

doi: 10. 6046/gtzyyg. 2016. 03. 12
引用格式: 鹿丰玲, 巩在武. 基于随机森林算法构建云 - 云阴影 - 水体掩模[J]. 国土资源遥感, 2016, 28(3): 73 - 79. (Lu F L, Gong Z W. Construction of cloud - shadow - water mask based on Random Forests algorithm [J]. Remote Sensing for Land and Resources, 2016, 28(3): 73 - 79.)

基于随机森林算法构建云 - 云阴影 - 水体掩模

鹿丰玲^{1,2}, 巩在武²

(1. 南京信息工程大学地理与遥感学院, 南京 210044; 2. 南京信息工程大学经管学院, 南京 210044)

摘要: 遥感图像数据中云和云阴影的存在是影响数据应用的主要原因, 专家已经研发了多种去除云及其阴影的方法。在对不同目标像元光谱曲线分析的基础上, 研究了基于随机森林(random forests, RF) 分类器的云 - 云阴影 - 水体掩模建立方法。由于云阴影是阴影与地表物体的叠加, 其光谱曲线与水体的光谱曲线之间存在细微的差别, 这使得决策树(decision tree, DT) 分类方法不能非常有效地应对这种细微差别。RF 分类器是建立在多个 DT 分类结果集成的基础上, 其算法原理保证了该算法的稳健性和有效性。研究结果表明: 在样本容量较少时, RF 算法比 DT 具有更好的分类效果; 而在样本容量增大到 250 ~ 400 个像元时, 2 种方法的分类效果没有明显区别。这表明 RF 算法可以成功地用于建立云 - 云阴影 - 水体掩模, 这将在遥感数据处理中得到更加广泛的应用。

关键词: 掩模; 遥感; 云 - 云阴影; 决策树(DT) 分类器; 随机森林(RF) 分类器

中图法分类号: TP 751.1 **文献标志码:** A **文章编号:** 1001 - 070X(2016)03 - 0073 - 07

0 引言

遥感图像数据在地表覆被变化、森林扰动及环境监测等研究中起着至关重要的作用。现有的多种卫星数据, 如 MODIS 数据、TM/ETM⁺ 数据以及 SPOT5 卫星数据等, 为揭示人类生存环境与气候变化之间的关系提供了可能。但是, 由于卫星数据在获取过程中不可避免地会受到大气和云的干扰, 许多卫星影像数据中存在部分云覆盖及云阴影现象, 这为准确提取地表信息带来困难, 直接影响了遥感数据的实际应用。

针对卫星遥感图像数据中存在云覆盖及其阴影的现象, 国内外学者都进行了相关研究, 并提供了部分解决方法。例如 Catalin^[1] 介绍了在其设计的软件中构造云 - 阴影掩模的一种方法, 其原理是把多光谱图像利用参数向量转换为灰度图像, 通过对比灰度值与给定的阈值将云及其阴影的掩模构建出来。事实上, 由于云对太阳光线具有较强的反射率, 其在多个波段图像上都表现为白色, 明显区别于其他地物, 因此云的掩模可以通过分析云的反射率和其他地物的反射率来建立。但是, 构造云阴影的掩模就比较困难, 主要因为云阴影受到云厚度、地表覆盖物和太阳入射角等多种因素的影响, 其在遥感图

像中表现为云阴影与其他地表物体重叠的阴影区。苏惠敏等^[2] 提出利用主成分变换对云和阴影分别去除的方法, 取得了较好的目视效果; 李炳燮等^[3] 提出利用多期 TM 图像数据对厚云及其阴影去除的方法, 首先通过分类识别出被厚云及其阴影覆盖的像元, 然后利用其他时相的、同一位置的、未被云和阴影覆盖的像元进行替换; 郑文武等^[4] 提出基于动态端元选择的线性光谱分解算法, 通过选择云、绿色植被、云阴影和不透水表面 4 种端元达到去除 TM 图像数据中的云和阴影的目的; 梁栋等^[5] 则利用支持向量机检测遥感图像中的云, 并结合太阳入射角判定云阴影区域, 在完成云及云阴影的初步去除后, 对未能去除的云及云阴影通过补偿的方法进行修复。

相关研究和所提出的算法在一定程度上实现了遥感图像中云和云阴影的检测与去除, 在目视效果上有了改善^[6-9]; 但对于遥感图像数据在地球环境变化监测方面, 更重要的是要看去除云和阴影后的数据是否可以继续使用(例如能否提供一致的归一化植被指数、能否用来反应森林扰动现象)。因此, 建立云和阴影的掩模, 以便在后续的遥感数据应用中把云和阴影去除, 也是非常关键的。本文针对 TM/ETM⁺ 图像中存在的云 - 云阴影 - 水体, 进行掩模构建分析。在相关的研究过程中发现, 当遥感数

据中存在少量的云或覆盖面积不大的云团时,这样的数据还是可以用于进一步信息提取的(如植被指数提取、地表温度分析及城市不透水层分析等);但是,当遥感数据中包含有大量的云团覆盖时,这样的数据基本上是不能使用的。本文研究的方法只针对前一种情况。

1 数据与方法

1.1 遥感数据

Landsat TM/ETM⁺ 遥感图像数据在城市热环境、植被物候变化及森林扰动等方面研究中得到了广泛应用,Landsat8 卫星数据的成功获取也为 Landsat 系列卫星数据的其他应用带来了机会。由于大气和云的影响,Landsat 图像数据应用不可避免地会受到云及其阴影的干扰。如图 1 所示,该图像中部分区域含有云及其阴影,但其他未被云覆盖的区域仍有较好的目视效果,在对云、阴影和水体进行掩模处理后,可用于其他相关研究。为此,本文以图 1 所示的 Landsat TM 多光谱图像作为研究数据,进行云-云阴影-水体的掩模建立研究。该卫星图像的时相

为 2010 年 8 月 19 日,轨道号/行号为 120/38,中心位置地理坐标为 E119°7'56",N31°51'7",波段组合为 TM5(R)TM4(G)TM3(B)。

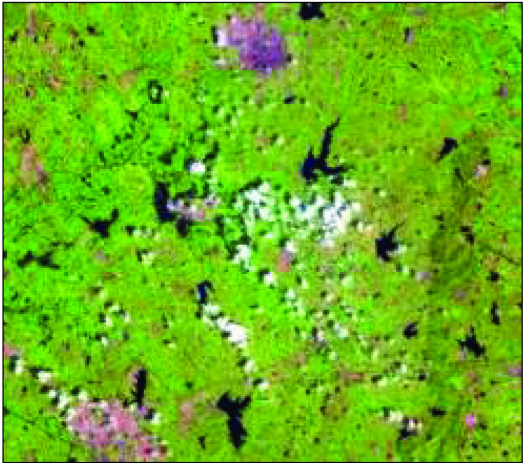


图 1 Landsat TM5(R)TM4(G)TM3(B) 假彩色合成影像(局部)
Fig.1 Landsat false color image composed of TM5(R)TM4(G)TM3(B)(part)

1.2 云-云阴影-水体光谱分析

利用 ENVI (Version4.7) 软件,提取了云、云阴影、水体以及其他地表物体的光谱曲线(图 2)。

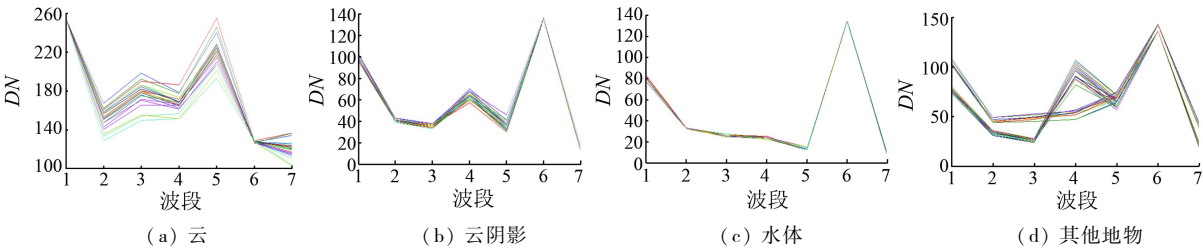


图 2 云、云阴影、水体和其他地物光谱曲线
Fig.2 Spectral curves of cloud, cloud shadow, water and other surface features

从图 2 可以看出,在可见光波段中云的光谱曲线明显区别于云阴影、水体和其他地物的光谱曲线,因此云很容易被区分出来;但也有些云(如薄卷云),可能因其光谱特征不明显而难以区分。水体和云阴影的光谱曲线极为相似,只是在近红外波段上云阴影的 DN 值略大于水体的 DN 值。这是因为近红外波段对绿色植被具有最大的反射峰值,能够用来探测健康绿色植被和区分水陆边界;而水体对近红外波段的吸收作用致使水体在该波段上 DN 值

较小;这使得近红外波段成为区分水体和云阴影的关键波段。其他地物类型,如植被、裸露土壤及不透水层等的光谱曲线分布比较杂乱,各有特点,但都与水体和云的光谱曲线有较大差别。云、云阴影、水体和其他地物的光谱曲线在热红外波段(TM6)均不具有可分性(光谱曲线汇集于一点),几乎没有区别。

将云、云阴影、水体和其他地物等 4 种地物类型各取 500 个样本,分别制作出各类地物在 6 个波段上的盒形图,如图 3 所示。

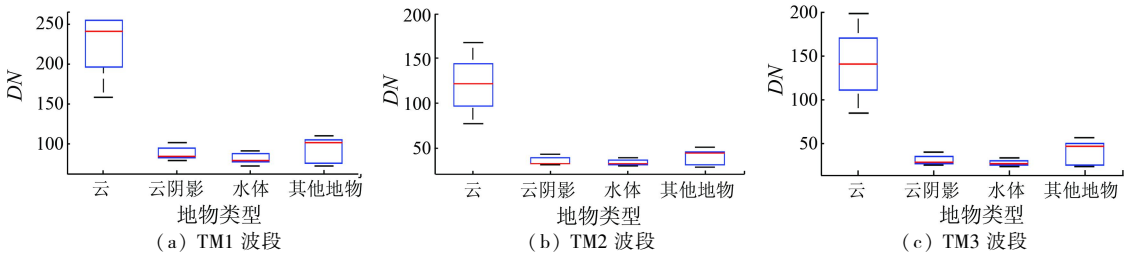


图 3-1 云、云阴影、水体和其他地物光谱盒形图
Fig.3-1 Boxplots of spectrum of cloud, cloud shadow, water and other surface features

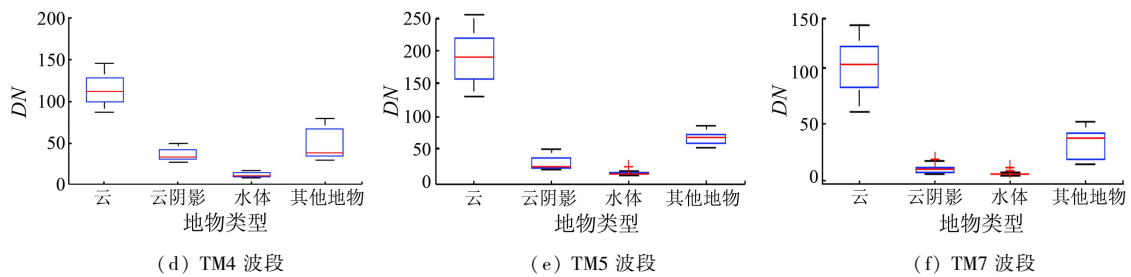


图 3-2 云、云阴影、水体和其他地物光谱盒形图

Fig. 3-2 Boxplots of spectrum of cloud, cloud shadow, water and other surface features

云在TM1 — 4波段上与其他3种类型地物都有着明显的可分性;其他地物类型与水体和云阴影在TM5和TM7波段上具有一定的可分性,但与云阴影存在一定的重叠;这是由于云阴影的光谱特征是阴影与地表覆盖物光谱特征的混合,因而导致云阴影与其他地物的光谱曲线没有非常明显的区别。云阴影和水体可以通过近红外波段来区分。

1.3 随机森林算法

1.3.1 决策树分类器

决策树(decision tree,DT)分类器是一种基于空间数据挖掘和知识发现的监督分类方法^[10],它不需要像最大似然分类器那样对数据进行正态分布的假设,因而该方法被广泛应用于空间数据分类和文本数据分类等方面。随着遥感技术的发展,DT分类方法也经常被应用于遥感数据的土地利用分类和信息提取研究中^[11-16]。

根据图3所示的盒形图,在ENVI软件中构造了图4所示的DT,并对DT进行了修剪;利用TM2, TM4和TM7波段的数据来构造DT,其中各个节点处的阈值是通过样本盒形图分析得出的。

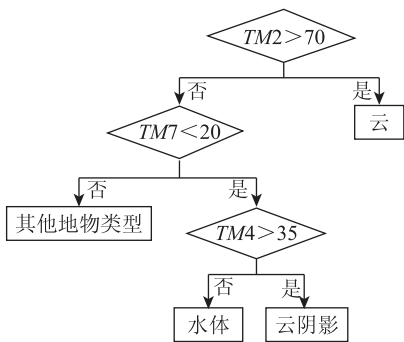


图 4 DT 分类器

Fig. 4 Decision tree classifier

在数值实验中发现,有部分颜色较深的绿色植被被误分为云阴影;同时,也有部分深度较浅的水体被误分为云阴影。出现这种情况的原因是因为云阴影与水体及其他地物类型在某些波段上的区分度不是太大,而人为选择的阈值可能不尽合理而影响

了对这几种地物类型的区分。为了避免这种人为选取阈值带来的误差,本文选用随机森林(random forests, RF)分类方法来建立云-云阴影-水体的掩模。

1.3.2 RF 分类器

RF是近几年来出现的一种集成学习算法^[17],所谓的集成是指把多个模型或多种不同类型模型所得到的结果通过一定的规则组合起来。RF的名称来自于DT分类器,当把很多棵DT放在一起时就构成了一个森林,对多棵DT的分类结果的集成通常优于单棵DT的分类结果。RF方法在人脸识别、指纹识别、纹理识别及分类中取得了较好的应用效果^[18-21]。

设 $h(x, \theta)$ 为一棵DT,其中 θ 表示DT中的参数向量。一个RF是由 N 棵随机产生的DT构成的。设随机产生的 N 个随机向量为 $\theta_1, \theta_2, \dots, \theta_N$,这些随机向量是独立同分布的,则由随机向量 θ_i 构造的DT为 $h(x, \theta_i)$,RF为这些DT的集合 $\{h(x, \theta_i)\}$ 。每棵DT是一个分类器,其中的每个节点都是一个弱分类器。RF的决策结果是所有DT分类结果的集成。最常用的集成方法是张量投票方法,即RF选择投票数最多的分类结果作为最终结果。图5为RF分类器的示意图。

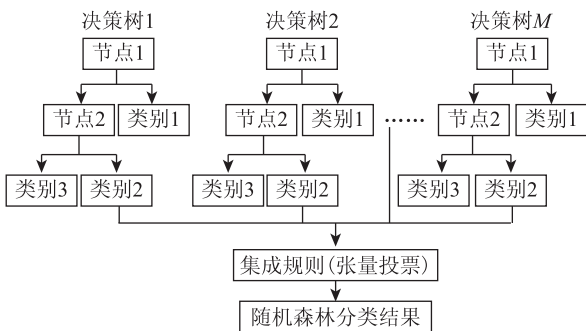


图 5 RF 分类器

Fig. 5 Random forests classifier

RF的算法设计使得DT的选取是随机的,样本的选择也是随机的。对每棵DT的分类结果采取投票规则进行集成,可以保证集成结果的客观性和稳

健性。同时,在 RF 方法中不需要人为设置每个节点的阈值,完全由算法自动实现对每支树枝的生长和剪枝过程,从而最大限度地避免了人为因素对最终结果的影响。

RF 的生长过程为^[18-21]:

- 1)从原始训练数据集中,应用 Bootstrap 方法随机抽取 M 个样本集,并由此构建 M 个 DT;
- 2)对每棵 DT 随机选取若干个样本特征,计算特征中蕴含的信息量,并以此为准则判断每个节点是否进行分裂;
- 3)让每棵 DT 最大限度地生长,不做任何裁剪;
- 4)将生成的 M 棵 DT 组成 RF,对 M 棵 DT 的结果进行集成。

事实上,RF 分类器和 DT 分类器都利用了各个类别的光谱特征信息;但是它们内部机制的不同决定了 RF 算法明显优于 DT 方法。

2 结果与分析

以遥感图像的 DN 值为输入数据,分别应用 RF 和 DT 算法进行了云-云阴影-水体掩模的提取。云、云阴影、水体和其他地物等 4 类样本的总数量为 681 个像元。

首先,从 681 个像元中随机选取 300 个像元作为训练样本,其余 381 个像元作为检验样本。在对 RF 和 DT 这 2 种分类器进行训练后,分别对检验样本进行分类。分类结果的混淆矩阵见表 1 和表 2。

表 1 RF 分类结果混淆矩阵
Tab. 1 Confusion matrix derived from random forest classifier

类别	云	其他地物	水体	云阴影	总计	用户精度/%
云	61				61	100
其他地物		96		2	98	98
水体			77		77	100
云阴影		1		144	145	99
总计	61	97	77	146	381	
制图精度/%	100	99	100	99		
总体精度/%	99.21		Kappa = 98.90			

表 2 DT 分类结果混淆矩阵

Tab. 2 Confusion matrix derived from decision tree classifier

类别	云	其他地物	水体	云阴影	总计	用户精度/%
云	61				61	100
其他地物		96		4	100	96
水体			77		77	100
云阴影		1		142	143	99
总计	61	97	77	146	381	
制图精度/%	100	99	100	97		
总体精度/%	98.6		Kappa = 98.18			

从表 1 和表 2 可以看出,RF 分类的总体精度略高于 DT 分类的总体精度,二者的 Kappa 系数差别不大。由于云和水体都具有明显区别于其他地物及云阴影的光谱特征,这使得 DT 分类器和 RF 分类器都能比较容易地把云和水体区分出来。但对于云阴影,有 2 个像元被 RF 分类器误判为其他地物,有 4 个像元被 DT 分类器误判为其他地物,这一细微差别显示了 2 种算法在处理光谱特征极为相似的不同地物类型时效果也有所差别。

其次,为了进一步分析 2 种算法对不同大小训练样本分类性能差别,分别建立了样本容量为 50, 100, 150, 200, 250, 300, 350 和 400 像元的训练样本集合,剩余的样本作为检验样本。以总体分类精度和 Kappa 系数作为评价指标,对 DT 和 RF 这 2 种算法的分类结果进行了对比。不同容量下的样本集合都是利用随机生成的随机数从样本总体中随机选取的,分类结果见表 3。

表 3 不同大小样本分类精度比较

Tab. 3 Comparison of accuracy of classification using different sizes of samples

训练样本大小/ 像元数	DT		RF	
	总体精度/%	Kappa	总体精度/%	Kappa
50	97	0.965	98	0.982
100	97	0.959	99	0.995
150	98	0.982	99	0.990
200	98	0.977	99	0.988
250	99	0.990	99	0.997
300	99	0.989	99	0.989
350	99	0.992	99	0.987
400	99	0.990	99	0.995

从表 3 可以看出,在处理较小的训练样本时,RF 算法给出了精度更高的分类结果。2 种算法在面对不同样本容量时具有比较明显的区别:当训练样本容量少于 100 个像元时,RF 算法的总体精度和 Kappa 系数均明显高于 DT 算法的总体精度和 Kappa 系数;当样本容量小于 200 个像元时,DT 算法的分类精度接近于 RF 算法的分类精度;当样本容量大于 200 个像元时,2 种算法的分类精度几乎没有差别。

在训练样本容量为 300 时,对 2 种算法进行训练后应用于整景图像,得到云-云阴影-水体掩模(图 6)。

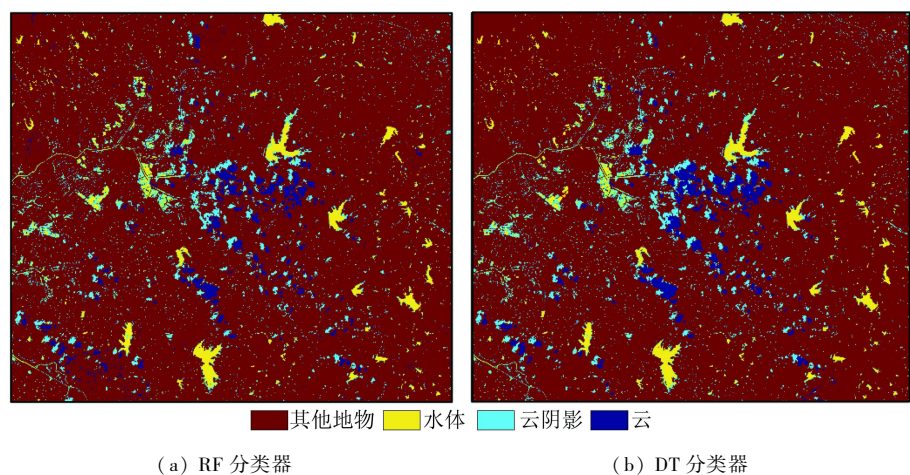


图6 由2种分类器得到的云-云阴影-水体掩模

Fig. 6 Cloud - shadow - water mask derived from two kinds of classifiers

对比图6和图1可以看出,2种算法均可将云和云阴影分离出来,范围较大的水体也能被区分出,总体上有3个特点:①2种算法把部分水体的边缘误分为云阴影,这是由于水体边缘部分是含有水草且颜色较浅的水体,其光谱特征与云阴影的光谱特征极为相似的缘故;②在图6(b)的左侧和下侧,DT分类器把部分其他地物误分为云;与图1对比可以看出,RF算法将其正确地分类为其他地物;③DT分类结果中被分类为云阴影的像元分布比较密集,这表明该方法把一部分非云阴影像元误分为云阴影;而这种现象在RF分类结果中相对较少。这表明了单一的DT及决策规则在处理光谱特征极为相似的像元时面临困难;而在RF分类器所得到的分类结果中,上述误分现象得到了明显改善。

从上述分析可以看出,训练样本的选择对算法的精度评价和掩模的提取有着重要影响。这种影响是由人为主观因素造成的,如对类别数的判断、不同类别典型光谱的提取、地物特征的增减等。但在同样容量的训练样本情况下,2种算法还是具有较明显区别的,RF算法要优于DT算法,这种优越性正是RF算法在理论上比DT算法先进的表现。RF是多棵DT的综合,是在训练样本随机抽取和DT随机训练的基础上对分类结果的综合,这种理论和机制决定了RF方法的优越性。

3 结论

1)遥感影像中云及云阴影的存在为数据的应用带来了较多的问题。如何更加有效地建立云及云阴影掩模,以便在遥感数据应用时把无用的像元排除在外是需要解决的一个重要问题。由于水体与云阴影具有极为相似的光谱曲线,利用DT分类器不

能较好地地区分云阴影与水体,故本文研究了利用RF算法构建云-云阴影-水体掩模。RF分类器是多个DT分类器结果的集成,可以较好地避免单一DT分类器带来的误分类现象。

2)相对于DT分类器,RF分类器能够更加有效地把云-云阴影-水体区分出来,建立相应的掩模。但是,由于浅水体与云阴影具有相似的光谱特征,部分深度不大的水体容易被误分为云阴影,这会给进一步应用带来一定问题,是今后需要深入研究的内容之一。

3)RF算法作为一种特征提取和分类方法已经在多个研究领域中得到成功应用。本文将RF算法用来建立云-云阴影-水体掩模,所得结果优于DT方法,但仍存在一些值得探讨的问题:①RF算法的鲁棒性,RF算法针对不同类型的云及其阴影建立掩模,其适用性和稳定性需要继续研究;②将遥感图像的DN值经定标及大气辐射校正后转换为反射率数据进行分析,其结果是否会优于本文结果;③本文算法是否可以推广到TM影像之外的其他类型遥感数据的掩模构建中。这些问题都有待今后进一步深入研究。

参考文献(References):

[1] Catalin - Dumitrescu C. A simple method of determining cloud - masks and cloud - shadow - masks from satellite imagery [J]. IEEE Geoscience and Remote Sensing Letters, 2014, 11(1): 10 - 13.

[2] 苏惠敏,薛亮. TM影像解译中去除云层及其阴影方法研究[J]. 西北大学学报:自然科学版, 2005, 35(6): 807 - 810. Su H M, Xue L. A study on the technology of removing clouds and their shadows during TM image interpretation [J]. Journal of Northwest University: Natural Science Edition, 2005, 35(6): 807 - 810.

[3] 李炳燮,马张宝,齐清文,等. Landsat TM遥感影像中厚云和阴影去除[J]. 遥感学报, 2010, 14(3): 534 - 545.

- Ri P, Ma Z B, Qi Q W, et al. Cloud and shadow removal from Landsat TM data [J]. *Journal of Remote Sensing*, 2010, 14 (3): 534 – 545.
- [4] 郑文武, 邓运员. 基于 LSMA 的 TM 影像云和阴影的检测与去除方法 [J]. *测绘科学*, 2010, 35 (3): 62 – 65.
- Zheng W W, Deng Y Y. Detecting and removing cloud and shade on TM imagery based on LSMA [J]. *Science of Surveying and Mapping*, 2010, 35 (3): 62 – 65.
- [5] 梁 栋, 孔 颖, 胡根生, 等. 基于支持向量机的遥感影像厚云及云阴影去除 [J]. *测绘学报*, 2012, 41 (2): 225 – 231, 238.
- Liang D, Kong J, Hu G S, et al. The removal of thick cloud and cloud shadow of remote sensing image based on support vector machine [J]. *Acta Geodaetica et Cartographica Sinica*, 2012, 41 (2): 225 – 231, 238.
- [6] Gabarda S, Cristóbal G. Cloud covering denoising through image fusion [J]. *Image and Vision Computing*, 2007, 25 (5): 523 – 530.
- [7] Maalouf A, Carré P, Augereau B, et al. A bandelet – based inpainting technique for clouds removal from remotely sensed images [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2009, 47 (7): 2363 – 2371.
- [8] Dybbroe A, Karlsson K G, Thoss A. AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling. Part I: Algorithm description [J]. *Journal of Applied Meteorology*, 2005, 44 (1): 39 – 54.
- [9] Dybbroe A, Karlsson K G, Thoss A. AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling. Part II: Tuning and validation [J]. *Journal of Applied Meteorology*, 2005, 44 (1): 55 – 71.
- [10] 申文明, 王文杰, 罗海江, 等. 基于决策树分类技术的遥感影像分类方法研究 [J]. *遥感技术与应用*, 2007, 22 (3): 333 – 338.
- Shen W M, Wang W J, Luo H J, et al. Classification methods of remote sensing image based on decision tree technologies [J]. *Remote Sensing Technology and Application*, 2007, 22 (3): 333 – 338.
- [11] 李 彤, 吴 骅. 采用决策树分类技术对北京市土地覆盖现状进行研究 [J]. *遥感技术与应用*, 2004, 19 (6): 485 – 487.
- Li T, Wu H. Application of decision tree classification to Peking land cover [J]. *Remote Sensing Technology and Application*, 2004, 19 (6): 485 – 487.
- [12] 齐红超, 祁 元, 徐 瑱. 基于 C5.0 决策树算法的西北干旱区土地覆盖分类研究——以甘肃省武威市为例 [J]. *遥感技术与应用*, 2009, 24 (5): 648 – 653.
- Qi H C, Qi Y, Xu Z. The study of the northwest arid zone land – cover classification based on C5.0 decision tree algorithm at Wu – wei City, Gansu Province [J]. *Remote Sensing Technology and Application*, 2009, 24 (5): 648 – 653.
- [13] 董金发, 刘庆生, 刘高焕, 等. 基于决策树分层分类法的粉煤灰堆场信息提取研究 [J]. *遥感技术与应用*, 2013, 28 (4): 576 – 581.
- Dong J F, Liu Q S, Liu G H, et al. Research on using the decision tree hierarchical classification method to extract fly ash information [J]. *Remote Sensing Technology and Application*, 2013, 28 (4): 576 – 581.
- [14] 陈静波, 刘顺喜, 汪承义, 等. 基于知识决策树的城市水体提取方法研究 [J]. *遥感信息*, 2013, 28 (1): 29 – 33, 37.
- Chen J B, Liu S X, Wang C Y, et al. Research on urban water body extraction using knowledge – based decision tree [J]. *Remote Sensing Information*, 2013, 28 (1): 29 – 33, 37.
- [15] 谢丽军, 张友静, 张子衡, 等. 结合 KPCA 和多尺度纹理的 IKONOS 遥感影像决策树分类 [J]. *遥感信息*, 2010 (3): 88 – 93.
- Xie L J, Zhang Y J, Zhang Z H, et al. The decision tree classification based on KPCA and multi – scale texture using IKONOS remote sensing image [J]. *Remote Sensing Information*, 2010 (3): 88 – 93.
- [16] 张晓娟, 杨英健, 盖利亚, 等. 基于 CART 决策树与最大似然比法的植被分类方法研究 [J]. *遥感信息*, 2010, 25 (2): 88 – 92.
- Zhang X J, Yang Y J, Gai L Y, et al. Research on vegetation classification method based on combined decision tree algorithm and maximum likelihood ratio [J]. *Remote Sensing Information*, 2010, 25 (2): 88 – 92.
- [17] Breiman L. Random forests [J]. *Machine Learning*, 2001, 45 (1): 5 – 32.
- [18] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest [J]. *BMC Medical Informatics and Decision Making*, 2011, 11: 51.
- [19] 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.
- [20] 姚登举, 杨 静, 詹晓娟. 基于随机森林的特征选择算法 [J]. *吉林大学学报: 工学版*, 2014, 44 (1): 137 – 141.
- Yao D J, Yang J, Zhan X J. Feature selection algorithm based on random forest [J]. *Journal of Jilin University: Engineering and Technology Edition*, 2014, 44 (1): 137 – 141.
- [21] 范士俊. 基于随机森林的全波形点云数据分类研究 [J]. *首都师范大学学报: 自然科学版*, 2013, 34 (5): 71 – 73, 78.
- Fan S J. Research on classification for airborne full waveform Li – DAR data based on random forest [J]. *Journal of Capital Normal University: Natural Science Edition*, 2013, 34 (5): 71 – 73, 78.

Construction of cloud – shadow – water mask based on Random Forests algorithm

LU Fengling^{1,2}, GONG Zaiwu²

(1. School of Geography and Remote Sensing, Nanjing University of Information Science and Technology, Nanjing 210044, China;
2. School of Economics and Management, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Clouds and their shadows in the remote sensing images are the key factors that influence the application of the data in many fields. Several methods, such as constructing cloud mask, replacement of the pixels, linear

mixture spectral analysis , and principal component analysis , have been proposed in the past decades to solve this problem. In this research , based on the analysis of spectral curve , the authors utilized Decision Tree (DT) classifier and Random Forest (RF) classifier to obtain the cloud - shadow - water mask. There was little difference between the spectral curve of shadow and water due to the mixture of shadow and other surface materials such as vegetation and impervious surface. In this case , the DT classifier could not effectively distinguish shadow and water because the decision rule and threshold were determined by analyzing the spectral curves of different samples . RF classifier was based on the ensemble of the results derived from multiple decision tree classifiers , which was more robust than one decision tree classifier. In this study , when there were only a few training samples , results that were more accurate were derived from RF classifier compared with the results from DT classifier . When the size of training samples lay in the range of 250 and 400 , no significant difference was found between the results derived from these two algorithms. This indicates that RF classifier could be used to deduce the cloud - cloud shadow - water mask successfully.

Keywords: mask ; remote sensing ; cloud - cloud shadow ; decision tree (DT) classifier ; random forests (RF) classifier

第一作者简介: 鹿丰玲(1973 -) ,女,硕士,讲师,主要从事公共气象与气象变化研究。Email: lfling63112@163.com。
(责任编辑: 邢 宇)