DOI:10.13203/j.whugis20140102 文章编号:1671-8860(2016)02-0228-07

基于随机森林的地理要素面向对象自动解译方法

顾海燕[1,2] 闫 利[1] 李海涛[2] 贾 莹[3]

1 武汉大学测绘学院,湖北 武汉,430079
2 中国测绘科学研究院,北京,100830
3 高德软件有限公司,北京,100080

摘 要:面向地理对象影像分析(GEOBIA)技术取得了显著的进展,代表了遥感影像解译的发展范式,其主要目标是发展智能化分析方法。随机森林机器学习方法是一种相对新的、数据驱动的非参数分类方法,具有自动特征优选、自动模型构建等优势,为智能化分析提供了有效手段。充分利用 GEOBIA 及随机森林机器学习的优势,提出了基于随机森林的地理要素面向对象自动解译方法,阐述了随机森林面向对象分类方法的技术流程,为设计和实现该方法提供了详细指导,有助于指导用户优选特征和构建分类模型。通过与支持向量机分类的对比实验证明,该方法可以自动进行特征优选及分类模型的构建,利用较少的特征得到较高的分类精度,在不损失性能的前提下减少了计算量和内存使用,能够为大范围、大区域地理要素自动解译提供先验知识及自动化的手段。

关键词:面向地理对象影像分析;随机森林;分类模型;特征选择

中图法分类号:P208;P237 **文献标志码**:A

面向地理对象影像分析(geographic object-based image analysis,GEOBIA)技术是地理信息科学中的一个新兴的迅速发展的研究领域,主要致力于研究如何分割遥感影像并产生有意义的地理影像对象,在一定的光谱、时-空尺度上计算这些对象的特征,最终生成与地理信息系统(GIS)兼容格式的地理信息。GEOBIA 的主要目标是发展与应用自动化或半自动化的影像解译理论、方法和工具,以提高影像解译的准确度及效率,减少人力、财力物力和主观性判断[1,2]。GEOBIA 具有以下优势:首先,综合利用了多源信息,如 GIS 数据、数字高程模型(DEM)以及景观生态、人文地理专题数据等;其次,充分利用了遥感影像的光谱、几何、纹理、拓扑、语义、时相等特征;同时,融合了主流影像分析方法,如监督分类、模糊逻辑、基于规则的分类等[3,4]。

GEOBIA 从提出到至今已有 10 余年的历史,受到了国内外众多学者及机构的关注,在城市监测、灾害监测、景观生态等领域得到广泛应用。GEOBIA 第一阶段主要发展的是影像分割、特征提取及分类的算法、工具及软件,目前最重要的发展趋势是面向对象分析的自动化与智能化[5-8]。然而,特征选择、规则集构建已成为制约 GEOBIA 自动化发展的关键因素。在特征选择和规则集构建方面,目前主要存在两大缺陷:一是难以确定哪些特征是非常重要的;二是不同的数据类型及不同的场景条件限制了分类规则集的应用[5]。因此,特征优选及构建分类规则集仍然是费时且具有挑战性的工作[9]。

随机森林(random forest,RF)是由数据驱动的非参数分类方法,它利用自助抽样技术和节点随机分裂技术构建多棵决策树,通过投票得到最终分类结果。该方法通过对给定样本进行学习训练形成分类规则,无需分类的先验知识,具有分析复杂地理信息系统分类特征的能力,对于噪声数据和存在缺失值的数据具有很好的鲁棒性,可以估计特征的重要性,具有较快的学习速度,相比当前流行的同类算法具有较高的准确性[9-11]。近年来,RF 算法已经应用到滑坡制图、城市树林制图、地表覆盖分类等领域。研究证明,该方法比传统方法更准确,运行速度更快,得到了研究者的广泛关注[12-16]。然而,目前研究主要局限于中低分辨率遥感影像像素级分类,关于面向对象的随机森林分类研究甚少,缺少对随机森林中特征优选

收稿日期:2014-08-04

项目资助:国家科技支撑计划(2012BAH28B03);地理空间信息工程国家测绘地理信息局重点实验室开放基金(201101)。

第一作者:顾海燕,博士生,主要从事遥感影像智能解译方法与应用研究。guhy@casm.ac.cn

策略的深入剖析以及利用随机森林进行面向对象分类的研究。

本文充分利用面向对象分析及随机森林机器学习的优势,提出基于随机森林的地理要素面向对象自动解译方法,通过特征自动优选及分类模型自动构建策略实现地表覆盖自动分类。

1　随机森林算法用于面向对象分类

1.1　算法原理

随机森林算法由 Leo Breiman 与 Adele Cutler 于 2001 年提出,是以多个决策树为基础分类器的集成分类器。它通过构造不同的训练子集来增加分类模型间的差异,从而提高组合分类模型的外推预测能力。随机森林方法主要包括训练与分类两个过程,方法流程图见图 1。

训练过程首先采用 bootstrap 自助抽样技术有放回地随机抽取训练样本,形成各个决策树的样本子集,未被选中的为 OOB (out-of-bag)样本;其次采用 CART (classification and regression trees)二元划分策略构建与样本子集对应的决策树,每个决策树的每个节点随机抽取 m 个特征(m 小于总特征数量 n),通过计算每个特征蕴含的信息量进行分裂生长;最后众多决策树构成一个随机森林。分类过程则是先对每个决策树进行分类,得到各自的分类结果,再利用简单投票法将所有分类结果进行综合,得到最终结果[10,11]。

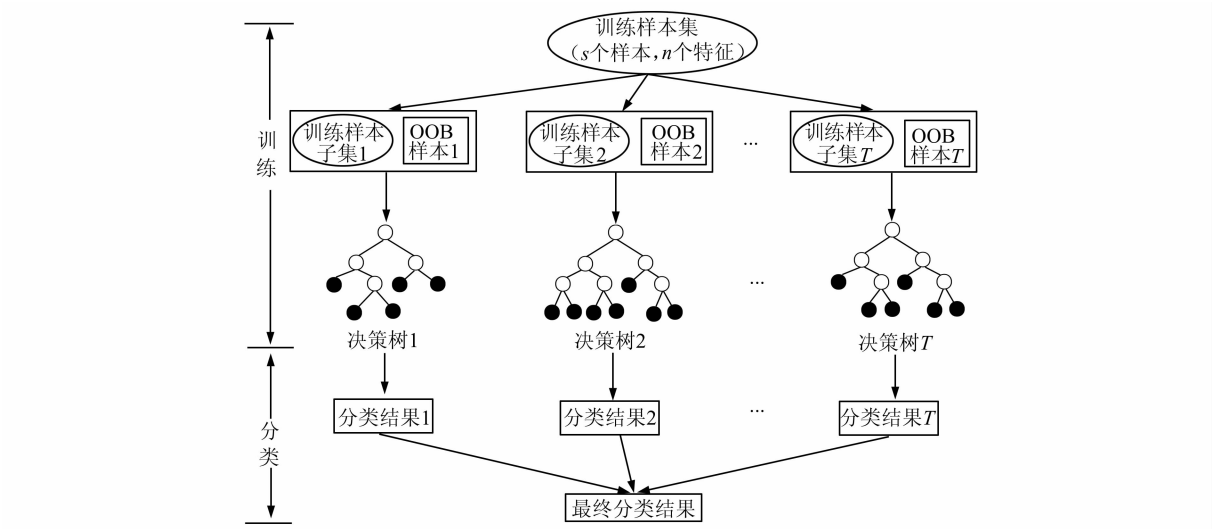

图 1　随机森林方法流程图

Fig. 1　The Flowchart of Random Forest

1.2　随机森林面向对象自动解译方法

面向对象分类的思想是,分类的最小单元不再是单个象素,而是“同质”多边形对象(图斑)。因此,首先通过分割技术得到多边形对象,接着统计对象的光谱、纹理、形状等特征,最后运用分类器实现面向对象的地理要素自动解译。

1) 影像分割。利用统计区域增长和异质性最小原则相结合分割方法(statistical region merging and minimum heterogeneity rule, SRMMHR)进行影像分割[17]。该方法充分利用了影像的光谱与形状信息,可以得到质量较好的多边形对象。

2) 特征提取。基于先验知识和用户知识进行特征选择,人工神经网络、支持向量机等分类器并不能进行特征优选,而随机森林算法能够分析大量的特征及少量的样本,计算特征值,值越大,说明所选的特征的重要性越高[18]。

3) 样本采集。根据准确性、代表性、统计性原则,采集地理要素的典型样本,形成样本集,一部分作为训练样本生成随机森林分类模型,一部分作为验证样本评价分类的精度。

4) 随机森林面向对象分类。该过程包括训练与分类过程。训练过程是根据训练样本及决策树理论得到分类模型,同时自动估算每个特征的重要性。分类过程是根据分类模型得到分类结果[19]。训练过程如下。

(1) 创建并初始化树集合、参与的样本序号、每个样本的测试分类等参数。

(2) 采用 bootstrap 自助抽样技术随机生成训练样本子集,利用递归方式训练单棵树。

① 计算当前节点样本中最大样本数量的类别,即为该节点的类别;

② 判断样本数量是否过少，或深度是否大于最大指定深度，或该节点是否只有一种类别，若是，则停止分裂；

③ 若步骤②中条件为否，则采用最优分裂策略针对某一变量进行左右分裂，最优分裂的依据是 $\frac{I}{N_l}\sum_i(|C_{i,l}|)^2+\frac{I}{N_r}\sum_i(|C_{i,r}|)^2$，其中 N_l 为左分裂的样本总数；N_r 为右分裂的样本总数；$C_{i,l}$ 为左分裂中类别 i 的样本个数；$C_{i,r}$ 为右分裂中类别 i 的样本个数；

④ 若不存在最优分裂或者无法分裂，则释放相关数据后返回；否则，处理代理分裂，分割左右分裂数据，调用左右后续分裂。

(3) 准确率判断及变量重要性计算

以准确率为终止条件或者计算变量的重要值时，则使用未参与当前树构建的样本测试当前树的预测准确率；当判断准确率达到标准或节点样本数少于设定值时，进入下一步；否则，返回前一步。若需计算变量的重要值，对于每一种变量，对每一个非参与样本替换该位置的变量值为另一随机样本的该变量值，再预测当前树的预测准确率，将其正确率的统计值与上一步当前树的预测准确率的差，累计到该变量的重要值中。

(4) 重复步骤(2)～(3)，直到准确率达到标准或者节点样本数过少为此，最终完成单个决策树的生成，加入到树集合中，形成随机森林模型。

(5) 精度评价。在随机森林中，可以用 OOB 误差估计作为泛化误差的无偏估计，不像其他方法通常需要交叉验证或利用分开的同分布的测试集[12]。

2 实验与分析

本文方法已集成到地理国情要素提取与解译系统 FeatureStation 中，并选取了三种典型区域(山区、平原、城区)不同数据源、不同分辨率的高分辨率遥感影像进行了实验，均取得较好效果。由于篇幅限制，本文只介绍城区实验情况，实验环境为：Intel(R) Core(TM), i7-2600 CPU@3.40 GHz, RAM 为 4 GB。

2.1 研究区数据

研究区为西安临潼城区，实验数据为 2011 年 7 月份的 WorldView-2 全色影像(分辨率为 0.5 m)与多光谱影像(分辨率为 2 m)。该数据已经过几何校正、融合处理，参与分类的波段为蓝、绿、红、近红外 1。该区域典型地类有耕地、林地、草地、房屋建筑区、道路、裸露地表、水体等，具体如图 2所示。

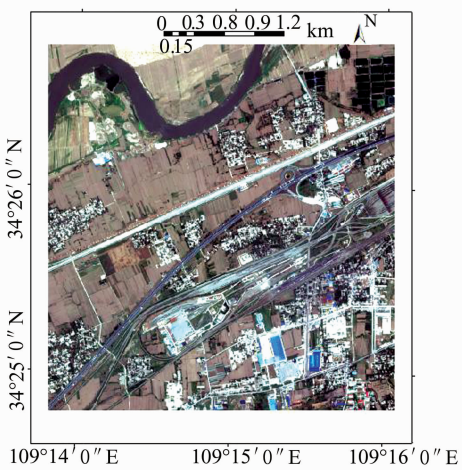

图 2 WorldView-2 融合影像

Fig. 2 Fused Image of WorldView-2

2.2 实验步骤

1) 影像分割

影像分割的原则是使影像对象内部异质性尽量小，以保证对象的纯度，而不同类别对象之间的异质性尽量大，以保证对象间的可分性。采用 SRMMHR 方法进行分割，利用分类反馈策略即以最高分类精度作为最优尺度选择标准。本实验在初始分割的基础上进行区域合并时，分割尺度为 200，光谱因子为 0.8，紧致度因子为 0.6 时，能得到相对较好的分割效果。具体见图 3。

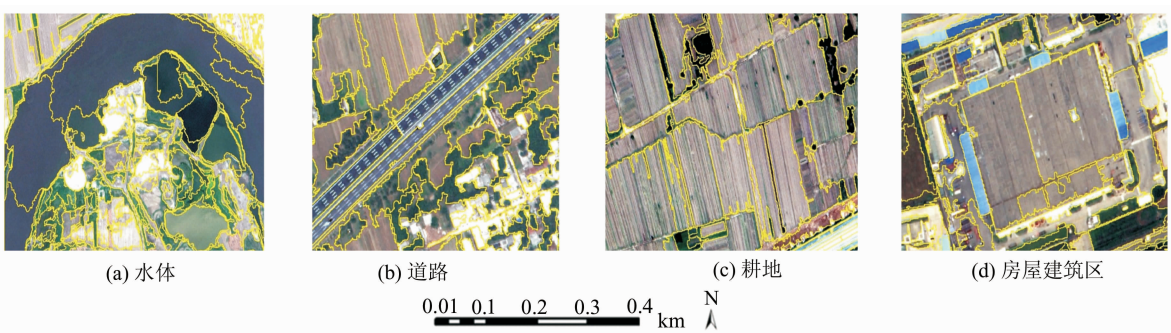

图 3 典型地类分割效果图

Fig. 3 Segmentation Result of Typical Land Groups

2) 特征提取

计算所有分割对象的光谱、纹理、形状等特征,根据经验与知识选择了 50 个特征,包括均值、方差、面积、周长、密度、同质性、对比度、熵、相关度等。可通过随机森林来计算特征的重要性,以此作为特征优选的策略,值越大,对分类的贡献越大。本实验部分特征重要度量值为 NDVI(30.3)>亮度(12.1)>密度(9.9)>红波段均值(5.9)>NDWI(5.4)>绿波段均值(4.7)。

3) 样本采集

针对耕地、林地、草地、房屋建筑区、道路、裸露地表、水体等 7 种地物类型,根据解译标志特征,采集典型样本,形成样本集。

4) 随机森林分类

采用 RF 方法进行面向对象分类,影响大的有两个参数,即组成森林的决策树个数及特征个数。当分裂特征个数为 5,决策树个数为 100 时,分类结果如图 4。

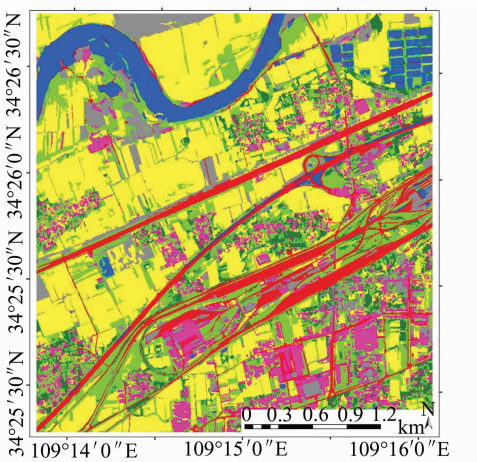

图 4　面向对象随机森林分类效果图

Fig. 4　Classification Result of RF GEOBIA Method

5)对比试验

支持向量机(support vector machine,SVM)是机器学习中的典型方法,分类性能超过决策树、神经网络、最大似然等方法[20],与随机森林方法效果相当[21],因此,本文利用 SVM 进行对比实验。SVM 类型有 C_SVC、NU_SVC 等,核函数有线性、多项式、径向基等。本实验 SVM 类型选择 C_SVC,核函数选择径向基函数,其中核参数 γ 值为 0.25,惩罚系数为 100,得到的分类结果如图 5 所示。

2.3　实验分析

1) 随机森林参数分析

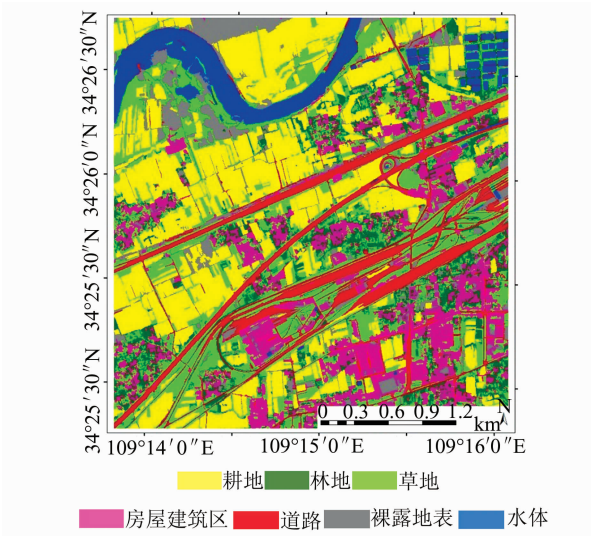

图 5　面向对象支持向量机分类效果图

Fig. 5　Classification Result of SVM GEOBIA Method

(1) 特征个数 M 对分类精度的影响

决策树个数 K 为 100,M 个特征用于分裂节点。随着 M 的变化,OOB 误差随特征个数 M 的变化趋势如图 6 所示。

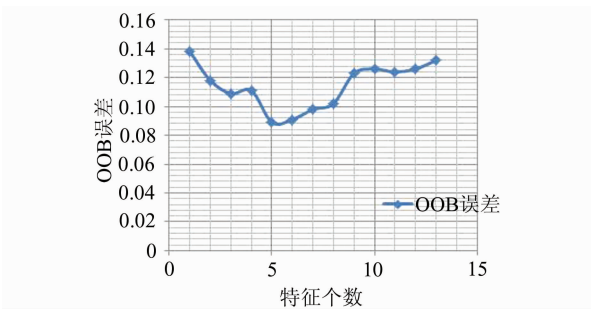

图 6　OOB 误差随特征个数变化的趋势图

Fig. 6　Effect of Feature Number(M) on OOB Error

可见,OOB 误差呈现先降低后上升的趋势,当 M 值为 5 时,OOB 误差最小,为 0.089,此时分类精度为 91.1%。

(2) 决策树个数 K 对分类精度的影响

为了评估决策树个数的理想值,M 保持常量为 5。随着 K 的变化,OOB 误差如图 7 所示。

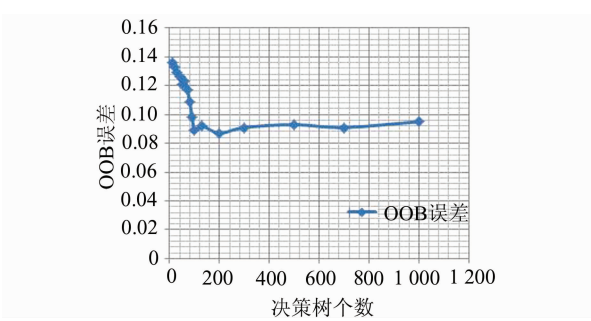

图 7　OOB 误差随决策树个数变化的趋势图

Fig. 7　Effect of Tree Number(K) on OOB Error

可见,随着 K 的增加,误差变小;当 K 为 200 时,误差达到最小,大于 200 时,误差差别很小,趋于稳定,计算时间随着 K 的增加而增加。当 K 为 200 时,OOB 误差为 0.087,此时,分类精度达到 91.3%。

2)特征对分类精度的影响

我们将 50 个特征分为 10 组,每组包含 5 个特征,首先考虑光谱,其次考虑形状,最后考虑纹理特征。运用这 10 组特征做两种方法的分类实验,总体精度如表 1 所示。对比图如图 8 所示。

表 1 10 组特征对应的 RF 和 SVM 分类总体精度

Tab. 1 Overall Accuracy of RF and SVM GEOBIA Methods for Ten Groups of Features

特征数量	RF GEOBIA 总体精度/%	SVMGEOBIA 总体精度/%
5	85.36	84.92
10	**91.08**	88.62
15	90.41	91.69
20	90.24	90.46
25	90.26	91.08
30	90.15	92.31
35	89.66	92
40	90.3	**93.23**
45	90.52	92.92
50	90.82	91.69

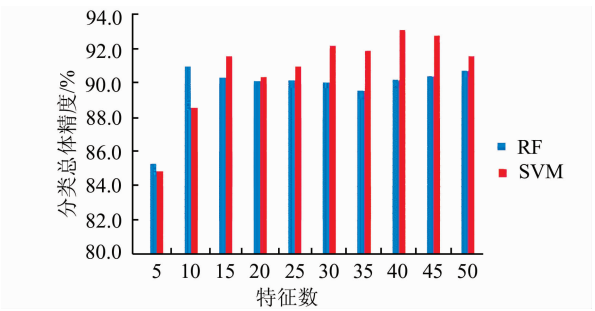

图 8 十组特征对应的 RF 和 SVM 分类总体精度比较图

Fig. 8 Comparison of RF and SVM GEOBIA Methods for Ten Groups with Different Numbers of Features

当特征数为 10 时,RF GEOBIA 方法能够达到最高的分类精度 91.08%,而当特征数为 15 时,SVM GEOBIA 方法才能达到类似的精度。可见,RF GEOBIA 方法利用较少的特征就能达到较高的精度,随着特征数量的增多,分类精度趋于稳定。而 SVM GEOBIA 方法利用较多的特征才能达到较高的精度,且分类精度呈现先上升后下降的趋势。由此证明了 RF GEOBIA 方法特征自动优选的重要性。

3) 计算时间分析

总体计算时间包括分割、特征计算、样本选择、分类等。在相同的计算环境下,由于两种方法分割、样本选择的时间是相同的,因此本实验只统计分析特征计算与分类时间。具体如表 2 所示。

随着特征数的增加,两种方法的计算时间都有较大增长,但在相同的特征组下,RF 方法的速度略快于 SVM 方法,这说明 RF 方法通过特征自动优选及分类模型自动构建,在不损失性能的前提下减少了计算量和内存使用。

表 2 特征计算与分类时间统计表

Tab. 2 Computing Time for Feature Calculation and Classification

特征数量	特征计算时间	RF 分类时间	SVM 分类时间
5	2′10″	20″	26″
10	4′06″	26″	34″
15	6′15″	29″	45″
20	8′56″	38″	52″
25	11′14″	43″	59″
30	17′20″	56″	1′18″
35	19′35″	1′08″	1′28″
40	24′16″	1′38″	1′56″
45	56′40″	1′57″	2′34″
50	1∶48′50″	2′11″	2′58″

3 结 语

本文利用面向地理对象分析及随机森林的优势,提出了随机森林面向对象分类方法;详细阐述了该方法技术流程。其具有自动进行特征优选及自动构建分类模型的优势,利用较少的特征就能得到较高的分类精度,相比当前流行的同类算法(如 SVM)具有显著的优越性。

随机森林方法是一种相对新的、数据驱动的非参数分类方法,只须通过对给定样本进行学习训练形成分类规则,具有分析复杂相互作用分类特征的能力;对于噪声数据和存在缺失值的数据具有很好的鲁棒性,可以估计特征的重要性,具有较快的学习速度。RF 内置的特征选择方法用于选择与分类模型密切相关的特征,自动分类模型能够减少人工解译的时间,为 GEOBIA 提供一种自动化的手段。

随机森林面向对象分类方法仅仅需要设置两个参数,一是决策树个数,二是随机分裂变量特征个数。200 个树之内,树的个数与分类精度成正比,一旦误差聚合,随机分裂变量的个数对分类精度的影响比较小。此外,无须通过交叉验证和训练样本集进行精度评价,内置的 OOB 误差可以评价分类精度。

不过,随机森林是一个新的机器学习方法,仍

处于不断发展中,许多问题需要深入研究,例如RF方法的基本理论及其改进方法;影响分类的因素,如空间尺度、分割方法、分割尺度、测试样本、对象特征、人类认知等。

参 考 文 献

[1] Wiki. GEO-Object-Based Image Analysis[OL]. http://wiki.ucalgary.ca/page/GEOBIA,2015

[2] Hay G J, Castilla G. Object-based Image Analysis: Strengths, Weaknesses, Opportunities and Threats (SWOT)[C]. The 1st International Conference on Object-based Image Analysis (OBIA), Salzburg, Austria, 2006

[3] Robertson L D, King D J. Comparison of Pixel-and Object-based Classification in Land-cover Change Mapping[J]. *International Journal of Remote Sensing*,2011,32(6):1 505-1 529

[4] Li Haitao, Gu Haiyan, Han Yanshun, et al. Object-oriented Classification of High-resolution Remote Sensing Imagery Based on an Improved Colour Structure Code and a Support Vector Machine[J]. *International Journal of Remote Sensing*,2010,31(6):1 453-1 470

[5] Blaschke T. Object Based Image Analysis for Remote Sensing[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*,2010, 65:2-16

[6] Definiens Imaging GmbH. Developer 8 Reference Book; Definiens Imaging GmbH[OL]. http://www.doc88.com/p-0963997861736.html

[7] Baraldi A, Boschetti L. Operational Automatic Remote Sensing Image Understanding Systems: Beyond Geographic Object-based and Object-oriented Image Analysis (GEOBIA/GEOOIA). Part 1: Introduction[J]. *Remote Sensing*,2012,4:2 694-2 735

[8] Baraldi A, Boschetti L. Operational Automatic Remote Sensing Image Understanding Systems: Beyond Geographic Object-based and Object-oriented Image Analysis (GEOBIA/GEOOIA). Part 2: Novel System Architecture, Information/Knowledge Representation, Algorithm Design and Implementation[J]. *Remote Sensing*,2012,4:2 768-2 817

[9] Stumpf A, Kerle N. Object-oriented Mapping of Landslides Using Random Forests[J]. *Remote Sensing of Environment*, 2011, 115(10): 2 564-2 577

[10] Breiman L, Cutler A. Random Forests[OL]. http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm,2015

[11] Breiman L. Random Forests[J]. *Machine Learning*, 2001,45(1):5-32

[12] Rodriguez-Galiano V F, Chica-Olmo M, Abarca-Hernandez F,et al. Random Forest Classification of Mediterranean Land Cover Using Multi-seasonal Imagery and Multi-seasonal Texture[J]. *Remote Sensing of Environment*,2012,121:93-107

[13] Guo L, Chehata N, Mallet C,et al. Relevance of Airborne LiDAR and Multispectral Image Data for Urban Scene Classification Using Random Forests[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*,2011,66(1):56-66

[14] Stumpf A, Kerle N. Combining Random Forests and Object-oriented Analysis for Landslide Mapping from very High Resolution Imagery[J]. *Procedia Environmental Sciences*, 2011,3:123-129

[15] Liu Yi,Du Peijun,Zheng Hui,et al. Classification of China Small Satellite Remote Sensing Image Based on Random Forests[J]. *Science of Surveying and Mapping*, 2012,37(4):194-196(刘毅,杜培军,郑辉,等.基于随机森林的国产小卫星遥感影像分类研究[J].测绘科学,2012,37(4):194-196)

[16] Lei Zhen. Random Forest and Its Application in Remote Sensing[D]. Shanghai: Shanghai Jiao Tong University,2012(雷震. 随机森林及其在遥感影像处理中应用研究[D].上海:上海交通大学, 2012)

[17] Li Haitao, Gu Haiyan, Han Yanshun, et al. An Efficient Multi-scale SRMMHR (Statistical Region Merging and Minimum Heterogeneity Rule) Segmentation Method for High-resolution Remote Sensing Imagery[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2009,2(2):67-73

[18] Verikas A, Gelzinis A, Bacauskiene M. Mining Data with Random Forests: A Survey and Results of New Tests[J]. *Pattern Recognition*, 2011,44(2):330-349

[19] OpenCV Documentation. OpenCV 2.4.6 Online Documentation[OL]. http://docs.opencv.org/modules/ml/doc/random_trees.html,2015

[20] Huang C, Davis L S, Townshend J R G. An Assessment of Support Vector Machines for Land Cover Classification[J]. *International Journal of Remote Sensing*,2002,23(4):725-749

[21] Pal M. Random Forest Classifier for Remote Sensing Classification[J]. *International Journal of Remote Sensing*, 2005,26(1):217-222

An Object-based Automatic Interpretation Method for Geographic Features Based on Random Forest Machine Learning

GU Haiyan[1,2] *YAN Li*[1] *LI Haitao*[2] *JIA Ying*[3]

1 School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

2 Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping, Chinese Academy of Surveying and Mapping, Beijing 100830, China

3 AutoNavi Software Co. Ltd., Beijing 100080, China

Abstract: Geographic object-based image analysis (GEOBIA) techniques have recently seen considerable development in comparison to traditional pixel-based image analysis, representing a paradigm shift in remote sensing interpretation. The main aim is to incorporate and develop geographic-based intelligence. The random forest (RF) machine learning method is a relatively new, non-parametric, data-driven classification method that can supply intelligent means for feature selection and classification modelling. This paper presents a novel RF GEOBIA method for land-cover classification that makes full use of the advantages of GEOBIA and RF. A detailed RF GEOBIA workflow is proposed to guide the design and implementation of the method, and to guide experts during elaboration of feature selection and classification modelling. Theoretical and experimental results are compared with the support vector machine (SVM) approach, demonstrating that it is a robust and intelligent method for land-cover classification with wrapper feature selection and classification modelling. The RF GEOBIA method reduces the number of features required, computing time, and memory requirements, with no associated reduction in performance. It also provides a priori knowledge for further classification and supports large scale applications where "big data" is involved.

Key words: geographic object-based image analysis (GEOBIA); random forest; classification model; feature selection

First author: GU Haiyan, PhD candidate, specializes in intelligent interpretation for remote sensing image. E-mail: guhy@casm.ac.cn

Foundation support: The National Science & Technology Pillar Program, No. 2012BAH28B03; Key Laboratory of Geoinformatics of National Administration of Surveying, Mapping and Geoinformation, No. 201101.