

DOI:10.19651/j.cnki.emt.2210852

基于超像素锚层收敛选点的高光谱图像聚类算法^{*}

杨滔 孙博 杨晓君

(广东工业大学信息工程学院 广州 510006)

摘要: 针对传统高光谱图像聚类算法难以有效处理数据量快速增长的高光谱图像的问题,提出基于超像素锚层收敛选点的高光谱聚类算法。采用 SuperPCA 对原始数据进行基于超像素切割的降维;利用 K-means 选取具有代表性的锚点,构建基于锚点的邻接矩阵;通过无核邻近分配的方法构建相似图,避免对热核参数的调整;最后进行谱聚类分析获得聚类结果。在 Indian Pines 和 Pavia Centre 高光谱数据集进行仿真实验,结果表明该算法获得的分类图所含错分点更少,地物分布更加平滑,与当前高光谱图像聚类算法相比具有更优的聚类效果。

关键词: 高光谱图像;图像分割;降维;锚点图;聚类算法

中图分类号: TP751 文献标识码: A 国家标准学科分类代码: 510.4050

Hyperspectral image clustering algorithm based on super-pixel anchor layer convergence point selection

Yang Tao Sun Bo Yang Xiaojun

(College of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Aiming at the problem that traditional hyperspectral image clustering algorithms are difficult to effectively deal with hyperspectral images with rapidly increasing data volume, a hyperspectral clustering algorithm based on hyper-pixel anchor layer convergence point selection is proposed. SuperPCA is used to reduce the dimension of original data based on super pixel cutting. Selecting representative anchor points by K-means, and constructing an adjacency matrix based on anchor points. The similarity graph is constructed by the method of non-nuclear neighbor assignment to avoid the adjustment of thermonuclear parameters. Finally, spectral clustering analysis is carried out to obtain clustering results. The simulation experiments on Indian Pines and Pavia Centre hyperspectral data sets show that the classification map obtained by this algorithm contains fewer false points, and the distribution of ground objects is smoother. Compared with the current hyperspectral image clustering algorithms, this algorithm has a better clustering effect.

Keywords: hyperspectral image; image segmentation; dimesion-reduction; anchor map; clustering algorithm

0 引言

利用星载、机载等多种机器平台获得的高光谱图像(hyperspectral image, HSI)通常记录了大量的空间信息和光谱信息,其在遥感领域有着广泛的应用前景。在不同波长的光谱信号下,HSI 中各类地物信息表现不同,在地质勘探、军事探测、环境监测等领域,聚类是 HSI 处理最常用的技术之一^[1]。

近年来,众多研究学者提出了大量的 HSI 聚类算法^[2-3]。在文献^[4]中,将现有的聚类算法大致分为以下

4类,具体包括:1)基于质心的聚类算法,如 K-均值(K-means)^[5]、模糊 C-均值(FCM)^[6]和 FCM_S1^[7]以及基于核的模糊聚类算法^[8]。这些算法常使用欧几里德距离进行相似性度量,对初始化和噪声具有高度敏感性;2)基于密度的聚类算法,如通过快速搜索和发现密度峰值进行聚类(CFSFDP)^[9];3)基于生物的聚类算法,如基于自适应多目标差分进化的自动模糊聚类算法(AFCMDE)^[10];4)基于图的聚类算法,如谱聚类(SC)^[11],稀疏子空间聚类(SSC)^[11],快速谱聚类^[12]等。由于高光谱图像存在数据量大、维数高、结构复杂的特点,现有的聚类算法普遍存在处理效率不

收稿日期:2022-07-27

^{*} 基金项目:科技部重大研发计划(2018YFB1802100)、广东省重大研发计划(2018B010115001)、广东省面上自然科学基金(2021A1515011141)、国家自然科学基金(61904041)项目资助

高的问题,如何快速有效地应对数据量日益膨胀的 HSI 是亟待研究的课题之一。在这种情况下基于锚点构建相似度矩阵的 HSI 聚类算法逐渐得到了运用。文献[13]从数据集中随机选点作为锚点,并构造关于数据点以及锚点的关系图,在此基础上进行谱聚类分析,该算法存在随机选点不能确保所选点一定能够覆盖总体样本数据的缺陷。文献[14]利用二叉树选取锚点,改善了随机选取锚点的问题,但高光图像高维的维度可能会给算法带来灾难。文献[15]提出了可伸缩的基于图的聚类算法,采用非负松弛(SGCNR),然而其聚类结果是不稳定的。上述结论表明传统的基于图的聚类算法在 HSI 数据集中依然有较大的提升空间。

基于降低算法复杂度、避免参数调整等亟待解决的问题,文献[16]提出了一种超像素主成分分析降维算法(SuperPCA),通过对不同区域进行差异化投影,产生具有区分性、紧凑性和抗噪性的特征,有效提高了 HSI 分类性能。文献[17]提出了一种大型图构建方法来利用所有数据点,这种方法高效且可扩展,降低了算法计算复杂度。文献[18]采用一种无核的方法构造邻近图,避免了对于热核参数的调整。受上述方法的启发,为了降低算法复杂度、解决随机选取锚点带来的不确定性以及避免灾难,本文提出了一种基于超像素锚层收敛选点的高光谱聚类算法,通过 Indian Pines 数据集和 Pavia Centre 数据集的仿真,验证了该算法能够有效处理高光图像。

1 相关理论

1.1 基于 SuperPCA 的降维算法

在高光谱图像分析中应用最广泛的无监督降维技术之一是主成分分析(principal component analysis, PCA)及其变体,然而 PCA 算法将所有频段数据视为一个整体进行统一投影,忽略了光谱特征的多样性,在一定程度上影响了降维的性能。本文采用 SuperPCA 算法来获取高光图像数据的低维特征,与基于整幅图像的 PCA 方法不同,SuperPCA 考虑了不同同质区域的多样性,对不同的区域进行了不同的投影。

1) 超像素切割

在文献[19]中,提出了一种基于图结构的 ERS 分割方法,通过构造无向图 $G = (V, E)$,在选择边的子集 $A \subseteq E$ 的同时将图划分为一个连通子图,使得生成的图 $G = (V, E)$ 由较小的连通子图组成,其中 V 是顶点集, E 是边集。ERS 的目标函数由熵率项 $O(A)$ 和平衡项 $B(A)$ 组成,熵率项有利于形成紧凑的同质区域,而平衡项有利于减少不平衡超像素的数量。熵率超像素的目标函数为定义如下:

$$A^* = \arg \max_A Tr\{O(A) + \alpha B(A)\} \quad (1)$$

式中: $O(A)$ 是熵率项, $B(A)$ 是平衡项, α 用来调节 $O(A)$ 和 $B(A)$ 的权重。

为了降低计算成本,在分割之前,对原始 HSI 应用 PCA^[16],获得了高光图像的第一个主成分 I_f , I_f 包含整

个 HSI 最重要的信息^[20]。

对第一主成分 I_f 进行区域切割可得到若干个同质区域。公式如下:

$$I_f = \bigcup_p^S \varphi_p, s. t. \quad \varphi_p \cap \varphi_g = \emptyset (p \neq g) \quad (2)$$

式中: S 表示超像素的数量, φ_p 指第 p 个超像素。

超像素大小决定着超像素切割的质量,超像素太大导致分割后的模糊标记的边界超像素需要再次划分,超像素太小则难以得出正确的标签,故本文使用多尺度分割策略^[21]。设第 k 级超像素的数目是 S_k , 定义如下:

$$S_k = (\sqrt{2})^k S_0; k = 0, \pm 1, \pm 2, \dots, \pm K \quad (3)$$

式中: S_0 是基本的超像素数,根据经验进行设置。

2) PCA 降维

PCA 算法通过计算在降维空间中最大化数据方差的低维表示来执行降维。总的来说,PCA 算法需要将 h 维的原始数据 $\mathbf{X} = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^{n \times h}$ 线性映射到低维空间 $\mathbf{Y} \in \mathbb{R}^{n \times \tilde{h}}$, 其中 $\tilde{h} < h$ 。可以定义如下:

$$y_i = \mathbf{P}^T x_i \quad (4)$$

式中:变换矩阵 \mathbf{P} 可由目标函数(5)解得,式(5)定义如下:

$$\mathbf{P}^* = \operatorname{argmax} Tr(\mathbf{P}^T \operatorname{Cov}(\mathbf{X}) \mathbf{P}); s. t. \quad \mathbf{P}^T \mathbf{P} = \mathbf{I} \quad (5)$$

式中: $\operatorname{Cov}(\mathbf{X})$ 是 HSI 数据矩阵 \mathbf{X} 的协方差, $Tr(\mathbf{X})$ 表示矩阵 \mathbf{X} 的迹, \mathbf{I} 表示单位矩阵。

3) 多尺度降维融合

利用基于多数投票的支持向量机(SVM)分类器进行分类,对分割后不同的分类结果进行融合^[19],定义如下:

$$l = \operatorname{argmax} N(i), N(i) = \sum_{j=1}^{2k+1} a_j J(l_j = i) \quad (6)$$

式中: $i \in \{1, 2, \dots, G\}$ 。 l 是 G 个可能类之一的类标签,对于测试像素, j 是分类索引, $N(i)$ 表示在第 i 类中预测的次数, J 为指示函数。 a_j 表示第 j 个分类器的投票权。本文使用等强度投票策略来进行分类器的投票, $a_j = \frac{1}{2k+1} (j = 1, 2, \dots, 2k+1)$ 。对各个降维后的矩阵进行组合,就可以得到降维后 HSI 数据,使得最终的降维矩阵能够包含对应区域的特征信息^[16]。

1.2 K-means 选取锚点

锚点是从降维后的 HSI 数据点中筛选出来能表征整个数据集的点集,锚点的选取很大程度上影响着聚类的性能。由于随机选取锚点存在不能够很好的代表整个数据集的风险,本文采用 K-means 选取锚点的算法并构造单层锚点图。设 $\mathbf{Y} = [y_1, y_2, \dots, y_n]^T \in \mathbb{R}^{n \times \tilde{d}}$ 是经过超像素处理后的降维矩阵,其中 \tilde{d} 代表数据降维后的特征维数。将 n 个数据样本划分为 m 类,其中 $m \ll n$, 每个数据点 $y_i \in \mathbb{R}^{\tilde{d}}$ 都属于 m 类中的一类, m 个子聚类中心就是所要选取的锚点^[22]。由式(7)可求得数据点之间的距离:

$$d(y_i, y_j) = \sum_{k=1}^{\tilde{d}} (y_{ik} - y_{jk})^2 = \|y_i - y_j\|^2 \quad (7)$$

式中: y_i 和 y_j 为矩阵 \mathbf{Y} 中的样本。接着计算类中心到该类各样本数据间的距离,使其值为代价函数,可由式(8)计算出:

$$W(C) = \sum_{l=1}^m \sum_{C(i)=l} \|y_i - \bar{y}_l\|^2 \quad (8)$$

式中: $\bar{y}_l = (\bar{y}_{1l}, \bar{y}_{2l}, \dots, \bar{y}_{ml})^T$ 表示第 l 个类的中心点处数值, $n_l = \sum_{i=1}^n I(C(i) = l)$, $I(C(i) = l)$ 表示指示函数,取值通常为 0 或 1, $W(C)$ 为同一类样本相似程度,其取值最小时,表明所选的类的中心点趋于稳定,则该点即为一个锚点。由下式可求解最优化问题:

$$C^* = \operatorname{argmin} W(C) = \operatorname{argmin} \sum_{l=1}^m \sum_{C(i)=l} \|y_i - y_l\|^2 \quad (9)$$

设 $\mathbf{U} = [u_1, \dots, u_m]^T \in \mathbb{R}^{m \times \tilde{d}}$ 表示生成的锚。

如图 1 所示,通过 K-means 选取点并构造的锚点图,黑色点为 10 000 个原始数据点,浅色点为所选取的 500 个锚点,层间连线为部分数据点与锚点间权重,基于锚点的算法复杂度为 $O(n\tilde{d}mt)$ 。

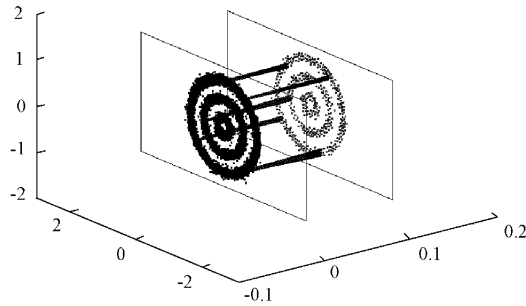


图 1 锚点图

1.3 构建相似度矩阵

构建相似图通常需要用高斯核函数计算两点间的距离,而高斯核函数中热核参数 σ 需手动调参,在设置不精确时易产生过拟合现象,为了避免出现该问题,本文采用一种无核邻近分配的算法来构建相似图。令 $\mathbf{A} = \{a_{ij}\} \in \mathbb{R}^{n \times n}$, $\forall i, j \in 1, \dots, n$ 表示亲和图的相似矩阵,定义 a_{ij} 为数据点 y_i 与 y_j 间邻接关系。用矩阵 $\mathbf{Z} \in \mathbb{R}^{n \times m}$ 表示数据集与所选锚点之间的邻接矩阵, z_{ij} 表示矩阵 \mathbf{Z} 的第 i 行第 j 列的元素,同时 z_{ij} 表示第 i 个数据点与第 j 个锚点的邻接关系^[22],定义如下:

$$\min_{z_i^T \mathbf{1} = 1, z_{ij} \geq 0} \|x_i - u_i\|_2^2 z_{ij} + \gamma z_{ij}^2 \quad (10)$$

式中: γ 为正则化参数。对于第 i 个数据点 x_i , 其他所有数据点 $\{x_1, x_2, \dots, x_n\}$ 连接到 x_i 的概率为。一般来说,距离 $\|x_i - x_j\|_2^2$ 越小,对应的概率 z_{ij} 越大,概率 $z_{ij} |_{j=1}^m$ 可由式(11)获得:

$$\min_{z_i^T \mathbf{1} = 1, 0 \leq z_i \leq 1} \sum_{j=1}^m \|x_i - x_j\|_2^2 z_{ij} \quad (11)$$

式中: $z_i \in \mathbb{R}^{n \times 1}$ 是第 j 个元素为 z_{ij} 的向量。式(11)有一个平凡解,即只有最近的数据点为 x_i 的临近点,其概率为 1。另一方面,如果不考虑两点间距离的情况下去解决以下问题:

$$\min_{z_i^T \mathbf{1} = 1, 0 \leq z_i \leq 1} \sum_{j=1}^m z_i^2 \quad (12)$$

最优解即为所有数据点都是 x_i 的邻近点,概率均为 $1/n$, 这种情况可以看作是邻元素的优先分配。结合式(11)和(12),概率 z_{ij} 可由式(12)求得:

$$\min_{z_i^T \mathbf{1} = 1, 0 \leq z_i \leq 1} \sum_{j=1}^m (\|x_i - x_j\|_2^2 z_{ij} + \gamma z_{ij}^2) \quad (13)$$

式中: $d_{ij}^x = \|x_i - x_j\|_2^2$ 表示第 j 个元素为 d_{ij} 的向量,则式(13)可写为向量形式,公式如下:

$$\min_{z_i^T \mathbf{1} = 1, 0 \leq z_i \leq 1} \|z_i + \frac{1}{2\gamma} d_i^x\|_2^2 \quad (14)$$

对于数据点 x_i , 可由式(14)来分配它的相邻数据。因此,可以通过求解式(15),为所有数据点分配相邻数据:

$$\min_{\forall i, z_i^T \mathbf{1} = 1, 0 \leq z_i \leq 1, j=1}^n \sum_{j=1}^n (\|x_i - x_j\|_2^2 z_{ij} + \gamma z_{ij}^2) \quad (15)$$

通过以上分析,可建立数据点 x_i 与锚点 u_i 之间的匹配关系,公式如下:

$$\frac{\min}{z_i^T \mathbf{1} = 1, z_{ij} \geq 0} \sum_{j=1}^m \|x_i - u_j\|_2^2 z_{ij} + \gamma z_{ij}^2 \quad (16)$$

式中: z_i^T 表示 \mathbf{Z} 的第 i 行,通过求解式(16)即可以得到式(10)中 \mathbf{Z} 的第 i 行, γ 取值通常在 0 到无穷之间,由文献[23]知, $\gamma_i = \frac{k}{2} d_{i,k+1}^x - \frac{1}{2} \sum_{j=1}^k d_{ij}^x$ 时式(14)有最优解, z_i 有精确的 k 个非零值,式(16)的解为:

$$z_{ij} = \frac{d_{i,k+1} - d_{i,j}}{k d_{i,k+1} - \sum_{j=1}^k d_{i,j}} \quad (17)$$

相似矩阵 $\mathbf{A}^{[17]}$ 为:

$$\mathbf{A} = \mathbf{Z} \mathbf{A}^{-1} \mathbf{Z}^T \quad (18)$$

式中: $\mathbf{A} \in \mathbb{R}^{m \times m}$ 为对角矩阵,其第 j 项可以表示为 $\mathbf{A}_{jj} = \sum_{i=1}^n z_{ij}$; \mathbf{A}^{-1} 为矩阵 \mathbf{A} 的逆。

1.4 谱聚类分析

谱聚类的目标函数^[4]为:

$$\min_{F^T F = I} \operatorname{Tr}(F^T \mathbf{L} F) \quad (19)$$

式中: $\mathbf{F} \in \mathbb{R}^{n \times c}$ 为类指引矩阵, n 为数据集数据点数, c 为标签类别数; $\mathbf{L} \in \mathbb{R}^{n \times n}$ 表示拉普拉斯矩阵;矩阵 \mathbf{L} 的 c 个最小的特征值对应的特征向量即为 \mathbf{F} 的最优解。拉普拉斯矩阵 \mathbf{L} 可由式(20)求出:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \quad (20)$$

式中:矩阵 $\mathbf{D} \in \mathbb{R}^{n \times n}$ 为对角矩阵,其第 i 项为 $d_{ii} = \sum_{j=1}^n \mathbf{A}_{ij}$ 。式(19)求解可以获得矩阵 \mathbf{F} , 矩阵 \mathbf{F} 为松弛的连

续解,需要进行 K-means 离散化才可以获得聚类结果。

1.5 算法步骤

步骤 1) 输入高光谱图像原始数据。

步骤 2) 利用 SuperPCA 算法对输入的高光谱数据集进行超像素切割,在其基础上对各同质区域进行 PCA 处理,最终获得降维后的高光谱数据。

步骤 3) 通过 K-means 选择具有代表性的锚点,并构建数据集与所选锚点之间的邻接矩阵。

步骤 4) 对谱聚类目标函数式(19)进行求解得到的松弛连续解,接着对所得松弛连续解进行 K-均值分析即可得到最终的聚类结果。

1.6 时间复杂度分析

输入的高光谱数据矩阵为 \mathbf{X} , 维度为 d , 数据点数为 n , 经过 SuperPCA 降维后的数据矩阵为 \mathbf{Y} , 维度为 \tilde{d} , 锚点数 m , 类别数 c , 簇类数 k 。CASAC 算法的计算复杂度可以分为:

- 1) 进行超像素切割降维,复杂度为 $O(n\tilde{d}d^2 + \tilde{d}d^3)$ 。
- 2) 基于 K-means 选取锚点的计算复杂度为 $O(n\tilde{d}mt)$, t 为生成锚点时 K-means 迭代次数。
- 3) 构建基于锚点的相似图的计算复杂度为 $O(n\tilde{d}m)$ 。
- 4) 获得松弛的连续解矩阵 \mathbf{F} 的计算复杂度为 $O(m^3 + m^2n)$ 。
- 5) 针对矩阵 \mathbf{F} 执行 K-means 离散化获得最终的聚类结果,其计算复杂度为 $O(nmr)$, r 为谱聚类中迭代次数。

该算法计算复杂度为 $O(n\tilde{d}d^2 + \tilde{d}d^3 + n\tilde{d}mt + n\tilde{d}m + m^3 + m^2n + nmr)$ 。

2 实验与分析

2.1 高光谱数据集

本文采用两个高光谱数据集进行验证。分别是 Indian Pines 数据集和 Pavia Centre 数据集。

Indian Pines:数据集是 1992 年在印度西北部由一个维利斯传感器采集。噪音和吸水带被丢弃,剩下有 145×145 pixels 和 220 个光谱带。数据集包含 16 个类别的 21 025 个数据样本用于实验研究。

Pavia Centre:数据集由反射光学系统成像光谱仪传感器所采集。去掉无用数据后,剩下有 $1 096 \times 715$ pixels,102 光谱带。数据集包含 9 个类别的 783 640 个数据样本。所以用于实验研究的光谱带为 102 个和总样本数为 783 640 个。

通过和其他 5 种流行的聚类算法进行比较,本文通过总准确度(OA),平均精度(AA),Kappa 系数,运行时间和聚类图对实验结果进行评估。聚类算法分别为:K-means、FCM、FCM_S1、SC 和 SGCNR。高光谱数据集分别是 Indian Pines、Pavia Centre。

2.2 Indian Pines 数据集实验结果分析

为了证明提出的算法的有效性,将其和 K-means、FCM、FCM_S1、SC、SGCNR 算法在两个高光谱数据集上进行了比较。本文通过总准确度(OA),平均精度(AA),Kappa 系数,运行时间和聚类图对实验结果进行评估。

在 Indian Pines 数据集上锚点的数量和近邻点数分别被设置为 $m=4 096$ 和 $k=5$ 。如表 1 所示,CASAC 的总聚类精度以及平均精度都是最高的,其中总聚类精度为 46.85%,平均精度为 39.54%。对于 Kappa 系数,CASAC 得到的 Kappa 系数是 0.350 7,比其他算法的 Kappa 系数高 0.062 5~0.108 2。与 K-means、FCM 和 FCM_S1、SC、SGCNR 相比,CASAC 提高了聚类性能。从聚类结果可以看出,K-means、FCM 和 FCM_S1、SGCNR 的聚类时间是同一数量级的,然而,SC 算法的消耗时间为 55.3 s,而 CASAC 为 18.323 6 s,比 SC 算法快了 3 倍。与其他算法相比,CASAC 在 AA,OA,Kappa 系数上都是最优的,时间比传统的 SC 算法要好。如图 2 所示,CASAC 得到的分类图中错分点更少,地物分布更加平滑,证明了 CASAC 在高光谱数据集下的高效性和可行性。

表 1 Indian Pines 数据集的实验结果

方法	K-means	FCM	FCM_S1	SC	SGCNR	CASAC
AA/%	36.33	35.30	33.16	31.43	28.26	46.85
OA/%	34.21	35.70	35.68	32.65	30.44	39.54
Kappa	0.268 9	0.284 4	0.288 2	0.267 3	0.242 5	0.350 7
时间/s	3.172 3	4.139 1	6.558 8	55.3	6.085	18.323 6

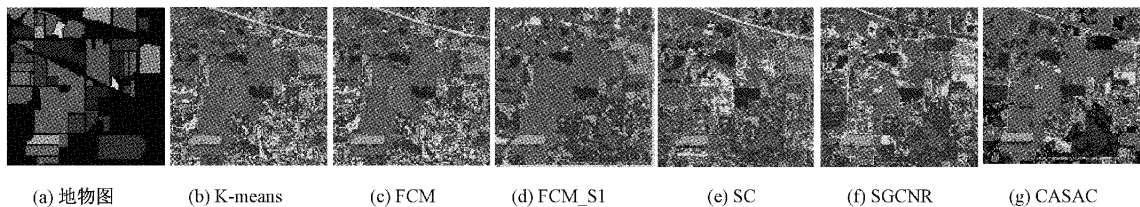


图 2 各算法在 Indian Pines 数据集下的聚类图

2.3 Pavia Centre 数据集实验结果分析

在 Pavia Centre 数据集上进行了实验验证,锚点数和近邻数分别设为 $m=256, k=6$ 。通过平均准确度、总体准确度和 Kappa 系数以及视觉聚类图来全面评估各种聚类算法的聚类性能。表 2 是所有算法在 Pavia Centre 数据集实验结果,为了便于观察,对最佳的结果进行了加粗标记。如表 2 所示,CASAC 的总聚类精度为 62.22,比其他算法都高。CASAC 的平均精度以及 Kappa 系数分别为 75.67、0.659 9,和其他算法的平均精度、Kappa 系数不相上下。与 K-means、FCM 和 FCM_S1、SC、SGCNR 相比,CASAC 提高了聚类性能。注意,Pavia Centre 属于大规模高光谱

数据集,样本数为 783 640。但由于“内存不足(OM)”的原因,SC 算法无法在 Pavia Centre 数据集上工作。从聚类结果可以看出,K-means 和 FCM 处于同一数量级,用时较短;FCM_S1、CASAC 以及 SGCNR 三种算法处于同一数量级。具体来看,CASAC 为 128.353 9 s,相比较于 FCM_S1 算法,其用时更长,但总聚类精度更好,与 SGCNR 算法相比,其聚类精度更好,用时更短。总的来说,CASAC 在 AA 系数上结果最优,在 OA,Kappa 系数也都有着不错的表现。如图 3 所示,CASAC 得到的分类图中错分点更少,地物分布更加平滑,证明了 CASAC 在 Pavia Centre 数据集下的高效性和可行性。

表 2 Pavia Centre 数据集的实验结果

方法	K-means	FCM	FCM_S1	SGCNR	CASAC
AA/%	53.79	53.76	54.06	61.38	62.22
OA/%	74.47	75.06	75.24	74.23	75.67
Kappa	0.650 6	0.646 5	0.650 6	0.653 7	0.655 9
时间/s	34.138 1	26.518 3	96.76	145.037 4	128.353 9

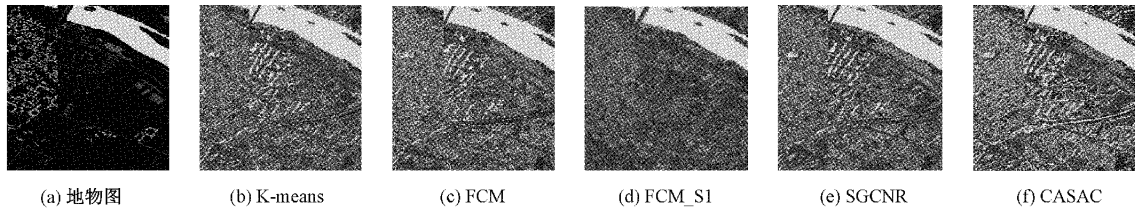


图 3 各算法在 Pavia Centre 数据集下的聚类图

2.4 参数实验

本文提出的 CASAC 算法需要用到的参数为 k 和 m ,其中 k 为近邻数, 2^m 为锚点数。为了取得最佳的实验结果,本文计算出了不同 m 和 k 下的 AA、OA 以及 Kappa 值, k 的取值范围为 2~26,在 Indian Pines 数据集上锚点的数量为 32~8 192,在 Pavia Centre 数据集上锚点数超过 1 024 时,算法耗时久且聚类精度提升较小,不予考虑,锚点取值范围设定为 8~512。在 Indian Pines 数据集上的实

验结果如图 4 所示,随着 m 的增加,AA、OA 以及 Kappa 曲线都呈现出先上升后下降再上升至最高点的一个过程,达到最高点后曲线开始回落,其中 AA 曲线最曲折,OA 曲线变化较小, $m=4 096$ 时,各项精度指标均达到最大。随着 k 的增加,AA、OA 以及 Kappa 曲线都呈现出先上升至最高点再下降再上升并逐渐回落的过程, $k=10$ 时各项精度指标达到最大。所以在 Indian Pines 数据集实验中选取 $k=10$ 和 $m=4 096$ 。

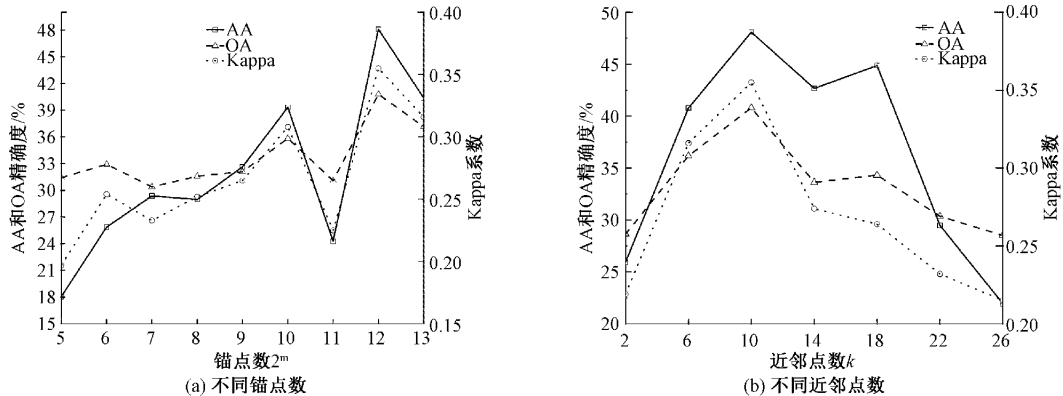
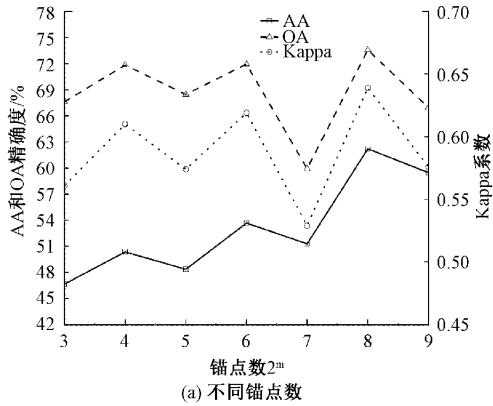


图 4 Indian Pines 数据集下的实验结果

在 Pavia Centre 数据集上实验结果如图 5 所示,随着 m 的增加,AA、OA 以及 Kappa 曲线呈现出波动上升后快速下降,再上升至最高点并逐渐下降的趋势, m 为 64~512 时曲线上升以及回落的幅度较大, $m=256$ 时各项精度数



据达到最大值。随着 k 的增加,曲线呈现出快速上升,随后缓慢回落的趋势, $k=6$ 时,AA、OA 以及 Kappa 系数取最大值。所以在 Pavia Centre 数据集实验中选取 $k=6$ 和 $m=256$ 。

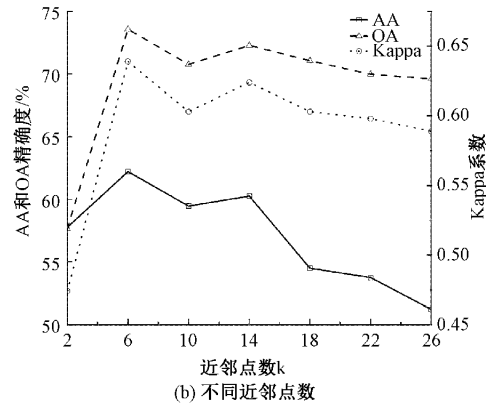


图 5 Pavia Centre 数据集下的实验结果

3 结 论

传统分类算法应用于高光谱图像分类时存在精度不佳、耗时久等难题,本文以提升高光谱图像分类性能为目标,创新性的在 K-means 选取锚点的基础上加入了超像素主成分分析并对数据进行处理,提出了一种新的基于超像素锚层收敛选点的高光谱聚类算法,该算法在有效降低了高光谱图像聚类复杂度的同时提高了聚类精度,实现了更好的高光谱图像分类效果。在两个常用高光谱图像数据集上进行对比实验,实验表明与传统的高光谱图像分类算法相比,本文所提高光谱分类算法所含错分点更少,地物分布更加平滑,在聚类精度、聚类耗时方面具有显著优势,充分说明了本文算法能够更好地应用于高光谱图像聚类的问题。

参考文献

- [1] 李铁,孙劲光,张新君,等. 高光谱遥感图像空谱联合分类方法研究[J]. 仪器仪表学报, 2016(6): 1379-1389.
- [2] 杨明莉,范玉刚,李宝芸. 基于 LDA 和 ELM 的高光谱图像降维与分类方法研究[J]. 电子测量与仪器学报, 2020, 34(5): 190-196.
- [3] 张琳. 基于 a-BvSBEM 主动学习的高光谱图像分类[J]. 国外电子测量技术, 2017(4): 17-20.
- [4] ZHANG H, ZHAI H, ZHANG L, et al. Spectral-spatial sparse subspace clustering for hyperspectral remote sensing images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54 (6): 3672-3684.
- [5] ZHANG B, LI S, WU C, et al. A neighbourhood-constrained K-means approach to classify very high spatial resolution hyperspectral imagery [J]. Remote

Sensing Letters, 2013, 4(2): 161-170.

- [6] GONG M, LIANG Y, SHI J, et al. Fuzzy C-means clustering with local information and kernel metric for image segmentation [J]. IEEE Transactions on Image Processing, 2013, 22(2): 573-584.
- [7] CHOUDHRY M S, KAPOOR R. Performance analysis of fuzzy C-means clustering methods for MRI image segmentation [J]. Procedia Computer Science, 2016, 89: 749-758.
- [8] NIAZMARDI S, HOMAYOUNI S, SAFARI A. An improved FCM algorithm based on the SVDD for unsupervised hyperspectral data classification [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2013, 6 (2): 831-839.
- [9] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344: 1492-1496.
- [10] ZHONG Y, ZHANG S, ZHANG L. Automatic fuzzy clustering based on adaptive multi-objective differential evolution for remote Sensing imagery [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2013, 6(5): 2290-2301.
- [11] ZHAI H, ZHANG H, ZHANG L, et al. A new sparse subspace clustering algorithm for hyperspectral remote sensing imagery [J]. IEEE Geoscience and Remote Sensing Letters, 2016, 14(1): 43-47.
- [12] YANG X, LIN G, LIU Y, et al. Fast spectral embedded clustering based on structured graph learning for large-scale hyperspectral image [J]. IEEE

- Geoscience and Remote Sensing Letters, 2020, PP(99): 1-5.
- [13] WANG S, ZHOU S, GUO X, et al. Anchor-based multi-view subspace clustering with diversity regularization[J]. IEEE Multimedia, 2020, 27(4): 91-101.
- [14] ZHU W, NIE F P, LI X. Fast spectral clustering with efficient large graph construction[C]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017: 2492-2496.
- [15] WANG R, NIE F, WANG Z, et al. Scalable graph based clustering with nonnegative relaxation for large hyperspectral image[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(10): 7352-7364.
- [16] JIANG J, MA J, CHEN C, et al. SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(8): 4581-4593.
- [17] QIAO L, ZHANG L, CHEN S, et al. Data-driven graph construction and graph learning: A review[J]. Neurocomputing, 2018, 312: 336-351.
- [18] NIE F, ZHU W, LI X. Unsupervised large graph embedding[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017.
- [19] LIU M Y, TUZEL O, RAMALINGAM S, et al. Entropy rate superpixel segmentation [C]. CVPR 2011, IEEE, 2011: 2097-2104.
- [20] LI J, ZHANG H, ZHANG L. Efficient superpixel-level multitask joint sparse representation for hyperspectral image classification [J]. IEEE Transactions on Geoscience & Remote Sensing, 2015, 53(10): 5338-5351.
- [21] DING M, ANTANI S, JAEGER S, et al. Local-global classifier fusion for screening chest radiographs [C]. Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications, SPIE, 2017, 10138: 64-69.
- [22] WANG R, NIE F, YU W. Fast spectral clustering with anchor graph for large hyperspectral images[J]. IEEE Geoscience & Remote Sensing Letters, 2017, 14(11): 2003-2007.
- [23] NIE F, WANG X, JORDAN M, et al. The constrained Laplacian rank algorithm for graph-based clustering[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 1969-1971.

作者简介

杨滔, 硕士研究生, 主要研究方向为机器学习与高光谱图像处理。

E-mail: 1214503613@qq.com

孙博, 讲师, 主要研究方向为集成电路封装设计、建模和仿真, 半导体器件多物理场可靠性建模与仿真, 电路-封装协同设计技术、可靠性评估及可控寿命技术。

E-mail: bosun@gdut.edu.cn

杨晓君(通信作者), 副教授, 主要研究方向为目标跟踪、无源定位、机器学习、数据聚类。

E-mail: yangxj18@gdut.edu.cn