

高光谱遥感影像降维的拉普拉斯特征映射方法

黄蕾

(湖北省基础地理信息中心, 武汉 430071)

摘要:针对高光谱遥感影像数据量大、数据冗余度高的特点,引入拉普拉斯特征映射方法对高光谱遥感数据进行非线性降维。为了解决传统流形学习方法不能处理大数据量遥感影像的问题,本文提出了基于多元线性回归的拉普拉斯特征映射线性解法。实验证明,本文提出的降维方法能够保持数据集在原始特征空间分布的局部几何属性,降维后的影像具有更好的分类精度。

关键词:高光谱;拉普拉斯特征映射;降维;分类

doi:10.3969/j.issn.1000-3177.2011.06.008

中图分类号:TP751 **文献标识码:**A **文章编号:**1000-3177(2011)115-0037-05

Laplacian Eigenmap for Hyperspectral Remote Sensing Image Dimensionality Reduction

HUANG Lei

(Provincial Geomatics Center of Hubei, Wuhan 430071)

Abstract: Feature extraction is an indispensable preprocessing step for large and high redundancy data of hyperspectral remote sensing image. In this paper, a Laplacian Eigenmap (LE) is introduced for dimensionality reduction. In order to overcome the shortcoming of conventional manifold learning which could not deal with large data, a linearization procedure for LE is proposed based on multiple linear regression analysis. The experiment demonstrates that the proposed dimensionality reduction method can preserve local geometry of samples in original feature space, the low dimensionality image could achieve a better classification accuracy.

Key words: hyperspectral; laplacian eigenmap; dimensionality reduction; classification

1 引言

随着现代传感器技术的发展,高光谱遥感影像能够在紫外到红外波长区的范围内提供细致丰富的地物光谱特性描述,从而大大地提高了遥感影像对地物的分类识别能力,已经广泛地应用于精细农业、城市规划、资源调查、环境监测等方面^[1~3]。然而,高光谱遥感影像在提供多波段、高量化级数的遥感影像同时,也给数据处理带来了新的挑战,主要表现在:数据量大、数据冗余度高和 Hughes 现象^[4]等方面。

为了充分利用高光谱的信息同时减少因数据冗余带来的复杂计算,必须挖掘高光谱特征空间的内

在性质,以更加有效的低维空间来表达原始高维特征空间中的样本分布,使样本更有利于分类。国际上针对减少数据维数同时不损失有意义的信息进行了一系列的研究,波段选择方法直接从原始的多个波段中选择具有意义的波段子集实现特征数目的减少^[5];以主成分分析(PCA)^[6]为代表的波段特征提取方法采用多变量线性变换方法将原始高维特征空间中的主成分信息组合到新的特征空间中。这些方法计算简单,对具有线性结构的数据集具有很好的效果,因此被广泛应用于高光谱遥感影像降维。

然而,在统计理论中,线性算子在将数据投影到由特征向量基组成的特征空间中时,忽略了数据在特征空间中的局部性质,主成分分析不能保持样本

收稿日期:2010-11-02 修订日期:2010-12-15

作者简介:黄蕾,硕士研究生。现工作单位:湖北省基础地理信息中心,主要研究方向:3S 技术应用。

E-mail: huanglei_gh@163.com

在新特征空间中同样具备原有的分布结构。为了解决这个问题,国际上提出了基于流形学习的降维方法^[7]。流形学习方法致力于找到样本在原始特征空间中分布的全局结构和局部结构,使得样本在降维后的新特征空间中能够基本保持原来的局部结构,即保持样本在特征空间中的流形。流形学习方法已经成功地应用于模式分析和机器智能领域^[8~9],但是目前还极少用于遥感影像处理领域,原因是流形学习方法需要考虑全部样本的邻域结构,因此计算量极大,在其中的特征值分解步骤中需要处理大小为样本数×样本数的方阵。如果直接将流形学习方法用于遥感影像降维,仅能快速处理 10³ 像素以内的影像,这使得流形学习方法目前没有在高光谱遥感影像降维中推广。

本文致力于解决上述问题,将流形学习方法用于高光谱遥感影像降维。为了保持样本在降维后的新特征集的局部几何属性,引入拉普拉斯特征映射方法对数据集进行非线性降维。为了解决流形学习方法不能处理大数据量遥感影像的问题,本文提出了一种拉普拉斯特征映射的线性解法。实验证明,本文提出的降维方法能够快速、高效地获得高光谱遥感影像的低维特征描述,基于流形学习的降维影像比原始影像和 PCA 方法的降维影像相比能够实现更优的监督分类精度。

2 拉普拉斯特征映射

拉普拉斯特征映射(Laplacian Eignmap)是一种非线性降维方法,假设原始数据集和它在低维空间的映射分别为 $X \in R^{m \times N}$ 和 $Y \in R^{d \times N}$,其中 N 为数据集的样本个数, m 为原始特征维数, d 为新特征维数($d < m$),原始数据集和映射集中的样本分别为 $x_i \in R^m$ 和 $y_i \in R^d$ 。拉普拉斯特征映射找到一种映射方式 $X \rightarrow Y$,同时保持数据集在特征空间分布的局部几何属性。拉普拉斯特征映射的目标函数是^[10]:

$$\arg \min_Y \sum_{i=1}^N \sum_{j=1}^N \|y_i - y_j\|^2 W(i, j) \quad (1)$$

其中 $W \in R^{N \times N}$ 是用于保持局部几何属性的样本关系矩阵,定义如下:

$$W(i, j) = \begin{cases} \exp(-\|x_i - x_j\|^2 / t) & x_i \text{ 和 } x_j \text{ 是邻域点} \\ 0 & x_i \text{ 和 } x_j \text{ 不是邻域点} \end{cases} \quad (2)$$

其中样本 x_i 的邻域点定义为数据集中与 x_i 距离最近的 k 个点的集合, t 是样本关系矩阵的参数。由公式(2)可知,在原始特征空间中距离近的点 x_i 、 x_j 在样本关系矩阵中具有较大的权值(i, j),这类点

在目标函数(1)中起着主要作用,拉普拉斯特征映射保证它们在低维空间中对应的 $\|y_i - y_j\|^2$ 最小。这样,在原始特征空间中的邻域点在新特征空间仍然是邻域点,因此拉普拉斯特征映射保持了数据集在特征空间分布的局部几何属性。

为了求解式(1),引入对角矩阵 M 和拉普拉斯矩阵 L ,定义如下:

$$M(i, i) = \sum_j W(i, j), L = M - W \quad (3)$$

这样,式(1)可以转化为^[11]:

$$\arg \min_Y \sum_{i=1}^N \sum_{j=1}^N \|y_i - y_j\|^2 W(i, j) = \arg \min_Y 2Y^T L Y \quad (4)$$

subject to: $Y^T Y = I_N$

式(4)的解 Y 就是过广义特征值问题如式(5)中 d 个最小的特征值对应的特征向量。

$$L v = \lambda M v \quad (5)$$

在高光谱遥感影像降维中,影像中全部像素的光谱向量组成原始高维特征空间集 $X \in R^{m \times N}$,其中 m 为高光谱影像波段数目, N 为影像的像素总数。实际应用中 N 值常常在数万以上,会导致式(2)中样本关系矩阵计算缓慢和式(5)的特征值分解无法进行。因此必须考虑将拉普拉斯特征映射进行线性化以解决样本数量过大的问题。具体方法如下:

(1)选取全部样本数的子集作为训练样本组成原始高维特征集 X ;

(2)利用拉普拉斯特征映射求出低维特征集 Y ;

①根据式(2)计算样本关系矩阵 $W \in R^{N \times N}$;

②根据式(3)计算对角矩阵 M 和拉普拉斯矩阵

L ;

③求解广义特征值问题 $L v = \lambda M v$,由 d 个最小的特征值对应的特征向量组成 Y 。

(3)利用多元线性回归方法,找到拉普拉斯特征映射的线性表达方式:

$$\arg \min_U \|Y - U^T X\|^2 \quad (6)$$

(4)利用投影矩阵 U 对原始高光谱遥感影像进行线性降维。

式(6)中, U 通过多元统计量的最小二乘法求解:

$$U = (X^T X)^{-1} X^T Y \quad (7)$$

3 实验与分析

3.1 实验数据描述

实验选用普渡大学工程学院公布的科学研究高光谱遥感数据,该数据是 AVIRIS 高光谱传感器获

取的印第安那州 Tippecanoe 郡西北部的影像,大小为 145×145 像素,220 波段。在附带的 GIS 地理数据文件中,一共包含了 15 种地物在影像中的真实位

置。本文研究其中 10 种主要地物,如图 1 所示。该影像是典型的精细农业高光谱影像,也是国际上公认适合高光谱遥感图像处理算法的实验数据^[12]。

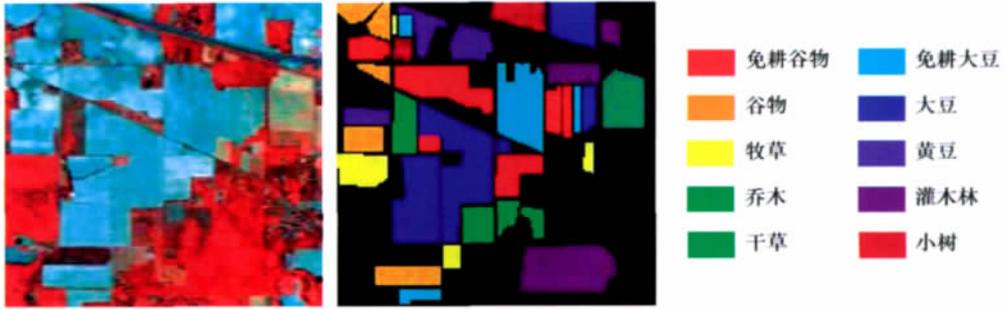


图 1 实验影像数据和真实 GIS 地理数据

3.2 实验方法

实验中,选取全部像素的 5% 作为训练样本进行拉普拉斯映射,并通过方法 1 找到拉普拉斯特征映射的线性表达方式,对原始高光谱影像降维,整个过程在 Matlab R2009a 软件环境中实现。同时,考虑原始影像和经过主成分分析降维的影像作为对比,选用分类精度作为高光谱影像降维的效果评定指标。在分类步骤中,选用支持向量机(SVM)方法进行监督分类,采用高斯径向核函数,在软件 ENVI 4.7 中实现。训练样本和测试样本从对应的真实 GIS 地理数据中随机产生,数量如表 1 所示。

3.3 结果分析

采用本文提出的方法和两种对比方法对影像中 10 类地物的分类结果和分类精度如图 2,图 3 所示。由图中数据可知,经过拉普拉斯特征映射降维后的影像分类精度明显提高,绝大多数地物的生产者精度都优于原始高光谱影像的分类精度和经过 PCA 方法降维后的分类精度。具体观察其中免耕谷物、牧草、免耕大豆和黄豆 4 类地物,经过 PCA 降维后分类精度反而略低于或明显低于原始影像直接分类精度,说明这种基于方差最大化的线性降维方法不

能保留全部样本降维后的局部判别信息。特别是 PCA 降维后免耕大豆的分类精度出现了极大的下降,说明尽管 PCA 降维后提高了影像的整体分类精度(OA),但它不是最优的降维方法。经过本文提出的拉普拉斯特征映射降维后,将原始高光谱影像的整体分类精度由 71.79% 提高到 81.02%,同时保证全部地物类别的分类精度都高于原始影像的分类精度。因此,拉普拉斯特征映射能够有效地用于高光谱遥感影像降维。

表 1 监督分类中的训练样本和测试样本数量

样本数量	训练样本	测试样本
免耕谷物	103	984
谷物	117	549
牧草	20	326
乔木	44	494
干草	36	357
免耕大豆	72	664
大豆	95	1759
黄豆	53	240
灌木林	117	882
小树	30	254

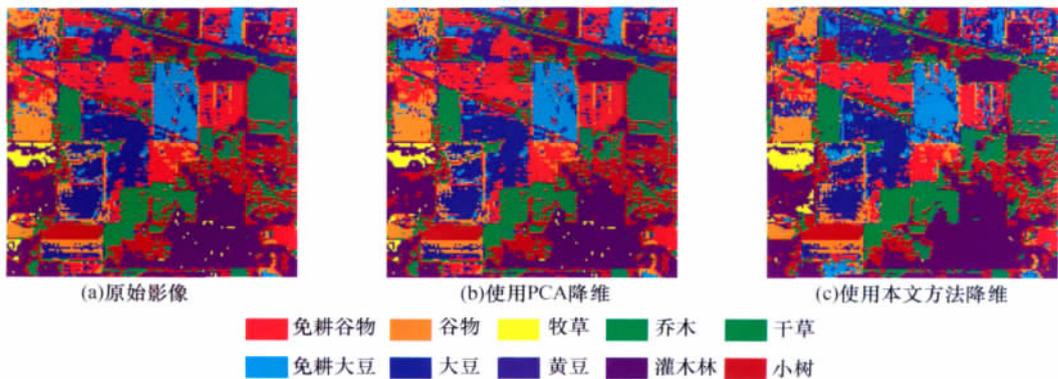


图 2 分类结果图

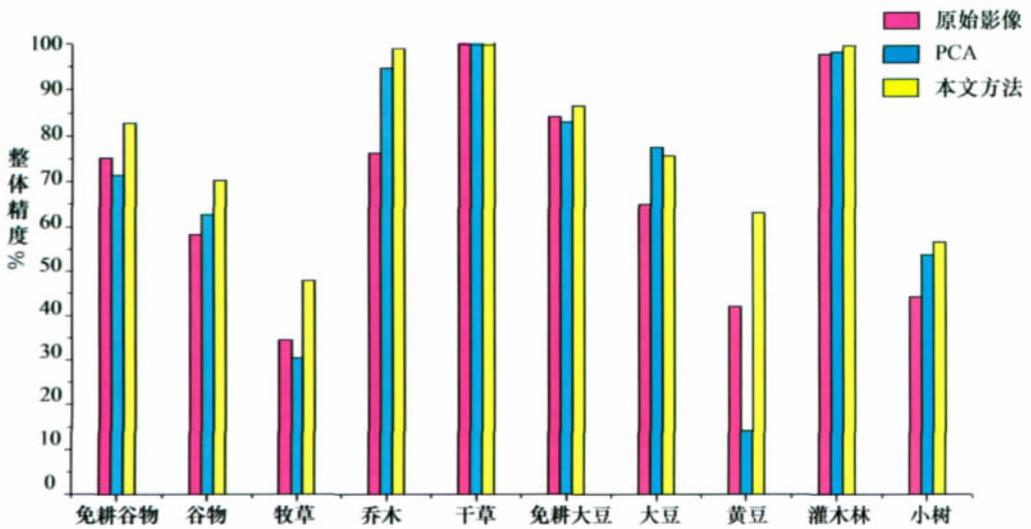


图3 每类地物生产者精度统计

在拉普拉斯特征映射中有两个重要的参数:邻域范围的大小 k 和降维后的维度 d 。图4,图5分别记录了不同大小的 k 和 d 的取值和降维后影像的分类结果的关系。图4中锁定 $d=10$,观察 k 值变化对最终分类总体精度和 kappa 系数的关系。由图4可知,分类结果对邻域范围大小 k 非常敏感,当 k 取值较小 ($k < 5$) 时,邻域点数量过少,受困于高光谱影像中存在的噪声影响,邻域范围没有表征整个样本流形的局部结构信息,目标函数(1)失去了其保持数据集在特征空间分布的局部几何属性的意义,因此此时分类精度较低;当 k 取值范围在 $8 \sim 15$ 时,邻域范围大小选取合适,此时分类精度最高;当 k 取值范围在 $15 \sim 35$ 时,邻域范围选取较大,此时目标函数(1)基本能保持局部几何属性,但是由于邻域范围较大,这种保持程度已经存在了较大的误差,最终导致降维后的影像分类精度只能达到 70% 附近,接近或略低于原始高光谱影像分类精度;当 k 取值过大 ($k > 40$) 时,邻域范围过大,目标函数(1)已经不再表征整体流形的局部结构信息,此时分类精度急剧下降。考虑到本例实验中选取全部样本数的子集作为训练样本组成原始高维特征集,进行拉普拉斯映射的样本数量约为 10^3 ,样本中主要地物类型为 10 种,选取邻域 $k=10$ 基本可以满足保持整个流形的局部结构信息。

图5中锁定 $k=10$,观察 d 值变化对最终分类总体精度和 kappa 系数的关系。由图5中可知,分类结果与 d 值的关系基本符合一般高光谱遥感影像特征选择的情况。当特征维数较小时,分类精度较低;当特征维数增大到接近高光谱遥感影像的本征维数,即高光谱影像中的端元的种类数或影像中存在的主要地物种类数时,经过合适的特征提取方法后分类精

度将会达到最高。在本例实验中,当 $k=10$ 、 $d=13$ 时分类精度达到 82.16% ,大大超过了原始高光谱影像的分类精度 71.79% 。当特征维数继续增大时,分类精度呈现缓慢下降的过程,原因是特征值排序较后的特征主要包含冗余信息和噪声,只包含少量的判别信息,因此,在训练样本数量有限的监督分类中,增加的特征反而会造成分类精度的下降。

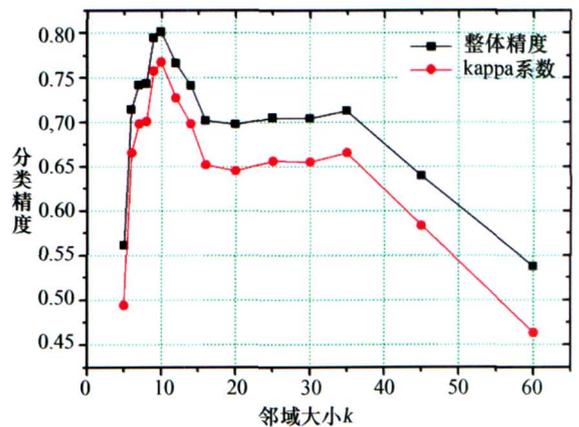


图4 k值变化对分类精度的影响

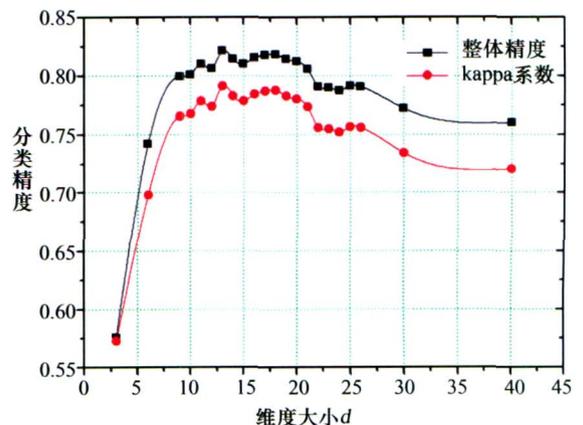


图5 d值变化对分类精度的影响

4 结束语

通过本文的实验证明,基于流形学习的方法能够用于高光谱遥感影像降维。本文提出的拉普拉斯特征映射的线性解法能够保持降维后的样本在原始特征空间中的局部几何属性。由于使用影像中全部像素的子集作为训练样本,拉普拉斯特征映射中的样本数量被限制在 10^3 像素左右,能够高效地对高光谱遥感影像降维。通过对降维后的影像进行监督分类实验,证明了经过拉普拉斯特征映射后图像的

分类总体精度能够达到 80% 以上,高于原始高光谱影像的分类精度和使用常规 PCA 降维方法后的分类精度。在本文的方法中,邻域大小 k 是影响降维后分类结果的关键因素,最合适的 k 值大小与样本流形的分布情况有关,即与样本的总数、样本类别的总数和样本在特征空间中的局部分布有关,本文通过实验找到最合适的 k 值为 8~12。如何根据高光谱遥感数据的特点自适应的选择合适的 k 值,是本文作者需要进一步研究的方向。

参考文献

- [1] T. C. Bau, S. Sarkar, G. Healey. Hyperspectral region classification using a three-dimensional gabor filterbank[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(9): 3457-3464.
- [2] N. Acito, M. Diani, and G. Corsini. Hyperspectral signal subspace identification in the presence of rare signal components [J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(4): 1940-1954.
- [3] J. M. Yang, B. C. Kuo, P. T. Yu, et al. A dynamic subspace method for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(7): 2840-2853.
- [4] G. Hughes. On the mean accuracy of statistical pattern recognizers[J]. IEEE Transactions on Information Theory, 1968, 14(1): 55-63.
- [5] M. Zortea, V. Haertel, R. Clarke. Feature extraction in remote sensing high-dimensional image data[J]. IEEE Geoscience and Remote Sensing Letters, 2007, 4(1): 107-111.
- [6] T. Zhang, D. Tao, X. Li, et al. Patch alignment for dimensionality reduction[J]. IEEE Transactions on Knowledge and Data Engineering, 2008: 1299-1313.
- [7] S. Roweis, L. Saul. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323.
- [8] L. Van der Maaten, E. Postma, H. Van den Herik. Dimensionality reduction: A comparative review[J]. Published Online, 2007(10): 1-41.
- [9] D. Cai, X. He, J. Han. SRDA: An efficient algorithm for large-scale discriminant analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2008: 1-12.
- [10] M. Belkin, P. Niyogi. Semi-supervised learning on Riemannian manifolds[J]. Machine Learning, 2004, 56(1): 209-239.
- [11] M. Belkin, P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Advances in Neural Information Processing Systems, 2002(1): 585-592.
- [12] L. Zhang, X. Huang. Object-oriented subspace analysis for airborne hyperspectral remote sensing imagery[J]. Neurocomputing, 2010, 73(4-6): 927-936.

(上接第 36 页)

参考文献

- [1] 刘诚, 李亚军, 赵长海, 等. 气象卫星亚像元火点面积和亮温估算方法[J]. 应用气象学报, 2004, 15(3): 273-280.
- [2] 毛克彪, 覃志豪, 官鹏, 等. 劈窗算法 LST 精度评价和参数敏感性分析[J]. 中国矿业大学学报, 2005, 34(3): 318-322.
- [3] 姚云军, 南鹏, 张照录, 等. 分裂窗算法在热红外遥感陆面温度反演中的应用[J]. 兰州理工大学学报, 2007, 33(6): 89-92.
- [4] BARDUCCI A, GUZZAI D, MARCOIONNI P, et al. Comparison of fire temperature retrieved from SWIR and TIR hyperspectral data[J]. Infrared Physics & Technology, 2004, 46(1-2): 1-9.
- [5] ABUELGASIM A, FRASER R. Day and night-time active fire detection over North America using NOAA-16 AVHRR data[J]. Geoscience and Remote Sensing Symposium, 2002(3): 1489-1491.
- [6] WRIGHT R, ROTHERY D A, BLAKE S, et al. Simulating the response of the EOS Terra ASTER sensor to high-temperature volcanic targets Geophys[J]. Geophysical Research Letters, 1999, 26(12): 1773-1776.
- [7] 孔祥生, 苗放, 刘鸿福, 等. 遥感技术在监测和评价土法炼焦污染源中的应用[J]. 成都理工大学学报(自然科学版), 2005, 32(1): 92-96.