系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
38(12) (2018, 12), 1417–1428

基于模糊 c 均值算法和人工蜂群算法的无监督波段选择*

谢福鼎[1] 雷存款[1] 李芳菲[1] 嵇 敏[2]

(1. 辽宁师范大学城市与环境学院, 大连 116029; 2. 辽宁师范大学计算机与信息技术学院, 大连 116081)

摘要 波段选择是高光谱影像处理中一种重要的降维方法. 在类标签不可获得的情况下, 如何选择出一个具有代表性的波段子集是一个挑战性的问题. 为了解决高光谱数据维数灾难以及光谱空间冗余的问题, 基于模糊 c 均值算法 (Fuzzy c-means, FCM), 人工蜂群算法 (Artificial Bee Colony, ABC) 与极大熵准则 (Maximum Entropy, ME), 文章提出了一种新的无监督波段选择方法. 该方法首先通过 FCM 算法将相似的波段划分到一个波段子集中, 然后以 ME 为 ABC 算法中的适应度函数, 寻找优化的波段子集. 为验证该算法的有效性, 在三个典型的高光谱数据集上, 将所提出的方法和其它一些有效的波段选择算法进行了分类精度和计算时间对比. 实验结果表明, 所提出的算法不但可以得到高的分类精度, 同时在计算时间上也具有明显的优势.

关键词 高光谱图像, 波段选择, 模糊 c 均值算法, 人工蜂群算法, 分类.

MR(2000) 主题分类号 62H30, 68T10

Unsupervised Band Selection Based on Fuzzy c-Means Algorithm and Artificial Bee Colony Algorithm

XIE Fuding[1] LEI Cunkuan[1] LI Fangfei[1] JI Min[2]

(1. *Department of Urban and Environmental Science, Liaoning Normal University, Dalian* 116029;
2. *Department of Computer Science and Technology, Liaoning Normal University, Dalian* 116081)

Abstract Band selection is an important dimension reduction technique in hyperspectral image processing. Under the condition that class labels are unavailable, how to select a representative band subset becomes a severe challenge. To address the problem of dimension disaster and spectral space redundancy, a novel unsupervised band selection method is introduced based on fuzzy c-means algorithm (FCM), artificial bee colony algorithm (ABC) and maximum Entropy (ME). In this method, the

* 国家自然科学基金 (41771178, 61772252) 资助课题.
收稿日期: 2018-12-03.
通信作者: 嵇敏, Email: jimin@lnnu.edu.cn.
编委: 杨争峰, 冯如勇.

similar bands are firstly divided into a band subset by FCM algorithm, and then ME is adopted as the fitness function in the ABC algorithm to find the optimized band subset. To verify the effectiveness of the proposal, the proposed method is compared with other efficient band selection algorithms on three typical hyperspectral datasets. Experimental results show that the proposed algorithm not only can achieve high classification accuracy, but also has obvious advantages in computing time.

Keywords Hyperspectral image, band selection, Fuzzy c-means algorithm, artificial bee colony algorithm, classification.

1 引 言

高光谱遥感是利用密集而大量的电磁波谱来获取感兴趣区域地物的相关信息. 高光谱遥感影像是将反映目标地物辐射的光谱信息与反映目标地物空间分布的影像信息有机地集于一体, 实现了 "图谱合一". 与传统的多光谱影像相比, 高光谱影像具有大量的更为密集的波段, 可以对同一目标地物连续成像, 因此对于不同土地覆盖类型具有更好的区分能力. 然而, 在对高光谱影像进行分类处理时, 丰富的波段信息同样导致了信息的冗余以及计算时间复杂度的增加. 因此, 如何在减少波段数量的同时, 保留波段分类性能就成为了高光谱图像分类中的关键问题.

一般来说, 高光谱数据的降维方法可以分为两类: 波段 (特征) 选择 [1,2] 和波段提取 [3]. 波段选择是通过一些规则或算法将原始高光谱数据中的关键和显著的波段挑选出来, 从而达到降低维数的目的. 而波段提取是通过变换将原始高光谱影像映射到另一空间, 并从该空间中提取出重要的波段作为原始数据的代表特征. 在高光谱遥感中, 高光谱影像的光谱信息和二维空间信息共同构成了高光谱数据立方体. 在光谱空间中, 每个像素点反映为一条连续光谱响应曲线, 不同的目标地物在高光谱影像中表现为不同的辐射强度. 在影像空间中, 每个波段则对应着一幅二维影像, 通常被视为一个特征. 相比波段提取, 波段选择能更好地保留原始波段的物理意义及光谱特征, 同时又达到了降维的目的, 因而深受广大高光谱领域工作者的青睐 [4].

无监督波段选择作为高光谱影像降维的重要手段, 一直是高光谱遥感研究领域的热点之一. 传统的无监督波段选择方法主要包含两类: 基于排序 [5] 的和基于聚类 [6−8] 的方法. 前者通过使用不同尺度 (如非高斯) 的度量准则对每个波段进行排序, 然后根据给定的维数或阈值来选择最具代表性的波段; 虽然可以根据每个波段的得分快速选择出具有代表性的波段, 但可能存在所选波段之间相关性较高的问题. 后者采用聚类的思想选择波段, 首先将原始波段划分为多个类簇, 使得属于同一类簇的波段尽可能相似, 然后从每个类簇中选取具有代表性的波段. 因此, 通过基于聚类的波段选择方法, 可以从整个波段空间中选择相似度尽可能小的波段子集. 传统的波段子空间分解方法 [7,8] 利用近邻波段相关系数中的极小点对高光谱波段空间进行分解, 虽然该方法易于理解, 但存在如下两个问题: 一是多个极小值点的存在导致波段子空间难以分解为恰当的个数; 二是只考虑相邻波段之间的相关系数, 而没有进行相关系数矩阵的全局性考虑, 可能会导致部分波段子空间划分结果不准确的问题. 因此, 本文提出了一种基于模糊 c 均值 (FCM) 聚类的波段子集分解方法. 通过 FCM 聚类自

适应地将波段划分为不相交的类簇 (波段子集), 使得同一类中的波段比不同类中的波段更为相似, 再从划分后的波段子集中选择具有代表性的波段, 构成一个波段组合.

从原始的波段空间中寻找 S 个最优波段是一个 NP-Hard 问题. 当原始数据空间维数较高时, 由于计算复杂度过高, 对所有可能的波段子集进行穷尽搜索通常是不可行的. 因此, 本文采用随机搜索策略作为最优波段组合的搜索方式.

近年来, 生物启发优化算法被广泛应用于高光谱波段选择的研究. 具有代表性的有: 遗传算法 [9], 蚁群算法 [10], 粒子群算法 [11] 以及人工蜂群算法 [12−14]. 其中, 人工蜂群算法是由 Karaboga 等人提出的一种新的元启发式优化算法, 它是在搜索过程中受蜜蜂自然觅食行为的启发, 模拟了自然界蜂群的自组织模型和群体智能, 最初用于多变量数值问题的优化 [12]. 该算法的突出优点是较好地协调了在原搜索范围进行精密搜索与搜索范围的扩展之间的矛盾, 从而大大提高了寻优的概率. 同时, 该算法全局搜索能力好, 收敛速度快, 鲁棒性强 [8]. 在文献 [12–14] 中, 大量的数值结果表明了 ABC 算法相比其他优化算法性能的优越性.

2 相关工作

2.1 FCM 算法

假设数据集 $DS = (x_1, x_2, \cdots, x_n) \subset R^d$, n 为样本点个数. 聚类就是将 DS 划分为 c 个互不相交的子集 (类), 使得在每个子集内的样本点尽可能的相似, 不同子集间的样本点尽可能的不相似.

聚类问题可以形式地表示为

$$DS = C_1 \cup C_2 \cdots \cup C_c, C_i \cap C_j = \emptyset, \quad i \neq j,\ i, j = 1, 2, \cdots, c, \tag{1}$$

其中 C_i 表示第 i 个类.

FCM 聚类算法就是通过极小化类内平方误差的加权和, 从而将 DS 分解为 c 个互不相交模糊子集.

其数学表达式为

$$\begin{aligned} &\text{Min} \quad J(U, V) = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^{m} (x_i - v_j)^2, \\ &\text{s.t.} \quad \mu_{ij} \in [0, 1], \quad \sum_{j=1}^{c} \mu_{ij} = 1, \end{aligned} \tag{2}$$

其中 m 为模糊因子 ($m > 1$), v_j 表示第 j 类的质心, μ_{ij} 代表第 i 个样本属于第 j 个类的隶属程度.

显然, 这是一个带有约束条件的极值优化问题. 通过拉格朗日乘子法求解方程 (2), 可以得到隶属度 μ_{ij} 和聚类中心 v_j 的迭代更新公式

$$\mu_{ij} = \left[\sum_{k=1}^{c} \left(\frac{(x_i - v_j)^2}{(x_i - v_k)^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \tag{3}$$

$$v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}}. \tag{4}$$

FCM 算法通过不断更新每个类的质心和每个样本点的隶属度, 直到所有类质心趋于稳定, 我们就得到了最终的聚类结果.

对于高光谱数据, 我们将每个波段看作一个样本点, 然后将所有初始波段聚类, 得到了一个波段子集的分解.

2.2 最大熵原理 (Maximum Entropy, ME)

在优化算法中, 准则函数的选择是一个关键的问题. 本文采用最大熵作为人工蜂群算法中的适应度函数, 以选择包含最丰富信息的波段子集.

在信息论中, 熵 (Entropy) 被定义为信息度量的基本单位. 设 X 是一个随机变量, 则它所包含的信息量可以用熵来量化, 如公式 (5) 所示

$$E(X) = -\sum_i p(x_i)\log_2 p(x_i), \tag{5}$$

其中, $E(X)$ 代表随机变量 X 的熵值; $p(x_i)$ 表示随机变量 X 第 i 个分量的概率密度函数. $p(x_i)$ 可以通过直方图方法求得.

波段选择是从原始波段集合 B 中选择一个最优波段子集 B', 为了最大限度地保留原始波段的信息, 一种有效的方法是选择包含信息最丰富的波段构成波段子集 [4]. 假设波段子集 $B' = \{B_1, B_2, \cdots, B_S\}$, 其信息丰富程度通过如公式 (6) 计算

$$\frac{1}{S}\sum_{i=1}^{S} E(B_i), \tag{6}$$

其中 S 为波段子集中包含的波段个数. 通过公式 (6), 我们得知: 波段子集 B' 的值越大, 表示其包含的 S 个波段越具有较好的地物分类能力.

2.3 人工蜂群算法 (Artificial Bee Colony, ABC)

Karaboga 等 [12] 提出的人工蜂群算法是一种模拟蜂群觅食行为的群体优化算法. 该算法模拟的人工蜂群由三种类型的蜜蜂构成, 分别是: 雇佣蜂, 跟随蜂和侦察蜂.

这种在觅食过程中出现的智能行为可以概括为

1) 在觅食过程的初始阶段, 蜜蜂开始随机探索周围环境以寻找蜜源. 雇佣蜂负责开采以前探索过的蜜源 (一个雇佣蜂对应一个蜜源), 并向蜂巢中等待的蜜蜂 (跟随蜂) 提供它们正在开采的蜜源地点的质量信息.

2) 跟随蜂根据雇佣蜂分享的信息, 通过轮盘赌选择的方式选择要开采的蜜源, 并在其周围进行贪婪搜索, 择优保留优质蜜源.

3) 对蜜源持续性的开采和利用会使得蜜源很快被消耗殆尽, 于是被消耗尽的蜜源所对应的雇佣蜂, 会转变成侦察蜂, 开始随机搜索蜂巢附近的新蜜源, 更新蜂群种群, 开始新的迭代.

在 ABC 算法中, 蜜源的位置代表了波段选择问题的一个可行解, 蜜源处的花蜜量 (收益度) 代表了可行解的质量, 花蜜量 (收益度) 越充足, 代表可行解的质量越好. 通过连续迭代, 该算法保持优良个体, 淘汰劣质个体, 并向全局最优解逼近 [8].

3 基于 FCM 和 ABC 算法的波段选择方法

在 FCM 算法中, 聚类数 c 通常需要预先指定. 在没有先验知识的情况下, 往往难以预先指定恰当的 c 值. 本文通过光谱数据集的光谱曲线图的可视化结果来确定聚类数. 波段的聚类还被称为波段子空间分解. 在波段聚类的基础上, 从每个类中随机选取 k 个不同的波段构成一个新的特征向量 (蜜源), 如图 1 所示, a_{ij} 表示从第 i 个类簇中选取的第 j 个特征. 这种选择方法有效地减少了特征之间的冗余, 但不能保证得到的蜜源 (波段组合) 具有良好的分类能力. 因此需要结合人工蜂群算法, 通过寻找最优蜜源的过程来优化该波段组合. 该方法能够在最大化保留原始波段信息的同时, 最小化所选波段之间的冗余信息.

a_{11}	$\cdots$	a_{1k}	$\cdots\ \cdots$	a_{c1}	$\cdots$	a_{ck}

图 1 包含 $S = k \times c$ 个波段的蜜源

(Figure 1 A nectar source consisted of $S = k \times c$ bands)

波段子集的优化过程如下.

蜜源初始化

按照公式 (7) 随机产生 N 个初始蜜源, 构成第 0 代初始种群, $A = \{a_i | i = 1, 2, \cdots, N\}$, 每个蜜源 a_i 代表了一种可能的波段组合, 由从不同子空间中选择的 S 个波段构成 $\{a_{ij} | j = 1, 2, \cdots, S\}$.

$$a_i^0 = a_i^{\min} + \text{ceil}[\text{Rand} \times (a_i^{\max} - a_i^{\min})], \quad i = 1, 2, \cdots, N, \tag{7}$$

其中, a_i^0 代表第 0 代种群中的第 i 个蜜源, $a_i^{\max}$ 和 $a_i^{\min}$ 分别是每个波段子集的上, 下界向量; Rand 是 (0,1) 范围内的均匀分布随机数.

蜜源的花蜜量 (收益度) 与可行解的质量相对应, 根据极大熵 (ME) 准则, 每个蜜源 a_i 的收益度 $F(a_i)$ 可由公式 (8) 计算,

$$F(a_i) = \frac{1}{S} \sum_{j=1}^{S} E(a_{ij}). \tag{8}$$

分别计算 N 个蜜源的收益度, 并按照从大到小进行排序, 前 $\frac{N}{2}$ 个蜜源对应引领蜂个体, 后 $\frac{N}{2}$ 个蜜源对应跟随蜂个体.

雇佣蜂阶段

对第 t 代蜜源 a_i^t 进行领域搜索, 随机选择个体 a_k^t, $k \in \{1, 2, \cdots, \frac{N}{2}\}$, $k \neq i$, 依照公式 (9) 随机选择 a_i^t 的某一维进行位置更新, 产生领域蜜源 V_i^t, 比较 a_i^t 和 V_i^t 的收益度值 F. 根据贪婪选择原则, 若 $F(a_i^t) < F(V_i^t)$, 则新蜜源 V_i^t 替换原蜜源 a_i^t; 否则, 保持 a_i^t 的位置不变.

$$V_{ij}^t = a_{ij}^t + \text{ceil}[(-1 + 2\text{Rand}) \times (a_{ij}^t - a_{kj}^t)]. \tag{9}$$

从公式 (9) 中可以看出, 如果 a_{ij}^t 和 a_{kj}^t 两个变量之间的差值减小, 那么对于 a_{ij}^t 位置上的扰动也会减小. 因此, 当搜索接近于全局最优解时, 步长就会自适应地减小 [12].

跟随蜂阶段

跟随蜂采用轮盘赌选择的方法, 选择较优蜜源进行开采. 根据公式 (10) 计算蜜源 a_i^t 被跟随蜂选择跟随的概率 P_i, 并按照公式 (9) 进行邻域搜索, 其中 $k \in \{\frac{N}{2} + 1, \frac{N}{2} + 2, \cdots, N\}$.

根据贪婪选择原则择优选择蜜源, 更新跟随蜂种群.

$$P_i = \frac{F(a_i)}{\sum_{i=1}^{\frac{N}{2}} F(a_i)}. \tag{10}$$

由公式 (10) 可知, 蜜源的收益度越高, 该蜜源被跟随蜂个体选择跟随的概率也随之越大.

侦察蜂阶段

记录每个蜜源被开采的次数, 若开采次数大于规定最大停留次数 Limit, 则舍弃该蜜源, 此时雇佣蜂的角色转变为侦察蜂, 根据公式 (7) 重新产生新蜜源.

本文所提出的 FCM-ABC 算法描述如下

输入: 高光谱数据集 DS; ABC 算法相关参数 (蜜源数量 N, 最大迭代次数 G, 蜜源最大开采次数 Limit).

输出: 最优波段子集 B'.

步骤 1 利用 FCM 聚类将波段分解为不同的波段子空间, 结合公式 (7) 随机从每个子空间中选择 k 个不同波段构成一个蜜源 a_i, 重复选择 N 次构成初始化种群 A, 并设置迭代次数 $t = 0$.

步骤 2 依照公式 (8) 计算 N 个蜜源的收益度 F_{a_i}, 并按照大小进行排序, 前 $\frac{N}{2}$ 个蜜源对应引领蜂个体, 后 $\frac{N}{2}$ 个蜜源对应跟随蜂个体.

步骤 3 根据公式 (9) 更新雇佣蜂种群.

步骤 4 根据公式 (9) 和 (10) 更新跟随蜂种群.

步骤 5 根据步骤 3 和 4 中的更新个体, 形成新的迭代种群.

步骤 6 判断每个蜜源被开采的次数是否达到最大停留次数 (Limit). 若是则舍弃该蜜源, 并重新产生新蜜源, 更新迭代种群. 同时记录当前蜂群搜索到的最优蜜源 (即最优波段组合 B').

步骤 7 令 $t = t + 1$, 判断当前迭代次数, 若 $t < G$, 则跳转至步骤 2, 开始新的迭代过程. 直至满足最大迭代次数, 此时停止搜索, 输出最优蜜源所对应的最优波段组合 B'.

4 实验结果与分析

为验证 FCM-ABC 算法的有效性, 我们在 Indian Pines, Pavia University 和 Salinas 3 个典型的高光谱数据集上进行了实验. 在遥感领域, 这 3 个数据集被广泛地用于测试高光谱数据分类方法的有效性. 实验环境为 AMD 四核处理器, CPU 3.99 Hz, 程序运行平台为 Matlab 2014a.

另外, 我们还采用监督波段选择方法 (TMI)[15], 半监督波段选择方法 (SDIR)[16] 以及三种无监督波段选择方法 (MIC)[17], FCM-GA[18] 和 WaluMI[19] 作为对比方法. ABC 算法中的初始种群由 $N = 30$ 个蜜源组成, 最大迭代次数为 $G = 150$, 蜜源最大停留次数 Limit = 5. 在 ABC 算法的研究中, 一般采用的种群规模为 20 到 50 之间 [21]. 为了分析 ABC 算法的有效性, 我们在 Indian Pines 数据集上运行了从 10 到 100 的不同的种群规模. 实验结果表明, 当种群规模 N 从 10 增加到 30 时, 分类精度 (总体精度 OA) 随着种群规模的增加而呈上升趋势, 随后呈现出下降趋势. 当种群规模为 30 时, 我们的算法产生了最好的分类结果. 种群规模的增长除了增加了时间复杂度外, 并不会显著影响 ABC 算法的性能 (约 1%). 对于最大迭

代次数 G, 通常设置为 50 到 300. 通过实验分析得出, 当迭代次数约为 120 次时, 算法明显收敛. 因此, 我们将这 3 个数据集的迭代次数设为 $G = 150$.

为了评价所选择波段的有效性, 采用了基于像素的高光谱图像分类方法, 使用支持向量机 (Support Vector Machine, SVM)[21] 作为分类器. 在 SVM 中, 我们采用一对多的方法来处理多分类问题, 以径向基函数 (RBF) 作为核函数, 并利用五层交叉验证对参数 C 和 γ 进行寻优. 对于每个数据集, 随机标记 20% 的样本构成测试集, 其余的 80% 作为测试集. 采用 3 个常用的指标来评价不同算法对高光谱影像的分类性能, 分别为总体精度 (OA), 平均精度 (AA) 和卡帕系数 (Kappa). 实验结果中所给出的分类精度为 SVM 分类器独立运行 30 次取均值和相应标准差的结果.

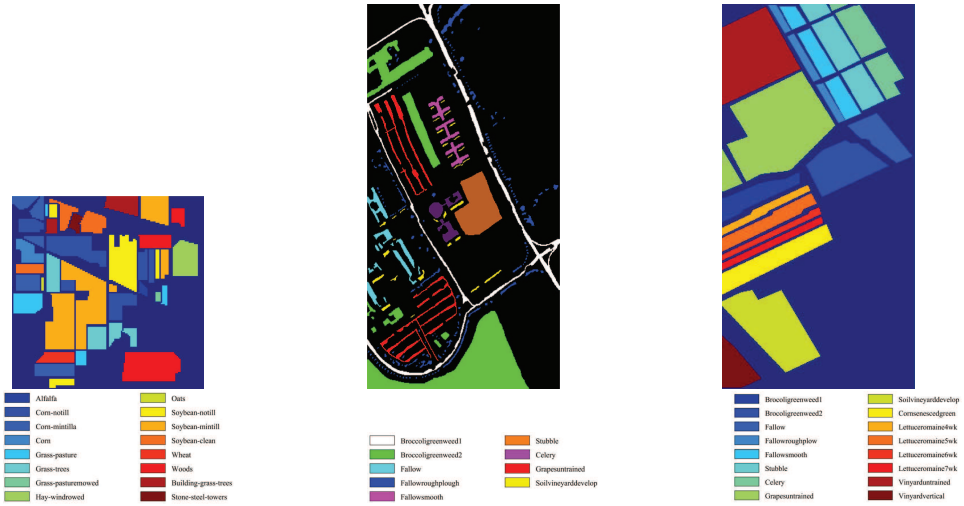
4.1 实验数据介绍

实验数据分别为由 AVIRIS 传感器采集的 Indian 和 Salinas 数据集以及 ROSIS 传感器采集的 Pavia University 数据集. 具体数据集描述如表 1 所示. 图 2 为 3 个高光谱数据集的地物分类图.

表 1 3 个高光谱数据集描述

(Table 1 The description of three hyperspectral datasets)

	Indian Pines	Pavia University	Salinas
类别数	16	9	16
波段数	200	103	204
影像尺寸	145 × 145	610 × 340	512 × 217
分辨率	20 m	1.3 m	3.7 m
样本数	10,249	42,776	54,129



(a) (b) (c)

图 2 3 个高光谱数据集的地物分类图. (a) Indian Pines 数据集. (b) Pavia University 数据集. (c) Salinas 数据集

(Figure 2 The ground truth of three hyperspectral datasets. (a) Indian Pines dataset. (b) Pavia University dataset. (c) Salinas dataset)

4.2 分类结果

表 2 记录了在 3 个数据集上的分类结果 (包含每类地物的平均分类精度和标准差, 以及 OA, AA, Kappa 系数和相应的标准差). 为了便于和其他 5 种算法的分类结果相比较, 我们采用了和这 5 种算法同样的波段数, 其中 Indian Pines 和 Salinas 数据集所选波段数均设置为 30, Pavia University 数据集所选波段数为 20. 由于本文的波段是基于 FCM 波段子集分解的基础上选择得出的, 我们无法得到与比较算法相同的波段数. 因此, Pavia University 数据集采用所选波段数为 18 时的分类结果与对比算法所选波段数为 20 时进行比较. 表 3 展示了 6 种算法在 3 个数据集上的分类精度对比结果.

表 2 3 个高光谱数据集的分类结果 (%)

(Table 2 The classification results of three hyperspectral datasets(%))

类数	Indian pines (30 Bands)	PU (18 Bands)	Salinas (30 Bands)
1	61.08±10.4	92.34±0.4	99.51±0.5
2	77.04±1.2	96.36±0.4	99.85±0.1
3	75.33±1.8	73.96±2.3	99.67±0.2
4	75.16±2.4	91.64±0.3	99.43±0.3
5	90.80±2.4	99.57±0.1	99.19±0.2
6	95.55±0.7	71.48±1.4	99.84±0.1
7	83.48±7.8	80.06±1.5	99.50±0.1
8	97.60±1.2	87.72±1.1	89.52±0.7
9	46.25±17.5	99.87±0.1	99.77±0.1
10	79.92±4.0	–	96.43±0.9
11	85.70±1.2	–	98.48±0.8
12	79.12±2.2	–	99.74±0.2
13	95.98±2.0	–	99.02±0.4
14	95.34±1.0	–	97.36±1.0
15	51.65±3.3	–	73.86±1.3
16	90.93±2.4	–	98.67±0.4
OA	83.94±0.6	90.30±0.1	93.70±0.1
AA	80.06±3.8	88.11±0.8	96.80±0.5
Kappa	81.65±0.7	87.02±0.1	93.00±0.1

如表 2 所示, 对于 Indian Pines 数据集, 本文算法在标记比例 20%, 选择波段数为 30 时, 获得 83.94%±0.6% 的总体分类精度, 对第 6, 8, 13 和 14 类地物的识别能力较好, 平均分类精度均达到 95% 以上; 对于 Pavia University 数据集, 在标记比例 20%, 选择波段数为 18 时, 取得 90.30%±0.1% 的总体分类精度, 对第 5 和 9 类地物的分辨能力较强, 平均分类精度接近于 100%; 对于 Salinas 数据集, 在标记比例 20%, 选择波段数为 30 时, 本文算法达到了 96.80%±0.5% 的平均分类精度. 其中, 除第 8 和 15 类地物外, 剩余 14 类地物均取得了良好的平均分类精度结果 (大于 90%), 其原因可能是由于第 8 类和 15 类地物的光谱特征过于相似而导致的. 此外, 通过对以上 3 个不同类型的高光谱数据集的实验结果, 可以看出本文方法对于实验数据的依赖程度较低, 具有普适性.

从表 3 可以看出, FCM-ABC 方法在 3 个数据集上的分类结果明显优于其他 5 种算法. 相比另外两个数据集, 6 种算法在 Indian Pines 数据集上没有获得较好的分类结果, 其原因可能是因为该数据集是严重的非均衡数据集. 然而, 这 6 种方法在 Salinas 数据集上均取得了理想的分类结果, 3 个精度指标均大于 90%. 四种无监督波段选择方法, 由于在选择波段时考虑了所选择波段的冗余性, 所以能够选择出了包含大量判别信息的低冗余波段组合, 从而提高了分类精度. WaLuMI 方法的分类精度优于 FCM-GA 方法, 其原因可能是 WaLuMI 方法中的互信息准则相比 FCM-GA 方法中的欧式距离能够更好地实现特征相关性的度量. 与 WaLuMI 方法和 FCM-GA 方法相比, MIC 方法得到了相对较低的结果, 这是因为 MIC 方法在第一次迭代过程中容易丢弃许多包含丰富信息的特征. 值得注意的是, Pavia University 数据集在缺失两个波段的情况下, 所提出的方法仍能取得良好的分类结果.

表 3 6 种方法在三个高光谱数据集上的分类结果对比 (%)

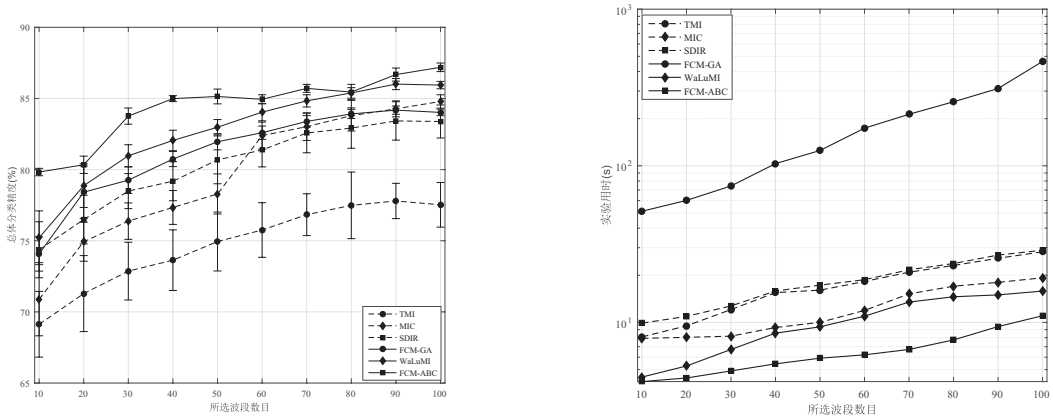
(Table 3 The comparison of classification results of six methods on three hyperspectral data sets (%))

数据集	精度指标	TMI	MIC	SDIR	FCM-GA	WaLuMI	FCM-ABC
Indian(30)	OA	72.9± 2.0	77.2±1.2	78.6±1.0	79.2±0.9	80.6±0.8	83.94±0.6
	AA	55.5± 2.3	64.5± 2.0	61.9±1.4	67.5±1.5	68.2±2.0	80.06±3.8
	Kappa	68.7± 2.3	64.5± 2.0	75.4± 1.1	76.2± 1.0	77.7± 1.0	81.65± 0.7
PU(18)	OA	86.2±1.6	88.7±0.9	90.5±0.7	89.3±0.6	90.1±1.0	90.30±0.1
	AA	79.4±2.7	84.6±1.8	85.5±1.7	86.4±1.0	86.7±1.2	88.11±0.8
	Kappa	81.4±2.2	85.0±1.2	87.3±1.0	85.7±0.9	86.7±1.4	87.02±0.1
Salinas(30)	OA	90.8±0.2	92.4±1.2	92.7±0.2	92.7±0.3	93.0±0.2	93.7±0.1
	AA	94.3±0.2	96.0±0.8	96.0±0.3	96.1±0.3	96.4±0.2	96.8±0.5
	Kappa	89.8±0.2	91.5±1.3	91.9±0.2	91.9±0.4	92.2±0.2	93.0±0.1

4.3 所选波段数和计算时间分析

下面我们以 Indian Pines 数据集为例, 分析了分类结果和选择波段所需要的计算时间随着所选波段数增加而变化的情况. 图 3(a) 给出了 6 种算法在所选波段个数从 10 增加到 100 时, 总体分类精度的变化情况. 如图 3(a) 所示, 本文方法与其余 5 种波段选择方法相比存在明显的优势, 这表明 FCM-ABC 算法能够有效的兼顾判别信息的保留和冗余信息的剔除两个方面. 同时, 这 6 种算法的分类精度均随着所选波段数的增加而呈上升趋势. 当所选波段数量大于 70 时, 由于波段间的冗余信息的叠加, 分类精度并没有出现显著的增长, 并逐渐呈现出趋于稳定的趋势.

图 3(b) 给出了 6 种方法在 Indian Pines 数据集上选择波段所需要的计算时间. 显然, 与其余 5 种算法相比, FCM-ABC 方法不仅获得了最高的分类精度, 而且波段选择的实验用时也远低于其他算法, 说明了所提出算法的有效性. 除了 FCM-GA 方法之外, 其他 5 种算法也显示了良好的实验用时表现. 这主要是由于 FCM-GA 方法通过基于种群的搜索机制来提高分类性能, 故所花费的时间较长. MIC 方法由于基于排序的搜索和无需进入迭代的特性, 在其余 5 种对比方法中实验用时表现最优. 总而言之, 本文提出的 FCM-ABC 方法能够兼顾波段选择的效果和实验用时. 展现了良好的分类能力和实现能力.



(a) (b)

图 3 (a) 6 种算法的分类精度随所选波段数的变化. (b) 6 种算法的实验用时随所选波段数的变化

(Figure 3 (a) The variation of classification accuracy of six algorithms with the number of selected bands. (b) The variation of execution time of six algorithms with the number of selected bands)

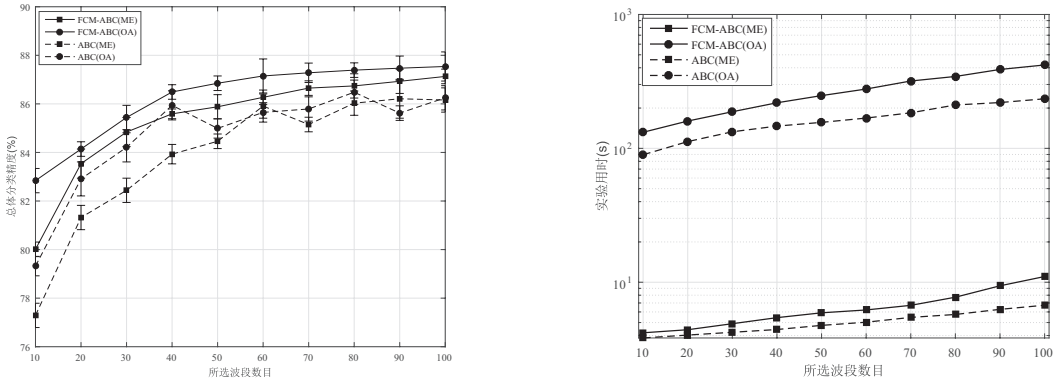
图 4(a) 给出了利用 FCM-ABC 方法和仅仅使用 ABC 方法进行波段选择, 并分别采用不同准则函 OA 和 ME 在 Indian Pines 数据集上的分类结果. 根据总体分类精度的曲线可以得出, 基于这两个准则函数, FCM-ABC 方法均取得了较好的结果. 采用 OA 作为准则函数, 由于在搜索过程中分类结果参与了对波段子集的评估, 所以其总体分类精度略高于采用极大熵 (ME) 准则的总体分类精度, 但这两种准则函数之间的分类精度相差较小. 此外, 不难发现, 采用 OA 作为适应度函数的 ABC 算法是缺乏稳定性的. 这可能是因为优化算法在找到最优波段组合之前就已经被终止了. 毕竟对于包含 200 个波段的 Indian Pines 数据集, 存在太多的波段组合. 以极大熵 (ME) 作为准则函数的 ABC 方法展现了不太理想的分类结果. 由于所选的波段组合存在较大的冗余, 从而导致了分类能力下降. 这充分说明了在该方法中进行波段子集分解的必要性.

在理论上, 如果使用 OA 作为适应度函数, 则 ABC 算法需要在每次迭代中执行分类算法 (SVM) 以获得总体精度, 这无疑增加了算法的耗时. 从图 4(b) 可以看出, 采用极大熵 (ME) 作为准则函数的实验用时比采用 OA 作为准则函数的实验用时要大大减少, 这与理论分析的结果具有一致性.

5 结 论

本文提出了一种基于波段聚类和波段组合优化的无监督波段选择算法. 通过 FCM 聚类将连续的波谱空间分解成相关性较小的波段子空间, 在一定程度上减少了所选择波段间的冗余性. 极大熵准则和人工蜂群优化算法的使用, 在最大化原始波段信息的同时, 降低了所选波段之间的冗余程度, 实现最优波段组合的选择. 在 3 个典型的高光谱数据集上的实验结果表明了所提出的无监督特征选择方法的有效性.

在高光谱数据降维的过程中, 如何确定最佳的波段数? 就目前的研究来看, 最佳波段数的确定是实验的结果. 是否可以定义一个有效性的降维指标来帮助我们确定最佳的维数? 这是一个十分值得研究的问题.



(a) (b)

图 4 (a) FCM-ABC 算法和 ABC 算法分别采取不同准则函数 (OA/ME) 时分类精度随所选波段数的变化. (b) FCM-ABC 算法和 ABC 算法分别采取不同准则函数 (OA/ME) 时实验用时随所选波段数的变化

(Figure 4 (a) The variation of classification accuracy of FCM-ABC and ABC with different criterion functions.
(b) The variation of execution time of FCM-ABC and ABC with different criterion functions)

参 考 文 献

[1] 樊利恒, 吕俊伟, 邓江生, 等. 基于分类器集成的高光谱遥感图像分类方法. 光学学报, 2014, **34**(9): 91–101.
(Fan L H, Lü J W, Deng J S, et al. Classification of hyperspectral remote sensing images based on bands grouping and classification ensembles. *Acta Optica Sinica*, 2014, **34**(9): 91–101.)

[2] 谢娟英, 谢维信. 基于特征子集区分度与支持向量机的特征选择算法. 计算机学报, 2014, **37**(8): 1704–1718.
(Xie J Y, Xie W X, et al. Several feature selection algorithms based on the discernibility of a feature subset and support vector machines. *Chinese Journal of Computers*, 2014, **37**(8): 1704–1718.)

[3] 杜培军, 王小美, 谭琨, 等. 利用流形学习进行高光谱遥感影像的降维与特征提取. 武汉大学学报 (信息科学版), 2011, **36**(2): 148–152.
(Du P J, Wang X M, Tan K, et al. Dimensionality reduction and feature extraction from hyperspectral remote sensing imagery based on manifold learning. *Geomatics and Information Science of Wuhan University*, 2011, **36**(2): 148–152.)

[4] Feng J, Jiao L, Liu F, et al. Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images. *Pattern Recognition*, 2016, **51**: 295–309.

[5] Solorio-Fernández S, Carrasco-Ochoa J A, Martínez-Trinidad J F. Hybrid feature selection method for supervised classification based on Laplacian score ranking, advances in pattern recognition — Second mexican conference on pattern, Puebla, Mexico, September, 2010.

[6] 唐贵华. 基于密度排序聚类和超像素分割的高光谱遥感影像降维方法研究. 硕士论文, 深圳大学, 2016.
(Tang G H. Ranking-based-clustering and superpixel segmentation for hyperspectral remote imagery dimensionality reduction. Master's thesis, Shenzhen University, 2016.)

[7] Zhang Y, Desai M D, Zhang J, et al. Adaptive subspace decomposition for hyperspectral data dimensionality reduction. *International conference on Image Processing, IEEE*, 1999, **2**: 326–329.

[8] 王立国, 赵亮, 刘丹凤. 基于人工蜂群算法高光谱图像波段选择. 哈尔滨工业大学学报, 2015, **47**(11): 82–88.
(Wang L G, Zhao L, Liu D F, et al. Artificial bee colony algorithm-based band selection for hyperspectral imagery. *Journal of Harbin Institute of Technology*, 2015, **47**(11): 82–88.)

[9] 赵冬, 赵光恒. 基于改进遗传算法的高光谱图像波段选择. 中国科学院大学学报, 2009, **26**(6): 795–802.
(Zhao D, Zhao G H. Band selection of hyperspectral image based on improved genetic algorithm. *Journal of University of Chinese Academy of Sciences*, 2009, **26**(6): 795–802.)

[10] 王立国, 魏芳洁. 结合遗传算法和蚁群算法的高光谱图像波段选择. 中国图象图形学报, 2013, **18**(2): 235–242.
(Wang L G, Wei F J. Band selection for hyperspectral imagery based on combination of genetic algorithm and ant colony algorithm. *Journal of Image and Graphics*, 2013, **18**(2): 235–242.)

[11] Ghamisi P, Benediktsson J. A. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *IEEE Transactions on Geoscience Remote Sensing Letter*, 2015, **12**: 309–313.

[12] Karaboga D, Basturk B A. powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. *Journal of Global Optimization*, 2007, **39**: 459–471.

[13] Karaboga D, Basturk B. On the performance of artificial bee colony (ABC) algorithm. *Applied Soft Computation*, 2008, **8**: 687–697.

[14] Karaboga D, Ozturk C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computation*, 2011, **11**: 652–657.

[15] Feng J, Jiao L. Zhang X, et al. Hyperspectral band selection based on trivariate mutual information and clonal selection. *IEEE Transactions on Geoscience Remote Sensing*, 2014, **52**(7): 4092–4105.

[16] Feng J, Jiao L, Liu F, et al. Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy. *IEEE Transactions on Geoscience Remote Sensing*, 2015, **53**(5): 2956–2969.

[17] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2002, **24**(3): 301–312.

[18] Wu J. Unsupervised intrusion feature selection based on genetic algorithm and FCM. *Lecture Notes in Electrical Engineering*, 2012, **154**: 1005–1012.

[19] Martínez-Usómartinez-Uso A, Pla F, Sotoca J M, et al. Clustering-based hyperspectral band selection using information measures. *IEEE Transactions on Geoscience Remote Sensing*, 2007, **45**(12): 4158–4171.

[20] Bazi Y, Melgani F. Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE Transactions on Geoscience Remote Sensing*, 2006, **44**(11): 3374–3385.