doi: 10.6046/gtzyyg.2014.02.15
引用格式: 张修远,刘修国. 基于随机森林算法的高维模糊分类研究[J]. 国土资源遥感,2014,26(2):87 - 92. (Zhang X Y, Liu X G. Study of high - dimensional fuzzy classification based on random forest algorithm[J]. Remote Sensing for Land and Resources, 2014,26(2):87 - 92.)

基于随机森林算法的高维模糊分类研究

张修远, 刘修国
(中国地质大学(武汉)信息工程学院,武汉 430074)

摘要: 高光谱数据的空间分辨率普遍偏低,混合像元分布广泛,故模糊分类方法常用于此类型数据的信息提取。针对模糊分类的精度常受限于特征维数和模糊样本选取等问题,提出了基于随机森林(random forest,RF)算法的高维模糊分类方法。首先将 RF 算法用于特征选择和模糊样本获取,然后在低维特征空间中利用模糊样本进行模糊分类,通过 2 步分类、遵循假设前提一致原则,实现 RF 和模糊分类 2 种分类器的融合; 并通过不同样本、不同实验区和分区优化前后的 3 个实验(包括 20 余次对比实验、60 多次子实验),验证了该方法不仅提高了模糊分类的精度,具有分类的有效性和可推广性,而且具有可优化性和对原始样本质量的鲁棒性。

关键词: 随机森林(RF); 模糊分类; 高维特征
中图法分类号: TP 751.1 **文献标志码**: A **文章编号**: 1001 - 070X(2014)02 - 0087 - 06

0 引言

自 20 世纪 60 年代模糊数学理论建立之后,模糊理论已被引入到许多行业,同时也被广泛应用于遥感图像分类中[1],已显示出在分析混合像元、提高分类精度等方面的优势; 但模糊分类应用于高维特征分类中易出现“维度灾难”,且高维矩阵运算的效率偏低。采用何种方式降维效果最佳,已成为模糊分类中待解决的问题[2]。另一方面,模糊分类法可以利用模糊样本作为训练样本,能够更真实地理解地类信息,生成更准确的分类图[3]; 但选取模糊样本很困难,故分类时常使用传统样本[4]。有的学者提出,为了确保分类的精度,可把选取的训练样本作为“准纯样本”进行模糊分类。这些样本经过一次模糊分类后就成为“模糊样本”,然后再对其进行模糊分类,在一定程度上可以改善分类的精度。但使用“准纯样本”进行第一次模糊分类与最大似然法分类相同,其理论基础是经典概率论的贝叶斯估计; 而将得到的“模糊样本”用于模糊分类显然没有遵循假设前提不变原则。

随机森林算法(random forest,RF)是一种分类和预测模型,具有很高的预测准确率, 对异常值和噪声具有很好的容忍度; 且不容易出现“过拟合”,在医学、生物信息及管理学等领域有着广泛的应用[5]。而在遥感图像信息提取中,RF 算法常用于特征选择和分类。

根据上述情况,本文提出基于 RF 算法的模糊分类方法: ①利用 RF 将高维特征空间映射至低维空间,并选取模糊样本; ②在低维特征空间中利用模糊样本进行模糊分类,获得最终分类结果。并通过实验表明该方法具有有效性、可推广性和可优化性等特点,可提高对高光谱数据的分类精度。

1 算法原理

1.1 模糊分类算法

模糊分类法与常规分类法不同,它认为一个像元是可分的,即一个像元可以在某种程度上属于某个类而同时在另一种程度上属于另一类,这种类属关系的程度用像元隶属度表示。确定像元的隶属度函数的数学模型和方法很多,在遥感图像模糊分类中一般采用最大似然估计来确定像元属于各类的隶属度函数。模糊分类算法的主要流程为:

1)构造训练样本的模糊分割矩阵。设 n 为训练样本总数,G 为预先确定的类别数,$f_i(X_j)$ 为像元 X_j 对第 i 类的隶属度,则训练样本的模糊分割矩阵 $\boldsymbol{A}$ 为

$$\boldsymbol{A}=\begin{bmatrix} f_1(X_1) & \cdots & f_1(X_n) \\ \vdots & & \vdots \\ f_G(X_1) & \cdots & f_G(X_n) \end{bmatrix}。 \quad (1)$$

收稿日期: 2013 - 06 - 04; 修订日期: 2013 - 07 - 15

2)利用获得的模糊分割矩阵求解模糊均值和模糊协方差矩阵,即

$$\boldsymbol{\mu}_c^* = \frac{\sum_{i=1}^{n} f_c(x_i) x_i}{\sum_{i=1}^{n} f_c(x_i)}, \quad (2)$$

$$\boldsymbol{V}_c^* = \frac{\sum_{i=1}^{n} f_c(x_i)(x_i - \boldsymbol{\mu}_c^*)(x_i - \boldsymbol{\mu}_c^*)^{\mathrm{T}}}{\sum_{i=1}^{n} f_c(x_i)}, \quad (3)$$

式中:$\boldsymbol{\mu}_c^*$ 为模糊均值矩阵;$\boldsymbol{V}_c^*$ 为模糊协方差矩阵;$f_c(x_i)$ 为像元 x_i 对第 i 类的隶属度;n 为训练样本总数。

3)为完成模糊特征空间的划分,本文采用最大似然估计,确定待分类像元对于各类别的隶属度。

1.2 RF 算法

随机森林(RF)算法基于统计学习理论,利用 Bagging(bootstrap aggregating)抽样方法(一种基于学习训练的有放回的抽样方法)从原始样本中抽取若干个样本,对抽取的样本使用基尼系数(Gini coefficient)作为属性度量,逐步回归建立分类回归决策树(classification and regression tree,CART)。CART 树的原理是使原本混乱的样本集通过决策树每层的划分变得相对纯净,反映到属性度量上就是使 Gini 系数下降。RF 算法使用多棵决策树对待分类像元进行预测,通过投票数决定类别的归属,并通过将不参与建树的样本作为"袋外数据"进行预测精度评价。

使用 RF 算法进行特征选择,是依据建立的 RF 算法计算每个特征对于分类结果的贡献度,然后根据贡献度自上而下选择若干特征。其方法可分为 3 种[6]:①最佳分类决策树特征选择法;②随机特征选取法[7-8];③Gini 系数下降法。这 3 种方法中,方法①只能定性地评价各特征贡献度的大小;方法②的算法较为复杂且理论基础不够牢靠;方法③依据统计学习理论,原理简单可靠,而且评价指标十分明确。因此,本文采取第三种 RF 特征选择方法,此方法的原理是获取每个特征使样本集 Gini 系数下降的程度。因为 RF 分类是通过对各个特征的逐层划分使原本混乱的样本集分裂为叶子节点纯净的样本集(即使 Gini 系数下降至最小),因此可将各特征使 Gini 系数下降的程度视为其对 RF 分类结果的贡献度。该方法通过遍历所有树节点,统计每个特征对应的 Gini 系数下降总和作为该特征的贡献度,再根据贡献度从高到低进行特征筛选。

1.3 基于 RF 算法的模糊分类

基于 RF 算法的模糊分类方法是基于分类器融合思想,其主要流程为:①选择纯净样本构建 RF,使用 Gini 系数下降法自高向低进行特征选择,保证所选特征贡献度之和大于 85%;②利用 RF 分类器进行分类,对分类结果中的每个类别随机选择若干个样本点,记录这些点的 RF 分类投票结果(即每个样本对每个类别的隶属度),构建模糊样本集;③使用模糊样本在所选特征维度上进行模糊分类。

RF 算法没有参照传统概率论的理论及方法,只通过统计学习的方法对数据进行评价投票,因此所选模糊样本可靠。该方法不直接将原始样本映射为模糊样本,这是因为原始样本较为纯净,会导致模糊协方差矩阵出现强相关甚至病态,影响分类精度。另外,模糊样本不直接取自原始样本,其精度只受 RF 训练误差影响;而 RF 对样本数据的噪声和异常值具有较强的稳定性,并且模糊分类对模糊样本的精度具有一定鲁棒性,因此该方法的分类精度对于原始样本的质量具有很强的稳定性。图 1 示出基于 RF 算法的模糊分类方法流程。

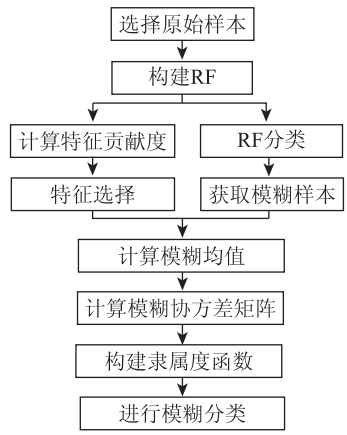


图 1 基于 RF 算法的模糊分类方法流程图

Fig. 1 Flow chart of fuzzy classification method based on RF algorithm

2 分类精度对比分析

实验使用的数据为 EO - 1 Hyperion 传感器数据,空间分辨率为 30 m,共有 242 个波段,光谱范围为 400 ~ 2 500 nm,光谱分辨率为 10 nm。本文设置了 2 个实验区,分别位于武汉东湖地区和武汉长江流域;2 个实验区均有较大面积的水体,在陆地部分植被覆盖较为密集,人工建筑与林地混杂分布不利于分类。

本文实验分类精度的计算方法是首先选择 n 个测试样区,然后利用所选择的测试样区评价各分类方法的整体分类精度,即

$$P_f = \frac{n_{\mathrm{hit}}}{n_{\mathrm{miss}}} \times 100\%, \quad (4)$$

式中:P_f 为分类方法 f 的整体分类精度;n_{hit} 为该分类方法中正确分类的测试样本个数;n_{miss} 为该分类方法中错误分类的测试样本个数。

在实际分类中,由于不同作业人员对不同区域的分类经验和认知水平有所差异,会出现样本可分离性较差、不同区域分类精度不一致等情况;而通过下述 3 则实验证明了新方法对于样本质量具有较高的鲁棒性(即算法稳定性),以及对不同区域都具有较强的分类能力(即算法的可推广性和可优化性),因而更符合实际分类的要求。

实验通过对比模糊分类、RF 分类和基于 RF 算法的模糊分类在不同样本条件下、不同实验区域及分区优化前后的分类精度,验证本文所提方法的有效性、可推广性和可优化性:①实验一选取 207 个测试样本对分类精度进行评价,选取 3 个训练样本集(618 个可分离性好的样本、622 个可分离性一般的样本、598 个可分离性较差的样本)分别构建 3 种分类器,对比不同样本条件下 3 种分类器的分类精度,用于验证本文方法的有效性;②实验二选择 187 个测试样本和 572 个训练样本进行 3 种分类,对比分类精度以验证本文方法的可推广性;③实验三利用实验一中可分离性好的 618 个训练样本和实验二中 572 个训练样本,一起构建基于 RF 算法的模糊分类器并进行分类,然后利用 2 个实验区的测试样本集评价分类结果的精度;并与样本分区训练所获得的分类结果进行对比,以验证本文方法的可优化性。

由于不可能在 242 维光谱空间进行模糊分类,需要利用 RF 算法进行特征选择,将特征贡献度计算结果从高到低排序(图 2),选择贡献度排名前 7 位的特征进行分类。

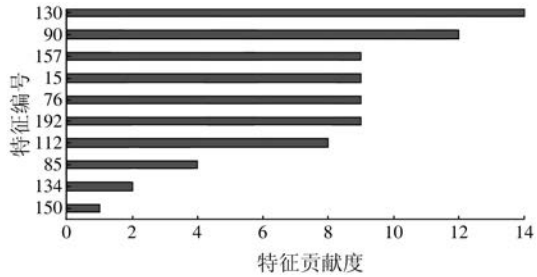
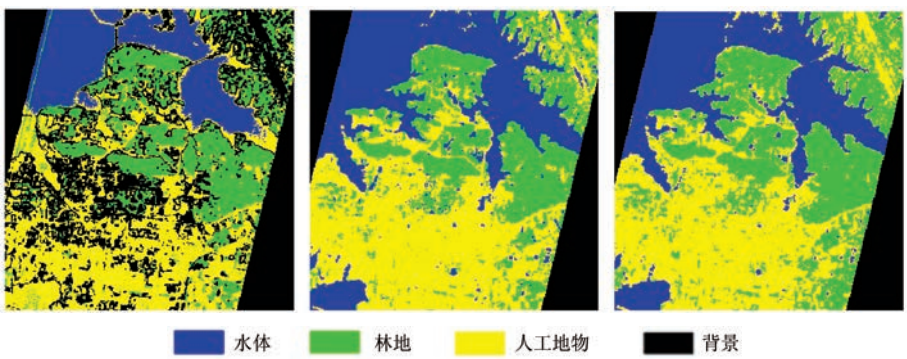


图 2　RF 算法提取的特征贡献度

Fig. 2　Feature contribution extracted with RF algorithm

2.1　有效性验证

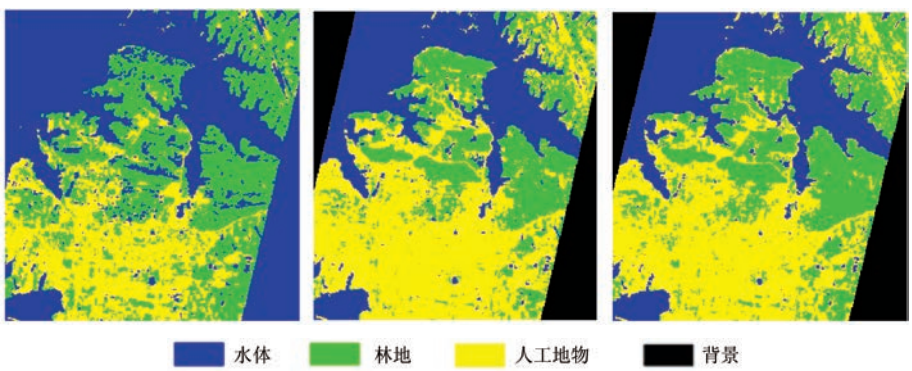
实验样本条件较好时、一般时和较差时 3 种算法的分类结果分别如图 3,4 和 5 所示。



(a) 传统模糊分类　　(b) RF 分类　　(c) 基于 RF 算法的模糊分类

图 3　样本质量较好时 3 种算法分类结果

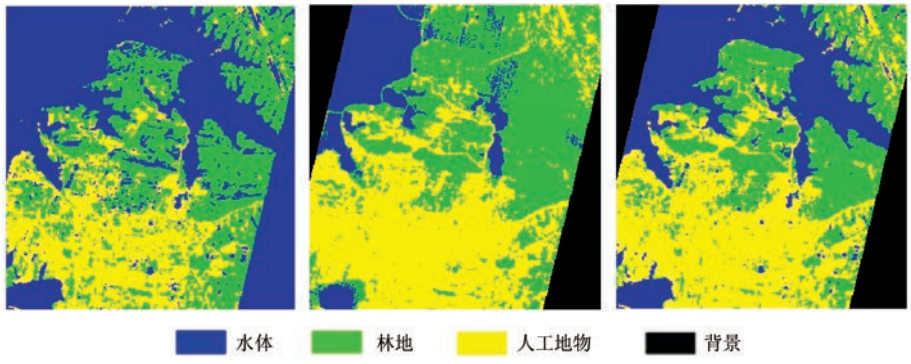
Fig. 3　Classification results of three algorithms when sample quality is good



(a) 传统模糊分类　　(b) RF 分类　　(c) 基于 RF 算法的模糊分类

图 4　样本质量一般时 3 种算法分类结果

Fig. 4　Classification results of three algorithms when sample quality is normal



(a) 传统模糊分类 (b) RF 分类 (c) 基于 RF 算法的模糊分类

图 5 样本质量较差时 3 种算法分类结果

Fig. 5 Classification results of three algorithms when sample quality is worse

表 1 列出基于不同质量样本的 3 种分类方法的分类精度。对分类精度最高的分类结果图进行像元数统计得知：水体占该区域面积的 26%，林地占 31%，人工建筑占 31%，背景占 12%。

表 1 基于不同质量样本的分类精度

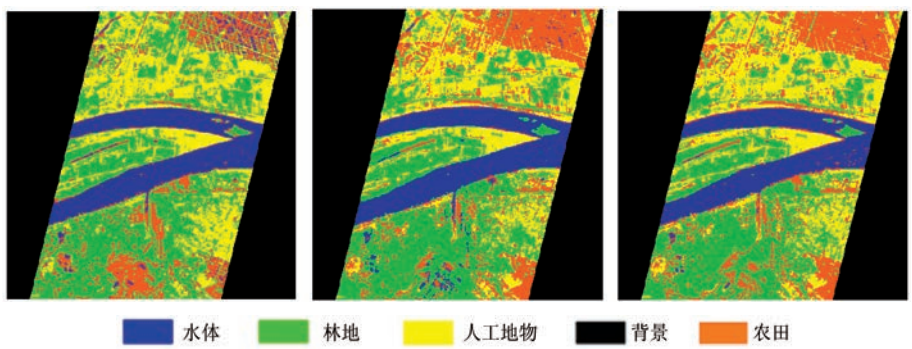
Tab. 1 Classification accuracy based on different samples with different quality (%)

样本质量	分类精度		
	传统模糊分类	RF 分类	基于 RF 算法的模糊分类
较好	79.0	88.9	90.1
一般	64.2	87.4	89.2
较差	63.1	71.2	87.0

通过上述实验可以看出：①基于纯净样本的传统模糊分类效果较差。这是因为其在本质上等同于最大似然法分类，而本实验的数据呈偏态分布，不满足其正态分布的假设前提，说明传统模糊分类不适用于本文研究区域的分类。②RF 算法得到合理的特征选择结果，而且表现出对数据质量一定的鲁棒性，但在样本较差的情况下精度明显下降。因此 RF 算法也不适合直接用于分类，可用于初分类获取模糊样本。③基于 RF 算法的模糊分类精度高于前 2 类算法分类精度，而且分类精度受样本质量的影响较小，分类精度均稳定在 87% 以上。因此针对本文研究区域，基于 RF 算法的模糊分类方法的分类效果最佳，证明该方法具有有效性。

2.2 可推广性验证

对于第二实验区，3 种算法分类结果如图 6 所示，表 2 列出 3 种分类方法的分类精度。



(a) 传统模糊分类 (b) RF 分类 (c) 基于 RF 算法的模糊分类

图 6 第二实验区分类结果

Fig. 6 Classification results of the second experimental area

表 2 第二实验区分类精度

Tab. 2 Classification accuracy of the second experimental area (%)

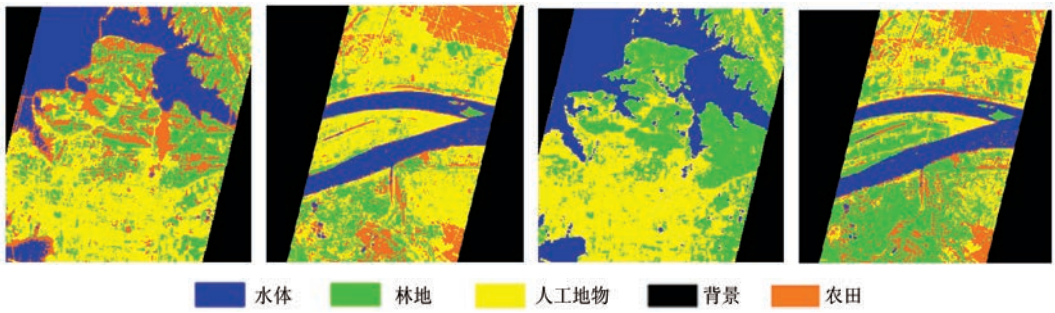
分类精度		
传统模糊分类	RF 分类	基于 RF 算法模糊分类
81.0	89.3	91.1

对分类精度最高的分类结果图进行像元数统计得知：水体占该区域面积 10%，林地占区域面积 27%，旱田占区域面积 17%，水田占区域面积 13%，背景占区域面积 33%。

通过实验二可以发现，在第二个实验区中基于 RF 算法的模糊分类精度依然高于另 2 类算法。与实验一相比，该实验区位于城市远郊，有大面积农田分布，且各类地物所占比例不同，可以认为 2 个实验区的地物分布具有一定差异；而 2 个实验区中基于 RF 算法的模糊分类精度都高于另 2 类算法，证明该方法具有可推广性。

2.3　可优化性验证

通过对比分区优化前、后分类结果以验证本文方法的可优化性。分区优化前的实验使用 2 个实验区的训练样本一起训练分类器,其分类结果如图 7(a)和(b)所示;分区优化后实验是通过分区进行分类器训练,其分类结果如图 7(c)和(d)所示。



(a) 第一实验区　(b) 第二实验区　(c) 第一实验区　(d) 第二实验区

(分区优化前)　(分区优化后)

图 7　分区优化前后分类结果

Fig. 7　Classification results before and after partition optimization

表 3 示出分区优化前、后分类精度。

表 3　分区优化前后分类精度

Tab. 3　Classification accuracy before and after partition optimization　(%)

分区优化	分类精度		
	第一实验区	第二实验区	2 区平均
优化前	86.4	82.3	84.4
优化后	90.1	91.1	90.6

通过实验三可以看出,分区优化前的分类在第一个实验区中许多林地被误分为农田,而第二个实验区的江心洲区域内林地被误分为人工地物。通过精度对比发现,分区优化前的分类精度均低于分区优化后的分类精度。因为 2 个实验区分别位于城市中心区和偏远郊区,同类样本的差异较大,分区处理可以增强区域内部的同质性,提高分类精度。上述结果验证了本文方法可以通过分区处理进一步提高分类精度,具有可优化性。

3　结论

1)本文提出的基于随机森林(RF)算法的高维模糊分类方法不仅克服了模糊分类的光谱维度局限性,而且提出了一种通过纯净样本构建模糊样本的解决方法。该方法的本质是利用分类器融合的思想,并保证了数理假设前提的一致性。

2) 通过 3 个实验(包括 20 余次对比实验、60 多次子实验)的结果证明,本文方法的分类精度不仅优于 RF 分类和模糊分类,而且对样本质量具有很强的鲁棒性,说明该方法具有有效性;该方法对于不同地物分布类型的区域都具有较好的分类效果,体现其具有可推广性;该方法还可以结合先验知识进行地理分层、分区,易于进一步提高其分类精度[9],体现出较强的可优化性。

3) 本文方法在特征贡献度分布均衡(即无法进行有效的特征选择)时失效。针对此问题,应首先进行特征提取工作。

参考文献(References):

[1] Wang F. Fuzzy supervised classification of remote sensing images [J]. IEEE Transactions on Geoscience and Remote Sensing, 1990,28(2):194 - 201.

[2] 张　永,吴晓蓓,徐志良,等. 基于多目标遗传算法的高维模糊分类系统的设计[C]//程代展,李　川. 第二十七届中国控制会议论文集. 北京:北京航空航天大学出版社,2008.
Zhang Y, Wu X P, Xu Z L. High - dimensional fuzzy classification system design based on multi - objective genetic algorithm[C]// Cheng D Z, Li C. Twenty - seventh Chinese Control Conference, Beijing: Beijing University Press, 2008.

[3] 杰　森. 遥感数字图像处理[M]. 北京:机械工业出版社,2007.
Jensen J R. Remote sensing digital image processing[M]. Beijing: Machinery Industry Publishing Society, 2007.

[4] Kumar U, Dasgupta A, Mukhopadhyay C, et al. Random Forest algorithm with derived geographical layers for improved classification of remote sensing data[J]. Machine Learning, 2012, 12(4): 1032 - 1043.

[5] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24(5): 123 - 124.

[6] Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting and randomization[J]. Machine Learning, 2000, 40(2): 139 - 157.

[7] 方匡南,吴见彬,朱建平,等. 随机森林方法研究综述[J]. 统计与信息论坛,2011,26(3):32 - 38.
Fang K N, Wu J B, Zhu J P, et al. A review of technologies on random forest[J]. Statistics and Information Forum, 2011, 26(3): 32 - 38.

[8] Breiman L. Randomizing outputs to increase prediction accuracy [J]. Machine Learning, 2000, 40(3): 229 - 242.

[9] Wilschut L I, Addink E A, Heesterbeek J A P, et al. Mapping the distribution of the main host for plague in a complex landscape in Kazakhstan: An object – based approach using SPOT – 5 XS, Landsat7 ETM$^+$, SRTM and multiple random forests[J]. International Journal of Applied Earth Observation and Geoinformation, 2013, 23:81 – 94.

[10] Shotton J, Johnson M, Cipolla R. Semantic texton forests for image categorization and segmentation[J]. IEEE Conference on Computer Vision and Pattern Recognition Anchorage, AK: IEEE, 2008:1 – 8.

[11] Ham J, Chen Y C, Crawford M M, et al. Investigation of the random forest framework for classification of hyperspectral data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2005, 43(3): 492 – 501.

[12] Miao X, Heaton J S, Zheng S F, et al. Applying tree – based ensemble algorithms to the classification of ecological zones using multi – temporal multi – source remote – sensing data[J]. International Journal of Remote Sensing, 2012, 33(6):1823 – 1849.

[13] Guan H Y, Li J, Chapman M, et al. Integration of orthoimagery and LiDAR data for object – based urban thematic mapping using random forests[J]. International Journal of Remote Sensing, 2013, 34(14):5166 – 5186.

[14] Palmer D S, O'Boyle N M, Glen R C, et al. Random Forest models to predict aqueous solubility[J]. Journal of Chemical Information and Modeling, 2007, 47(1):150 – 158.

[15] Richard C D, Edwards T C J, Beard K H, et al. Random Forests for classification in ecology[J]. Ecology, 2007, 88(11): 2783 – 2792.

Study of high – dimensional fuzzy classification based on random forest algorithm

ZHANG Xiuyuan, LIU Xiuguo

(*College of Information Engineering, China University of Geosciences, Wuhan* 430074, *China*)

Abstract: The spatial resolution of hyperspectral data is generally very low, the mixed pixels are extensively distributed, and hence fuzzy classification is commonly used in the mixed pixel analysis. As the accuracy of fuzzy classification is often limited by the feature dimensions and fuzzy samples selection, the random forest (RF) algorithm is put forward in this paper to select features and obtain fuzzy samples; in the low – dimensional feature space, fuzzy samples are used to make fuzzy classification. Fuzzy classification and RF are merged by using two – step classification, following the principle of unanimity assumption. Using different samples, different experimental areas and different partition optimization situations, the authors conducted three comparative experiments, and the results show that the method proposed in this paper solves the limitation of fuzzy classification and improves its accuracy. It is also proved that the classification accuracy of the method is robust for the original sample.

Key words: random forest(RF); fuzzy classification; high dimensional features

第一作者简介: 张修远(1991 –),男,本科生,主要研究方向为高分辨率遥感信息提取。Email: zh – x – y2008@163.com。

通信作者: 刘修国(1969 –),男,中国地质大学(武汉)信息工程学院教授。Email: liuxg318@163.com。

(责任编辑: 刘心季)