

基于随机森林的大姚县 TM 遥感影像分类研究

王 栋 岳彩荣 田传召 范怀刚 王跃辉

(西南林业大学林学院 云南 昆明 650224)

摘要: 随机森林(Random Forest)是一种组合多棵决策树分类器的新的分类算法。以楚雄州大姚县为例,采用 Landsat-TM 数据,通过最大似然、支持向量机、随机森林 3 种分类器进行分类对比研究。结果表明,支持向量机和随机森林的分类精度明显优于最大似然法,两者分类精度相差不大;在分类时间上,最大似然法明显比随机森林和支持向量机快,支持向量机最慢。综合分析,随机森林算法表现更优,它在保证分类精度的前提下,也能保证一定的时间效率,更适宜实际生产应用。

关键词: 随机森林; 遥感影像; 分类精度; 大姚县

中图分类号: S771.8 文献标识码: A 文章编号: 1671-3168(2014)04-0001-05

Classification of TM Remote Sensing Image Based on Random Forests of Dayao County

WANG Dong, YUE Cai-rong, TIAN Chuan-zhao, FAN Huai-gang, WANG Yue-hui

(Forestry College, Southwest Forestry University, Kunming 650224, China)

Abstract: Random Forest is a new classification algorithm, which is combined with a classifier of multiple decision trees. By using Landsat-TM data of Dayao County, Chuxiong Province, and taking three classification methods including maximum likelihood, support vector machine and random forest, the superiority of random forest classifier are analyzed. The results show that the classification precision of support vector machine and random forest classification is obviously superior to the maximum likelihood, and they have little difference in the precision. At the classified time, maximum likelihood is significantly faster than random forests and support vector machine, and support vector machine is the slowest one. According to comprehensive analysis, random forest method is the best one, it does not only ensure the classified precision, but also guarantee efficiency; it is more suitable for actual production applications.

Key words: random forests; remote sensing image; classification; precision assessment; Dayao County

遥感影像分类是遥感信息提取的重要手段,是目前遥感技术中的热点研究内容^[1]。遥感影像分类实际上就是将图像中每个像元点或每块区域根据其在不同波段的光谱特征、空间结构特征或其他信

息按照某种规则划分为不同的类别^[2]。因此,分类方法是遥感影像分类研究的重要内容之一。研究^[3-5]表明,没有哪一种分类方法或分类器能够在各种情况下都作为最佳分类器,选择和分类实施都

收稿日期: 2013-12-23.

基金项目: 国家自然科学基金(31260156); 亚太森林网络(APFNET/2011/PA004); 西南林业大学科技创新基金(1339)。

作者简介: 王 栋(1987-),男,黑龙江大庆人,硕士研究生。主要从事地理信息系统的应用研究。

通信作者: 岳彩荣(1964-),男,云南建水人,教授,博士生导师。从事遥感与 GIS 的教学与研究。

需要首先针对研究区域和研究目标进行大量试验和比较,在此基础上选择最有效的分类方法。遥感数据分类的精度决定于分类过程的各个环节,包括分类特征选择和提取、适当的数据先验知识以及合适的分类方法等^[5]。所以,选择合适的分类方法能够有效提高分类结果的精度。

目前的分类方法有很多,包括最小距离法、最大似然法等传统分类方法和支持向量机、神经网络和随机森林等机器学习分类方法。随机森林(Random Forests)算法是 Breiman 等人在 2001 年提出的一种基于分类与回归决策树(Classification And Regression Tree, CART)的组合算法。随机森林中的每一棵决策树依赖于一个由训练确定的参数组成的随机向量,通过为每棵树用 Bagging 方法生成有独立同分布的训练样本集并在它们上训练,森林中生成的树的参数随机向量也是独立同分布的^[6]。相比于其他分类器,随机森林在遥感影像分类中能够获得较好的精度和较快的速度^[7]。本研究以楚雄州大姚县为例,基于大姚县 2006 年 Landsat TM 数据和小班外业调查数据,分别采用最大似然、支持向量机、随机森林 3 种分类器进行分类研究,通过对比分析 3 种分类器的分类结果,研究随机森林分类算法的优越性。

1 研究方法简介

1.1 随机森林分类算法原理

随机森林(Random Forest, RF)是一类专门为决策树分类器设计的组合方法。它组合多棵决策树作出的预测,其中每棵树都是基于随机向量的一个独立集合的值产生的,如图 1 所示^[8]。

随机森林使用决策树装袋,首先通过随机从原训练数据 D 中有放回地选取 t 个样本(每次约有 $1/3$ 的样本未被抽中),生成原样本集的 $2/3$ 大小的训练样本集 $\{D_t, t=1, 2, \dots, t-1, t\}$,再针对每个训练样本集创建 t 个随机变量 $\{T_t, t=1, 2, \dots, t-1, t\}$ 控制所有决策树。于是,训练数据 D_t 和随机变量 T_t 生成第 t 棵决策树,相当于产生一个分类器 $h(X, T_t)$,其中 X 是一个输入变量。

由于生成决策树的时候,随机有放回抽取样本的过程是独立的,所以其中 $\{T_t, t=1, 2, \dots, t-1, t\}$ 是一组独立且具有相同分布的随机向量序列。也正是因为训练过程是相互独立的,所以通过并行方法生成每个决策树可以实现效率的提高。生成所有的决策树之后,用多数投票的方法对所有决策树的

分类结果进行综合,然后得出最终结果。

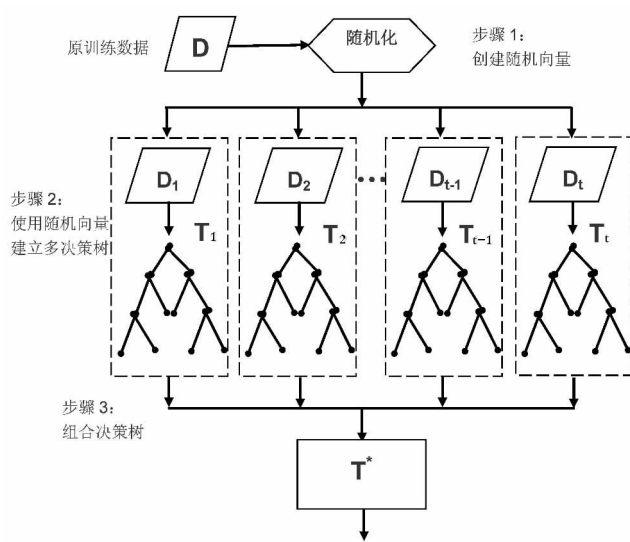


图 1 随机森林

Fig. 1 Random forests

1.2 支持向量机分类算法原理

支持向量机(Support Vector Machine, SVM)是由 Vapnik V 等^[9]提出来的建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的一种机器学习方法。支持向量机的基本思想是首先通过非线性变换将输入空间变换到一个高维空间,使样本线性可分;然后在线性可分的情况下求取最优分类面,而这种非线性变换是通过适当的内积实现的^[10]。基本理论的数学描述为:在条件

$$\begin{cases} y_i [(\omega \cdot x_i) + b] \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad (i=1, 2, \dots, l) \quad (1)$$

的约束下,求函数

$$\phi(\omega, \xi) = \frac{1}{2}(\omega \cdot \omega) + C(\sum_{i=1}^n \xi_i) \quad (2)$$

的极小值。通过求其对偶问题,归结为一个二次函数极值问题,即在约束条件

$$\begin{cases} \sum_{i=1}^l y_i a_i = 0 \\ 0 \leq a_i \leq c \end{cases} \quad (i=1, 2, \dots, l) \quad (3)$$

下求解函数

$$W(a) = \sum_{j=1}^l a_j - \frac{1}{2} \sum_{i,j=1}^l a_i a_j y_i y_j k(x_i - x_j)$$

最大值。

求解上述问题后得到的最优分类判别函数为

$$f(x) = \text{sgn}\left\{\sum_{i=1}^l a_i y_i k(x_i \cdot x) + b\right\} \quad (5)$$

支持向量机的关键在于选择恰当的核函数,采用不同的核函数将导致不同的算法。常用的核函数有线性核函数,多项式核函数,径向基核函数,Sigmoid核函数和复合核函数。研究^[11]表明,径向基核函数的分类精度最高。

1.3 最大似然法分类算法原理

最大似然分类法(Maximum Likelihood Classifier, MLC)是经常采用的监督分类方法之一,是通过求出每个像素对于各类别的似然度,把该像素分到似然度最大的类别中去的方法。最大似然法假定训练区地物的光谱特征和自然界大部分随机现象一样,近似服从正态分布,利用训练区可求出均值、方差以及协方差等特征参数,从而求出总体的先验概率密度函数^[12]。最后,根据函数对影像进行判别分类。

2 研究数据的获取与处理

2.1 研究数据的获取

本研究所用遥感影像数据为2006年11月18日Landsat 5卫星专题成像仪(Thematic Mapper, TM)获取的数据。Landsat 5卫星是美国于1984年3月1日发射的光学对地观测卫星,卫星所获得的影像是当前在全球应用最为广泛的卫星遥感信息源。其TM传感器有7个波段,其中TM1~TM5, TM7共6个波段的空间分辨率为30 m,远红外波段的TM6空间分辨率为120 m。研究所用的影像范围为25°3′42.63″N~26°3′27.50″N, 100°3′2.51″E~101°1′7.89″E(图2),辅助数据为森林资源调查小班数据。

采用的软件平台是ENVI 4.8和EnMAP-Box(Environmental Mapping and Analysis Program)插件。EnMAP-Box是一个采用IDL语言设计开发的包含随机森林算法和支持向量机的插件,用于对遥感影像的自动分类。

2.2 数据处理

首先,根据常用的分类系统与大姚县小班数据的实际情况,将土地覆盖利用分为6个类别——农地、针叶林、阔叶林、灌木林、建筑用地和水域。利用ArcGIS从小班数据中分别提取这6个类别的图斑。然后,对影像提取植被归一化指数NDVI和影像的6个波段合成,这样可以较好地地区分植被与非植被,有助于影像的分类。

然后,依据已提取出的小班数据,建立分类的训

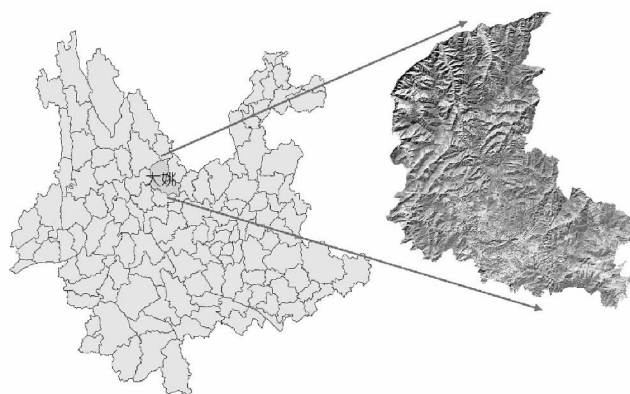


图2 研究区合成图像(R:5,G:4,B:3)的位置

Fig. 2 Composite image in the study area
(R: 5, G: 4, and B: 3)

练样本数据。建立过程主要靠目视解译,利用影像的Band 5、NDVI、Band 3合成RGB显示为模拟真彩色图像,可以更好地解译地物类别。由于每一类别在影像中的地物特征不同,根据地物显示的不同类型将每一类地物的兴趣区分为对应类别。农地对应分为农地粉、农地紫和农地绿3类;针叶林对应分为针叶林深绿和针叶林浅绿2类,阔叶林同针叶林一样分为2类;灌木林对应分为灌木林粉、灌木林绿;建筑用地为建筑用地深紫;水域为水域蓝。依据小班数据对应影像绘制多边形感兴趣区,最终建立训练样本(图3)。



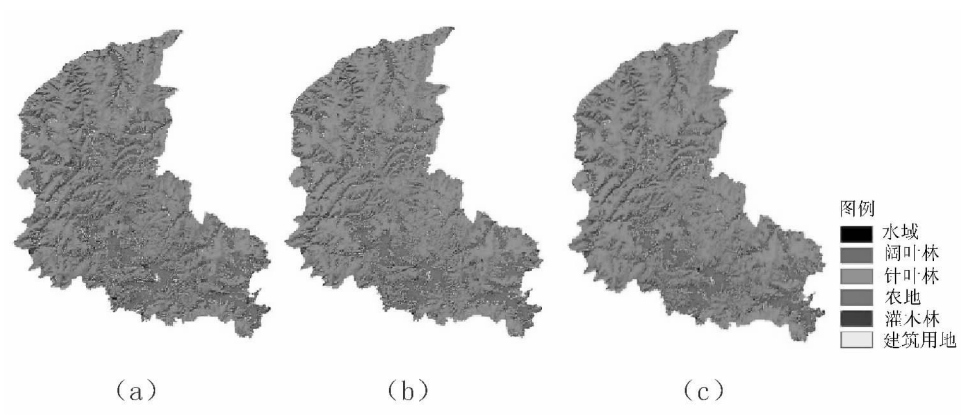
图3 训练样本

Fig. 3 Training sample

基于建立的训练样本分别采用最大似然法(MLC)、支持向量机(SVM)和随机森林(RF)分类器对所研究区的TM影像进行土地利用分类。最大似然法采用ENVI中的最大似然分类器对影像进行

分类,分类结果如图 4(a) 。支持向量机采用 En-
MAP – Box 插件对影像进行分类,首先,通过训练样
地进行核函数的 γ 和 C 值的参数寻优,结果分别为
0.1 和 10。然后,根据计算出的参数值对影像分类,
分类结果如图 4(b) 。随机森林同样采用 EnMAP –

Box 插件进行分类,其中只有分类树数目一个参数。
通过对结果的分析比较,当分类树数目为 100 时,分
类结果相对稳定,所以按照分析选择 100 进行分类,
分类结果如图 4(c) 。



(a) 最大似然分类器; (b) 支持向量机分类器; (c) 随机森林分类器

图 4 单个分类器的分类结果
Fig. 4 Classification of a single classifier

从图 4 中可以看出,各分类器的分类结果总体
差别不大,但是各个类别之间的分类差别较大,主要
是农地与灌木林、针叶林与阔叶林,以及城镇与灌木
林的区分方面。

3 结果与精度评价分析

基于之前提取出的各个地类的小班数据,随机
选取部分作为检验分类结果的真实样本数据,分别
对 3 种方法的分类结果建立混淆矩阵进行精度检
验,计算出 3 种分类器的总体精度和 Kappa 系数(表
1)。同时,统计 3 种分类器的分类结果,各地类面积,
用以评价 3 种分类器对各地类的分类效果(表 2)。

表 1 3 种分类器的总体分类精度和 Kappa 系数
Tab. 1 Three overall classification accuracy and Kappa
coefficient of classifier

分类方法	MLC	SVM	RF
分类精度 / %	66. 12	75. 69	75. 39
Kappa	0. 5665	0. 675	0. 6732

从表 1 中可以看出,整体分类结果,支持向量机
的分类总体精度和 Kappa 系数最高,支持向量机的
整体分类效果最好。随机森林的精度相对支持向量

表 2 3 种分类器分类结果的各地类面积
Tab. 2 Three classifier classification results of each type
of area m²

地类	MLC	SVM	RF
水域	11975400	11144700	11709900
阔叶林	595328400	375492600	426258900
针叶林	2041491600	2396144700	2347138800
农地	587979000	638614800	747175500
灌木林	734066100	582130800	476730000
建筑用地	57963600	25502400	19881000

机较低,但是仅相差 0.3%。最大似然在 3 种分类
方法中效果相对最差。从表 2 可以看出,水域的分
类结果相差不多,但是在图 4 中,3 种方法对于较大
的湖泊分类效果较好,而对于较窄河流的区分都不
是很好;在阔叶林和针叶林的区分上,随机森林的分
类效果表现相对更优,最大似然分类结果存在针叶
林漏分的现象;随机森林对于农地和灌木林的分类
效果相对更接近实际数值,支持向量机和最大似然
都存在农地错分为灌木林类别;最大似然的分类结
果在建筑用地上存在明显过分类现象,而随机森林
则存在漏分现象,支持向量机的分类结果相对较好。

对比各分类器的分类速度,随机森林在分类树

数目为 100 的情况下分类时间为 285 s, 而最大似然分类仅用了 14 s, 支持向量机分类时间最长, 为 1 972 s。最大似然分类虽然分类精度相对最差, 但是分类速度是最快的; 支持向量机分类的总体精度最高, 但是分类时间却是最大似然的约 141 倍; 随机森林的总体精度比最大似然的高 9.27%, 分类时间是最大似然的约 20 倍。

根据随机森林的定义可以得知, 分类树的数目对于分类结果的总体精度会有一定的影响。所以, 将随机森林分类器的分类树数目选择为 1, 测试分类时间约为 19 s, 同时分类精度下降到 71.67%。

为了研究其变化规律, 分别试验分类树数目从小到大的, 分类所用的时间及分类结果的总体精度如图 5 所示。

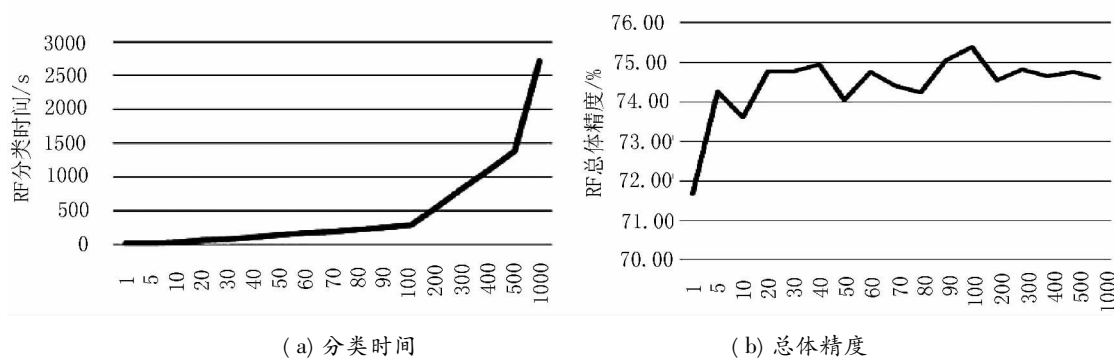


图 5 随机森林的分类时间和总体精度变化

Fig. 5 Classification of the random forest time and the overall accuracy

图 5(a) 中, 分类所用时间随着分类树数目的增加而增加; 图 5(b) 显示分类总体精度在分类树的数目为 100 时最高, 整体没有明显的规律。在数目小于 100 时, 分类总体精度很不稳定。而在超过 100 时, 分类时间大幅增加的同时, 精度却相对减少, 但变化幅度不大。所以, 从整体角度来看, 随机森林相对于最大似然法和支持向量机的分类效果较好。

4 结语

随机森林算法是目前比较流行的一种数据挖掘算法, 由于其性能表现较好, 被广泛应用于各个领域。研究表明, 最大似然法分类方法在分类速度上有较好优势, 但是精度相对较低; 支持向量机在分类速度上相对最慢, 但分类精度高; 随机森林综合表现优秀, 它在保证分类精度的前提下, 也能保证一定的时间效率, 更适宜实际生产应用。同时, 对于随机森林分类树的数目选择, 如何能在保证分类精度的同时提高分类速度, 还有待进一步地研究和探讨。

参考文献:

- [1] 贾坤, 李强子, 田亦陈, 等. 遥感影像分类方法研究进展[J]. 光谱学与光谱分析, 2011, 31(10): 2618-2623.
- [2] ZHAO Ying-shi, et al(赵英时, 等). The Principles and Methods for Analysis and Application of Remote Sensing(遥感应用分析原理与方法). Beijing: Science Press(北京: 科学出版社) 2003.
- [3] 陈绍杰, 单丹丹, 赵卫常. 土地覆盖遥感分类方法的比较与分析[J]. 辽宁工程技术大学学报: 自然科学版, 2010, 29(4): 567-570.
- [4] 史泽鹏, 马友华, 王玉佳, 等. 遥感影像土地利用/覆盖分类方法研究进展[J]. 中国农学通报, 2012, 28(12): 273-278.
- [5] 柏延臣, 王劲峰. 结合多分类器的遥感数据专题分类方法研究[J]. 遥感学报, 2005, 9(5): 555-563.
- [6] Breiman, L. Random Forests[J]. Machine learning, 2001, 45(1): 5-32.
- [7] 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.
- [8] Pang-Ning Tan, Michael Steinbach, Vipin Kumar(范明, 范宏建等译). Introduction to Data Mining(数据挖掘导论)[M]. 北京: 人民邮电出版社, 2006: 178-179.
- [9] V. Vapnik. The Nature of Statistical Learning Theory[M]. New York: Springer-Vedag, 1995.
- [10] 潘屹峰, 黄晶. 基于 SVM 和 Random Forest 的高光谱遥感影像分类[J]. 华东科技, 2012, 10: 27-29.
- [11] 谭琨, 杜培军. 基于支持向量机的高光谱影像分类[J]. 红外与毫米波学报, 2008, 27(2): 123-128.
- [12] 梅安新, 彭望琰, 秦其明, 等. 遥感导论[M]. 北京: 高等教育出版社, 2001: 198-199.