李奇生,赵成萍,尹子琴,等. 均衡 FCM 算法在农作物遥感影像解译中的应用[J].江苏农业学报,2020,36(5):1163-1168.
doi:10.3969/j.issn.1000-4440.2020.05.013

均衡 FCM 算法在农作物遥感影像解译中的应用

李奇生[1], 赵成萍[1], 尹子琴[2], 李 博[3], 周新志[1]
(1.四川大学电子信息学院,四川 成都 610065; 2.云南省云计算学会,云南 昆明 650032; 3.四川大学水利水电学院,四川 成都 610065)

摘要: 针对传统的模糊 C-均值聚类算法(FCM 算法)对大数据集收敛速度慢,聚类不均匀类别样本时出现大类吃小类现象以及对初始聚类中心点要求高等问题,提出了一种基于均衡样本集思想的模糊 C-均值聚类算法(均衡 FCM 算法)。选取 Landsat8、Sentinel2A 遥感卫星采集获得的哈尔滨市宾县 2018 年遥感图像,验证方法的有效性。结果显示,提出的均衡 FCM 算法可以改善传统 FCM 算法存在的问题,验证了均衡 FCM 算法的有效性。

关键词: 均衡 C-均值聚类算法(均衡 FCM 算法); 混合像元; 面积提取; 图像分类
中图分类号: S127 **文献标识码:** A **文章编号:** 1000-4440(2020)05-1163-06

Application of balanced FCM algorithm on the interpretation of crops remote sensing image

LI Qi-sheng[1], ZHAO Cheng-ping[1], YIN Zi-qin[2], LI Bo[3], ZHOU Xin-zhi[1]
(1.*College of Electronics and Information Engineering, Sichuan University, Chengdu* 610065, *China*; 2.*Cloud-computing Academic Society of Yunnan Province, Kunming* 650032, *China*; 3.*College of Water Resource & Hydropower, Sichuan University, Chengdu* 610065, *China*)

Abstract: To solve the conventional fuzzy C-means clustering algorithm(FCM algorithm) problems including slow convergence speed for large data sets, the occurrence of neglect of smaller clustered groups when the clustering categories are uneven, and high requirement on the initial clustering center points, this paper proposed a fuzzy clustering algorithm model based on balanced data sets (BDS-FCM algorithm). To verify the effectiveness, the remote sensing images of Bin County, Harbin City collected by Landsat8 and Sentinel2A remote sensing satellites in 2018 was selected as experimental subjects. Results of the experiment show that the proposed BDS-FCM algorithm can improve the conventional FCM algorithm and verify the effectiveness of BDS-FCM.

Key words: fuzzy C-means clustering algorithm based on balanced data sets(BDS-FCM algorithm); mixed pixel; area extraction; image classification

由于中低分辨率遥感卫星空间分辨率较低,单个像元中可能会存在多种地物信息(混合像元)。该问题是进行遥感解译的难点之一,许多研究者进行过混合像元分解研究[1-3]。世界各国现已将遥感技术融入农作物分类、面积估算等工作中,发展至今已经取得不错的成就[4-6]。在多作物分类以及面积估算工作中,能否改善或解决混合像元问题常常可以决定解译工作能否达到理想精度。常见的混合像元分解模型有线性模型、非线性模型、几何光学模型以及模糊模型等。Kai 等在决策树分类后加入基于线性光谱混合模型用于分解混合像元,提升了分类和面积提取精度[7]。由 Bezdek 等提出的基于模糊集思想的模糊 C-均值算法(FCM),操作简单,将每

收稿日期:2019-11-08
基金项目:国家自然科学基金项目(U1933123)
作者简介:李奇生(1996-),男,河南洛阳人,硕士研究生,研究方向为模式识别与智能系统。(E-mail)563692411@qq.com
通讯作者:赵成萍,(E-mail)sc_zcp@scu.edu.cn

个样本点用一个隶属度来反映数据的关联程度,不同于其他的硬分类算法,它建立了样本类属的不确定性,客观反映现实世界,被广泛用于各领域[8-9]。

模糊 C-均值算法被很多研究者用于遥感研究。Kaur 等[10]利用了模糊算法隶属度函数对混合像元进行分解,但该算法对噪声敏感、鲁棒性较差,初始聚类中心的选择对最终的聚类结果影响很大,若选取不当可能会陷入局部最优,对运算速度有很大影响。同时在处理样本集差距较大的多类别聚类中,很可能会出现大类吞并小类等情况。很多研究人员在改善其鲁棒性上提出了优化方法,主流的方法有2 种。一种是对模糊算法自身目标函数的优化,例如将区间 2 型模糊理论引入模糊算法改进目标函数[11],利用局部空间信息和灰色信息给出新的目标函数[12],引入一种加权因子同时考虑居中像素与其相邻像素之间的空间距离和隶属关系,以此优化目标函数[13];另一种方法是将其他算法和 FCM 算法相融合,解决 FCM 的局限性,例如在 FCM 前引入 SSO 算法优化目标函数,进一步寻找最优聚类中心[14]。Honglei 等[15]提出一种将模糊 C 均值聚类与马尔可夫随机场相结合的聚类算法,算法本身鲁棒性很强,分类精度高,但在农业遥感面积估算工作中,样本集的分布一般并不平均,以上方法在处理类别间样本差距较大的情况时效果仍不够理想。

鉴于此,本研究提出一种均衡模糊 C-均值聚类算法(均衡 FCM 算法),将模糊算法与支持向量机算法(SVM 算法)相融合,在 SVM 算法对样本集粗分类后的规则文件中选取分类把握较大的点作为纯净样本点,采用小样本类过采样方法平衡数据集,再将纯净样本点各维度特征值的平均值作为初始中心点输入 FCM 算法,并将该方法用于多类别农作物的解译。

1 材料与方法

1.1 研究区域与数据

宾县是黑龙江省哈尔滨市的下辖县,位于黑龙江省南部(图 1)。其主要农作物有玉米、水稻等。主要粮食作物空间变化呈现较强的规律性。土地利用率高,其中耕地面积比例为 59.56%,林地面积比例为 31.64%,水域面积比例为 4.97%,居民用地面积比例为 3.22%,而其他类型用地如草地、未利用土地等面积比例为仅占 0.61%[16]。宾县年降雨量少,适合进行遥感研究。宾县统计局 2016 年农作物播种面积统计结果显示玉米和水稻的播种面积占总面积的 94.5%,因此本研究对主要农作物玉米和水稻进行解译。宾县玉米种植面积分布较均衡,水稻种植面积集中在北部和西部地区。玉米播种时间在 4 月 20 日至 5 月 10 日之间,收获在 10 月中旬。水稻 于 4 月育苗,5 月插秧,9 月末至 10 月上中旬收获,主要作物生长期基本同步。在 5 月底至 6 月中旬左右,玉米还未完全长出,水稻处于泡田整地期,在该时期作物田块易于识别。为了方便方法验证,在无云或少云天气下选定了 2 种 6 幅遥感图像,即 2018 年 6 月 1 日、10 月 10 日的 Landsat8 OLI 图像和 2018 年 5 月 31 日、7 月 25 日各 2 幅 Sentinel 2A 图像。为了方便研究,对 5 月 31 日以及 7 月 25 日各 2 幅 Sentinel2A 图像进行拼接,得到完整的宾县区域。但在 5 月底 6 月初时,玉米尚未长出,其光谱信息更接近于裸地,在此情况下不易区分玉米和建筑两种地物,而处于泡田期的水稻与水域相近。为此利用 10 月 10 日的 Landsat8 数据以及 7 月 25 日的 Sentinel2A 数据进行掩膜,在此时期植被已经完全长出,根据其光谱信息将水域与建筑地物掩膜并裁剪 5 月底 6 月初数据。在选定特征值时,利用多波段信息进行波段运算,计算归一化植被指数(*NDVI*)以及陆表水指数(*LSWI*),计算公式如下:

$$NDVI=(NIR-RED)/(NIR+RED) \quad (1)$$

$$LSWI=(NIR1-SWIR1)/(NIR1+SWIR1) \quad (2)$$

同时选取 Landsat8 的第 6 波段以及 Sentinel 2A 的第 11 波段 *SWIR1* 作为特征值输入模型。

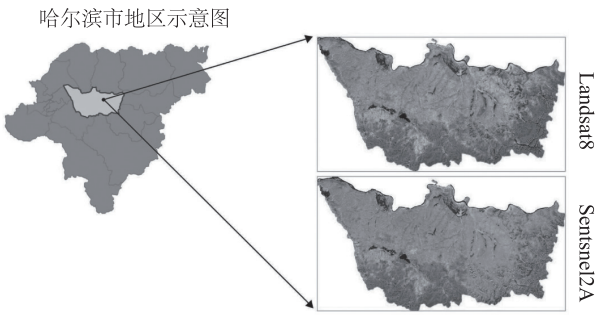


图 1 研究区域位置图

Fig.1 Location of areas studied

1.2 研究方法

将 BDS-FCM 算法应用于不同空间分辨率和光谱分辨率的卫星数据。具体工作模型如图 2 所示,共分为 4 个模块。

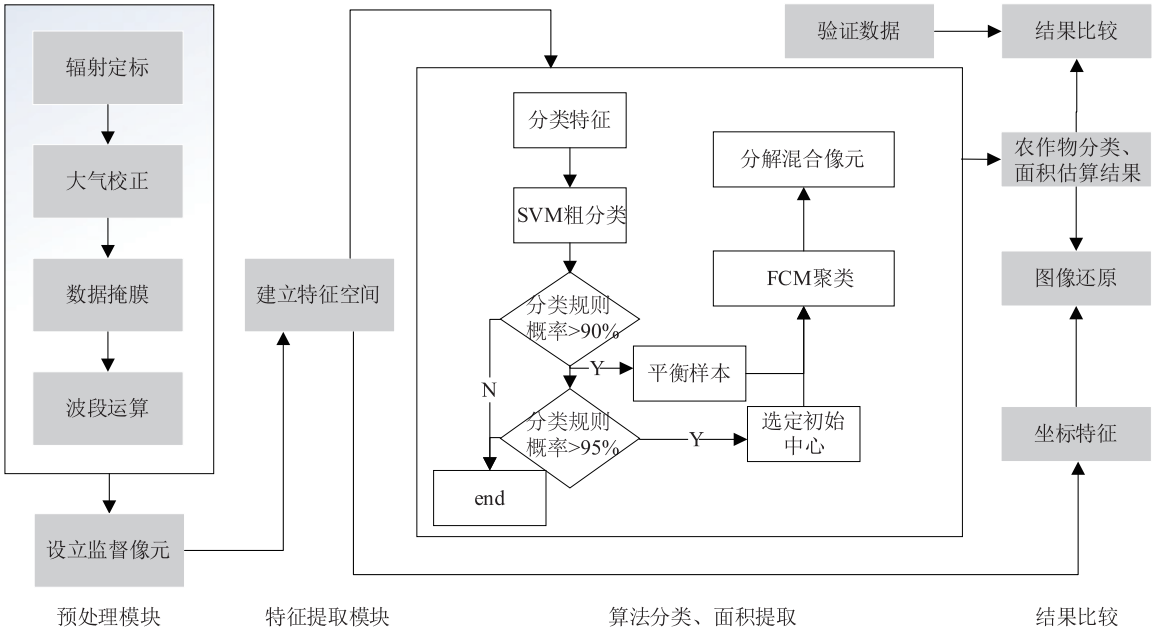


图 2 BDS-FCM 工作模型

Fig.2 Working model of fuzzy C-means clustering algorithm model based on balanced data sets(BDS-FCM)

1.2.1 遥感数据预处理模块 预处理是进行遥感研究的必备流程。卫星原始遥感数据无法直接用于图像解译,对其进行预处理的目的主要是对遥感图像进行噪声滤除,更好地利用预处理后的遥感图像所反映的地物目标波谱特性进行反演、统计和分析。包括辐射定标、大气校正、几何校正等。

1.2.2 特征提取模块 以人工解译的方式选定各类监督像元,设定样本集中共 C 类子样本。对其进行特征提取后组成一个多维特征空间。该特征空间分为 2 个部分,用于分类输入的多维特征以及地理空间坐标特征。特征空间可由一个特征矩阵来表示,矩阵中元素为 $A_{x,y}=\{a_1,a_2,\cdots,a_m\}$,其中 x、y 表示其空间地理坐标,以便最后进行图像还原。a_m 为该点的第 m 个特征值。

1.2.3 算法处理模块 BDS-FCM 算法执行步骤如下:

(1)选用 SVM 工具箱中 libsvm 方法进行粗分类。选择参数训练 SVM 模型,将监督像元打上标签输入训练函数得到结构模型 model。其中在进行 SVM 算法选择时,分类处理可选择模型 C-支持向量分类机(C-SVC)和 V-支持向量分类机(V-SVC)。面对不同的应用场景应选择不同的分类方法以达到最优效果。同时核函数有线性、多项式、RBF 等,选定合适的核函数将特征合理地映射至高维空间也是影响分类结果的重要因素之一。

(2)SVM 分类预测,将方法 1.2 中的特征矩阵输入第 1 步训练的 model 中进行分类,得到粗分类结果文件 decision_values,由其统计分类结果并确定小样本集。该文件为一个矩阵,可表示为 $D=[d_{x,y,c}]$,其中 $d_{x,y,c}$ 表示对横、纵坐标为 x、y 的点分类结果第 c 类的决策度,以百分数表示。

(3)扩充小样本集,选用的方法为基于线性直插的过采样方法(Synthetic minority oversampling technique,SMOTE)。SMOTE 算法是由 Nitesh 等提出的面对小样本的采样方法[17],其原理如图 3 所示。

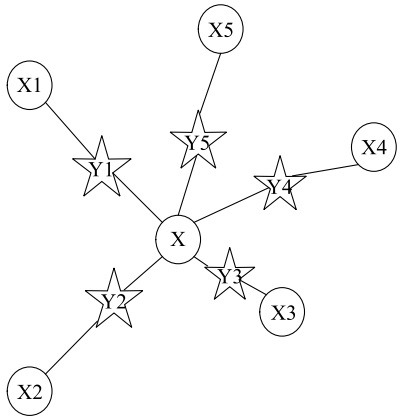


图 3 基于线性直插的过采样方法(SMOTE)示意图

Fig.3 Schematic diagram of synthetic minority oversampling technique (SMOTE)

其中 X 为小样本集中的一个随机样本，X_1、X_2、X_3、X_4、X_5 是其相邻样本，人工合成样本点为 Y_1、Y_2、Y_3、Y_4、Y_5。按照以下公式随机插入在 X 与它相邻样本连线上的某一不确定位置。

$$Y_i = X_i + \text{rand}(0,1) \times (X - X_i) \quad (3)$$

为了减少混合像元点对结果的影响。扩充小样本集时选择决策度($d_{x,y,c}$)>0.9 的点作为样本进行扩充，扩充倍数为第 2 步中统计分类结果中大样本与小样本的比值，并将人工合成样本并入原始样本。将决策度($d_{x,y,c}$)>0.95 的点进行分类别平均，所得各类的平均值作为下一步模糊算法的初始中心点输入。

(4)进行模糊聚类。$A_{x,y}$ 点对各类地物的隶属度表示为集合 $U_{x,y} = \{u_{x,y,1}, u_{x,y,2}, \cdots, u_{x,y,C}\}$。模糊像元点的隶属度矩阵满足：

$$\text{Max}(U_{x,y}) < \delta \quad (4)$$

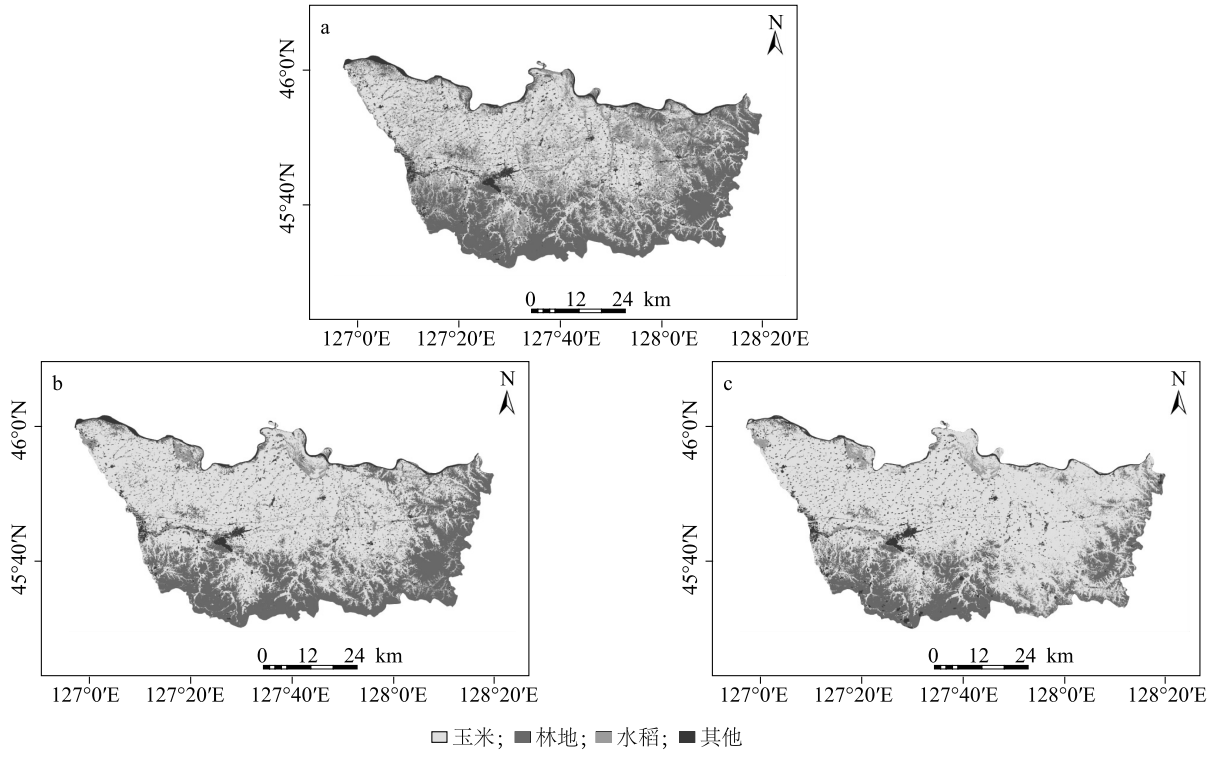
对于阈值(δ)的选择可以根据不同遥感数据源来调整，不同的空间分辨率形成模糊像元的概率不同，分辨率越高的数据阈值(δ)可设置越高。根据隶属度矩阵将一个像元面积(S)进行分解，第 i 类作物占第 j 个混合像元的比例面积表示为：

$$S_{j,i} = S \times u_{j,i} \quad (5)$$

1.2.4 结果验证模块 将方法 1.3 的分类结果以及分类后的面积进行统计。一方面将各像元分类结果与坐标特征相结合还原图像，另一方面将试验结果和验证数据进行数据比对以验证算法的准确度。

2 结果与分析

数据特征空间建立后输入模型，SVM 方法选取参数更好设定的 V-支持向量分类机，其中惩罚因子设置为 0.5，设定对特征维度低、较大数据集分类效果较好的 RBF 核函数处理数据集。在模糊算法参数选择中设定 c=3，m=2，$\varepsilon = 1 \times 10^{-5}$。Landsat8 以及 Sentinel 2A 的分类效果如图 4 所示，同时加入使用传统 FCM 对 Landsat8 数据处理的分类结果进行对比。为了验证本试验的分类精度，采取混淆矩阵进行统计分析。为了保证选取验证参照点的科学性，对 3 类地物采取分层抽样的方法共选取 500 个参照点，同时利用与全色波段融合后的 15 m 分辨率 Landsat8 数据确定参照点的地物归属。验证结果如表 1 所示。



a：传统 FCM 算法对 Landsat8 数据处理的分类图；b：均衡 FCM 算法对 Landsat8 数据处理的分类图；c：均衡 FCM 算法对 Sentinel2A 数据处理的分类图。

图 4 不同算法分类效果示意图

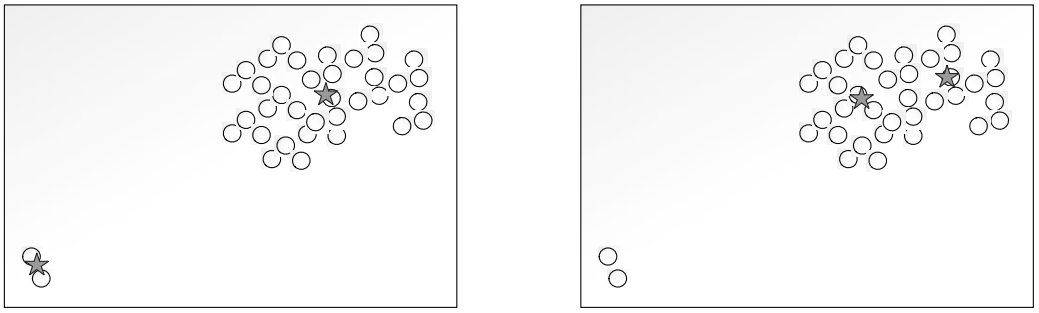
Fig.4 Schematic diagram of classification effects using different algorithms

表 1 均衡 FCM 算法对多种数据集分类的验证

Table 1 Verification of classification of various data sets by fuzzy C-means clustering algorithm model based on balanced data sets(BDS-FCM)

数据集	分类后类别	参考数据			总体分类精度(%)	卡帕系数(%)
		玉米(像元)	水稻(像元)	林地(像元)		
Landsat8	玉米	294	4	2	96.2	92.98
	水稻	3	45	6		
	林地	3	1	142		
Sentinel2A	玉米	292	7	6	94.8	90.34
	水稻	5	43	5		
	林地	3	0	139		

从分类结果图可以看到水稻样本点没有被正常识别,而一些玉米地和林地交界处的样本点被识别为水稻。由于 FCM 本质上是一种聚类算法,实现原理为优化目标函数以达到类内距离最小化。图 5a 表示分类所要达到的效果,两类别分类并不均匀。当 FCM 算法用于该数据分类时,会出现图 5b 的问题,即为了达到距离最小化将类别中心向大类靠近,甚至将小类看作噪声点,导致小类被吞并。因此在试验数据中处于大类别的玉米样本和林地样本将小样本水稻样本吞噬,而交界处的混合样本点数量多于水稻样本,因此被识别为第 3 类。在本研究算法(均衡 FCM 算法)中,Landsat8 数据处理的总体分类精度达到了 96.2%,卡帕系数为 92.98%;Sentinel2A 数据处理的分类精度达到了 94.8%,卡帕系数为 90.34%。可以看出,本研究算法(均衡 FCM 算法)对 Landsat8 数据处理的分类精度高于 Sentinel2A 数据处理。但是 Sentinel2A 的空间分辨率高于 Landsat8,这是由于在处理 Sentinel2A 数据时由两景数据拼接时出现的色差问题影响了图像解译过程,进而导致分类精度下降。



a:正确分类 b:FCM实际聚类

图 5 FCM 算法的聚类对比图

Fig.5 Comparison of cluster images of fuzzy C-means(FCM) algorithm

阈值(δ)的选择影响混合像元分解的精度。对宾县 Landsat8 和 Sentinel2A 数据分别选取 0.80、0.82 的阈值,此参数是根据大量手动调试结果挑选出的较优参数。选定 Landsat8 数据源验证面积统计精度,将 30 m 分辨率的多光谱数据与 15 m 分辨率的全色波段融合得到 15 m 分辨率数据,并通过人工先验知识设立监督像元点后进行监督分类和面积统计用于验证(表 2、表 3)。对各作物的面积统计精度(表 3)进行加权平均可得总精度,受图像拼接时的误差影响,Sentinel2A 的总精度略低于 Landsat8。由于分解了混合像元,各类作物的面积统计精度获得了提升。

表 2 混合像元分解前后像元统计结果对比

Table 2 Comparison of statistical results of mixed pixels before and after decomposition

数据	玉米(像元)	水稻(像元)	林地(像元)
Landsat8 原始数据	160 462	22 393	122 624
Landsat8 原始数据分解后数据	109 424	21 367	174 689
Sentinel2A 原始数据	238 841	29 897	208 449
Sentinel2A 原始数据分解后数据	222 009	31 227	223 951

表 3 混合像元分解前后面积统计精度对比

Table 3 Comparison of area statistical accuracy before and after mixed pixels decomposition

数据	各作物的面积统计精度(%)			总精度(%)
	玉米	水稻	林地	
Landsat8 原始数据	91.203	98.05	96.11	93.17
Landsat8 原始数据分解后数据	93.35	99.1	99.12	95.54
Sentinel2A 原始数据	94.84	94.03	95.42	94.99
Sentinel2A 原始数据分解后数据	95.17	94.65	95.91	95.38

3 讨 论

本研究提出了均衡模糊 C-均值聚类算法(均衡 FCM 算法),并用于农作物遥感解译。优化了模糊算法对初始中心点的选择,解决了当样本集不平衡时大类吞并小类的问题。该算法利用 SVM 算法对数据源的特征值进行粗分类,根据粗分类的结果确定 FCM 算法的初始类中心点并扩充小样本数据集以达到数据平衡的效果。选取哈尔滨市宾县不同数据源的图像进行试验,分类结果和卡帕系数表明该算法可以达到较好的分类效果。通过分解混合像元提升了各类地物的面积统计精度。通过本试验得到如下结论:(1)聚类算法在处理类别分布不均匀样本时,常常会将小样本当作噪声而影响聚类精度,这对于农作物遥感解译过程是难以接受的。(2)由于地形、土壤等因素的影响,农作物遥感样本常出现分布不均情况,因此在处理类似问题时要对样本集进行均衡化处理。(3)模糊算法将像元分解至多类,利于处理中低精度遥感数据问题,而在农作物遥感解译时利用此算法可以提高解译精度,尤其在作物面积统计上。

研究中存在的问题:混合像元和纯净像元最高隶属度阈值的选择问题上本研究未给出明确的方法,选择的主要依据是试验结果和经验。但此阈值的选择影响农作物面积统计精度,需要提出一套选择理论。此外,由于本研究算法(均衡 FCM 算法)在分类前要进行粗分类以确定样本类别分布,耗时较长。因此建议在解译前目视粗略判断各类别分布情况,若分布相对均衡,可直接利用 FCM 算法进行解译。

参考文献:

[1] LI Q, LAN H, ZHAO X, et al. River centerline extraction using the multiple direction integration algorithm for mixed and pure water pixels[J]. GIScience & Remote Sensing, 2019, 56(2): 256-281.

[2] XIAN-CHUAN Y, XIAO-FENG C, HENG-ZHI C, et al. Mixed-Pixel decomposition of SAR images based on single-pixel ICA with selective members[J]. GIScience & Remote Sensing, 2011, 48(1): 130-140.

[3] KAVZOGLU T, REIS S. Performance analysis of maximum likelihood and artificial neural network classifiers for training sets with mixed pixels[J]. GIScience & Remote Sensing, 2008, 45(3): 330-342.

[4] 孟令奎,李晓香,张 文. 植被覆盖区 VIIRS 与 MODIS 遥感指数的相关性[J]. 江苏农业学报, 2018,34(3):570-577.

[5] SON N T, CHEN C F, CHEN C R, et al. AssBDSment of Sentinel-1A data for rice crop classification using random forests and support vector machines[J]. Geocarto International, 2018, 33(6): 587-601.

[6] 何瑞银,沈明霞,从静华,等. 植被信息提取过程中 ETM+遥感影像的分类方法[J]. 江苏农业学报, 2008, 24(1):29-32.

[7] KAI W, JUN Z, GUOFENG Z. Early estimation of winter wheat planting area in Qingyang city by decision tree and pixel Unmixing methods based on GF-1 satellite data[J]. Remote Sensing Technology and Application, 2018, 33(1): 158-167.

[8] MAHELA O P, SHAIK A G. Recognition of power quality disturbances using S-transform based ruled decision tree and fuzzy C-means clustering classifiers[J]. Applied Soft Computing, 2017, 59: 243-257.

[9] LIANG-QUN L, WEI-XIN X, ZONG-XIANG L. A novel quadrature particle filtering based on fuzzy c-means clustering[J]. Knowledge-Based Systems, 2016, 106: 105-115.

[10] KAUR S, BANSAL R K, MITTAL M, et al. Mixed pixel decomposition based on extended fuzzy clustering for single spectral value remote sensing images[J]. Journal of the Indian Society of Remote Sensing, 2019, 47(3): 427-437.

[11] QIU C, XIAO J, HAN L, et al. Enhanced interval type-2 fuzzy c-means algorithm with improved initial center[J]. Pattern Recognition Letters, 2014, 38: 86-92.

[12] KRINIDIS S, CHATZIS V. A robust fuzzy local information C-means clustering algorithm[J]. IEEE Transactions on Image ProcBDSing, 2010, 19(5): 1328-1337.

[13] ZHANG H, SHI W, HAO M, et al. An adaptive spatially constrained fuzzy c-means algorithm for multispectral remotely sensed imagery clustering[J]. International Journal of Remote Sensing, 2018, 39(8): 2207-2237.

[14] BUI Q T, NGUYEN Q H, PHAM V M, et al. A novel method for multispectral image classification by using social spider optimization algorithm integrated to fuzzy C-mean clustering[J]. Canadian Journal of Remote Sensing, 2019, 45(1): 42-53.

[15] HONGLEI Y, JUNHUAN P, BAIRU X, et al. Remote sensing classification using fuzzy C-means clustering with spatial constraints based on Markov random field[J]. European Journal of Remote Sensing, 2013, 46(1): 305-316.

[16] 成胜权. 基于 RS 和 GIS 的宾县土地利用和土壤侵蚀的定量研究[J]. 水利科技与经济, 2012, 18(9):100.

(责任编辑:张震林)