

# 基于 SSMFA 与 $k$ NNS 算法的高光谱遥感影像分类

王立志, 黄 鸿, 冯海亮

(重庆大学光电技术及系统教育部重点实验室, 重庆 400044)

**摘 要:** 为了研究高光谱影像数据的维数约简和分类问题, 提出了一种基于半监督边际费希尔分析(SSMFA)和  $k$ NNS 的高光谱遥感影像数据分类算法. 该方法利用有标记数据和无标记数据的信息获得数据的内在流形结构, 通过 SSMFA 将高光谱数据从高维观测空间投影到低维流形空间, 然后利用邻域内多个近邻点的信息通过  $k$ NNS 分类器对低维空间中的数据进行分类. 在 Urban, Washington 和 Indian Pine 数据集上的分类识别实验表明, 该方法能够较为有效地发现高维空间中数据的内蕴结构, 在每类随机选取 4, 6, 8 个有类别标记的样本 10 个无类别标记的样本的情况下, 该方法的总体分类精度能够比 MFA +  $k$ NNS 提高 0.8% ~ 2.5%, 比 MFA +  $k$ NN 提高 2.8% ~ 4.5%, 比其他算法提高 4.0% ~ 7.0%, 分类精度有了明显的提高.

**关键词:** 高光谱影像; 地物分类; 图嵌入框架; 最近邻

**中图分类号:** TP751.1 TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2012)04-0780-08

**电子学报 URL:** <http://www.ejournal.org.cn>

**DOI:** 10.3969/j.issn.0372-2112.2012.04.026

## Hyperspectral Remote Sensing Image Classification Based on SSMFA and $k$ NNS

WANG Li-zhi, HUANG Hong, FENG Hai-liang

(Key Laboratory on Opto-Electronic Technique and Systems, Ministry of Education, Chongqing University, Chongqing 400044, China)

**Abstract:** In order to explore dimensionality reduction and classification in hyperspectral remote sensing image, an algorithm based on semi-supervised marginal Fisher analysis (SSMFA) and  $k$ -nearest-neighbor simplex ( $k$ NNS) is proposed in this paper. First, the data are projected from a high-dimensional space onto low-dimensional space by SSMFA combined with the information of different classes. Then, classification is performed under the  $k$ NNS classifier by using a few neighbors from each class. The experimental results on the Urban data set, Washington DC Mall data set and Indian Pine data set show the effectiveness of the proposed algorithm, when  $i(i=4, 6, 8)$  labeled samples and 10 unlabeled samples of each class are randomly selected for training and 100 samples of each class for testing, the overall accuracy of our proposed algorithm is improved by 0.8% - 2.5%, 2.8% - 4.5% and 4.0% - 7.0%, respectively, as compared with MFA +  $k$ NNS, MFA +  $k$ NN and other methods.

**Key words:** hyperspectral images; land cover classification; graph embedding framework; nearest neighbor

### 1 引言

高光谱遥感技术是 20 世纪末遥感领域的重大成就之一. 高光谱影像分辨率高, 且地物光谱曲线是连续的, 能够区分同一种地物的不同类别, 为研究地物的细微特征提供了有效的途径, 目前高光谱遥感技术已被广泛应用于环境监测、海洋遥感、生态调查、精准农业、地质矿岩识别等诸多领域<sup>[1~3]</sup>. 然而, 高光谱数据的采样间距一般在纳米级, 使得相邻波段的相关性很强, 并且光谱维数很高, 传统的分类识别方法容易陷入维数灾难现象而无法获得理想的分类结果<sup>[4,5]</sup>. 因此, 如何对高光谱

数据实现有效的维数约简, 并建立合适的分类模型, 成为高光谱遥感影像分类的关键. 研究发现高维数据集往往具有一些低维的非线性结构, 这种低维结构数学上称为流形<sup>[6]</sup>. 如果能够从高维采样数据中有效地展开数据的低维流形结构, 发现高维数据分布的内在规律, 就可以在维数约简的同时保持事物的本质.

国内外学者在处理高维数据的研究中, 已产生了一系列算法, 经典技术主要有: 主成分分析 (Principal Component Analysis, 简称 PCA)<sup>[7]</sup>, 线性鉴别分析 (Linear Discriminant Analysis, 简称 LDA)<sup>[8]</sup>等. PCA 和 LDA 都是建立在“全局线性”这一基本假设基础上的, 不能发现隐藏

在高维数据中的低维流形,无法揭示高光谱数据的内蕴结构<sup>[9,10]</sup>.而用于非线性数据分析的算法有局部线性嵌入(Local Linear Embedding,简称 LLE)<sup>[11]</sup>、等距映射(Isometric Feature Mapping,简称 ISOMAP)<sup>[12]</sup>、拉普拉斯特征映射(Laplacian Eigenmap,简称 LE)<sup>[13]</sup>等.但是上述算法只能得到训练样本的低维嵌入,没有明晰的投影矩阵,不能直接处理新样本,因此难以直接用于分类.He 等提出的邻域保持嵌入算法(Neighborhood Preserving Embedding,简称 NPE)<sup>[14]</sup>和局部保持投影算法(Locality Preserving Projection,简称 LPP)<sup>[15]</sup>可以有效地解决新样本问题.但 NPE 与 LPP 等方法并没有利用数据的类别信息,属于非监督学习算法.Yan 等提出了边际费希尔分析算法(Marginal Fisher Analysis,简称 MFA)<sup>[16]</sup>,引入数据的类别作为监督信息,能够提供显式的线性映射,在提取高维数据的特征时能够达到较好的效果.但是 MFA 是全监督算法,而实际应用中获取训练样本数据类别标签的代价较大,因此限制了其应用.由此,半监督学习应运而生,它同时利用无标记数据和有标记数据,只需较少的人工参与就能获得更高的学习精度,成为解决此类问题的发展趋势.

传统的最近邻分类算法(Nearest-Neighbor,简称 NN)已被广泛应用于解决图像分类的问题中.在特征空间里,它通过度量新样本与每个已知类别的训练样本之间的距离(一般利用欧氏距离或马氏距离等)或相似度进行判别分析,以确定查询对象所属的类别.但是面对多分类问题时,数据的分布特征不确定,使得仅通过搜索到的单个最近邻点所获得的分类结果常常不准确.针对这个问题,Fu 等提出了  $k$ NNS 算法( $k$ -Nearest-Neighbor Simplex,简称  $k$ NNS)<sup>[17,18]</sup>,利用同类别邻域内多个数据点的信息进行分类鉴别,在人脸识别问题中取得了较好的效果.

针对上述问题,本文提出了一种基于半监督边际费希尔分析(Semi-Supervised Marginal Fisher Analysis,简称 SSMFA)和  $k$ NNS 的高光谱遥感影像分类算法,该方法利用 SSMFA 将高光谱数据从高维观测空间投影到低维空间,以提取高光谱数据的鉴别特征,然后通过  $k$ NNS 分类器对低维空间的数据实现分类.在 Urban、Washington 和 Indian Pine 高光谱影像数据集上的实验结果表明本文算法能够较为有效地解决高光谱遥感影像的分类问题.

## 2 算法介绍

### 2.1 半监督边际费希尔分析算法(SSMFA)

MFA 算法是以图嵌入框架为基础,对数据点的分布不作任何假设,而是通过近邻数据点之间的几何关系挖掘数据分布的统计特性,因此 MFA 可以看作是基于局部几何关系的线性鉴别分析算法.

在 LDA 算法中如果在同类数据中存在少数相距较远的样本(多模分布),会造成 Fisher 鉴别准则因类内散布矩阵奇异而无法求解,使得 LDA 算法性能下降.而 MFA 重新定义了描述类内数据紧密性和类间数据分离性的准则,类内紧密性利用同类别数据点与其近邻点之间的距离之和描述;类间分离性利用不同类别的边际数据点与其近邻点之间的距离描述.虽然 MFA 在一定程度上克服了 LDA 无法解决的样本多模分布问题,但是 MFA 是全监督算法,且需要调整的参数较多,本文提出的 SSMFA 算法是一种基于图嵌入的半监督学习方法,对只有少数样本有类别标记的情况和减少调整参数方面进行了改进.

假设高维欧氏空间  $R^D$  中一组含  $N$  个样本的数据集为  $\mathbf{X} = \{x_1, \dots, x_N\}$ ,SSMFA 方法希望找到一个映射函数  $\mathbf{Y} = \mathbf{X}^T \mathbf{A}$  将  $X$  投影到低维空间  $R^d (d \ll D)$  中,得低维构造点  $\mathbf{Y} = \{y_1, \dots, y_N\}$ .SSMFA 构建了两种图,在邻域内如果两个样本点属于同类或近邻,则连接起来构成本征图  $G_i$ ;如果两个样本点属于不同的类,则连接起来构成惩罚图  $G_p$ ,以保持样本点的类间关系.其可分性准则为:将同类或近邻的数据点拉近,同时将不同类的数据点推远.SSMFA 首先采用  $k$ -nn 算法对每个数据点  $x_i (i = 1, \dots, N)$  计算其  $k$  近邻  $x_1^i, \dots, x_k^i$ ,并定义点  $x_i$  与近邻点  $x_j$  之间的权值为  $\mathbf{W}_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$ ,与其他点的权值为  $\mathbf{W}_{ij} = 0$ .

① 通过以下目标函数使同类或近邻数据点保持紧密:

$$\begin{aligned} & \arg \min \sum_{j \text{ 与 } i \text{ 在 } G_i \text{ 中相连}} \|y_i - y_j\|^2 \mathbf{W}_{ij}^l \\ & = \arg \min \sum_{j \text{ 与 } i \text{ 在 } G_i \text{ 中相连}} \text{tr}(y_i^2 + y_j^2 - 2y_i y_j) \mathbf{W}_{ij}^l \\ & = \arg \min \text{tr}(2\mathbf{Y}^T \mathbf{D}^l \mathbf{Y} - 2\mathbf{Y}^T \mathbf{W}^l \mathbf{Y}) \\ & = \arg \min \text{tr}\{2\mathbf{Y}^T (\mathbf{D}^l - \mathbf{W}^l) \mathbf{Y}\} \\ & = \arg \min \text{tr}\{2\mathbf{a}^T \mathbf{X} (\mathbf{D}^l - \mathbf{W}^l) \mathbf{X}^T \mathbf{a}\} \\ & = \arg \min \text{tr}(2\mathbf{a}^T \mathbf{X} \mathbf{L}^l \mathbf{X}^T \mathbf{a}) \end{aligned} \quad (1)$$

$$\mathbf{W}_{ij}^l = \begin{cases} \beta \cdot \mathbf{W}_{ij}, & x_i \text{ 与 } x_j \text{ 属于相同的类别} \\ \mathbf{W}_{ij}, & x_i \text{ 与 } x_j \text{ 近邻} \\ 0, & \text{其他} \end{cases} \quad (2)$$

式中  $\text{tr}$  表示矩阵的迹,  $\mathbf{W}^l$  是本征图  $G_i$  对应的权重矩阵,  $\mathbf{D}^l$  是对角矩阵,  $\mathbf{D}_{ii}^l = \sum_j \mathbf{W}_{ij}^l$ ,  $\mathbf{L}^l = \mathbf{D}^l - \mathbf{W}^l$  是 Laplacian 矩阵,  $\mathbf{a}$  为待求的投影向量.系数  $\beta$  为大于 1 的常数.

② 通过以下目标函数使不同类数据点保持远离:

$$\arg \max \sum_{j \text{ 与 } i \text{ 在 } G_p \text{ 中相连}} \|y_i - y_j\|^2 \mathbf{W}_{ij}^p$$

$$\begin{aligned}
&= \arg \max \sum_{j \text{ 与 } i \text{ 中相选}} \text{tr}(y_i^2 + y_j^2 - 2y_i y_j) \mathbf{W}_{ij}^p \\
&= \arg \max \text{tr}(\mathbf{2Y}^T \mathbf{D}^p \mathbf{Y} - \mathbf{2Y}^T \mathbf{W}^p \mathbf{Y}) \\
&= \arg \max \text{tr}\{\mathbf{2a}^T \mathbf{X}(\mathbf{D}^p - \mathbf{W}^p) \mathbf{X}^T \mathbf{a}\} \\
&= \arg \max \text{tr}(\mathbf{2a}^T \mathbf{X} \mathbf{L}^p \mathbf{X}^T \mathbf{a}) \quad (3)
\end{aligned}$$

$$\mathbf{W}_{ij}^p = \begin{cases} \mathbf{W}_{ij}, & x_i \text{ 与 } x_j \text{ 属于不同的类别} \\ 0, & \text{其他} \end{cases} \quad (4)$$

式中  $\mathbf{W}^p$  是惩罚图  $G_p$  对应的权重矩阵,  $\mathbf{D}^p$  是对角矩阵,  $\mathbf{D}_{ii}^p = \sum_j \mathbf{W}_{ij}^p$ ,  $\mathbf{L}^p = \mathbf{D}^p - \mathbf{W}^p$  是 Laplacian 矩阵.

SSMFA 算法的目标函数定义为:

$$\mathbf{a}^* = \arg \min_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{X} \mathbf{L}^p \mathbf{X}^T \mathbf{a}}{\mathbf{a}^T \mathbf{X} \mathbf{L}^p \mathbf{X}^T \mathbf{a}} \quad (5)$$

利用 Lagrange 乘子法,上述最优化问题可以转化为以下方程的特征值求解:

$$\mathbf{X} \mathbf{L}^p \mathbf{X}^T \mathbf{a} = \lambda \mathbf{X} \mathbf{L}^p \mathbf{X}^T \mathbf{a} \quad (6)$$

解得式(6)的  $d$  个最大特征值  $\lambda = \{\lambda_0, \lambda_1, \dots, \lambda_{d-1}\}$  所对应的特征向量  $\{\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_{d-1}\}$  即为所求的从高维空间到低维空间的映射  $\mathbf{A} = \{\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_{d-1}\}$ .

## 2.2 $k$ NNS 算法

传统的 NN 算法在解决分类问题中应用广泛,它通过度量新样本与单独的最近邻点之间的距离判别样本所属的类别.但是对于多分类问题,数据分布不规则,仅依靠单个近邻点得到的分类结果的准确性下降.而在样本数据充足的情况下,每个数据点与其近邻点可以构成或近似构成一个局部线性的欧氏空间.并且 Roweis 和 Saul 在《科学》杂志中提出了用近邻点的线性重构系数表征数据点邻域内的几何结构.在此基础上得出了  $k$ NNS 算法,利用样本与其邻域内多个近邻点的重构误差衡量样本与对应类别的吻合程度,减少了因单个不同类近邻点引起的误分,比 NN 算法有更好的鲁棒性.

对于数据集  $\mathbf{X} = \{x_i: x_i \in R^D\}_{i=1}^N$ ,在局部空间内,  $k$ NNS 用数据点  $x_0$  的  $k$  个局部近邻点的线性重构描述  $x_0$ ,

$$S(x_1, x_2, \dots, x_k) = \left\{ \sum_{i=1}^k \omega_i x_i \mid \sum_{i=1}^k \omega_i = 1, \omega_i \geq 0 \right\} \quad (7)$$

$S(x_1, x_2, \dots, x_k)$  表示  $x_0$  的重构函数,  $\omega_i$  表示近邻点  $x_i$  的权重系数.则数据点  $x_i$  与其近邻重构结果之间的重构误差  $d_{k\text{NNS}}$  可通过如下目标函数获得:

$$d_{k\text{NNS}} = \arg \min \left\| x_i - \sum_{j=1}^k \omega_{N(j)}^{(i)} x_{N(j)}^{(i)} \right\| \quad (8)$$

$\{N(j)\}_{j=1}^k$  表示与  $x_i$  近邻的  $k$  个样本点的索引值.重构误差包含了待分类样本与同类或非同类样本在分布上的距离关系,重构误差越小,表明样本与该类数据的距离越近,与该类数据的相似程度也越高,其属于此类别

的可能性就越大.由此解决多分类问题.

假设  $\mathbf{L} = \{l_i: l_i \in \{1, 2, \dots, q\}\}_{i=1}^N$  为  $X$  对应的类别标签,共含有  $q$  种类别.对于新样本数据  $x_i$ ,首先从每类数据中通过计算获得与之近邻的  $k$  个数据点的集合  $\{x_{N(j)}^{(l)}: l \in \{1, 2, \dots, q\}\}_{j=1}^k$ ,然后利用式(8)计算样本  $x_i$  与每类  $k$  近邻数据集之间的重构误差(或“距离”)  $\{d_{k\text{NNS}}^{(l)}: l \in \{1, 2, \dots, q\}\}$ ,最后利用以下目标函数即可获得样本  $x_i$  所属的类别  $l_i$ :

$$l_i^* = \arg \min_l d_{k\text{NNS}}^{(l)} \quad (9)$$

相比传统的 NN 算法,  $k$ NNS 分类方法的优越性可以二分类为例进行说明.如图 1 所示,位于两个三角形中间的点 P 属于圆点类,  $k$ NNS 方法分别用两类中与 P 点近邻的数据点线性重构点 P, P 点与方块类的距离(或“重构误差”)为  $a$ ,与圆点类的重构误差为  $b$ ,因为  $b < a$ ,所以将 P 归为重构误差小(或“距离近”)的圆点类.而在简单利用欧氏距离分类的 NN 方法中, P 与方块类近邻点的最短距离为  $a$ ,与圆点类近邻点的最短距离为  $c$ ,  $a < c$ ,导致将 P 误分成方块类.

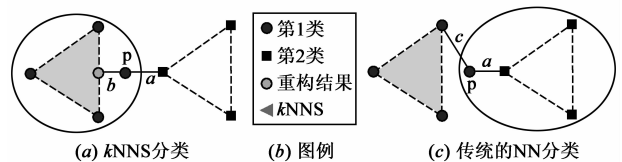


图1  $k$ NNS 与传统 NN 方法的对比

## 2.3 算法步骤

表 1 给出了本文所提出的算法的具体实现步骤.

表 1 本文算法的实现步骤

输入: 高维数据集  $\mathbf{X}$  和对应的类别标签  $\mathbf{L}$

输出: 从高维空间到低维空间的映射函数  $\mathbf{A}$ , 低维空间构造点  $\mathbf{Y}$

- (1) 从数据集中随机选取一定数量的训练样本和测试样本;
- (2) 构建训练样本点的权重矩阵  $\mathbf{W}$ ;
- (3) 利用 SSMFA 定义的  $\mathbf{W}^l$  和  $\mathbf{W}^p$  构建类内近邻图和类间远离图;
- (4) 利用式(6)边际鉴别准则计算从高维观测空间到低维空间的映射函数;
- (5) 利用映射函数将训练样本和测试样本投影到低维空间,获得低维空间构造点;
- (6) 利用  $k$ NNS 方法对测试样本数据进行分类

## 3 实验与分析

为了验证本文算法的有效性,采用了 Urban、Washington 和 Indian Pine 数据进行实验,并将本文算法和 PCA、LDA、LPP、NPE 以及 MFA 进行比较.

### 3.1 Urban 数据集

Urban 数据集是由 HYDICE 传感器获取的高光谱影像,空间分辨率为 3m,覆盖了某城郊居住区.该影像大小为  $307 \times 307$  像素,波谱范围从 400nm 到 2500nm,共包

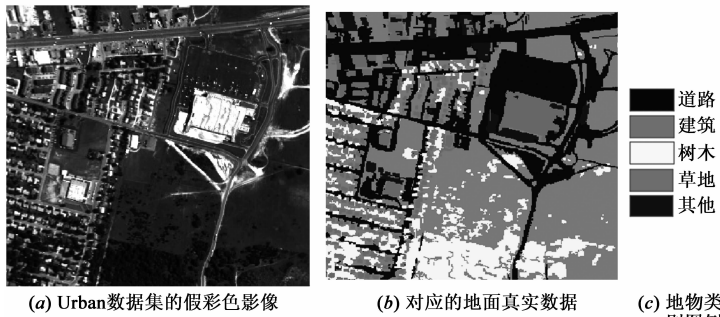


图2 Urban数据集的地物影像和与之对应的地面真实数据

含 210 个波段.图 2 给出了实验所使用的高光谱影像和与之对应的地物类别信息.剔除 48 个(1~4,76,87,101~111,136~153,198~210)受大气吸收影响的波段,将剩下的 162 个波段用于实验研究.本实验分别从“道路”、“建筑”、“树木”和“草地”等 4 类已知地物中随机选取 14 (其中 4 个有类别标记,10 个没有),16(其中 6 个有类别标记,10 个没有),18(其中 8 个有类别标记,10 个没有)个数据点作为训练样本,100 个数据点作为测试样本,分别以 4 labels、6 labels、8 labels 表示,每种方法都重复实验 10 次,最终的分类精度取 10 次实验的平均值.

实验设置如下:分别用 PCA、LDA、LPP、NPE 以及 MFA 对 Urban 数据集进行降维,然后利用传统的  $k$  近邻分类器进行分类,为了比较  $k$ NN 和  $k$ NNS 的分类效果,采用 MFA 降维利用  $k$ NNS 分类,同时采用本文的 SSMFA 算法降维并用  $k$ NNS 进行分类.图 3 给出了采用各种方法对高维数据降维到不同维数下进行分类的总体分类精度.表 2 给出了各种方法在各维数下总体精度的最高值.本文还以 6 labels 的数据为例利用各种算法将高维数据降维到 30 维(LDA 降维到 3 维)并进行分类,分析其对每类地物的分类效果.表 3 给出了各算法对 6 labels 的数据在 30 维下(LDA 为 3 维)对每类地物分类的分类精度.

由图 3、表 2 和表 3 可得,所有算法随着训练样本数量的增加,其总体分类精度也随之增加,因为训练样本的数量越多,能够提供的具有代表性的先验信息就越多,就越能表征数据的整体分布情况,通过各种算法学习得到的样本鉴别特征就更有效,进而提高分类精度.PCA 和 LDA 的分类效果优于 LPP 和 NPE 算法,其原因是 LPP 和 NPE 只利用了训练数据的局部近邻结构信息,忽略了数据的整体结构信息.而 PCA 算法利用了所有训练样本的整体分布信息,LDA 算法为监督算法,利用所有训练样本的类别信息提取鉴别特征,所以这两种方法的识别性能反而更好.

表 2 各种方法分类的最高总体精度(%)

算法	最高总体精度(%)		
	4 labels	6 labels	8 labels
PCA + $k$ NN	71.8	74.5	76.8
LDA + $k$ NN	71.8	74.5	77.0
LPP + $k$ NN	68.0	72.3	75.5
NPE + $k$ NN	69.5	72.0	75.3
MFA + $k$ NN	74.8	77.0	79.5
MFA + $k$ NNS	77.0	79.5	81.3
SSMFA + $k$ NNS	78.8	81.5	83.8

表 3 各种方法在 30 维下的每类分类精度和总体精度(%)

算法	每一类的分类精度(%)				总体精度(%)
	道路	建筑	树木	草地	
PCA + $k$ NN	75.0	72.0	81.0	70.0	74.5
LDA + $k$ NN(3)	72.0	77.0	81.0	68.0	74.5
LPP + $k$ NN	69.0	75.0	75.0	65.0	71.0
NPE + $k$ NN	65.0	69.0	80.0	66.0	70.0
MFA + $k$ NN	80.0	79.0	80.0	68.0	76.8
MFA + $k$ NNS	78.0	80.0	82.0	72.0	78.0
SSMFA + $k$ NNS	84.0	83.0	85.0	73.0	81.3

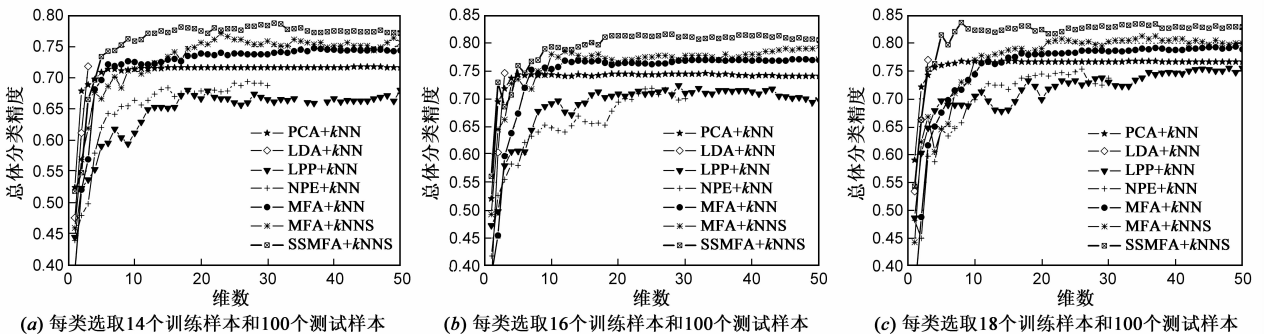


图3 采用不同的方法在不同的维数下对Urban数据集进行分类的总体分类精度

MFA 通过近邻样本之间的几何关系表征数据的分布特征,利用样本的类别信息,在拉近同类样本的同时推远异类样本,在一定程度上克服了 LDA 无法解决的样本多模分布问题,因此在都采用传统  $k$  近邻分类的

情况下,MFA 算法的分类精度高于其他算法.但  $k$  近邻分类仅利用单个最近邻点的信息,而  $k$ NNS 度量样本点与其邻域内多个数据点的相似度,减少了因少数异类数据点引起的误分,所以采用 MFA 降维  $k$ NNS 分类比

MFA 降维  $k$ NN 分类的分类精度高. MFA 是全监督算法, 对无类别标签的数据不具有很好的泛化性. 在有类别标记的样本点很少的情况下, 大量无类别标记的样本点也包含对分类有用的信息, SSMFA 不但利用了样本空间的判别信息和局部几何结构, 而且能够同时利用少量有标记数据和大量无标记数据帮助估计数据的流形结构, 充分利用了数据中的信息, 保持了数据的分布结构信息, 提高了算法的性能, 获得更高的学习精度, 因此本文算法在各种情形下的总体分类精度都是最高的, 且对每类的分类精度也优于其他算法.

在 4 个样本有类别标记 10 个样本没有类别标记的情况下, 本文算法比 MFA +  $k$ NNS 的总体分类精度提高了约 1.8%, 比 MFA +  $k$ NN 提高了约 4%, 比其他算法提高了约 7.0%; 在 6 个样本有类别标记 10 个样本没有类别标记的情况下, SSMFA +  $k$ NNS 的总体分类精度比 MFA +  $k$ NNS 提高了约 2.0%, 比 MFA +  $k$ NN 提高了约 4.5%, 比其他算法提高了约 7.0%; 在 8 个样本有类别标记 10 个样本没有类别标记的情况下, 本文算法比 MFA +  $k$ NNS 的总体分类精度提高了约 2.5%, 比 MFA +  $k$ NN 提高了约 4.3%, 比其他算法提高了约 6.8%.

### 3.2 Washington 数据集

Washington 数据集也是由 HYDICE 传感器获取的高光谱影像, 获取时间为 1995 年 8 月 23 日, 空间分辨率为 4m, 覆盖了华盛顿特区国家广场的局部. 该影像大小为  $1280 \times 307$  像素, 波谱范围从 400nm 到 2400nm, 共包含 210 个波段, 覆盖了建筑、道路、小路(广场中的砾石小径)、草地、树木和水体等 6 类确定类别的地物. 图 4 给出了实验所使用的高光谱影像和与之对应的地面真实数据. 剔除 19 个(103 ~ 106, 138 ~ 148, 207 ~ 210)受大气吸收影响的波段, 将剩下的 191 个波段用于实验研究. 本实验分别从 6 类已知地物中随机选取 14, 16, 18 个数据点作为训练样本, 100 个数据点作为测试样本, 每种方法都重复实验 10 次, 最终分类精度取 10 次实验的平均值. 图 5 给出了采用各种方法对高维数据降维到不同维数下进行分类的总体分类精度. 表 4 给出了各

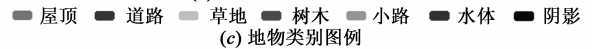
种方法在各维数下总体精度的最高值. 表 5 给出了各算法对 6 labels 的数据在 30 维下(LDA 为 5 维)对每类地物分类的分类精度, 表中 C1-C6 分别对应建筑-水体等 6 类地物.



(a) 华盛顿国家广场数据集的假彩色影像



(b) 对应的地面真实数据



(c) 地物类别图例

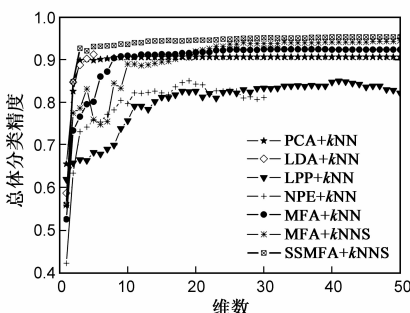
图 4 华盛顿国家广场数据集的地物影像和与之对应的地面真实数据

表 4 各种方法分类的最高总体精度 (%)

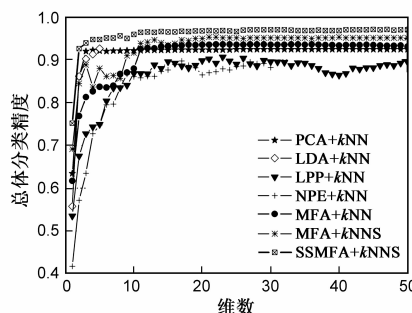
算法	最高总体精度 (%)		
	4 labels	6 labels	8 labels
PCA + $k$ NN	90.8	92.5	94.3
LDA + $k$ NN	91.3	92.5	93.8
LPP + $k$ NN	85.0	90.8	92.0
NPE + $k$ NN	85.0	89.8	91.5
MFA + $k$ NN	92.5	93.5	96.0
MFA + $k$ NNS	94.5	95.3	97.3
SSMFA + $k$ NNS	95.3	97.3	99.3

表 5 各种方法在 30 维下的每类分类精度和总体精度 (%)

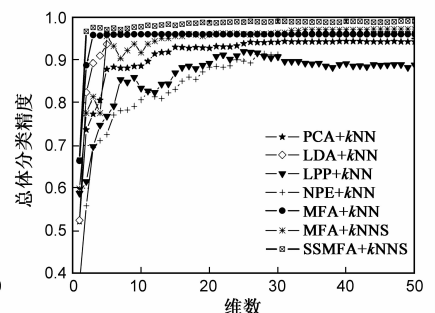
算法	每一类的分类精度 (%)						总体精度
	C1	C2	C3	C4	C5	C6	
PCA + $k$ NN	75.0	98.0	95.0	97.0	93.0	97.0	92.5
LDA + $k$ NN(5)	72.0	97.0	98.0	99.0	94.0	95.0	92.5
LPP + $k$ NN	76.0	92.0	91.0	93.0	94.0	91.0	89.5
NPE + $k$ NN	83.0	89.0	92.0	90.0	89.0	87.0	88.3
MFA + $k$ NN	81.0	88.0	99.0	96.0	99.0	95.0	93.0
MFA + $k$ NNS	82.0	99.0	98.0	99.0	97.0	95.0	95.0
SSMFA + $k$ NNS	84.0	100	100	99.0	100	99.0	97.0



(a) 每类选取 14 个训练样本和 100 个测试样本



(b) 每类选取 16 个训练样本和 100 个测试样本



(c) 每类选取 18 个训练样本和 100 个测试样本

图 5 采用不同的方法在不同的维数下对华盛顿国家广场数据集进行分类的总体分类精度

从图 5、表 4 和表 5 可以看出,高光谱数据处在一个具有较低维数的流形上,随着训练样本数量的增加,从样本中获得的先验信息就越全面,就越能够有效地展开数据的流形结构,发现高维数据分布的内在规律,各算法的总体分类精度也随之升高.而 SSMFA +  $k$ NNS 算法能够充分利用少量有类别标记样本和大量无标记样本发现高光谱数据的低维流形结构,使得其分类效果优于其他算法,在 4,6,8 个样本有类别标记 10 个样本没有类别标记的情况下,其总体分类精度比 MFA +  $k$ NNS 分别提高了约 0.8%,2.0%,2.0%,比 MFA +  $k$ NN 分别提高了约 2.8%,3.8%,3.3%,比其他算法分别提高了约 4.0%,4.8%,5.0%,且对 6 labels 数据中 C2、C3 和 C5 的分类精度达到了 100%.

### 3.3 Indian Pine 数据集

Indian Pine 数据集是由美国国家航空航天局在 1992 年 6 月 12 日通过 AVIRIS 传感器获取的美国 Indiana 州西北 100km<sup>2</sup> 区域的高光谱影像,其影像大小为

145 × 145 像素,波谱范围从 400nm 到 2450nm,共包含 220 个波段,覆盖了 16 类确定类别的地物.图 6 给出了实验所使用的高光谱影像和与之对应的地面真实数据.将第 104 至 108、第 150 至 163 以及第 220 波段等 20 个波段剔除,以消除大气吸收的影响,剩下的 200 个波段用于实验研究.本实验从地物数据点较多且类与类之间波段相关度较小的类别中选出了 4 类: Corn-notill、Corn、Grass/Trees、Woods,表 6 给出了所选地物之间的波段相关度.从这 4 类地物中随机选取 14,16,18 个数据点作为训练样本,100 个数据点作为测试样本,每种方法都重复实验 10 次,最终的分类精度取 10 次实验的平均值.图 7 给出了采用各种方法对高维数据降维到不同维数下进行分类的总体分类精度.表 7 给出了各种方法在各维数下总体精度的最高值.表 8 给出了各算法对 6 labels 的数据在 30 维下(LDA 为 3 维)对每类地物分类的分类精度.



图 6 Indian Pine 数据集的地物影像和与之对应的地面真实数据

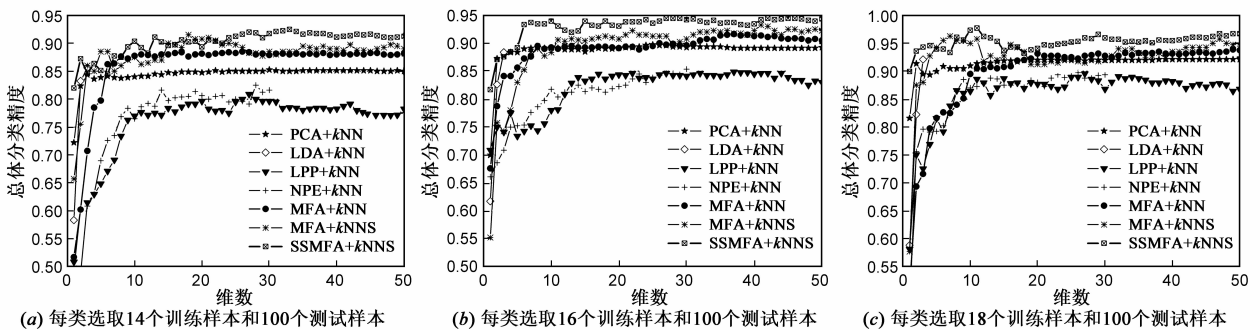


图 7 采用不同的方法在不同的维数下对 Indiana 数据集进行分类的总体分类精度

由图 7、表 7 和表 8 可得,由于选择的是相关度较低的 4 类地物,所以各算法的分类效果比对所有类别地物分类的效果好.Grass/Trees 和 Woods 都能够被较好地区分,但是 Corn-notill 和 Corn 的分类精度较低.SSMFA +  $k$ NNS 算法在有标记样本很少的情况下,还可以利用大量无标记样本中的信息估计数据的流形结构以提高分类精度.本文算法的总体分类精度最高,比 MFA +  $k$ NNS 分别提高了约 1.0%,1.2%,1.5%,比 MFA +  $k$ NN 分别

提高了约 4.0%,3.0%,4.0%,比其他算法分别提高了约 7.0%,4.7%,5.8%.

表 6 所选地物之间的波段相关度

	Corn-notill	Corn	Grass/Trees	Woods
Corn-notill	1	0.8953	0.3255	0.4380
Corn	0.8953	1	0.0539	0.5682
Grass/Trees	0.3255	0.0539	1	0.6714
Woods	0.4380	0.5682	0.6714	1

表 7 各种方法分类的最高总体精度(%)

算法	最高总体精度(%)		
	4 labels	6 labels	8 labels
PCA + $k$ NN	85.3	89.8	92.0
LDA + $k$ NN	85.5	88.5	92.0
LPP + $k$ NN	80.8	84.8	89.5
NPE + $k$ NN	82.5	85.3	89.8
MFA + $k$ NN	88.5	91.5	93.8
MFA + $k$ NNS	91.5	93.3	96.3
SSMFA + $k$ NNS	92.5	94.5	97.8

表 8 各种方法在 30 维下的每类分类精度和总体精度(%)

算法	每一类的分类精度(%)				总体精度(%)
	Corn-notill	Corn	Grass/Trees	Woods	
PCA + $k$ NN	86.0	76.0	94.0	97.0	88.3
LDA + $k$ NN(3)	73.0	90.0	91.0	99.0	88.3
LPP + $k$ NN	71.0	73.0	94.0	98.0	84.0
NPE + $k$ NN	70.0	84.0	88.0	99.0	85.3
MFA + $k$ NN	86.0	75.0	96.0	99.0	89.0
MFA + $k$ NNS	84.0	87.0	94.0	98.0	90.8
SSMFA + $k$ NNS	86.0	89.0	99.0	98.0	93.0

## 4 总结与展望

本文通过对高光谱数据的维数约简和分类模型中常用算法的分析,提出了一种基于 SSMFA 和  $k$ NNS 的高光谱遥感影像分类算法.该方法将高维数据投影到低维空间,充分利用数据的类别信息,使同类数据点更加紧密而不同类数据点更加稀疏,并利用  $k$ NNS 分类模型,通过度量新样本与各类别邻域之间的距离以判定其所属类别.在 Urban、Washington 和 Indian Pine 数据集上的分类识别实验说明了本文的方法能够较为有效地解决高光谱遥感影像分类的问题.在随机选取 4,6,8 个有类别标记的样本 10 个无类别标记的样本的情况下,本文方法的总体分类精度能够比 MFA +  $k$ NNS 提高 0.8% ~ 2.5%,比 MFA +  $k$ NN 的分类精度提高 2.8% ~ 4.5%,比其他算法提高 4.0% ~ 7.0%.

但是本文的算法需要计算新样本与每个邻域的欧氏距离,计算复杂度偏大.而且当已知样本数量很少时,这种欧氏距离度量标准也可能会失效,如何解决这些问题,需要进一步研究.

## 参考文献

[1] 梁亮,杨敏华,李英芳.基于 ICA 与 SVM 算法的高光谱遥感影像分类[J].光谱学与光谱分析,2010,30(10):2724-2728.  
Liang Liang, Yang Min-hua, Li Ying-fang. Hyperspectral remote sensing image classification based on ICA and SVM algorithm [J]. Spectroscopy and Spectral Analysis, 2010, 30(10): 2724-2728. (in Chinese)

[2] 梁亮,刘志霄,杨敏华等.基于可见/近红外反射光谱的稻米品种与真伪鉴别[J].红外与毫米波学报,2009,28(5):353-356.  
Liang Liang, Liu Zhi-xiao, Yang Min-hua, et al. Discrimination of variety and authenticity for rice based on visual/near infrared reflection spectra [J]. Journal of Infrared and Millimeter Waves, 2009, 28(5): 353-356. (in Chinese)

[3] 杨国鹏,周欣,余旭初等.基于相关向量机的高光谱影像混合像元分解[J].电子学报,2010,38(12):2751-2756.  
Yang Guo-peng, Zhou Xin, Yu Xu-chu, et al. Relevance vector machine for hyperspectral imagery unmixing [J]. Acta Electronica Sinica, 2010, 38(12): 2751-2756. (in Chinese)

[4] 王志辉,丁丽霞.基于叶片高光谱特性分析的树种识别[J].光谱学与光谱分析,2010,30(7):1825-1829.  
Wang Zhi-hui, Ding Li-xia. Tree species discrimination based on leaf-level hyperspectral characteristic analysis [J]. Spectroscopy and Spectral Analysis, 2010, 30(7): 1825-1829. (in Chinese)

[5] 尹继豪,王艳,王义松.一种改进的高光谱图像中多小目标检测算法[J].电子学报,2010,38(9):1975-1978.  
Yin Ji-hao, Wang Yan, Wang Yi-song. A revised multi-target detection approach in hyperspectral image [J]. Acta Electronica Sinica, 2010, 38(9): 1975-1978. (in Chinese)

[6] 罗四维,赵连伟.基于谱图理论的流形学习算法[J].计算机研究与发展,2006,43(7):1173-1179.  
Luo Si-wei, Zhao Lian-wei. Manifold learning algorithms based on spectral graph theory [J]. Journal of Computer Research and Development, 2006, 43(7): 1173-1179. (in Chinese)

[7] M Turk, A Pentland. Eigenfaces for recognition [J]. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.

[8] P Belhumeur, J Hespanha, D Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 20(7): 711-720.

[9] Ma Li, M M Crawford, Tian Jin-wen. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(11): 4099-4109.

[10] Ma Li, M M Crawford, Tian Jin-wen. Generalized supervised local tangent space alignment for hyperspectral image classification [J]. Electronics Letters, 2010, 46(7): 497-498.

[11] S Roweis, L Saul. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326.

[12] J Tenenbaum, V D Silva, J C Langford. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(5500): 2319-2323.

[13] M Belkin, P Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation [J]. Neural Computation,

2003, 15(6): 1373 – 1396.

- [14] He Xiao-fei, Cai Deng, Yan Shui-cheng, et al. Neighborhood preserving embedding [A]. Proceedings of the 10th IEEE International Conference Computer Vision (ICCV'05)[C]. Beijing, 2005. 1208 – 1213.
- [15] He Xiao-fei, Yan Shui-cheng, Hu Yu-xiao, et al. Face recognition using Laplacianfaces [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 328 – 340.
- [16] Yan Shui-cheng, Xu Dong, Zhang Ben-yu, et al. Graph embedding and extensions: A general framework for dimensionality reduction [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1): 40 – 51.
- [17] Fu Yun, Yan Shui-cheng, T S Huang. Classification and feature extraction by simplexization [J]. IEEE Transactions on Information Forensics and Security, 2008, 3(1): 91 – 100.
- [18] Fu Yun, Yan Shui-cheng, T S Huang. Discriminant simplex analysis [A]. Proceedings of the IEEE Conference ICASSP [C]. Las Vegas, NV, 2008. 3333 – 3336.

#### 作者简介



王立志 男, 1984 年生于河北衡水. 2007 年获得工学学士学位, 现为重庆大学硕士研究生. 主要研究方向为模式识别, 图像处理.



黄 鸿(通讯作者) 男, 1980 年生于湖南新宁. 2008 年于重庆大学获得博士学位, 现为重庆大学硕士生导师. 主要研究方向为流形学习, 模式识别, 图像处理.

E-mail: hhuang@cqu.edu.cn