# 基于新型光谱相似度量的高光谱影像谱聚类算法

陈 伟[1] 余旭初[1] 张立福[2] 张鹏强[1]
(信息工程大学测绘学院 郑州 450052)[1]
(中科院遥感应用研究所 遥感科学国家重点实验室 北京 100101)[2]

**摘 要** 高斯径向基函数是基于光谱向量间欧氏距离的度量,其对于同种地物光谱变化的适应性较弱,使得基于高斯径向基函数的高光谱影像谱聚类算法的性能下降。为了解决该问题,从光谱曲线形状描述出发,基于光谱角度余弦提出了一种新型光谱相似度量,并将其用于构建谱聚类算法的亲和度矩阵。最后利用多组高光谱数据进行了实验分析,结果证明了该算法的有效性。

**关键词** 高光谱影像,谱聚类,规范割准则,光谱相似度量

## Novel Spectral Similarity Measurement Based Spectral Clustering Algorithm in Hyperspectral Imagery

CHEN Wei[1] YU Xu-chu[1] ZHANG Li-fu[2] ZHANG Peng-qiang[1]
(Institute of Surveying and Mapping, Information Engineering University, Zhengzhou 450052, China)[1]
(The State Key Laboratory of Remote Sensing Sciences, Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101, China)[2]

**Abstract** As the gaussian radial basis function (RBF) is based on the Euclidean distance of two spectral vectors, it is not sensitive for variation of spectral curves of a material, which results in decrease of the performance of the RBF based spectral clustering for hyperspectral imagery degenerate. In order to solve this problem, according to the spectral curves similarity description, a novel spectral similarity measurement based on spectral angle cosine was proposed, and the measurement was used to build the affinity matrix used by spectral clustering algorithms. Finally, the experiments carried on with several hyperspectral data. The results of the experiments prove the validity of the proposed method.

**Keywords** Hyperspectral image, Spectral clustering, Normalized cut, Spectral similarity measurement

## 1 引言

高光谱遥感将反映目标辐射属性的光谱与反映目标空间和几何关系的图像有机地结合在一起,续写和完善了光学遥感从黑白全色影像通过多光谱到高光谱的全部影像信息链[1],其图谱合一的特点为分类、探测及目标识别提供了极大的便利。其应用领域已涵盖地球科学的各个方面,在地质找矿和制图、大气和环境监测、农业和森林调查、海洋生物和物理研究等领域发挥着越来越重要的作用。非监督分类是高光谱影像处理的重要技术手段之一,由于不需各类地物的样本信息,因此其对摄影区域先验知识的依赖较少,它是监督分类的重要补充手段。然而,以 k-means 为代表的传统聚类算法,要求样本在特征空间中的分布为近似凸球形,即每类中各分量的方差接近相等时才可能有较好的分类效果[2],否则算法极易陷入局部最优。谱聚类算法(Spectral Clustering Algorithms, SCA)克服了传统聚类算法的这一缺点,可在任意形状的样本空间中聚类,目前已经在计算机视觉、影像分割和数据挖掘等领域得到了广泛应用。本文将从图划分的角度出发对谱聚类算法的原理进行深入的分析,并将其应用于高光谱影像的非监督分类。

亲和矩阵的构造是谱聚类算法的关键一步,通常采用高斯径向基函数(Radial Basis Function, RBF)来衡量数据点对间的相似性测度。RBF 的基础是光谱向量间的欧式距离,这种度量对于光谱向量在欧氏空间中的距离变化较为敏感,但是对于同种地物的光谱变化这一在真实高光谱影像中十分普遍的现象,容易引起错误判别,从而降低谱聚类在高光谱影像处理中的效果。针对该问题,提出了一种以光谱角度余弦为基础的相似性度量,并将其应用于高光谱影像的谱聚类分析。

## 2 新型光谱相似度量

谱聚类算法的基本步骤可以总结为:根据待聚类的数据集,生成描述两两数据点对之间相似性程度的亲和矩阵(Affinity Matrix),然后对根据亲和矩阵生成的图拉普拉斯矩阵(Graph Laplacian matrices)进行谱分解,并利用谱分解得到的合适的特征向量来描述数据的低维结构,最后在低维空间中利用 k-means 等经典方法得到最后的聚类结果。由此可见,亲和矩阵的构造是谱聚类算法的关键,如式(1)所示的 RBF 是谱聚类中最为常用的相似性度量函数。

到稿日期:2011-12-13 返修日期:2012-03-20

陈 伟(1983—),男,博士生,主要研究方向为模式识别、高光谱遥感技术,E-mail:oliver8383@sina.com。

$$S_{ij}=\exp\left(-\frac{\| x_i-x_j \|^2}{2\sigma^2}\right) \tag{1}$$

由于 RBF 基于光谱向量的欧氏距离，因此对于光照强度变化导致的同种地物的光谱变化情况，容易发生错误识别。针对该问题，本文提出如下所示的新型光谱相似度量：

$$k_c=\begin{cases}\exp\left(-\dfrac{1-\dfrac{1}{C_{ij}}}{\theta}\right), & C_{ij}\in(0,1]\\ 0, & C_{ij}=0\end{cases} \tag{2}$$

式中，$C_{ij}$ 是按照式(3)计算的光谱向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的光谱角度余弦，$\theta>0$ 为需要设置的参数，显然 $k_c$ 满足非负性和对称性。

$$C_{ij}=\frac{\boldsymbol{x}_i'\boldsymbol{x}_j}{\| \boldsymbol{x}_i \| \| \boldsymbol{x}_j \|} \tag{3}$$

参数 $\theta$ 的作用类似于 RBF 中的尺度参数 $\sigma^2$，其取值对聚类结果的好坏有较大的影响。文献[14]中提出了一种在一组给定的参数值下反复运行谱聚类算法，并通过对聚类结果的评价来自动选择合适值的方法。然而，这种方法显著增加了计算量，并且参数的取值范围仍需提前人工设定。此外，当待聚类数据的局部统计值不相同时，对所有数据采用相同的参数值很难得到满意的结果。所以在此参考 RBF 自动调整局部尺度参数的方法[13]，采用参数自适应选择策略计算每个样本合适的参数 $\theta_i$，以代替一个固定的全局参数 $\theta$。

有以下不等式成立：

$$d_{ij}^2=1/C_{ij}-1\geqslant 0 \tag{4}$$

式中，$d_{ij}^2$ 可视为样本 $i$ 和 $j$ 之间的某种距离的平方。考虑各像素自适应参数的影响，从像素 $i$ 观测到的它与像素 $j$ 的距离可以表示为 $D_{ij}=d_{ij}/\theta_i$；反之，从像素 $j$ 观测到的它到像素 $i$ 的距离可记为 $D_{ji}=d_{ij}/\theta_j$，将 $D_{ij}$ 和 $D_{ji}$ 代入式(2)，则可获得自适应参数的相似性测度的表达式：

$$k_{Ac}=\begin{cases}\exp\left(\dfrac{1-1/C_{ij}}{\theta_i\theta_j}\right), & C_{ij}\in(0,1]\\ 0, & C_{ij}=0\end{cases} \tag{5}$$

$\theta_i$ 可以通过计算样本 $i$ 所处邻域的局部统计值得到。在本文中，如式(6)所示，$\theta_i$ 通过计算样本 $i$ 与其 $t$ 最近邻个像素的距离 $d_{ij}$ 的均值来得到：

$$\theta_i=\frac{1}{t}\sum_{j=1}^{t}d_{ij} \tag{6}$$

为了比较本文提出的新型光谱相似度量和 RBF 的效果，从 ENVI 软件中自带的、美国喷气推进实验室(JPL)的典型地物光谱数据库(jpl1. sli)中选择方解石(calcite)、绿脱石(nontronite)和石英(quartz)这 3 种类型矿物的光谱。为了更好地模拟真实影像中普遍存在的、由于各种因素造成的同种地物间的光谱差异，对于每种类型矿物都从数据库中选择 3 种不同的光谱参与比较。光谱曲线在数据库中的名称及其对应的矿物类型如表 3 所列。原始数据的波长范围为 0.4～2.5μm，共有 826 波段，对比分析中采用对区分矿物比较有利的波长范围为 1.9～2.5μm 的 151 个波段的数据，9 种矿物的光谱反射率曲线如图 1 所示。

表 1 和表 2 分别为本文相似度量与参数自适应的 RBF 对于这 3 种地物 9 条光谱曲线的相似度。由于石英与方解石在光谱强度上非常接近，因此 RBF 在它们两者之间产生了错分(错误以加粗斜体表示)。由此可见，本文提出的光谱相似度量对于同种地物光谱变化的适应性要好于 RBF。

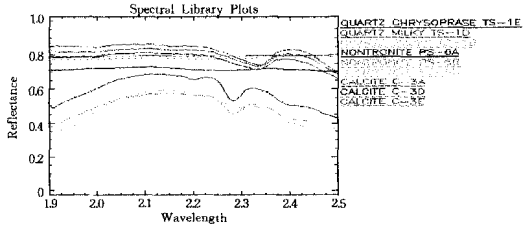



图 1　9 种矿物的光谱曲线

表 1　本文测度度量结果

| | Ⅰ | Ⅱ | Ⅲ | Ⅳ | Ⅴ | Ⅵ | Ⅶ | Ⅷ | Ⅸ |
|---|---|---|---|---|---|---|---|---|---|
| Ⅰ | 1 | 0.9773 | 0.9756 | 0.1191 | 0.1018 | 0.0521 | 0.7159 | 0.7030 | 0.6620 |
| Ⅱ | | 1 | 0.9998 | 0.1041 | 0.0928 | 0.0498 | 0.6315 | 0.6099 | 0.5821 |
| Ⅲ | | | 1 | 0.1035 | 0.0924 | 0.0498 | 0.6286 | 0.6067 | 0.5798 |
| Ⅳ | | | | 1 | 0.9541 | 0.7384 | 0.1807 | 0.2073 | 0.1835 |
| Ⅴ | | | | | 1 | 0.8287 | 0.1381 | 0.1582 | 0.1382 |
| Ⅵ | | | | | | 1 | 0.0671 | 0.0760 | 0.0705 |
| Ⅶ | | | | | | | 1 | 0.9916 | 0.9863 |
| Ⅷ | | | | | | | | 1 | 0.9723 |
| Ⅸ | | | | | | | | | 1 |

表 2　RBF 度量结果

| | Ⅰ | Ⅱ | Ⅲ | Ⅳ | Ⅴ | Ⅵ | Ⅶ | Ⅷ | Ⅸ |
|---|---|---|---|---|---|---|---|---|---|
| Ⅰ | 1 | ***0.4156*** | 0.5535 | 0.0385 | 0.0002 | 0.0001 | 0.3027 | ***0.5265*** | 0.1431 |
| Ⅱ | | 1 | 0.9571 | 0.0013 | 0.0000 | 0.0000 | 0.7093 | 0.7093 | 0.7125 |
| Ⅲ | | | 1 | 0.0009 | 0.0002 | 0.0010 | 0.7121 | 0.7234 | 0.6017 |
| Ⅳ | | | | 1 | 0.4154 | 0.2377 | 0.0000 | 0.0015 | 0.0003 |
| Ⅴ | | | | | 1 | 0.6894 | 0.0000 | 0.0000 | 0.0000 |
| Ⅵ | | | | | | 1 | 0.0000 | 0.0000 | 0.0000 |
| Ⅶ | | | | | | | 1 | 0.9052 | 0.9187 |
| Ⅷ | | | | | | | | 1 | 0.7647 |
| Ⅸ | | | | | | | | | 1 |

表 3　矿物光谱曲线名称及其类型

| 编号 | 光谱曲线名称 | 矿物类型 |
|---|---|---|
| Ⅰ | QUARTZ CHRYSOPRASE TS-1E | 石英 |
| Ⅱ | QUARTZ MILKY TS-1D | 石英 |
| Ⅲ | QUARTZ SMOKY TS-1B | 石英 |
| Ⅳ | NONTRONITE PS-6A | 绿脱石 |
| Ⅴ | NONTRONITE PS-6B | 绿脱石 |
| Ⅵ | NONTRONITE PS-6D | 绿脱石 |
| Ⅶ | CALCITE C-3A | 方解石 |
| Ⅷ | CALCITE C-3D | 方解石 |
| Ⅸ | CALCITE C-3E | 方解石 |

## 3　高光谱影像谱聚类算法

本节从图划分的角度出发对基于多类划分的规范割准则的谱聚类算法原理进行分析；在此基础上，给出了高光谱影像谱聚类算法流程，并根据高光谱影像的数据特点对算法细节进行了研究。

### 3.1　谱聚类算法原理

利用数据点对的相似性，构建无向加权图 $G=(V,E)$，那么第 $i$ 个待聚类的数据就是图 $G$ 中的一个顶点 $V_i$，$E$ 代表两两顶点之间的边，第 $i$ 个和第 $j$ 个顶点之间的边表示为 $E_{ij}$，其权重为成对数据点 $i$、$j$ 之间的相似性 $w_{ij}$，并且有 $w_{ij}=w_{ji}$。如此，就将聚类问题转化为图划分问题。基于图划分准则对图进行分割就可完成原始数据的聚类。最小化绝大多数图划分准则是一系列的 NP 完备(Non-deterministic Ploynomial Complete)问题，难以得到精确解，但是可以通过问题的连续放松形式，采用谱聚类方法对图划分准则进行逼近[9]，从而得

到近似解。

主要的图划分准则包括：最小割准则(Minimum cut)、比例割准则(Ratio cut)、规范割准则(Normalized cut)等[3]。虽然两类划分最小割准则不是一个 NP 完备问题，且存在高效的求解方法[10]，但是其容易发生歪斜分割，即容易分割出孤立点。比例割准则的定义平衡了子集包含的顶点个数，因此在避免歪斜分割的发生上相比最小割准则有所进步，但是仍然没有考虑同一子集内部的相似性程度。相比之下，两类划分规范割准则[8]定义如下：

$$\mathrm{Ncut}(P,\bar{P})=\frac{\mathrm{cut}(P,\bar{P})}{\mathrm{vol}(P)}+\frac{\mathrm{cut}(P,\bar{P})}{\mathrm{vol}(\bar{P})} \tag{7}$$

式中，$\mathrm{cut}(P,\bar{P})$为顶点子集 $P$ 与其补集 $\bar{P}$ 的割，$\mathrm{vol}(P)$为 $P$ 的容量。最小化归一割，不仅考虑到 $P$ 与 $\bar{P}$ 之间的相似性最小，而且考虑了 $P$ 与 $\bar{P}$ 自身内部顶点的相似性程度最大。多类划分的规范割准则是两类划分规范割准则的扩展，其定义如下所示：

$$\mathrm{Ncut}(P_1,\cdots,P_m)=\sum_{i=1}^{m}\frac{\mathrm{cut}(P_i,\bar{P}_i)}{\mathrm{vol}(P_i)} \tag{8}$$

设 $\boldsymbol{X}=\{x_1,x_2,\cdots x_n\}$为 $n$ 个待聚类的样本，根据相似度量函数可构建一个 $n\times n$ 大小的亲和度矩阵 $\boldsymbol{S}$。将 $\boldsymbol{S}$ 的各行元素相加得到度矩阵 $\boldsymbol{D}$，度矩阵是对角矩阵，进而可以得到图拉普拉斯矩阵。有两类图拉普拉斯矩阵，分别是非规范化图拉普拉斯矩阵，其形式为：

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{S} \tag{9}$$

以及两种形式的规范化图拉普拉斯矩阵，其形式分别为：

$$\boldsymbol{L}_s=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}=\boldsymbol{I}-\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}} \tag{10}$$

$$\boldsymbol{L}_r=\boldsymbol{D}^{-1}\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{S} \tag{11}$$

将 $n$ 个待聚类的样本视为图 $G$ 的 $n$ 个顶点，给定图 $G$ 的一个划分 $P_1,P_2,\cdots,P_m$，定义 $m$ 个 $n$ 维列向量，第 $j$ 个列向量记为：

$$\boldsymbol{\gamma}_j=(\gamma_{1,j},\gamma_{2,j},\cdots,\gamma_{n,j})' \tag{12}$$

向量中的元素 $\gamma_{i,j}$定义为：

$$\gamma_{i,j}=\begin{cases}\dfrac{1}{\sqrt{\mathrm{vol}(P_j)}}, & \text{if } v_i\in P_j\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

定义矩阵 $\boldsymbol{M}\in\mathbb{R}^{n\times m}$，其第 $j$ 列为 $\boldsymbol{\gamma}_j$，可以证明有以下等式成立：

$$\boldsymbol{M}'\boldsymbol{M}=\boldsymbol{I} \tag{14}$$

$$\boldsymbol{\gamma}_j'\boldsymbol{D}\boldsymbol{\gamma}_j=1 \tag{15}$$

$$\boldsymbol{\gamma}_j'\boldsymbol{L}\boldsymbol{\gamma}_j=\frac{\mathrm{cut}(P_j,\bar{P}_j)}{\mathrm{vol}(P_j)} \tag{16}$$

结合式(8)、式(14)－式(16)可将最小化规范割准则表示成如下形式：

$$\min \mathrm{Tr}(\boldsymbol{M}'\boldsymbol{L}\boldsymbol{M}) \tag{17}$$

式(17)的限制条件为 $\boldsymbol{M}'\boldsymbol{D}\boldsymbol{M}=\boldsymbol{I}$($\boldsymbol{I}$ 为单位阵)，且 $\boldsymbol{M}$ 中的元素仅能取如式(13)所示的离散形式。因此，式(17)所示的是一个离散优化问题，仍是一个 NP 难问题。求解上述 NP 难问题的近似解方法就是忽略关于 $\boldsymbol{M}$ 中元素离散的限制条件，使其可取任意实数解，放松离散条件并且设 $\boldsymbol{Q}=\boldsymbol{D}^{1/2}\boldsymbol{M}$，则式(17)的连续放松形式可表示为：

$$\begin{cases}\min\limits_{\boldsymbol{Q}\in\mathbb{R}^{n\times m}} \mathrm{Tr}(\boldsymbol{Q}'\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}\boldsymbol{Q})\\ \text{subject to } \boldsymbol{Q}'\boldsymbol{Q}=\boldsymbol{I}\end{cases} \tag{18}$$

式(18)是一个标准的迹最小化问题。根据瑞利商原理[12]，该优化问题最优解 $\boldsymbol{Q}_S$ 是由 $\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$即 $\boldsymbol{L}_s$ 最小的 $m$ 个特征值所对应的特征向量构成的。最后将 $\boldsymbol{Q}_S$ 的每行作为样本，利用 k-means 进行聚类，就能得到最终的划分结果。

### 3.2 高光谱影像谱聚类算法流程

基于上述谱聚类算法原理，结合高光谱影像聚类问题的实际情况，给出以下算法流程：

① 利用空间邻域聚类方法对高光谱影像进行过分割，并计算每一块的光谱均值；

② 利用所有光谱均值作为待聚类的样本，根据式(5)所示的光谱相似度量(或参数自适应的高斯径向基核函数[13])，按照 $t$ 最近邻法构造稀疏亲和矩阵 $\boldsymbol{S}$；

③ 依据亲和矩阵 $\boldsymbol{S}$ 计算规范化图拉普拉斯矩阵 $\boldsymbol{L}_s$；

④ 计算得到 $\boldsymbol{L}_s$ 最小的 $m$ 个特征值对应的特征向量 $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_m$；

⑤ 构建矩阵 $\boldsymbol{U}\in\mathbb{R}^{n\times m}$，其第 $i$ 列为 $\boldsymbol{u}_i$；

⑥ 记 $\boldsymbol{U}$ 中的元素为 $u_{ij}$，构建矩阵 $\boldsymbol{T}\in\mathbb{R}^{n\times m}$，其中的元素记为 $t_{ij}$，$t_{ij}$ 的计算方法按照式 $t_{ij}\in u_{ij}/(\sum_{k=1}^{m}u_{ik}^2)^{1/2}$进行；

⑦ 将矩阵 $\boldsymbol{T}$ 每一行转置视为一个样本，得到 $\boldsymbol{y}_i\in\mathbb{R}^m$，$i=1,\cdots,n$；

⑧ 利用 k-means 算法将 $\boldsymbol{y}_i$ 聚成 $m$ 类，从而得到每块子影像光谱均值的谱聚类结果。

⑨ 结合每块子影像光谱均值的聚类结果和步骤①的分块信息得到最终的聚类结果影像。

谱聚类算法需要构建 $n\times n$ 大小的亲和矩阵，$n$ 为待聚类样本个数。而高光谱影像像素数目众多，因此，谱聚类算法要成功应用于高光谱影像的聚类，除了要设计合适的相似性度量，还需要考虑算法的效率与内存的压力。针对这一问题，本文采用 $t$ 最近邻法构造稀疏亲和矩阵，并采用空间邻域聚类对高光谱影像进行预分割处理。

### 3.3 稀疏亲和矩阵的构造

采用不同的亲和矩阵构造方法，可得全连接图、$\varepsilon$ 邻近图以及 $t$ 最近邻图。考虑到亲和矩阵的大小为 $n\times n$，$n$ 为样本个数，因而全连接图不适用于高光谱影像非监督分类这种大规模聚类问题。比较现实的解决方案是采用稀疏亲和矩阵或者通过 Nyström 采样用原始亲和矩阵的子矩阵来代替整个原始亲和矩阵。$\varepsilon$ 邻近法不适合样本间距离尺度变化较大的情况[11]，而对于高光谱影像而言，由于地物的复杂多样和成像过程中的干扰因素等多种原因的共同作用，地物光谱在特征空间中距离的变化尺度可能较大，同一幅影像中的光谱也可能不服从同一分布。基于 Nyström 采样[16]的方法是一种随机的方法，其基本思想是从原始亲和矩阵中随机选择一些位置的值构成子矩阵，然后通过矩阵计算，达到用子矩阵来代替整个原始的亲和矩阵的目的。其效率虽然较高，但由于记录的并不是数据间最主要的关系，因此其最终聚类精度与 $t$ 最近邻图相比较差[17]。在面临样本间距离尺度变化较大的情况时，$t$ 最近邻法可使得密度不同的各部分间存在较好的连接，此时，互 $t$ 最近邻法的表现处于 $\varepsilon$ 邻近法和 $t$ 最近邻法之间，倾向于将图 $G$ 划分为密度不同的几个连通分支，然而，这样的连通分支并不能与真实聚类相对应。综合上述分析，采用 $t$ 最近邻法的风险相对较小，另外，经验证明 $t$ 最近邻法对参数选择的敏感性要低于其它方法，因此，本文采用 $t$ 最近

邻法构建稀疏亲和矩阵。

$t$ 最近邻法的最邻近样本个数 $t$ 的选择尚乏理论指导，只能凭经验做出一些定性的分析。$t$ 的选择不宜过小，应尽量确保相似图是整体相连的或至少仅包含少量的连通分支和孤立的顶点，但如果过大又会导致内存和计算的开销无意义地大幅增长。有一些理论可以用于分析样本数量趋于无穷大时相似图的连接情况[18]，对于 $t$ 最近邻法和互 $t$ 最近邻法而言，将参数 $t$ 选择为 $\log(n)$（$n$ 为样本个数）可使得到的相似图是整体连接的。对于高光谱影像聚类而言，样本数量是有限的，因此上述理论的指导意义是有限的，但是其至少为实际应用中参数 $t$ 的设置提供了一个可供参考的下限。

### 3.4 空间领域聚类分割

如果采用 $t$ 最近邻法构造稀疏亲和矩阵，那么其计算的复杂度将达到 $O(N^2d)$，其中 $N$ 为像素总数，$d$ 为高光谱影像的维数。如果处理的是影像分割问题，则可以设定一个邻域半径 $r$，当前像素就在该邻域内寻找最近似的 $t$ 个邻居，如此便将构造亲和矩阵的计算复杂度降低到了 $O(Nnd)$，其中 $n=(2r+1)^2$。但是，影像分割与聚类存在着明显的不同，聚类要求同种地物即便在空间上是不连续的也应在结果中标记为一类，而影像分割显然要求标记为一类的地物既是同种类型又在空间上是连续的，因此在这里不能采用上述方法来构造亲和矩阵。针对该问题，本文采用的方案是首先采用空间邻域聚类方法[4]对影像进行过分割，然后对分割后的结果求取每一块的光谱均值，用这些均值参与谱聚类，最后利用聚类结果和分割信息得到整幅高光谱影像的最终聚类结果。

高光谱影像作为地理环境的记录，在一定尺度下地物的连续性也反映在高光谱影像中，因此与其他记录自然景观的影像一样，高光谱影像的邻域像元之间也存在相关性。空间邻域聚类正是基于这种相邻像素间的强相关性实现同质区域的分割。设待聚类的高光谱影像为 $\boldsymbol{X}$，影像高为 $L$，宽为 $S$，那么第 $i$ 行第 $j$ 列像素对应的光谱向量可记为 $\boldsymbol{x}_{ij}$，其中 $i=1,2,\cdots,L$；$j=1,2,\cdots,S$。令 $\boldsymbol{P}$ 为经过空间邻域聚类后得到的标号影像，其中的元素为 $p_{ij}=\{1,2,\cdots,n\}$，$n$ 是最终得到的影像的分割块数。$T$ 为预先设定好的分割阈值，当相邻像素光谱向量的相似度大于这一阈值时将它们分为一类，反之将其视为两类，并在当前总块数 $n_C$ 的基础上加 1。

如图 2 所示，其中深灰色代表当前像素，浅灰色代表需与其比较的邻域像素，空间邻域聚类采用自左向右、自上而下的顺序遍历影像，对影像中的每个像素又按照左、左上、上、右上的顺序与其邻居比较相似性程度。

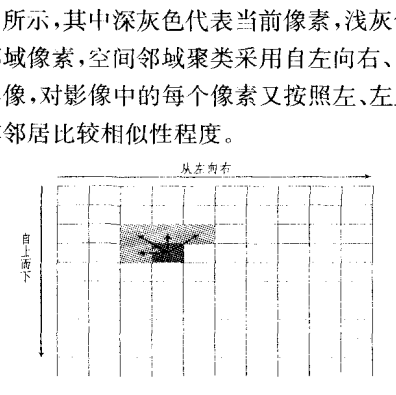



图 2 空间邻域聚类空间扫描顺序示意图

基于本文新型光谱相似性度量的空间邻域聚类方法的步骤如下：

① 首先按参数自适应方法计算高光谱影像每个像元相应的 $\theta$；

② 令 $\boldsymbol{P}$ 中第 1 行第 1 列的像素为第一类，记 $p_{11}=1$，当前总类别数 $n_C=1$；

③ 对于 $\boldsymbol{x}_{ij}$，按照左、左上、上、右上的顺序利用式(5)计算与其邻域光谱向量的相似性程度。对于 $i=1$ 的情况，$\boldsymbol{x}_{1j}$ 只需与其左边像素进行比较，如果 $k_{Ac}(\boldsymbol{x}_{1j},\boldsymbol{x}_{1j-1})\geqslant T$，则有 $p_{1j}=p_{1j-1}$，反之有 $n_C=n_C+1$，$p_{1j}=n_C$；对于 $j=1$ 的情况，$\boldsymbol{x}_{i1}$ 只需与其上面、右上的邻居进行比较；对于 $j=S$ 的情况，$\boldsymbol{x}_{i1}$ 只需与其左面、左上、上面的邻居进行比较；剩余情况的像素则需和 4 个方向的邻居进行比较，对于第 1 行像素以外的其他像素而言，如果有 $\max k_{Ac}\geqslant T$，则将 $p_{ij}$ 赋值为与 $\boldsymbol{x}_{ij}$ 最为相似的邻居的标号，反之则有 $n_C=n_C+1$，$p_{ij}=n_C$。

④ 按照③的原则从左至右、自上而下遍历整幅影像，最后输出结果 $\boldsymbol{P}$。

## 4 实验

为了验证本文算法的有效性，对不同的高光谱影像聚类算法进行定量与定性的分析实验。本实验采用的计算机硬件环境为：Intel Core2 CPU 3.0GHz、2.99GHz；内存为3.25GB。操作系统为 Microsoft Windows XP，算法利用 MATLAB 7.5 编程实现。

### 4.1 精度评价方法

谱聚类是一种非监督的分类方法，最终聚类结果的类别标签号与检验样本的标签号不一致，如聚类结果的第 1 类可能对应的是检验样本的第 3 类。如果采取遍历法对每种可能的情况进行比较并取最优结果，则需要 $CNum!$ 次，$CNum$ 为聚类类别数。因此，采用匈牙利算法（Hungarian Algorithm）调整聚类结果每类的标签号[15]，然后依据下式计算聚类精度 $P$。

$$P=\frac{\sum_{i=1}^{n}\phi(real\_labels_i,map(test\_labels_i))}{n} \tag{19}$$

式中，$\phi(a,b)$ 为：

$$\phi(a,b)=\begin{cases}1, & a=b\\ 0, & a\neq b\end{cases} \tag{20}$$

式中，$real\_labels_i$ 为第 $i$ 个检验样本的真实标签号，$test\_labels_i$ 是其聚类后的标签号，$map(test\_labels_i)$ 代表利用匈牙利算法将聚类后的标签号置换为合适的类别标签。

### 4.2 高光谱影像聚类实验

本文算法采用了 $t$-最近邻法构造稀疏亲和度矩阵。实验分析了 $t$ 的不同取值对聚类精度的影响，并与原始数据 k-means[7]结果、PCA 降维后的 k-means 结果和模糊 C 均值（FCM）[6]的结果进行了比较。由于 FCM 是软分类方法，得到的结果是样本对于每一类的隶属度，为便于比较，将样本划分到隶属度或概率最大的一类中来得到聚类结果。由于本文算法最终调用 k-means 方法在低维空间中完成聚类，而 k-means 方法带有一定随机性，同时 FCM 也存在随机性，因此对于每幅影像的每种算法都连续进行 20 次试验，并记录了它们的平均精度。实验用到了 3 组实验数据，由于篇幅限制，文中仅作了简要的介绍，更详细的信息可参考文献[5]。

(1)实验数据一

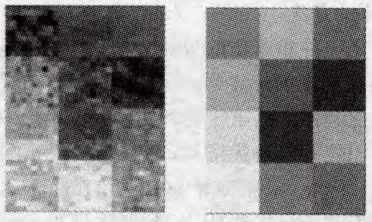
2001 年 5 月 31 日，由 NASA 的 EO-1 卫星上的 Hyperion 传感器获取的南非博茨瓦纳（Botswana）Okavango 三角洲地区影像，其光谱分辨率大约为 10nm，光谱范围 400～2500nm，共

242 波段，经过辐射校正，去除噪声和大气吸收波段。实验采用原数据中的 10～55、82～97、102～119、134～164、187～220 波段，共 145 波段。地面覆盖类型的样本采集根据植被测量和航空摄影测量获得，样本信息如表 4 所列。

表 4　Hyperion Botswana 数据样本信息

| 标号 | 名称 | 标号 | 名称 |
|---|---|---|---|
| 1 | Water | 2 | Hippo grass |
| 3 | Floodplain grasses 1 | 4 | Floodplain grasses 2 |
| 5 | Reeds1 | 6 | Riparian |
| 7 | Firescar 2 | 8 | Island interior |
| 9 | Acacia woodlands | 10 | Acacia scrublands |
| 11 | Acacia grasslands | 12 | Short mopane |
| 13 | Mixed mopane | 14 | Exposed soils |

其中第 2－13 类这 12 种类型都为植被，植被类型之间的区分比较困难，有利于验证算法的有效性。由于本文算法首先采用空间邻域聚类的方法对影像进行预分割，因此从每种植被类型的样本中随机选择 100 个，共计 1200 个光谱向量构成一幅宽为 30 像元、高为 40 像元的形如调色板的影像，如图 3(a)所示；图 3(b)所示的是调色板模拟影像对应的类别标号影像，其中每种颜色代表一种地物，每种植被类型的 100 个光谱向量构成了一个 10×10 像元大小的子影像。



(a) 调色板影像一　(b) 标号影像一

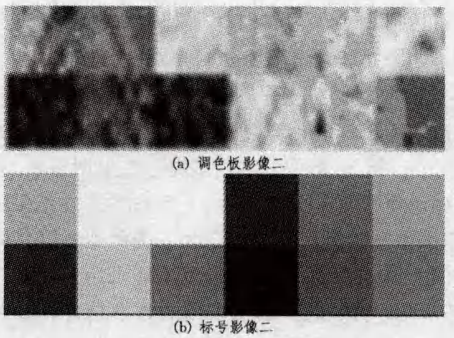
图 3　Hyperion Botswana 样本构成的调色板影像

(2)实验数据二

AVIRIS 于 1996 年 3 月 23 日获取了佛罗里达州肯尼迪空间中心(Kennedy Space Center，KSC)附近的影像，实验选择了其中信噪比较高的 120 个波段，分别是 5～97 和 105～131 波段。样本选择的依据是成像区域地面覆盖图与 Landsat 专题制图仪(TM)影像，样本信息如表 5 所列。与实验数据一的方式类似，对于第 1－12 类，从每类样本中随机选择 100 个，共计 1200 个光谱向量构成一幅宽为 60 像元、高为 20 像元的形如调色板的影像，如图 4(a)所示；图 4(b)所示的是调色板模拟影像对应的类别标号影像。

表 5　AVIRIS KSC 数据样本信息

| 标号 | 名称 | 标号 | 名称 |
|---|---|---|---|
| 1 | Scrub | 2 | Willow |
| 3 | CP Hammock | 4 | CP Oak |
| 5 | Slash Pine | 6 | Oak Broadleaf |
| 7 | Hardwood swamp | 8 | Graminoid marsh |
| 9 | Spartina marsh | 10 | Cattail Marsh |
| 11 | Salt marsh | 12 | Mud flats |
| 13 | water | | |

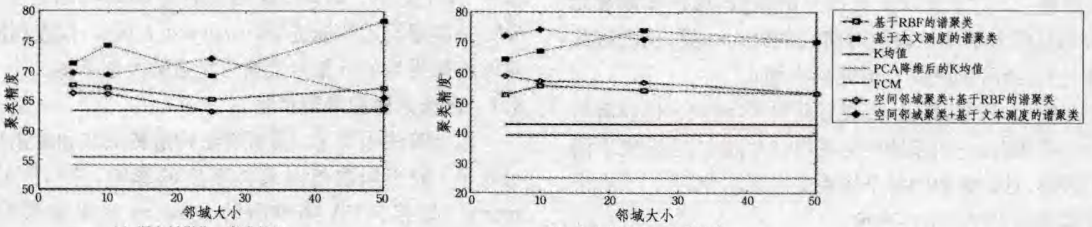


(a) 调色板影像二

(b) 标号影像二

图 4　AVIRIS KSC 样本构成的调色板影像

根据实现细节的不同，共有 4 种谱聚类算法参与比较，它们分别是：利用本文测度构造亲和矩阵的谱聚类方法(记为 $k_{Ac}$)；在 $k_{Ac}$ 之前利用空间邻域聚类进行过分割的谱聚类方法(记为 $Sk_{Ac}$)；利用 RBF 构造亲和矩阵的谱聚类方法(记为 $k_{Arbf}$)以及在 $k_{Arbf}$ 之前利用空间邻域聚类进行过分割的谱聚类方法(记为 $Sk_{Arbf}$)。对于两组试验数据，这 4 种谱聚类方法在不同邻域大小情况下的聚类精度变化情况分别如图 5 所示，具体的精度值记录在表 6 中。图 5 中的实线是 k-means、PCA 降维后的 k-means 和 FCM 这些参与比较的方法的精度，它们的精度值则记录在表 7 中。




(a) 调色板影像一聚类精度　(b) 调色板影像二聚类精度

图 5　两组调色板影像聚类精度比较

表 6　两组模拟数据谱聚类精度

| 邻域大小 | 调色板影像一 | | | | 调色板影像二 | | | |
|---|---|---|---|---|---|---|---|---|
| | $k_{Ac}$ 精度(%) | $Sk_{Ac}$ 精度(%) | $k_{Arbf}$ 精度(%) | $Sk_{Arbf}$ 精度(%) | $k_{Ac}$ 精度(%) | $Sk_{Ac}$ 精度(%) | $k_{Arbf}$ 精度(%) | $Sk_{Arbf}$ 精度(%) |
| 5 | 71.24 | 69.64 | 67.53 | 66.21 | 64.30 | 72.01 | 52.45 | 58.75 |
| 15 | 74.25 | 69.34 | 67.15 | 66.13 | 67.10 | 74.08 | 55.35 | 56.97 |
| 25 | 69.16 | 72.03 | 65.23 | 63.19 | 73.88 | 70.83 | 53.88 | 56.45 |
| 50 | 78.39 | 65.56 | 67.15 | 63.68 | 70.15 | 70.14 | 52.74 | 53.05 |

表 7　两组模拟数据的其他聚类算法精度

| 实验数据 | K 均值(%) | PCA 降维后的 K 均值(%) | FCM(%) |
|---|---|---|---|
| 调色板影像一 | 55.50 | 54.17 | 63.27 |
| 调色板影像二 | 38.98 | 24.85 | 44.78 |

此外，为了便于对聚类的结果进行目视判读，同时为了验证本文算法处理大影像的能力，选择地物分布较为清晰、空间分辨率相对较高的、由 OMIS 成像光谱仪获取的太湖沿岸高光谱影像。该影像如图 6(a)所示，光谱覆盖范围 0.46～

12.85μm，共128波段，影像宽为347像元、高为513像元，实验采用受噪声影响比较小的6～64、113～128的75个波段。对其进行基于两种亲和矩阵的谱聚类分析，邻域大小为50，聚类结果如图6(b)、(c)所示，基于其它算法的聚类结果如图6(d)－(f)所示。

(a) OMIS太湖沿岸影像　(b) $Sk_{Ac}$结果影像　(c) $Sk_{Arbf}$结果影像

(d)k-means结果影像　(e)降维后k-means结果影像　(f)FCM结果影像

图6　OMIS太湖沿岸影像聚类结果图

### 4.3　实验结果分析

以上实验结果证明，由于谱聚类算法可在任意形状的特征空间中进行聚类，而不要求数据在特征空间中呈凸球形分布，因此其在精度上优于k均值和FCM。而基于本文提出的新型光谱相似度量的谱聚类方法在精度上又明显优于基于RBF构造亲和矩阵的谱聚类方法，这是由于本文提出的相似度量与RBF相比，不容易受到由于光照、阴影、地形起伏等因素引起的同种地物光谱变化的影响。此外，本文采取的空间邻域聚类预分割策略并没有降低最终的聚类精度，通过对如图6(b)和(c)所示的谱聚类结果进行目视判读也可以发现，其没有遗失影像中的主要结构信息。谱聚类的精度受到邻域大小的影响，并不是邻域越大精度就越高，相反，过大的邻域在某些情况下会降低最终的聚类精度，例如调色板影像一选择50邻域的情况。这是因为通过空间邻域聚类后，参与聚类的样本数从1200分别下降到了142和156，显然，50的邻域对于如此少的样本数来说引入了许多不必要的信息，反而造成了精度的下降。但对于真实高光谱影像而言，即便经过空间邻域聚类的预处理，其样本数量仍然较为庞大(如太湖数据，预处理后的样本数量仍然达到9047)，因此选择50大小的邻域并不为过。

**结束语**　本文对谱聚类算法的原理进行了深入分析，其本质上是一种图划分准则的近似，具体的途径是图拉普拉斯矩阵的谱分解。谱聚类利用的基本信息是数据点对的相似性，因此亲和矩阵的构造对其精度有较大影响。针对常用的RBF在识别光谱变化的同种地物时存在的不足，提出了一种基于光谱角度余弦的新型光谱相似性测度，用于构造亲和矩阵。针对高光谱影像像素众多，逐个像素处理计算量过大的问题，本文采用空间邻域聚类方法对原始影像进行预分割，然后对分割结果进行谱聚类，有效减轻了计算的压力。最后利用光谱数据库中的标准光谱和真实高光谱影像数据进行了实验分析，实验结果一方面证实了本文提出的相似性测度对于同种类型物质光谱变化情况的适应性，另一方面验证了谱聚类算法相对于其他聚类算法对非监督分类精度的提高作用，证明了基于新型相似性测度的谱聚类算法在处理高光谱影像时的优势，此外也验证了空间邻域聚类预分割策略的有效性。

## 参考文献


[1] 杨国鹏，余旭初. 高光谱遥感影像的广义判别分析特征提取[J]. 测绘科学技术学报，2007，24(2)：101-105

[2] 边肇祺，张学工. 模式识别[M]. 北京：清华大学出版社，2001：230-248

[3] 蔡晓妍，戴冠中，杨黎斌. 谱聚类算法综述[J]. 计算机科学，2008，35(7)：14-18

[4] 耿修瑞，张霞，陈正超，等. 一种基于空间连续性的高光谱图像分类方法[J]. 红外与毫米波学报，2004，23(4)：299-302

[5] 杨国鹏. 基于机器学习方法的高光谱影像分类研究[D]. 郑州：郑州测绘学院，2010

[6] 袁金国. 遥感图像数字处理[M]. 北京：中国环境科学出版社，2006：233-238

[7] 童庆禧，张兵，郑兰芬. 高光谱遥感——原理、技术与应用[M]. 北京：高等教育出版社，2006：190-191

[8] Shi J，Malik J. Normalized cuts and image segmentation[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence，2000，22(8)：888-905

[9] Zhou D，Bousquet O，Lal T N，et al. Learning with Loacl and Global Consistency[C]// Advances in Neural Information Processing Systems. Cambridge，MA，USA，MIT Press，2004，16：321-328

[10] Stoer M，Wagner F. A simple min-cut algorithm [J]. ACM，1997，44(5)：585-591

[11] Luxburg U V. A tutorial on spectral clustering[J]. Statistics and Computing，2007，17(4)：395-416

[12] Lütkepohl H. Handbook of Matrices [M]. Chichester：Wiley，1997

[13] Zelnik-Manor L，Perona P. Self-tuning spectral clustering[C]// Advances in Neural Information Processing Systems. Cambridge，MA，USA，MIT Press，2005，17：1601 -1608

[14] Ng A，Jordan M，Weiss Y. On spectral clustering：analysis and an algoritm[C]// Advances in Neural Information Processing Systems. Cambridge，MA，USA，MIT Press，2002，14：849-856

[15] Wu M，Schölkopf B. A Local Learning Approach for clustering [C]// Advances in Neural Information Processing Systems. Cambridge，MA，USA，MIT Press，2007，19：1529-1536

[16] Fowlkes C，Belongie S，Chung F，et al. Spectral grouping using the Nyström method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2004，26(2)：214-225

[17] Chen W Y，Song Y Q，Bai H J，et al. Parallel Spectral Clustering in Distributed Systems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2010，29(6)：1002-1031

[18] Brito M，Chavez E，Quiroz A，et al. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection[J]. Statistic and Probability Letter，1997，35：33-42