	0.9762	0.995	0.9977	0.9995	1.0000

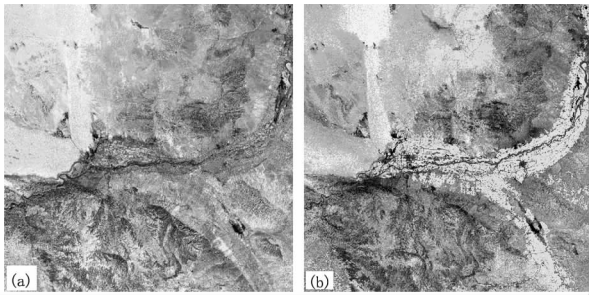

图 2　PCA 变换前后对比图

(a)原始图(R5,G4,B3),(b)PCA 变换(R1,G2,B3)

3　结论

随着遥感数据获取技术的进展，多光谱和高光谱数据急剧增加，并因为研究目的的不同，导致数据冗余信息的出现，进而影响对需要信息的提取，因此，在研究和数据预处理过程中，通过主成分变换，对数据进行压缩、图像增加和信息提取就显示出其特殊意义。

通过对实验区的主成分变换应用实例研究，发现主成分变换在遥感影像数据处理过程中既可以减少数据量，也可以去除冗余信息而且不损失绝大部分信息，能够提高遥感影像数据处理速度，增强影像，有选择的突出重点信息。该文为数据预处理过程提供了示范应用研究，将在遥感特征提取中得到广泛的应用。

参考文献：

[1] 龙永红. 概率论与数理统计[M]. 北京：高等教育出版社，2001.

[2] 日本遥感研究会编. 刘勇卫，贺雪鸿译. 遥感精解[M]. 北京：测绘出版社，1993.

[3] 梅安新. 遥感导论[M]. 北京：高等教育出版社，2001.

[4] 张玉君，曾朝铭，陈薇. ETM+(TM)蚀变遥感异常提取方法研究与应用——方法选择和技术流程[J]. 国土资源遥感，2003，2(56)：44-51.

[5] 甘淑，袁希平，何大明. 澜沧江流域山区土地覆盖遥感监测中 PCA 特征变换处理[J]. 昆明理工大学学报，2002，25(6)：85-89.

Practical Application of Principal Component Analysis in Remote Sensing Image Processing

MI Changlin[1], MA Aigong[1], ZHANG Xiaodong[2], SUN Jingguan[1], YANG Xuelian[1]

(1. Linyi Bureau of Land and Resources, Shandong Linyi 276001, China; 2. NingXia Geological Surveying Institute, NingXia Yinchuan 750003, China)

Abstract: Principal component analysis is a method to express several multi-variable factors without losing information as far as possible based on relations of variables. In multi-spectral images and hyper-spectrum images, there is much redundant information because of relaion of the data in different bands, including many redundant information. Through principal component analysis, most information of remote sensing images can be expressed with fewer bands. Thus, it can not only reduce data size, but also eliminate redundant information. Principal component analysis method is always used to conduct data pre-processing in order to depress data and enhance images. In this paper, practical application of principal component analysis in remote sensing image processing has been studied.

Key words: Principal component analysiss; characteristics values; remote sensing