基于决策树的高光谱遥感影像分类方法研究

华 晔,张 涛,奚后玮,王玉斐,黄秀丽
(中国电力科学研究院,江苏 南京 210003)

摘 要:为了验证将决策树算法用于高光谱遥感影像分类的可行性,提出了一种二叉决策树自动构建算法用于高光谱遥感影像分类。通过对高光谱遥感影像进行现场采样、对样本进行统计和训练,生成了一棵二叉决策树,从决策树中提取出分类规则,并对高光谱遥感影像进行分类。生成的决策树简单明了,分类规则易于理解,分类效率和精度都比较高,实现了高光谱遥感影像从数据降维、样本选择、样本训练、决策树生成、影像分类的"一体化"和"自动化"。

关键词:二叉决策树;高光谱遥感影像;分类;最佳阈值;自动构建
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2012)06-0198-05

Research on Method of Hyperspectral Remote Sensing Image Classification Based on Decision Tree

HUA Ye, ZHANG Tao, XI Hou-wei, WANG Yu-fei, HUANG Xiu-li
(China Electric Power Research Institute, Nanjing 210003, China)

Abstract: In order to validate the feasibility of using decision tree algorithm for hyperspectral remote sensing image classification, it proposes a method of building decision tree automatically for hyperspectral remote sensing image classification. Based on hyperspectral remote sensing image on-site sampling, sample statistics and training, generate a binary decision tree, extract classification rule from the decision tree and classify the hyperspectral remote sensing image. The whole tree is simple and the classification rules are easy to understand. Both classification efficiency and accuracy are satisfactory. The study makes it "integration" and "automation" to reduce the dimensionality of hyperspectral data, sample selection, sample training, decision tree generation and image classification.

Key words: binary decision tree; hyperspectral remote sensing image; classification; best threshold; automatic building

0 引 言

高光谱遥感影像记录了地物目标的连续光谱,包含的信息更丰富,具备了识别更多种类的地物目标以及以更高的精度进行目标分类的能力。但高光谱遥感影像的巨大数据量,也会给信息提取造成一定的困难。

针对高光谱遥感影像的分类,很多传统的分类方法都显示出了它们的不足。如今,计算机技术和遥感技术正飞速发展,遥感影像的信息提取和土地利用、土地覆盖分类等研究已开始运用决策树分类技术[1]。通过各类研究与实验显示出:决策树分类算法较其他的遥感影像分类算法而言具有以下优势:分类精度高,生成的分类规则直观、易于理解[2]。决策树分类算法是从一组无序、没有规则的事例中推理生成一套分类规则,利用分类规则对遥感影像数据进行特征空间分割,结果简单明了[3,4]。其中二叉决策树结构直观,便于分析与理解,对输入数据空间特征和分类标志,有着很好的弹性和鲁棒性,构造起来简单、灵活,具有很好的分类效果[5]。

目前大部分利用决策树对遥感影像进行分类的研究之中,决策树的构建并不是完全自动的,一般情况下是在选取好训练样本之后,进行人工分析,再结合操作者自身的经验和影像的其他资料信息人工构建决策树,之后再利用其他程序、商业软件等根据从决策树中提取的分类规则对影像进行分类。当待分类的影像和分类条件等因素改变时使用这种方法,会使得分类的效率低下,并且不能充分利用影像信息以及地学知识等辅助信息。高光谱遥感是目前遥感技术研究的一个主要趋势之一,将决策树用于高光谱遥感影像的分类研究则更为少见。研究设计一个系统,在不对高光谱遥感影像进行预先降维的情况下,让决策树来选择对分类最有利的波段并构建分类规则,最后对影像进行分类,从而实现波段选择,分类规则的生成和影像分类的一体化是十分有必要的。这样在保持了高光谱影像光谱信息丰富的优势下,更是充分显示了决策树用于

收稿日期:2011-11-11;**修回日期**:2012-02-16
基金项目:国家电网科技项目(SG11075-1)
作者简介:华 晔(1985-),男,江苏南京人,硕士,助理工程师,主要研究方向为信息安全。

遥感影像分类的优势。

高光谱遥感影像中,光谱信息是最直接的信息源,利用光谱的统计信息进行分类,避免了目视判读过程中的主观性和低效的情况,将识别问题转化为光谱特征空间的定量求解,具有客观性和高效性等优势[6,7]。将决策树分类技术与高光谱遥感影像分类相结合,可以利用决策树的优势,充分挖掘高光谱数据中最有用的信息,对于研究高光谱遥感影像分类新方法,对于提高地物目标的识别能力和分类精度,都具有重大的理论和实际意义。针对目前遥感影像的决策树分类算法尚不完善以及将决策树算法用于高光谱遥感影像的研究成果较少,文中提出了一种基于决策树的高光谱遥感影像分类方法。

1　二叉决策树分类原理

二叉决策树是决策树中的一个简单形式。二叉树除叶子节点之外,每个节点仅有两个分支,即每个节点 N_i 都有且只有两个子节点 N_{il} 和 N_{ir}。使用二叉决策树分类器,能够把复杂的多类分类问题转化为多级多个两类分类问题。在每个节点 N_i 处,分类器都把样本集分为左和右两个子集,每一部分可能依旧包含着多类别的样本,需要把每一部分再划分成两个子集,以此类推,直到所分成的每一部分只包含同一个类别的样本,或某一类样本占优势而不可再分为止[8]。

二叉决策树的概念简单,生成的规则直观、便于理解和分析,在每个树节点之上可以根据需求使用不同的决策特征,采取不同的决策规则。二叉决策树的设计方式灵活多变,便于利用先验知识,十分适合于遥感影像的分类研究。

图1是一个二叉决策树的实例。在这个实例中,树的每一个节点上只选择一个特征,并提供了相应的决策阈值。对未知样本 X 而言,只需从根节点到叶节点,依次把 X 的某个特征值与相应的决策阈值做比较,即可做出决策,把 x 划分到对应的分支,最后分到合适的类别。

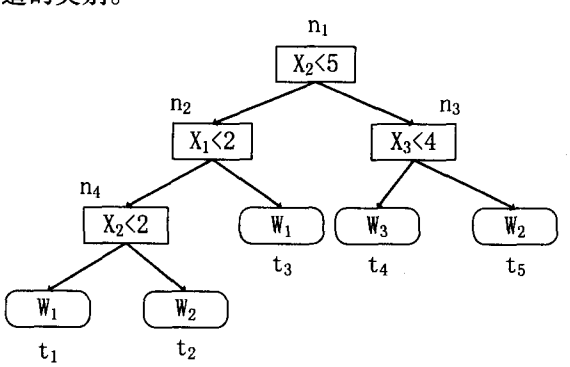

图1　二叉决策树的实例

将二叉决策树应用于遥感影像分类具有以下优势:

(1)生成的规则易于理解;

(2)整体计算过程并不是太过复杂;

(3)比较适用于处理遥感信息;

(4)直观,可以清楚地看出哪些属性较为重要;

(5)分类精度较高;

(6)二叉决策树也是一种较好的波段选择方法。

2　分类试验

2.1　试验数据

文中所用的试验数据为某地区的 OMIS 高光谱遥感影像,原始影像共有 128 个波段。由于传感器自身原因,65~96 波段的图像的信噪比很低、图像模糊,基本没有利用价值,可以去除以减小数据量;可见光 1~3 波段处受大气散射影响严重,影响图像使用价值,需去除;还需去除水汽吸收峰影响严重的波段和其它一些噪音严重,会影响分类的波段。文中根据上述要点,选择了其中的 45 个波段参与试验。将这些波段重新赋以波段 1~45 的序号,其中波段 1~9 为可见光波段,波段 10~31 为近红外波段,波段 32、波段 33 为热红外波段,波段 34~波段 45 为短波红外波段。

文中选取的试验区域影像大小为 340 * 512,根据目视判读和研究区域的实地调查图,确定了水体、农田、草地、房屋和道路五类具有代表性的地物。

2.2　试验原理

2.2.1　分类依据

在高光谱遥感影像中,各波段的灰度信息是最直接的信息源,因此文中将高光谱影像的各波段的灰度作为决策特征,决策树的构建过程也是波段选择的过程。在二叉决策树分类器的设计之中,需要考虑的关键问题是在各个分支上选取哪些决策特征分量,及如何选取各个分支处的决策阈值。通过对待分类地物进行采样、训练就可以获得所需要的先验知识。

文中在获取构建决策树所需的先验知识时,主要利用了高光谱遥感影像的光谱信息。通常情况下地物类别在一维特征空间满足正态分布的特征,根据类别在特征空间中的均值 μ 和标准差 σ 就可以确定每个类别在各波段处的特征分布区间。在构建决策树之前,通过对每个类别的地物进行灰度值的采样,分别计算出每类别在各波段上的均值和标准差,在插入阈值时,选择在所选类对的交点处插入。在正态分布的情况下,各波段上两两类别间的交点就是两个正态概率密度曲线的交点,其计算公式为;其中 μ_1、σ_1、μ_2、σ_2 分别表示两类地物的样本均值和标准差[9]。待判断点如果位于交点左侧,则将它归为左类,若位于交点右侧,则将他归为右类。错分概率可以用两条曲线所覆盖的公

共面积来表示，其中 x 为交点，F_1、F_2 分别为左、右两边曲线覆盖的面积函数，函数的两个参数表示覆盖区域的起止坐标。

在正态分布中，样本特征值落入 $[\mu-2\sigma,\mu+2\sigma]$、$[\mu-\sigma,\mu+\sigma]$ 范围的概率分别是 95.4%、68.3%[10]。两类地物间的可分离性可以以两类地物所对应的两条正态分布曲线的重叠程度来衡量，其主要包括以下三种情况：

(1)两类地物间可分性很好。如图 2 所示，将地物类别的概率密度曲线分为 A、B、C、D、E 五个区域。A 区域小于 $\mu-2\sigma$，B 区域为 $[\mu-2\sigma,\mu-\sigma]$，C 区域为 $[\mu-\sigma,\mu+\sigma]$，D 区域为 $[\mu+\sigma,\mu+2\sigma]$，E 区域大于 $\mu+2\sigma$。若两类地物的概率密度曲线的交点位于两类地物概率密度曲线的 A 或者 E 区域时，类间重叠较少，两类地物的可分性很好。

(2)两类地物间可分性一般。当两类地物的概率密度曲线的交点位于两类地物概率密度曲线的 B 或者 D 区域时，两类地物存在着轻度重叠现象，可分性一般。

(3)两类地物间可分性较差。如果两类地物的概率密度曲线交点位于 C 区间，则两类地物重叠现象严重，可分性较差。

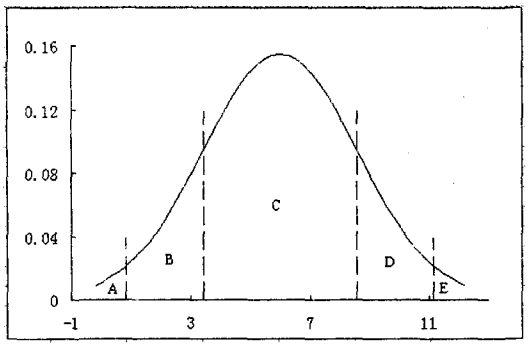

图 2 正态分布曲线区域划分示意图

2.2.2 决策树节点数据结构设计

数据结构设计在程序设计中显得至关重要，一个好的数据结构可以使算法更加精炼，从而提高开发效率[11]。决策树数据结构设计主要在于对决策树各个节点进行数据结构设计。有了足够的树节点信息才能构建合理的决策树，因此树节点的数据结构十分重要[12]。树的每一个节点信息都储存在自定义的结构体中。树节点的结构如下所示：

```
struct node                    //树节点结构
{
int number;                    //待分类的类别个数
int category[classNum];        //待分类类别编号
int mode;                      //决策特征量编号
float value;                   //判别阈值
int left;                      //左子节点指针
int right;                     //右子节点指针
int father;                    //父节点指针
};
```

2.2.3 分类算法设计

构建决策树之前，根据已有知识和经验确定存在的地物类别，确定训练区，对每个类别进行一定数量的样本采集，通过统计计算获得建树所需要的先验知识；在建立二叉决策树时，从根节点开始，对每一个节点判断其所含类别的可分性，确定在该节点处选用哪个波段作为决策特征以及决策阈值的大小；若可分，则完善该节点的信息。

●具体算法思路如下：

(1)在一个决策树的节点处，获得待区分类别数 number 和待区分类别的信息 category[classNum]；

(2)根据各个类别的样本像素灰度值进行统计计算，得出各类在各波段之中的像素灰度平均值及标准差，并在各波段上根据像素灰度平均值的大小对类别进行排序；

(3)依据公式得出各波段中相邻两个类别之间的交点，若交点符合标准，则将该交点作为当前候选阈值记录下来；

(4)在当前节点的所有候选阈值之中，根据最佳阈值选择的标准选择出最佳阈值，通过此阈值能够将当前的特征空间进行划分；

(5)以这个阈值为准当前节点可以生成两个子节点，特征值小于阈值的类别被左子节点包括，特征值大于阈值的类别被右子节点包括，同时统计子节点的 number 和 category[classNum]信息；

(6)重复上述步骤，根据节点信息，依次处理所有节点，完成决策树的建立。

●关键技术点如下：

(1)考虑两类地物正态分布曲线交点与当前节点处所有待分类别之间的关系，必须都满足类别可分的条件，才能插入有效的阈值。

(2)对所有的候选阈值，计算均值间标准距离，μ_1、μ_2 差距越大，类间距离越大；σ_1、σ_2 越小，则类内聚性越好，取 d 最大时所对应的阈值作为当前节点的判断规则。这样可以保证所选的阈值可以很好地区分两种类别，所用的波段的优势大于其他波段。

(3)规定每个类别只被一个叶子节点所包括，即在处理某一个子节点时，之前已经被区分出的类别就不再参与判决。这样可以有效地控制决策树的深度和节点的数量，使树的结构简单明了，决策规则易于理解。

文中在构建二叉决策树时，首先依据的标准是类

别间不重叠,遍历整棵树,若还存在包含了混合类别的叶子节点,则降低标准,允许类别的轻度重叠,继续构建决策树。决策树的建立过程即是由根节点开始,尽最大可能地将各个节点划分为两个子节点,直到没有合适的阈值或节点中待分类别数为一而导致不能再划分为止。

完成决策树的建树后,遍历整棵树,如果叶节点数等于类别数,则这棵决策树可以区分出所有的地物类别;如果叶节点数比类别数小,则这棵决策树没有将所有的类别区分出来。根据叶节点中的类别信息构建的叶节点所包含的类别与现实中的地物类别的对应关系,建立起分类判别规则。

试验时,先通过对地物类别进行现场采样,得到各类别在每个波段中的样本像素灰度信息,之后根据各类的样本像素灰度统计信息建立决策树。统计信息是类别的整体属性,不会受某个样本误差的影响,这是统计信息的可用保证。决策树算法在一个新节点处判断其可分性时,依据的是当前节点待分类别的样本像素灰度均值和标准差,并不是每个待分类别的所有样本[13]。这样做,可以同时兼顾算法的效率和分类的质量。

2.3　试验过程

分类所用的软件为基于 Microsoft Visual C++6.0 自主开发的决策树分类系统。试验操作过程如下:

(1)选取三个波段,分别作为 R、G、B 层做假彩色合成,这样会使影像呈现出鲜明的颜色,便于目视判读、选取地物的训练样本。文中分别将第 13 波段、第 5 波段和第 2 波段作为 R、G、B 层。

(2)根据自己的经验和知识,以试验区域的实地调查图为辅助,确定可将地物分为:水体、农田、草地、房屋、道路五类。

(3)根据试验的主要目的,直接使用高光谱影像的灰度信息指导影像分类。在影像上,分别对五类地物进行灰度样本采集,每个地物类别采集 192 个像素样本。通过计算得到样本的均值和标准差,从而得到每类地物在各波段上的特征分布,指导决策树的构建。

(4)从构建出来的决策树中提取分类规则,对高光谱遥感影像进行分类。

(5)对分类结果进行精度评价,精度评价使用分类正确率和 Kappa 系数两项指标。

2.4　试验结果

通过分类系统生成的二叉分类决策树如图 3 所示:

构建决策树的过程也是波段选择的过程。根据自动构建的决策树,可以看到只有波段 1、波段 5、波段 32、波段 45 四个波段参与了决策,也就是说在对高光谱遥感影像进行分类的时候只需要用到这四个波段,这四个波段对待区分地物的区分能力最强。提取出分类规则,运用这四个波段的数据进行影像分类,生成分类结果图像,既提高了计算的效率又保证了较高的分类精度。

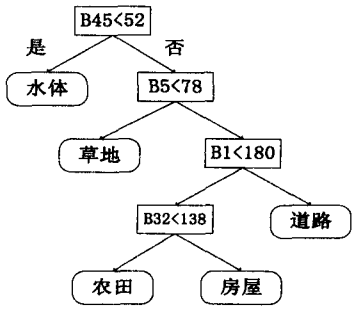

图 3　构建出来的决策树

根据研究区域实地调查图进行目视判读,为每个类别选取一定数目像素点作为评价样本,对分类结果进行精度评价。分类结果的混淆矩阵如表 1 所示。

表 1　决策树分类结果混淆矩阵

参考 / 分类	水体	农田	草地	房屋	道路	总计
水体	247	0	1	1	0	249
农田	0	219	4	13	10	246
草地	3	21	245	2	10	271
房屋	0	10	0	232	29	271
道路	0	0	0	2	211	213
总计	250	250	250	250	250	1250
正确率(%)	98.80	87.60	98.00	92.80	84.40	92.32
总的正确率 =92.32%,kappa = 0.9040						

从表中可以看出,文中提出的决策树算法用于高光谱遥感影像分类,总的分类精度达到 92.32%,Kappa 系数为 0.904;而采用最大似然法对影像分类,在不进行波段选择的情况下,分类精度为 79.04%,Kappa 系数为 0.7380,使用经过最佳指数因子法(OIF)选择的排名第一的波段组合进行最大似然法分类,分类精度为 80.64%,Kappa 系数为 0.7580。经过对比发现,文中提出的决策树分类算法具有较高的分类精度,并且运行时间明显小于最大似然法分类所需要的时间。

3　结束语

文中利用计算机自动构建出二叉决策树用于高光谱遥感影像分类,验证了将决策树算法应用于高光谱遥感影像分类的可行性;运用决策树分类法,实现了高光谱遥感影像分类中波段选择与影像分类的统一,实现了影像分类的"一体化"和"自动化"。经验证,文中提出的二叉决策树分类算法具有较高的分类精度和分类效率。

同时文中也存在着一些不足,在进一步研究中,主要可以从以下几点深入:

(1)文中在构建决策树以及分类时,使用到的信息只是高光谱遥感影像的光谱信息(即灰度)。在进一步研究中,可以加入其它信息,如:纹理特征、波段间相互关系等;要充分发挥决策树优势,加入地理信息、DEM 等高程的、属性的特征作为分类参考。

(2)高光谱影像所包含的信息非常丰富,具有识别更多地物、将地物大类再细化的能力。文中由于辅助资料的不足以及主要目的是探索是否可以将决策树算法应用于高光谱遥感影像所以未设定过多的类别。在今后的研究和实际应用之中,根据需要可以设定更多的类别,或将大类别继续细化以体现出高光谱遥感影像的优势。

(3)文中所包括的试验在构建决策树时,选用最佳阈值只考虑了当前的节点分类情况,不妨称为当前节点的最佳阈值(局部最优)。在进一步研究中,加入当前节点的所有子节点及孙节点的整体分类情况,研究其确定代价,选择整棵决策树的最佳阈值(全局最优);文中所在寻找最佳阈值时,仅用交点对应的两个类别在特征空间中的均值和标准差(μ_1、μ_2、σ_1、σ_2)的几何运算即作为评价标准。该标准只能体现该阈值对当前两个类别的影响,未考虑对其它待区分类别的影响。归一化的正态曲线在范围上,总的覆盖面积是 1。所以,在进一步研究中,可以考虑以阈值对应的所有类别的交叉面积值(即重叠区域面积)作为阈值性能评价标准。当前阈值对应的所有类别交叉面积总和越小,就能表明其越适合做最佳阈值。

(4)文中只用了 45 个波段参与试验,今后可以用包含波段数更多的影像进行试验以检验决策树波段选和分类的能力。

(5)可以同时考虑基于信息量和类间可分性的波段选择方法,优化决策树的决策规则设定,获取更高的分类精度和效率。

参考文献:

[1] 王大鹏,王周龙,崔青春,等. 基于决策树的龙口市土地利用/覆盖分类研究[J]. 水土保存研究,2007,8(4):250-252.

[2] 边肇祺. 模式识别[M]. 北京:清华大学出版社,1998.

[3] Sreerama K M,Steven S,Simon K. A system for induction of oblique decision trees[J]. Artificial Intelligence Research,1994(2):1-33.

[4] 李 爽,张二勋. 基于决策树的遥感影像分类方法研究[J]. 地域研究与开发,2003,2(1):17-21.

[5] Muchoney D,Borak J,Chi C,et al. Application of the MODIS global supervised classification model to vegetation and land cover mapping of central America[J]. International Journal of Remote Sensing,2000,21(6):1115-1138.

[6] 李 爽,丁圣彦,钱乐祥. 决策树分类法及其在土地覆盖分类中的应用[J]. 地感技术与应用,2002,2(1):6-11.

[7] 周成虎,骆剑承,杨晓梅,等. 遥感影像地学理解与分析[M]. 北京:科学出版社,1999.

[8] 闫培洁,于子凡,王勇军. 基于遥感影像光谱信息的二叉决策分类树自动生成方法研究[J]. 测绘科学,2009,34(6):184-186.

[9] 于子凡,林宗坚. 遥感影像分类的一种二叉决策树自动生成方法[J]. 测绘信息与工程,2006,31(4):42-44.

[10] 陈锡康. 经济数学方法与模型[M]. 北京:中国财政经济出版社,1982.

[11] 罗来平,宫辉力,赵文吉,等. 遥感图像决策树分类器研究与实现[J]. 遥感信息,2006(3):13-15.

[12] 于子凡. 面向地物目标的中分辨率遥感影像信息提取研究[D]. 武汉:武汉大学,2007.

[13] Patterson A,Niblett T. ACLS user manual[M]. Glasgow:Intelligent Terminals Ltd,1983:35-44.

(上接第 197 页)

[5] 杨德志,扈玉莲. 储油罐的纵向变位识别与罐容表标定[J]. 赤峰学院学报,2011,27(2):19-22.

[6] 郜欣春,王申重. 储油罐发生纵向倾斜时罐容表的重新标定[J]. 河南科学,2011,29(3):354-356.

[7] 伍人暾. 关于储油罐内液体数量的技术分析[J]. 广州化工,2011,39(7):144-146.

[8] 王妍玲,李 明. 椭圆形封头卧式贮罐液位与容积对应关系的建立[J]. 齐齐哈尔大学学报,2002,18(1):88-90.

[9] 任永良,高 胜,张瑞杰,等. 基于 GeoMedia 的油田注水系统建模软件设计[J]. 计算机技术与发展,2007,17(9):231-233.

[10] 肖 甫,王汝传,孙力娟. 多关联性虚拟现实系统的设计与实现[J]. 计算机技术与发展,2009,19(12):36-39.

[11] Teng Yingyan,Zheng Junsheng,Gao Zhijun. Design and Implementation of Interactive 3D Scenes Based on Virtools[C]//2009 International Forum on Computer Science-Technology and Applications. [s. l.]:[s. n.],2009:87-89.

[12] Li Jizu,Zhang Shaohong. Application of Virtual Reality Technologies to the Simulation of Coal Miners Safety Behaviors[C]//Open-source Software for Scientific Computation (OSSC). Guiyang,China:[s. n.],2009:60-62.

[13] Li Xunxiang,Li Anding,Chen Dingfang. Research on Distributed Multi-screen Display Technique Based on Virtools[C]//Proceedings of 7th International Conference on Computer-Aided Industrial Design and Conceptual Design. [s. l.]:IEEE Press,2006:756-761.