高光谱数据降维技术研究

王旭红, 肖 平, 郭建明
(西北大学 城市资源学系, 陕西 西安 710069)

摘 要: 高光谱数据对地物具有更高的光谱分辨率,但是由于高光谱数据巨大的数据量以及相邻波段之间的强相关性,导致了对这种数据的许多分类方法达不到应有的效果,从而在某种程度上制约了其广泛的应用。研究表明,特征提取的理论与方法对高光谱信息的优化处理是十分有效的。实验结果表明,在一定的分类精度范围内,减低维数而不丢失信息,可以提高分类器的效能,实现高维遥感数据的优化处理和高效利用。

关键词: 特征提取; 数据降维; 高光谱数据; 影像分类
文献标识码: B **文章编号**: 1000—288X(2006)06—0089—03 **中图分类号**: TP751.1

Research on Dimensionality Reduction Technology of Hyperspectral Data

WANG Xu-hong, XIAO-ping, GUO Jian-ming
(Department of City and Resource Science, Northwest University, Xi'an, Shaanxi 710069, China)

Abstract: Hyperspectral data have a high spectral resolution for the objects of the earth. However, many analysis approaches of hyperspectral data do not provide a promising result because of its great data volume and strong correlation between its neighboring bands. Consequently, it restricts the efficiency and broad application of high resolution data. The research indicates that feature extraction is the highly effective theory and method to optimize hyperspectral data and information. The result of experiment shows that with a given precision of classification, the reduction in dimensionality without loss of information improves the classifier performance, and helps to achieve the aims of optimal process and effective utilization of hyperspectral remote sensing data.

Keywords: feature extraction; reduction of data dimensionality; hyperspectral data; image classification

与常规的遥感数据相比,高光谱数据的光谱分辨率高(一般可达到 nm 级),从理论上讲对地表地物的分辨能力得到了极大的增加,但是由于高光谱数据波段数多(往往在 100 个波段以上)、数据量大,相邻波段间相关性强,在某种程度上也阻碍了高光谱数据的广泛应用。因此,在不降低分类精度的前提下,选择有效的特征信息参加分类以降低高光谱数据的维数,是这一高维信息空间在遥感领域内应用研究的一个重要问题。

1 高光谱数据降维的方法

近年来,国内外对高光谱数据降维方式的研究不断增多。Jimenez 等从理论上对数据维减小的可能性进行了论证[1],为高光谱数据降维奠定了理论基础。目前常用的高光谱数据降维的方法有波段选择、数据源划分、特征提取和数据融合等 4 种类型[1—3]。其中,特征提取是高效的降维方法之一,其中具代表性的有以下几种[1,3]: (1) 主成分分析(PCA, Principal Component Analysis); (2) 分段主成分变换(SPCT, Segmented Principal Components Transform); (3) 投影法(Jimenez)、判别式分析方法(DAFE, Discriminant Analysis Feature Extraction)、决策边界特征提取(DBFE, Decision Boundary Feature Extraction)和非参数加权特征提取方法(Kuo. etc)等方法。

1.1 主成分分析(PCA)

主成分分析(PCA)[1,3],又叫 $K—L$ 变换,是特征提取中最经典的方法之一。采用多变量线性变换的方法对高维数据进行特征提取,产生一个按信噪比大小(MNF, Minimum Noise Fraction)进行排列的低维子空间,消除图像之间的相关性,用几个主成分就可以代表原始数据中的大多数信息。目的是要把原来多个波段中的有用信息集中到数目尽可能少的新组分图像中,并使这些组分图像之间互不相交,从而大大降低了数据量。

收稿日期:2006-05-24 **修稿日期**:2006-08-21
基金项目:西北大学 2005 校内基金和科研启动基金,基础测绘科技项目(146014020201-05)
作者简介:王旭红(1968—),女(汉族),陕西咸阳市人,工学博士,高级工程师,从事地理信息系统、遥感图像处理、空间数据挖掘等方面的研究。E-mail:Jqy-wxh@163.com

1.2 判别式分析方法(DAFE)

判别式分析方法(Discriminant Analysis Feature Extraction,DAFE)也是一种特征提取的算法[4-5]。其基本原理是构建一个原始影像数据(向量)的线性组合,即任意数据集的类间方差与类内方差的比值,并使之最大化,从而保证了变换后的向量具有最大的可分离性(Richards,1993)。

不失一般性,设目标分类数为 2 类;d 为线性变换矩阵。

$$\sigma_A^2/\sigma_w^2 = \text{类间距离}/\text{类内距离} \quad (1)$$

式中:σ_A^2 ——是变换后类型 1 和类型 2 之间的标准差;σ_w^2 ——是变换后类型 1 和类型 2 的类内标准差的平均值。

由特征选择的一般式可得,变换后方差与变换前方差之间的关系为:

$$\sigma_w^2 = d^T \sum_w d \quad (2)$$

$$\sigma_A^2 = d^T \sum_A d \quad (3)$$

判别式分析方法的目标就是寻找到使类间方差与类内方差的比值 σ_A^2/σ_w^2 达到最大值时的映射变换 d,亦使下式最大化。

$$\lambda = \sigma_A^2/\sigma_w^2 = d^T \sum_A d / d^T \sum_w d \quad (4)$$

新特征提取标准是使(4)中两者比率最大;类间分离度越大,提取的特征空间越好。

1.3 决策边界特征提取(DBFE)

Lee and Landgrebe(1993)提出[4-5],可以定义一个决策边界特征矩阵 DBFM (Decision Boundary Feature Matrix)来分离出对分类有用的特征和多余的特征。并且,决策边界特征矩阵 DBFM 的秩是维持在原始特征空间中同样分类精度所需要的最小特征维数。也就是说,对于某一种分类器而言,通过 DBFM 可以得到进行分类所需要的最小特征维数,与非零特征值对应的特征向量是最优特征向量的必要组分。

对于两类问题,决策边界特征矩阵(DBFM)为:

$$\sum_{\text{DBFM}} = \frac{1}{K}\int_s N(X)\, N^t(X)\, p(X)\, \mathrm{d}X \quad (5)$$

式中:$N(X)$ ——为法向量集合;$p(X)$ ——似然密度函数;$K = \int_s p(X)\,\mathrm{d}X$;$s$ ——决策界,函数的积分在决策界上进行。

公式(6)是针对两类问题,但是很容易扩展到多类问题。

$$\sum_{\text{DBFM}} = \sum_{i=1}^{M}\sum_{j,j\neq i}^{M} p(\omega_i)\, p(\omega_j) \sum^{ij}\text{DBFM} \quad (6)$$

式中:M ——类别数;$p(\omega_i)$ ——ω_i 的先验概率;$\sum^{ij} DBFM$ ——ω_i 和 ω_j 间 DBFM。

2 算法在分类中的实际应用

2.1 实验区选择

实验区为太湖,波段数为 124 波段、图像大小为 1826×536,总数据量为 241 M;剔除有条纹的质量差的波段外,实际参加实验波段为 63 个(有效波段数),空间分辨率为 4 m,光谱分辨率为 10 nm。

2.2 实验流程

在实际中,实验流程可分为分类目标确定及训练样本选取、特征提取和初始分类、完善分类目标与训练样本、最终分类/精度评估等步骤。

(1) 分类目标的确定及训练样本的选取。首先调入以 56,40,16 三波段合成显示的高光谱影像,如附图 5a 所示;目标类确定为:植被(Veg)、海水(Seawater)、渔塘(Fishery)、农作物(crop)、建筑物(Building)5 类。

(2) 特征提取和初始分类。初始分类目的是在既定算法下,判别分类目标确定的是否正确,是否存在被忽略的地类。所制定的分类类别之间是否都可以区分,如果不区分,应该进一步精确选取合适的训练样本。

(3) 完善分类目标与训练样本。对初始分类结果分析与比较,调整分类目标以及训练样本,重新进行特征提取,形成用于分类的新的特征子集。

(4) 最终分类/精度评估。进行最终分类,并利用各种精度评估方法对结果进行评估。

2.3 实验结果

2.3.1 判别式分析方法(DAFE) 在确定了 5 个分类目标和训练样本以后,由 DAFE 算法,计算出 4 个特征值对应的特征向量的矩阵,如表 1 所示,分类如附图 5b 所示。

2.3.2 决策边界特征提取(DBFE) 在确定了 5 个分类目标和训练样本以后,由 DBFE 算法,计算出 14 个特征值对应的特征向量的矩阵,如表 2 所示,分类如附图 5c 所示。

2.3.3 精度对比 用同一实验数据分别对主成分分析、判别式分析方法、决策边界特征提取 3 种算法得到的土地利用/土地覆盖类型图进行精度评估,结果如表 3 所示。

表 1　DAFE 算法的特征提取转换矩阵和特征向量

成分	转换矩阵			特征向量					
	特征值	方差百分比	累加百分比	通道 1	通道 2	通道 3	通道 4	通道 5	…
1	12.699 9	71.907 0	1.907 0	0.173 39	0.170 47	0.061 31	- 0.018 78	0.018 78	…
2	3.537 4	20.028 9	91.935 8	0.107 39	0.030 33	0.084 59	0.078 50	0.061 17	…
3	1.059 5	5.998 9	97.934 7	0.262 63	0.201 84	0.236 37	0.023 46	- 0.142 08	…
4	0.364 8	2.065 3	100.000 0	0.040 81	0.031 18	0.166 65	0.044 72	0.062 67	…

表 2　DBFE 算法的特征提取转换矩阵和特征向量

成分	转换矩阵			特征向量					
	特征值	方差百分比	累加百分比	通道 1	通道 2	通道 3	通道 4	通道 5	…
1	2.955 2	39.217 0	39.217 0	0.043 19	0.052 57	0.017 39	- 0.018 16	- 0.063 23	…
2	1.310 3	17.387 9	56.604 9	- 0.040 41	- 0.131 11	- 0.111 47	0.008 25	- 0.071 96	…
3	0.910 4	12.081 1	68.685 9	- 0.092 47	- 0.055 96	0.071 92	0.022 30	0.074 33	…
4	0.585 4	7.768 6	76.454 6	0.098 07	0.008 68	- 0.032 50	- 0.029 85	0.009 99	…
…	…	…	…	…	…	…	…	…	…
14	0.018 3	0.243 2	100.000 0	0.012 06	0.091 01	- 0.013 86	0.003 36	- 0.117 46	…

表 3　遥感图像分类结果精度分析表

算子类别	植被/ %	海域/ %	建筑用地/ %	鱼池/ %	农田/ %	Kappa 系数
PAC 分类	73.32	83.28	75.37	86.43	73.12	0.826
DAFE 分类	76.52	83.54	76.69	87.36	74.57	0.839
DBFE 分类	77.63	84.32	78.36	89.12	76.28	0.847

3　结论与讨论

实验证明了特征提取方法对高光谱数据原始信息的优化处理、高维信息数据降维的可行性,为高光谱数据的推广应用提供了有利的保证。在实际应用中,可以通过特征提取的方法来达到数据简约的目的,以保证在充分保证利用原始图像信息的基础上,达到信息快速提取的目的。

在理论上(具体处理算法上),目前较常应用的依然是主成分分析(PCA 算法),但 PCA 算法起源于 TM 等低维数据的处理,而非高维数据的特有算法,因此,研究和探索新的压缩算法始终是信号处理、图像处理、计算机和遥感界专家矢志不移的研究目标。判别式分析方法(DAFE)优点是不必知道样本的分布形式,缺点是仅仅在类别分布正常时有效,当类别呈现多模型混合分布的时候,DAFE 的效果就不是很理想了。

决策边界特征提取(DBFE)算法直接利用样本向量值,从而避免了 DAFE 算法的局限性;但是,这种算法要求有较大的样本数,较小的或不充足的样本数不能保证算法的有效性。

因此,在高光谱数据处理中,应根据高维数据的特点、训练样本的数量、数据量等多个方面考虑数据简约、降维的方法。

[参 考 文 献]

[1] 赵春晖,刘春红. 超谱遥感图像降维方法研究现状与分析[J]. 中国空间科学术,2004,5(10):28—36.

[2] 赵英时. 遥感应用分析原理与方法[M]. 北京:科学出版社,2003.

[3] Jia Xiuping, Richards J A. Segmented Principal Components Transformation for Efficient Hyperspectral Remote Sensing Image Display and Classification[J]. IEEE Trans. on Geoscience and Remote Sensing, 1999, 37(1): 538—542.

[4] Lee C, Landgrebe D A. Decision Boundary Feature Selection for Non-Parametric Classifications[J]. IEEE Transactions on System, Man, and Cybernetics, 1993, 23(2): 433—444.

[5] Lee C, Landgrebe D A. Feature Extraction and Classification Algorithms for High Dimensional Data[D], Electrical Engineering, Purdue University, West Lafayette, Indiana, 1992.

(部分参考文献略)