文章编号 1004-924X(2011)12-3025-09

半监督流形学习及其在遥感影像分类中的应用

黄 鸿*,秦高峰,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:为了有效利用已标记与未标记样本提高遥感影像分类精度,提出了一种新的半监督流形学习方法—半监督流形鉴别嵌入法(SSMDE)。该方法利用标记样本的类别信息构建类内图和类间图来表征样本数据的类别联系,并计算相应的权重矩阵;利用标记和未标记数据构建全局散度矩阵来表征数据的整体结构。在此基础上,通过优化目标函数得到投影矩阵,在保持特征空间中数据整体结构的前提下,使同类数据点之间保持近邻关系、不同类数据点的距离尽可能大。在人工数据集和遥感影像上的实验结果表明,SSMDE 分类率为 92.36%,且分类结果与政府统计数据之间的误差均小于 5%。该方法通过有效利用少量标记样本和大量无标记样本实现半监督学习,有效提高了遥感影像的分类精度。

关 键 词:遥感影像;土地分类;图像分类;特征提取;半监督流形学习
中图分类号:TP391.4;TP73 **文献标识码**:A **doi**:10.3788/OPE.20111912.3025

Semi-supervised manifold learning and its application to remote sensing image classification

HUANG Hong*, QIN Gao-feng, FENG Hai-liang

(*Key Laboratory of Optoelectronic Technique Systems, Ministry of Education, Chongqing University, Chongqing* 400044, *China*)
* *Corresponding author, E-mail: hhuang.cqu@gmail.com*

Abstract: To improve the remote sensing image classification accuracy by incorporating labeled and unlabeled samples, this paper proposes a new manifold learning method called Semi-supervised Manifold Discriminant Embedding (SSMDE). This method uses data point labels to construct two relational graphs, within-class graph and between-class graph, they then are taken to encode the class relation information indicated in the labeled data points and to construct two weighted matrices. The labeled and unlabeled data points are utilized to construct the total scatter matrix to describe all the data points. Finally, the projection matrix of SSMDE is obtained by solving an optimization problem. The SSMDE method can not only take into account the discriminant information of labeled data, but also preserve the global structure of all data points. The experimental results on both synthetic and remote sensing images show that the proposed method can achieve the classification accuracy of 92.32% and the error between the classification results by the SSMDE and the government statistics is less than

收稿日期:2011-04-06;**修订日期**:2011-06-21.
基金项目:国家自然科学基金资助项目(No. 61101168);重庆市科技攻关重点项目(No. CSTC2009AB2231);重庆市自然科学基金资助项目(No. CSTC2009BB2195)

5%, which demonstrates the effectiveness of SSMDE.

Key words: remote sensing; land classification; image classification; feature extraction; Semi-supervised Manifold Learning

1 引　言

随着遥感技术的不断进步与发展，遥感信息在科学研究和国民经济中的应用越来越受到各行各业的重视，遥感影像分类则是遥感应用中重要的信息处理手段之一。然而，由于“同物异谱”、“同谱异物”现象的存在，直接利用光谱反射(辐射)特性或影像亮度值提取地物类别，尤其对于两类反射特性相似的地物，导致分类精度下降[1]。因此，如何利用遥感影像中丰富的空间和光谱信息来提取鉴别特征并提高遥感影像分类的精度，已成为遥感影像研究中的热点和前沿问题。

在遥感影像分类研究中，已产生一系列理论和算法上的研究进展，主要可分为3类：(1)子空间方法。其目的是寻求一个能保持数据集在原始欧式空间所呈现的几何结构的线性子空间，如主成分分析(Principal Component Analysis，PCA)、线性判别分析(Linear Discriminant Analysis，LDA)、最大边界准则(Maximum Margin Criterion，MMC)[2]、局部保持投影(Locality Preserving Projections，LPP)[3]等；(2) 核方法。该方法将欧式空间的数据点映射到高维再生核希尔伯特空间，然后在高维核空间中采用表征理论找到子空间，如核主成分分析(Kernel PCA，KPCA)[4]、核线性判别分析(Kernel Discriminant Analysis，KDA)等[5-6]；(3) 流形学习方法。近年来研究发现遥感影像可由一些连续的变量来参数化，本质上属于低维子流形。C. M. Bachmann[7-8]等明确提出遥感影像数据可描述为低维嵌入空间上呈几何结构的流形。流形学习可有效发现遥感影像中的流形结构，如等距映射(ISOMAP)[9]等。

上述研究对遥感影像特征提取和分类进行了积极的探索。但是，目前应用于遥感影像分类中的流形学习要么是非监督学习方法，不能有效利用标记样本中的有用信息；要么是监督学习方法，要求标记所有的训练样本，成本过高，有时是不现实的。在实际问题中，人们常常面对着大量的未标记数据以及相对很少的有标记数据。因此，如何有效融合流形学习和半监督学习方法，从标记数据和未标记数据中学习出有用的知识来改善学习性能是近年来研究的热点。

L. Capobianco 和 A. Garzelli 等提出一种新的半监督核正交子空间投影方法[10]，并应用于高光谱数据目标检测中，取得了较好的效果。G. Camps-Vall 和 T. V. Bandos 等提出一种基于谱图理论的半监督学习方法[11]，但需通过迭代来得到低维嵌入。任广波和张杰等提出基于生成模型学习的遥感影像半监督分类方法[12]，同样需要通过递归计算的方式对分类器进行优化。杨伟和方涛等提出采用朴素贝叶斯的半监督学习遥感影像分类方法[13]，但初始训练样本选取严格。D. Cai 和 X. F. He 等提出一种新的半监督鉴别分析方法[14] (Semi-supervised Discriminant Analysis，SDA)，提高了图像检索的精度。Y. Q. Song 等构建了半监督学习的统一框架，并提出了半监督最大边界准则(Semi-supervised Maximum Margin Criterion，SSMMC)[15]和半监督流形保持嵌入 (Semi-supervised Sub-manifold Preserving Embedding，S^3MPE)算法[16]，在图像识别中取得了较好效果，但算法中需对多个参数进行选择和优化。

针对上述问题，本文提出一种融合半监督学习和流形学习的遥感图像特征提取方法-半监督流形鉴别嵌入法(Semi-Supervised Manifold Discriminant Embedding，SSMDE)，并应用于遥感影像分类。该算法通过将用户先验知识提供的标注信息融入半监督流形学习方法，不仅可以揭示隐藏在遥感影像中的低维流形结构，且在一定程度上融入了人的视觉解译信息，得到了一个结合了用户语义理解的遥感影像低维鉴别子流形特征，提高遥感影像分类精度。

2 相关工作

本文首先对 MMC 和 SSMMC 进行简要介

绍。为方便表述，首先介绍特征提取问题。

2.1　特征提取

假定数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\in\mathbf{R}^{m\times n}$ 采样自本征维数为 $d(d\ll m)$ 的低维子流形 $\boldsymbol{M}$，$l_i\in(1,2,\cdots,c)$ 表示样本 $\boldsymbol{x}_i$ 的类别标签。特征提取是指将样本数据从原始输入空间通过线性或非线性映射投影到一个低维特征空间，找出隐藏在高维观测数据中有意义的低维特征。

2.2　极大边界准则(MMC)

MMC 的目的是寻求一组最佳鉴别矢量进行投影变换，实现特征空间样本的类间散度最大、类内散度最小。其目标函数可表示为：

$$\boldsymbol{J}(\boldsymbol{A})=\arg\max_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I}}\operatorname{tr}(\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{S}_{\mathrm{b}}-\boldsymbol{S}_{\mathrm{w}})\boldsymbol{A}),\quad(1)$$

式中：tr(·)表述矩阵的迹，$\boldsymbol{I}$ 是单位阵，$\boldsymbol{S}_{\mathrm{b}}$ 表示类间散布矩阵，$\boldsymbol{S}_{\mathrm{w}}$ 为类内散布矩阵，具体定义如下：

$$\boldsymbol{S}_{\mathrm{b}}=\sum_{i=1}^{c}n_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^{\mathrm{T}},\quad(2)$$

$$\boldsymbol{S}_{\mathrm{w}}=\sum_{i=1}^{c}\sum_{j=1}^{n_i}(\boldsymbol{x}_i^j-\boldsymbol{\mu}_i)(\boldsymbol{x}_i^j-\boldsymbol{\mu}_i)^{\mathrm{T}},\quad(3)$$

式中：$\boldsymbol{\mu}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_i^j$ 表示第 i 类样本均值，$\boldsymbol{\mu}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$ 表示样本总体均值，$\boldsymbol{x}_i^j$ 表示第 i 类中第 j 个样本。

从式(1)中可得知，MMC 的思路是使基于特征空间的类间散度与类内散度之差最大化，且 MMC 为监督学习方法，要求标记所有的训练样本。

2.3　半监督极大边界准则(SSMMC)

SSMMC 旨在找到一个线性投影子空间，样本数据投影到特征空间后，不仅具有好的可分性，而且保持了样本的局部结构。SSMMC 的目标函数表示为：

$$\boldsymbol{J}(\boldsymbol{A})=\arg\max_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I}}\operatorname{tr}(\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{S}_b-\lambda_1\boldsymbol{S}_w-\lambda_2\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{A}),\quad(4)$$

式中：λ_1 与 λ_2 是一个常数，$0\leqslant\lambda_1,\lambda_2\leqslant1.0$。$\boldsymbol{L}$ 是归一化的图 Laplacian 矩阵，用来表征无标记数据的局部结构。下面对 $\boldsymbol{L}$ 的构建进行简单介绍。

$\boldsymbol{L}$ 是基于谱图理论的，即通过构建近邻图 G 来度量无标记数据之间的相似性。针对样本数据集 $\boldsymbol{X}$ 中的每个无类别标签的数据点 $\boldsymbol{x}_i$，找到其个最近邻点，记为 $\mathrm{N}(\boldsymbol{x}_i)$。考虑每一点 $\boldsymbol{x}_j$ 对 $\boldsymbol{x}_i$，若 $\boldsymbol{x}_j\in\mathrm{N}(\boldsymbol{x}_i)$，则连接图 G 中 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 两点。根据图 G，即可计算权重矩阵 $\boldsymbol{W}$，具体定义如下：

$$\boldsymbol{W}_{ij}=\begin{cases}\exp\left(-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sigma^2}\right) & \boldsymbol{x}_i\in\mathrm{N}(\boldsymbol{x}_j)\ \text{or}\ \boldsymbol{x}_j\in\mathrm{N}(\boldsymbol{x}_i)\\ 0 & \text{otherwise}\end{cases}.\quad(5)$$

根据权重矩阵 $\boldsymbol{W}$，即可计算 $\boldsymbol{L}$

$$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2},\quad(6)$$

式中：$\boldsymbol{D}$ 是对角矩阵，$\boldsymbol{D}_{ii}=\sum_j\boldsymbol{W}_{ij}$。

采用 Lagrange 函数，式(4)中的最优化问题可转换为广义的特征值求解问题

$$(\boldsymbol{S}_b-\lambda_1\boldsymbol{S}_w-\lambda_2\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{v}=\lambda\boldsymbol{v},\quad(7)$$

式(7)中对应的 d 个最大特征值 $\lambda=(\lambda_0,\lambda_1,\cdots,\lambda_{d-1})$ 对应的特征向量 $\mathbf{V}=(\boldsymbol{v}_0,\boldsymbol{v}_1,\cdots,\boldsymbol{v}_{d-1})$，即为所求的投影矩阵 $\mathbf{A}=(\boldsymbol{v}_0,\boldsymbol{v}_1,\cdots,\boldsymbol{v}_{d-1})$。

SSMMC 通过构建图 Laplacian 矩阵，有效地融入未标记数据的局部结构信息，实现了半监督学习，在一定程度上提升了分类精度[17-19]。但是，SSMMC 只考虑未标记数据的局部信息，忽略了全局结构信息。同时，存在多个参数(σ，λ_1，λ_2)，参数的优化选择较为困难。

3　半监督流形鉴别嵌入(SSMDE)

针对 SSMMC 中存在的忽视全局结构以及参数过多的问题，本文提出一种融合数据整体结构信息的半监督流形学习方法-SSMDE。该方法首先将多光谱遥感图像根据不同波段对地物的光谱反射特性，将遥感图像数据生成为 $m\times n\times b$ 的数据集，即 $\boldsymbol{X}_{\mathrm{orig}}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{m\times n\times b}\}^{\mathrm{T}}$，其中 $m\times n$ 是图像的空间尺寸，b 是波段数，然后将 $\boldsymbol{X}_{\mathrm{orig}}$ 转换为 2 维数据 $\boldsymbol{X}_{\mathrm{all}}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{t\times b}\}^{\mathrm{T}}$，其中 $t=m\times n$，b 是波段数。从 $\boldsymbol{X}_{\mathrm{all}}$ 中选取不同地物数据点作为训练样本集，然后通过构建类内图和类间图来表征标记样本中的鉴别地物类别信息。同时，利用样本数据来表征样本数据的整体结构，进而寻求一组最佳鉴别矢量进行投影变换，实现特征空间中不仅保持数据整体结构，而且同类地物数据点之间保持近邻关系、不同地物数据点的距离

尽可能大，从而达到提高分类精度的目的。

从数据集 $\boldsymbol{X}_{\text{all}}$ 中选取部分样本数据，根据先验知识标注不同地物样本类别标签，同时随机选取部分样本数据作为无标记数据，组成样本训练集 $\boldsymbol{X}$，即 $\boldsymbol{X}=\{(\boldsymbol{x}_1,l_1),(\boldsymbol{x}_2,l_2),\cdots,(\boldsymbol{x}_l,l_l),\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n\}$，$l_i$ 为对数据点 $\boldsymbol{x}_i$ 标注的地物类别标签，前 l 个点具有地物类别信息，其余的 $n-l$ 个为无类别标记样本。

在部分样本数据地物类别信息已知的前提下，SSMDE 通过样本数据集构建类内图 G_{w} 和类间图 G_{b} 来度量数据点之间的相似性。对数据集 $\boldsymbol{X}$ 中的每个数据点 $\boldsymbol{x}_i$，找到其 k 个最近邻点，记为 $\text{N}(\boldsymbol{x}_i)$。然后对每一点 x_i，近邻数据集 $\text{N}(\boldsymbol{x}_i)$ 可以分为两部分：$\text{N}_{\text{w}}(\boldsymbol{x}_i)$ 与 $\text{N}_{\text{b}}(\boldsymbol{x}_i)$，其中 $\text{N}_{\text{b}}(\boldsymbol{x}_i)$ 表示近邻点来自于不同类地物样本点，$\text{N}_{\text{w}}(\boldsymbol{x}_i)$ 则表示近邻点来自于同类地物样本点。考虑每一点 x_j 对 $\boldsymbol{x}_i$，若 $x_j\in\text{N}_{\text{w}}(\boldsymbol{x}_i)$，则连接图 G_{w} 中 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 两点；若 $\boldsymbol{x}_j\in\text{N}_{\text{b}}(\boldsymbol{x}_i)$，则用一条边连接图 G_{b} 中 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 两点。根据图 G_{w} 与 G_{b}，即可计算权重矩阵 $\boldsymbol{W}_{\text{w}}$ 和 $\boldsymbol{W}_{\text{b}}$，具体定义如下：

$$\begin{cases}\boldsymbol{W}_{\text{w},ij}=\begin{cases}1 & \boldsymbol{x}_i\in\text{N}_{\text{w}}(\boldsymbol{x}_j)\ \text{or}\ \boldsymbol{x}_j\in\text{N}_{\text{w}}(\boldsymbol{x}_i)\\ 0 & \text{otherwise}\end{cases}\\ \boldsymbol{W}_{\text{b},ij}=\begin{cases}1 & \boldsymbol{x}_i\in\text{N}_{\text{b}}(\boldsymbol{x}_j)\ \text{or}\ \boldsymbol{x}_j\in\text{N}_{\text{b}}(\boldsymbol{x}_i)\\ 0 & \text{otherwise}\end{cases}\end{cases}. \tag{8}$$

多光谱遥感图像数据投影后最理想的目标是使同类地物样本数据间散度尽可能小、不同类地物数据间散度尽可能大，基于上述考虑，寻求最佳鉴别矢量的目标函数可表示为：

$$\begin{cases}\min\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{\text{w},ij}\\ \max\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{\text{b},ij}\end{cases}. \tag{9}$$

但在遥感图像半监督学习问题中，已知地物类别的样本数据往往非常有限，直接以式(9)为目标函数得到的投影空间可能会导致“过训练”或“过拟合”问题，即投影空间对训练数据非常有效，但是对于测试样本其分类精度会急剧下降。因此，在利用已知地物类别的样本数据提取鉴别特征的同时，需要保持样本数据的整体结构，避免“过训练”的情形。本文中引入总体散布矩阵 $\boldsymbol{S}_t$ 来表征不同地物数据的整体信息，具体定义如下：

$$\boldsymbol{S}_t=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\text{T}}=\frac{1}{n}\boldsymbol{X}(\boldsymbol{I}-\frac{1}{n}\boldsymbol{e}\boldsymbol{e}^{\text{T}})\boldsymbol{X}^{\text{T}}, \tag{10}$$

式中：$\boldsymbol{e}=(1,1,\cdots,1)^{\text{T}}$。

$\boldsymbol{S}_t$ 不仅表征具有不同地物类别样本数据点之间的联系，而且融合了所有训练数据(包括标注数据和无标注数据)的分布信息，可在一定程度上弥补标注样本数据过少的问题。SSMDE 寻求实现特征空间中保持不同地物数据的整体结构，而且最小化同类地物样本点之间的类内散度、最大化不同类样本点之间的类间散度。因此，SSMDE 算法的特征空间可以通过以下的优化问题求得，即：

$$\begin{cases}\min\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{\text{w},ij}\\ \max\sum\limits_{ij}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{\text{b},ij}\\ s.\,t.\ \ \boldsymbol{A}^{\text{T}}\boldsymbol{S}_t\boldsymbol{A}=\boldsymbol{I}\end{cases}, \tag{11}$$

通过代数变换，式(11)可表示为

$$\begin{cases}\min\ tr(\boldsymbol{A}^{\text{T}}\boldsymbol{L}_{\text{b}}\boldsymbol{X}^{\text{T}}\boldsymbol{A})\\ \max\ tr(\boldsymbol{A}^{\text{T}}\boldsymbol{L}_{\text{w}}\boldsymbol{X}^{\text{T}}\boldsymbol{A})\\ s.\,t.\ \ \boldsymbol{A}^{\text{T}}\boldsymbol{S}_t\boldsymbol{A}=\boldsymbol{I}\end{cases}, \tag{12}$$

式中：$\boldsymbol{L}_{\text{w}}=\boldsymbol{D}_{\text{w}}-\boldsymbol{W}_{\text{w}}$ 为类内 Laplacian 矩阵，$\boldsymbol{L}_{\text{b}}=\boldsymbol{D}_{\text{b}}-\boldsymbol{W}_{\text{b}}$ 为类间 Laplacian 矩阵，$\boldsymbol{D}_{\text{w},ii}=\sum\limits_{j}\boldsymbol{W}_{\text{w},ij}$，$\boldsymbol{D}_{\text{b},ii}=\sum\limits_{j}\boldsymbol{W}_{\text{b},ij}$，皆为对角矩阵。

通过代数变换与采用 Lagrange 函数，式(12)中的最优化问题可转换为广义的特征值求解问题

$$\boldsymbol{X}(\boldsymbol{L}_{\text{b}}+\boldsymbol{S}_t)\boldsymbol{X}^{\text{T}}\boldsymbol{v}=\lambda\boldsymbol{X}\boldsymbol{L}_{\text{w}}\boldsymbol{X}^{\text{T}}\boldsymbol{v}, \tag{13}$$

式(13)中 d 个最大特征值 $\lambda=[\lambda_0,\lambda_1,\cdots,\lambda_{d-1}]$ 对应的特征向量 $\boldsymbol{V}=[\boldsymbol{v}_0,\boldsymbol{v}_1,\cdots,\boldsymbol{v}_{d-1}]$，即为所求的投影矩阵 $\boldsymbol{A}=[\boldsymbol{v}_0,\boldsymbol{v}_1,\cdots,\boldsymbol{v}_{d-1}]$。

SSMDE 算法的具体步骤如表 1 所示。

表 1　SSMDE 算法

Tab. 1　SSMDE algorithm

输入：N 个训练样本的数据集 $\boldsymbol{X}=\{(\boldsymbol{x}_1,l_1),(\boldsymbol{x}_2,l_2),\cdots,(\boldsymbol{x}_l,l_l),\boldsymbol{x}_{l+1},\cdots,\boldsymbol{x}_n\}$，$\boldsymbol{x}_i\in\boldsymbol{R}^m$，嵌入特征维数 d
输出：投影矩阵 $\boldsymbol{A}$，嵌入特征 $\boldsymbol{Y}$
1. 利用部分已知地物类别的标记数据集 $\boldsymbol{X}$ 构建类内图 G_{w} 和类间图 G_{b}；
2. 构造 G_{w} 的权重矩阵 $\boldsymbol{W}_{\text{w}}$ 和 G_{b} 的权重矩阵 $\boldsymbol{W}_b$；
3. 计算总体散布矩阵 $\boldsymbol{S}_t$，同类地物数据点间 Laplacian 矩阵 $\boldsymbol{L}_{\text{w}}$ 及不同地物数据点间 Laplacian 矩阵 $\boldsymbol{L}_{\text{b}}$；
4. 计算 d 维嵌入：求解式(13)特征方程中的 d 个最大特征值对应的特征向量，构成投影矩阵 $\boldsymbol{A}=[\boldsymbol{v}_0,\boldsymbol{v}_1,\cdots,\boldsymbol{v}_{d-1}]$，即可计算嵌入特征 $\boldsymbol{Y}=\boldsymbol{A}^{\text{T}}\boldsymbol{X}$。

4　实验与讨论

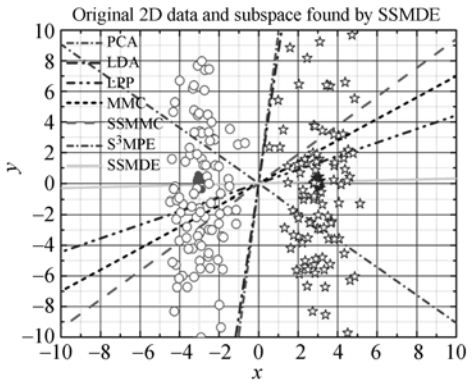
为了验证本文提出的 SSMDE 算法的有效性，分别在人工数据集和重庆市大渡口区 2000 年多光谱遥感影像数据集上进行实验，将 SSMDE 与 PCA、LDA、LPP、MMC、SSMMC、S^3MPE 进行比较。对于这些识别算法，调整它们的参数到最佳，实验使用最近邻分类器（K-NN）（$k=1$）完成最后的分类。

本文利用的多光谱数据来自 2000 年 7 月份美国陆地卫星 7 号（Landsat-7）获取的 ETM＋影像，首先根据地面控制点对遥感影像进行几何校正、辐射校正以及影像增强等数据预处理，去除影像中存在的部分噪声对图像分类的干扰。根据大渡口区的行政区划裁剪出本文需要的研究区域。

实验首先通过训练样本集计算得到特征空间的投影矩阵，然后对测试样本数据集进行投影，得到测试样本的低维特征，最后用 K-NN 分类器进行分类，并把分类后的各种地物的土地面积与当年统计局的统计数据相比较分析误差。

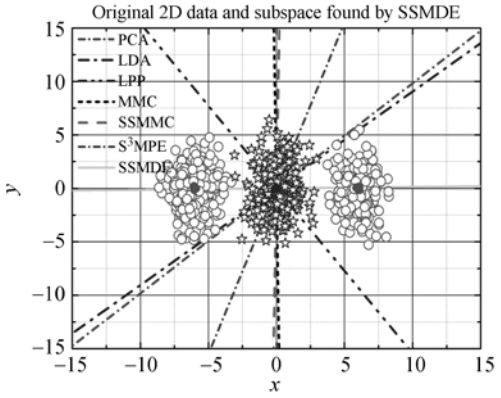
4.1　人工数据集（Synthetic datasets）实验

图 1(a)中表示第一个人工数据集，在这个数据集中，两类分别成高斯分布。图中的线表示每种算法把数据样本投影到一维空间上的投影面。结果表明，PCA 和 LDA 方法找到的投影面很接近，显然这两个投影面不能很好地把这两类分开，LPP 和 MMC，SSMMC，S^3MPE 方法比传统的 PCA 和 LDA 算法效果好很多，但也不能很好地把这两类分开，从图中可知 SSMDE 方法能很好地把这两类分开。



(a)合成数据集 1

(a)Synthetic data set 1



(b)合成数据集 2

(b)Synthetic data set 2

图 1　不同算法在人工数据集上的二维嵌入结果

Fig. 1　Two-dimensional embedding results of different algorithms on synthetic data

当同一类的样本聚集成多个群体时，如图 1(b)中第 2 个人工数据集，第一类成单高斯分布，第二类成两个分开的高斯分布。在此类数据集上 SSMDE 方法分类效果就更能凸显出来，它既考虑了标记数据的鉴别信息又考虑了全局结构信息在特征空间中不仅能保持数据整体结构，且能实现同类数据点之间保持近邻关系、不同类数据点的距离尽可能大。因此，在两个合成数据集上实验都验证了 SSMDE 算法的有效性效果。

4.2　遥感图像实际数据集实验

大渡口区 2000 年多光谱遥感影像包含 7 个波段，空间分辨率为 12.5 m，分为建筑、河流、耕地、湖泊、森林、疏叶林、绿地 7 类不同地物。在遥感影像中选取实验样本，根据先验知识从每类中选取 150 个数据点与另外随机选取的 700 个数据

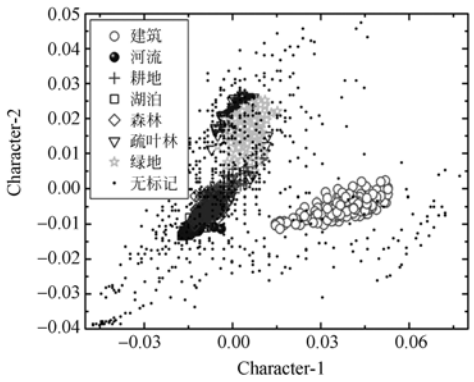


图 2　原始数据二维空间分布图

Fig. 2　Distribution of original data in two-dimensional space

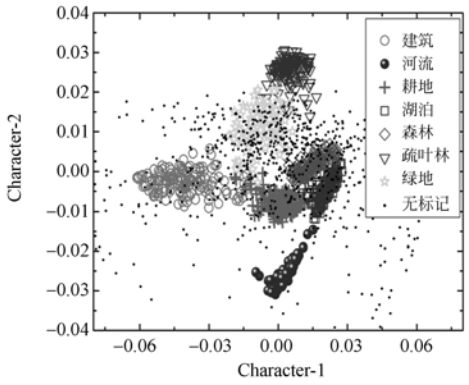


图 3　原始数据通过 SSMDE 二维嵌入结果

Fig. 3　Two-dimensional embedding results by SSMDE

点一起组成训练数据集。图 2 为直接利用训练数据集的前两维示意图，用来表示各类数据样本点在二维空间的分布情况。图 3 为采用 SSMDE 算法将训练数据集投影到二维嵌入空间的嵌入结果。

从图 2 和图 3 可得知，样本数据集在投影前不同类别之间出现不同程度的混叠，但经过 SSMDE 学习并投影到二维空间后，增大了类间距离，减小了类内距离，使不同类别样本之间的可分性显著增强，有利于分类器分类。

4.3　实验结果评价

实验的评价标准分为两部分，首先从有类别标签的训练样本中选取部分样本用于识别率测试，与传统算法的识别率进行比较，并同时计算各种算法的复杂度；其次，为了更好的验证 SSMDE 算法的性能，对整幅大渡口区遥感影像数据进行投影与分类，然后根据遥感影像的空间分辨率等地理空间信息得出不同地物的面积，把这些数值与 2000 年统计数据相比较计算误差率。

4.3.1　不同算法的识别率及算法复杂度的比较

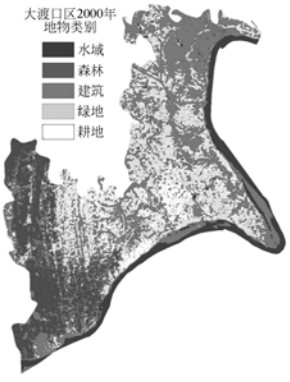
由于整幅遥感影像数据量过大，本文从每类地物中选取 100 个标记点，并随机选取的 700 个无类别点组成实验数据集。进行实验时，每次在每类样本数据中随机选取 80 个样本点作为训练集，剩下的作为测试集。为了检验各种算法的识别率和计算复杂度，本文对各种算法的识别率与运算时间进行了对比分析，使用的计算机配置为：2.91G 的 AMD Athlon Ⅱ X3 CPU，2G 的 DDR2 内存。每种方法的实验结果为 10 次实验的平均值，具体结果如表 2 所示。

表 2　在实际数据集上不同算法精度及复杂度比较(均值方差)

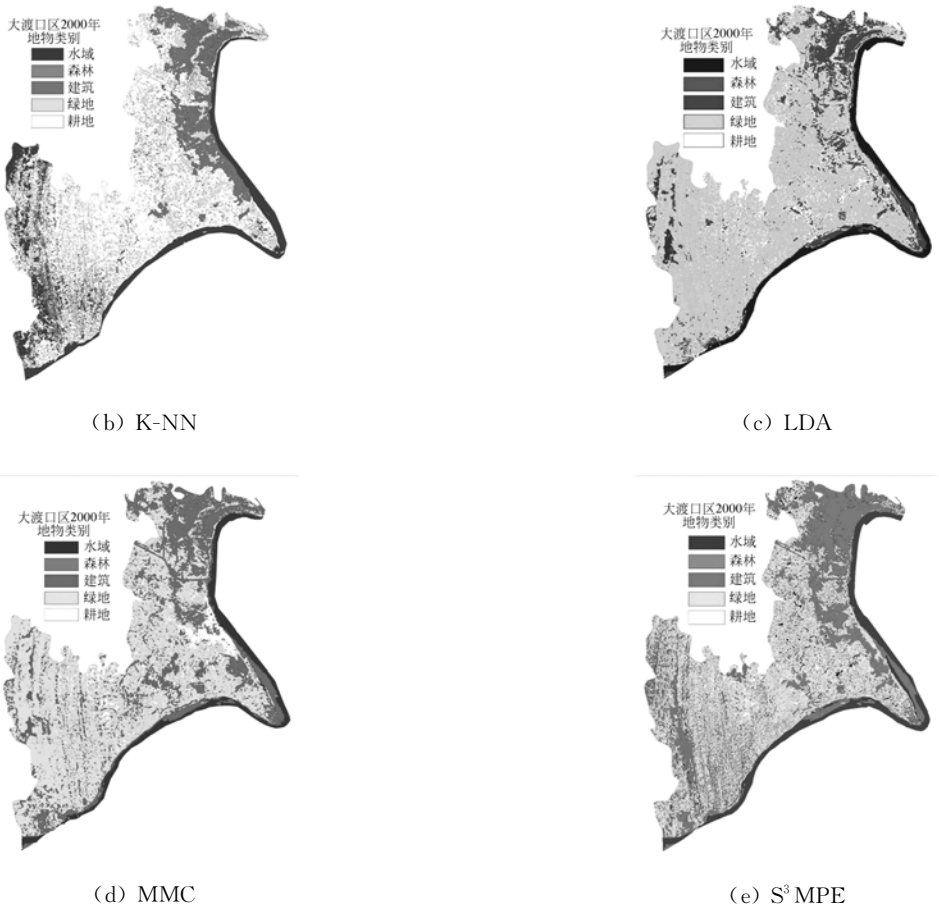
Tab. 2　Comparison of different methods on actual dataset (mean±std)

	算法	精度 /(%)	计算复杂度 /s
传统算法	K-NN	57.27±3.2	0.438
子空间法	PCA	66.36±3.7	0.844
	LDA	74.94±1.7	0.426
传统流形	LPP	80.62±3.1	0.859
学习法	MMC	75.11±2.3	0.828
	SSMMC	80.84±3.9	0.860
半监督流形学习	S^3MPE	85.23±2.7	0.844
	SSMDE	92.36±1.3	0.622

实验结果如表 2 所示，SSMDE 算法在识别性能上优于其它几种算法。通过对比分类器在与流形学习结合前后的分类精度看出，分类器的分类精度也得到了提高，效果没有本文提出的算法效果识别率高。因为 SSMMC 忽略了全局结构信息，分类的性能受到影响。SSMDE 算法在利用标记数据的类别信息的基础上，考虑了未标记数据的局部信息和全局结构信息，不会因有限的标记样本导致过训练的情形，所以 SSMDE 分类性能要优于 SSMMC。同时，SSMMC 和 S^3MPE 性能易受到参数的影响。本文提出的 SSMDE 算法在整个遥感影像的分类结果如图 4(a)所示，通过图 4 对比可以看出，本文算法分类精度效果总体要优于其他几个传统的算法。



(a) SSMDE



(b) K-NN　　(c) LDA

(d) MMC　　(e) S^3MPE

图 4　不同算法在 2000 年大渡口区上的分类结果图

Fig. 4　Classification results of different methods on Dadukou

4.3.2　分类结果总体分类精度与 Kappa 系数

本文分别从总体分类精度和 Kappa 系数两个方面对分类结果的精度进行评价。总体分类精度等于被正确分类的像元总和除以总像元数。被正确分类的像元沿着混淆矩阵的对角线分布,它显示出被分类到正确地表真实分类中的像元数。另一种计算分类精度的方法是 Kappa 系数,针对各种算法分类结果计算 Kappa 系数,该值越大说明分类精度越好。本文分类结果的总体分类精度和 Kappa 系数如表 3 所示。

表 3　分类精度评价

Tab. 3　Evaluation of classification accuracy

	总体精度/(%)	Kappa coefficient
K-NN	45.359	0.421 5
PCA	60.268	0.612 3
LDA	70.827	0.694 1
LPP	75.413	0.736 9
MMC	71.247	0.701 2
SSMMC	75.744	0.732 8
S^3MPE	80.028	0.778 2
SSMDE	91.268	0.884 5

4.3.3　分类结果与统计数据相比的误差率

把分类后的结果图转成矢量格式(矢量图)利于计算各类地形的面积,转成矢量图后,根据遥感影像的空间分辨率等地理空间信息得出不同土地类型面积如表 4,本文通过查找相关统计资料得出 2000 年底大渡口区总的行政区划面积以及大渡口区林地和耕地面积的统计数据见表 4。

表4　SSMDE获得的各土地类型面积与统计数据的误差

Tab. 4　Errors between statistic data and results by proposed method　(hm^2)

土地类型	遥感图像得出面积	2000年实际统计数据	误差率
建筑用地	3 785.682 16	—	—
森林用地	1 657.477 35	1 582.27	4.75%
河流	911.498 67	—	—
绿地用地	1 727.858 81	—	—
耕地用地	2 073.024 45	2 056	0.83%
总面积	10 155.541 44	10 300	1.40%

由于从重庆市及大渡口区统计年鉴中，只查到2000年大渡口区的耕地和林地以及行政区划面积的统计数据。因此，本文只对上面两种地形面积和总面积进行了精度分析。从表4可以看出，误差率均低于5%，尤其是耕地面积误差很小，为0.83%。可见本文提出的方法最终输出效果误差在实际应用中是可以接受的。

5　结　论

本文提出一种新的半监督流形学习算法-半监督流形鉴别嵌入法(SSMDE)。该方法通过构建类内图 G_w 和类间图 G_b 来表征标记数据之间的类别联系信息，并引入全局散度矩阵来表征标记与未标记数据的整体分布信息，然后通过优化目标函数来得到投影矩阵，在特征空间中不仅能保持数据整体结构，且能使同类数据点之间保持近邻关系、不同类数据点的距离尽可能大，从而有效提高分类精度。该算法既考虑了标记数据的鉴别信息又考虑了全局结构信息，避免了“过训练”情形。在2000年重庆市大渡口区进行的遥感影像实验表明，SSMDE算法的识别率为92.36%，比SSMMC提高了12%，且与统计数据之间的误差均小于5%。

目前土地分类误差评价只是单方向的，双向误分的情况没有考虑进去，如部分林地被错分成耕地，部分耕地被错分成林地，而在误差分析中只是单纯的考虑林地总的面积等。因此在下一步研究中要增强误差评价体系方面的研究。

参考文献：

[1] 赵英时. 遥感应用分析原理与方法[M]. 北京：科学出版社，2003.
ZHAO Y S. *The Principle and Method of Analysis of Remote Sensing Application* [M]. Beijing: Science Press, 2003. (in Chinese)

[2] WAN M H, LAI Z H, JIN Z. Feature extraction using two-dimensional local graph embedding based on maximum margin criterion[J]. *Applied Mathematics and Computation*, 2011, 217(23): 9659-9668.

[3] 董超，赵慧洁，王维，等. 采用局部正交子空间投影的高光谱图像异常检测[J]. 光学 精密工程，2009，17(8)：2004-2010.
DONG CH, ZHAO H J, WANG W, *et al.*. Hyper spectral image anomaly detection based on local orthogonal subspace projection [J]. *Opt. Precision Eng.*, 2009, 17(8): 2004-2010. (in Chinese)

[4] FORERO S V, ANGULO J, CHANUSSOT J. Morphological image distances for hyperspectral dimensionality exploration using Kernel-PCA and ISOMAP [J]. *Proceedings of* 2009 *IEEE International Geosci. Rem. Sens. Symposium*, 2009, 3(12): 109-112.

[5] 黄鸿，李见为，冯海亮. 基于有监督核局部线性嵌入的面部表情识别[J]. 光学 精密工程，2008，16(8)：1471-1477.
HUANG H, LI J W, FENG H L. Facial expression recognition based on supervised kernel local linear embedding[J]. *Opt. Precision Eng.*, 2008, 16(8): 1471-1477. (in Chinese)

[6] 李粉兰，唐文彦，段海峰，等. 分数次幂多项式核函数在核直接判别式分析中的应用[J]. 光学 精密工程，2007，15(9)：1410-1414.
LI F L, TANG W Y, DUAN H F, *et al.*. Application of fractional power polynomial kernel function to kernel direct discriminant analysis [J]. *Opt. Precision Eng.*, 2007, 15(9): 1410-1414. (in Chinese)

[7] BACHMANN C M, AINSWORTH T L, FUSINA R A. Bathymetric retrieval from hyperspectral imagery using manifold coordinate representations [J]. *IEEE Trans. Geosci. Rem. Sens*, 2009, 47(3):

884-897.

[8] BACHMANN C M, AINSWORTH T L, FUSINA R A, *et al.*. Manifold coordinates representations of hyperspectral imagery: Improvements in algorithm performance and computational efficiency [J]. *IEEE Trans. Geosci. Rem. Sens*, 2010, 7(12): 4244-4247.

[9] 叶强. 基于流形学习像素分布流的高光谱图像数据分割方法[D]. 西安：西安电子科技大学，2010.
YE Q. *A method of segmentation for hyperspectral image base on manifold learning pixel distribution flow*[D]. Xi'an: Xi'an University, 2010. (in Chinese)

[10] CAPOBIANCO L, GARZELLI A, CAMPS V G. Target detection with semi-supervised kernel orthogonal subspace projection [J]. *IEEE Trans. Geosci. Rem. Sens*, 2009, 47(11): 3822-3833.

[11] BANDOS T V, BRUZZONE L, CAMPS V G. Classification of hyperspectral images with regularized linear discriminant analysis [J]. *IEEE Trans. Geosci. Rem. Sens*, 2009, 47(3): 862-873.

[12] REN G B, ZHANG J, MA Y, *et al.*. Generative model based semi-supervised learning method of remote sensing image classification [J]. *Journal of Remote Sensing*, 2010, 14(6): 1090-1104.

[13] 杨伟，方涛，许刚. 基于朴素贝叶斯的半监督学习遥感影像分类[J]. 计算机工程，2010，36(20)：167-169.
YANG W, FANG T, XU G. Semi-supervised learning remote sensing image classification based on naive Bayesian [J]. *Computer Engineering*, 2010, 36(20): 167-169.

[14] CAI D, HE X F, HAN J W. Semi-supervised discriminant analysis[C]. *IEEE International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil*, 2007: 1-7.

[15] SONG Y Q, NIE F P, ZHANG C S, *et al.*. A unified framework for semi-supervised dimensionality reduction [J]. *Pattern Recognition*, 2008, 41(9): 2789-2799.

[16] SONG Y Q, NIE F P, ZHANG C S. Semi-supervised sub-manifold discriminant analysis [J]. *Pattern Recognition*, 2008, 29(13): 1806-1813.

[17] 黄鸿，李见为，冯海亮. 融合局部和全局结构的流形学习[J]. 光学 精密工程，2009，17(3)：626-632.
HUANG H, LI J W, FENG H L. Fusion of local and global structures for manifold learning [J]. *Opt. Precision Eng.*, 2009, 17(3): 626-632. (in Chinese)

[18] YANG Y, LI X, PAN Y, *et al.*. Binary sparse nonnegative matrix factorization[J]. *IEEE Trans on Circuits and Systems for Video Technology*, 2009, 19(5): 772-777.

[19] NIE F P, XU D, TSANG I W, *et al.*. Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction [J]. *IEEE Trans on Image Processing*, 2010, 19(7): 1057-7149.

作者简介：

黄　鸿(1980－)，男，湖南新宁人，博士，讲师，硕士生导师，2003 年、2005 年、2008 年于重庆大学分别获得学士、硕士、博士学位，主要从事流形学习、模式识别、遥感影像智能处理等方面的研究。E-mail：hhuang.cqu@gmail.com

秦高峰(1985－)，男，河南鹿邑人，硕士研究生，2009 年于重庆大学获得学士学位，主要从事模式识别、遥感影像分类、地理信息空间等方面的研究。E-mail：qingaofeng1006@gmail.com

冯海亮(1962－)男，陕西西安人，博士，教授，1985 年于陕西师范大学获得学士学位，1992 年、2008 年于重庆大学分别获硕士、博士学位，主要从事应用数学、流形学习、人脸识别等方面的研究。E-mail：fhliang@cqu.edu.cn