决策树方法在遥感地质填图中的应用

孙 赜[1],白志强[1],樊光明[2],施 彬[2]

1. 北京大学地空学院,北京 100871
2. 中国地质大学地球科学学院,湖北武汉 430074

摘要: 决策树理论在遥感分类中,分类准确、高效. 依据其理论方法,对青海省民和地区的遥感数据——ETM+(enhanced thematic mapper plus)进行了分类,选用的 ETM+数据为 1999 年 10 月份数据,数字高程(DEM)数据来自于 1∶25 万民和幅地形图,数据格式为 MapInfo 通用格式 MIF,数据进行了坐标转换(地理坐标),对原始数据进行了处理,从等高线中提取数字高程. 对遥感数据进行地形及光照矫正,计算植被因子及缨帽变换的 3 个分量,同其他 5 个遥感波段结合形成原始分类图层,同时确定目标分类结果. 原始数据的采样基于目视,首先采用不同的彩色合成方案突出不同的目标地物,交互式进行采样,使用 IDL 语言编制程序从原始数据中提取地物数字信息,使用 Clementine7. 2 对数据进行处理,其中 10%的采样数据验证模型准确率,其余数据用来推算模型,对数据进行 10 次迭代,同时给予 75%的剪枝,得到区分不同地物(如红层、黄土等)的最合适图层(band 1 & band 3)和具体数值,形成决策树模型,将决策树模型导入 Envi4. 0 中,对原始数据(9 个图层)进行计算形成初步分类结果图,对初步分类结果图进行一定的碎片合并,最终形成分类结果图. 该图同 1∶25 万地质图进行对比确认分类的效果,同传统分类图比较确认决策树分类方法优于传统分类. 另外来自于决策树所提取的信息,有利于地学知识的归纳总结.

关键词: 决策树;遥感解译;地质填图.

中图分类号: P623;P627 **文章编号**: 1000-2383(2004)06-0753-06 **收稿日期**:2004-08-28

Application of Decision Tree Method in Remote Sensing Geological Mapping

SUN Ze[1], BAI Zhi-qiang[1], FAN Guang-ming[2], SHI Bin[2]

1. *Faculty of Earth and Space Sciences, Peking University, Beijing* 100871, *China*
2. *Faculty of Earth Sciences, China University of Geosciences, Wuhan* 430074, *China*

Abstract: The decision tree is an effective and accurate method in remote sensing classification. We use this method to classify the remote sensing data—ETM+(enhanced thematic mapper plus), which covers most of the area of Minhe County, Qinghai Province. The acquisition date of ETM+ is October 29, 1999. We get digital elevation model (DEM) data from 1∶250 000 topography map of Minhe area. The format of DEM is MapInfo exchange format *. mif which converted to geography coordinate. After primary treatment of the raw data, the DEM data is derived from the contour line. The ETM+ scenes are rectified using the DEM and sun-illumination model. NDVI and other three indexes from Tasseled Cap transform were calculated from RS data. All these indexes are stacked with five RS bands. The target objects are selected. The sampling of the target object is based on visual observation. First, false color composed imagery is essential, and the sampling process is interactive. The digital information of the target object is derived from the program, which compiled by IDL. The decision tree model was calculated by Clementine 7. 2 software suite. About 10% raw data were used to validate the accuracy of the model. Meantime others were used to build the model. Ten iterative numbers and 75% trim radio are the suit parameters for this model. Then we get the most suitable layer and numerical value for distinguishing different target objects. For instance, distinguishing Tertiary red clastic and loess's best layer is band 1 and band 3. In the next step, we import the model to Envi 4. 0 and classify the raw data into different target objects. After some basic treatments, for example clump and

基金项目:中国地质调查局"民和县幅 1∶25 万数字地质填图项目"(No. 200213000016)和"数字填图过程、多元数据整合及成果表达方式研究"项目(No. 基[2003]009-02).

作者简介:孙赜(1974-),男,博士生,从事区域地质调查多元数据整合研究. E-mail: sun_ze@sina. com

assign class color, we get the final result map. The map is contrasted with 1∶250 000 geology map of Minhe area and achieved the accuracy of classification. The result is that the decision tree method is better than traditional classify methods. Another conclusion is that the rules from decision tree could help geologists to gain a appropriate geological conclusion.

Key words: decision tree; interpretation of remote sensing; geological mapping.

遥感 ETM+数据应用于地质填图的研究在国内还不是很多,多数停留在彩色波段合成等目视解译上,各种彩色合成图像用于地质研究的方案也很多,如 TM7(R) TM5(G) TM3(B),根据 Anderson (1994)研究成果,该波段组合反映了最多的土壤信息;TM4(R) TM5(G) TM7(B)波段用于构造地质的研究(Ricchetti, 2001),同时该波段组合可以用来检测土壤的含盐情况(Mulders and Epema,1986; Menenti *et al*.,1986);许多国外研究者也提出多种方法用于冰雪地带填图,Rott(1994)使用比值波段 3/5 的比值范围和波段 3,5 的取值区分冰和雪,Jacobs (1997)提出使用比值波段 4/5 区分冰帽. 这些研究方法多基于经验性的结果,部分可能考虑了不同离子对于不同波长光的吸收、反射性,显然我们需要一种新的研究方法从图像中提取有用的知识,而不是仅仅局限于经验性的知识,如何从图像中提取更多信息,如何进行成功的分类是自动化解译的难题. 在这一方面,国外进行了大量的研究,根据目前国内、国外的进展,神经网络和决策树成为分类的热点工具. 神经网络相对来说更为复杂,整个操作过程可以比喻成"黑箱",部分研究表明常规情况它的分类准确率略高于决策树,但是需要耗费大量的时间建立解译模型,需要使用者输入部分参数,在训练过度的情况下会造成很大的偏差;而在知识支持的条件下决策树容易取得更好的分类效果,更易于理解和推广,本文主要讨论决策树方法进行遥感图像的计算机分类.

决策树理论广泛应用于数据分类领域,决策树方法将复杂的决策形成过程归纳为简单的易于理解的规则(Rasoul and David, 1991). 在遥感分类中,特别是土地覆盖分类研究(Friedl and Brodley, 1997;Friedl and Strahler,1999;Swain and Hauska, 1997),分类准确、高效. 决策树优于传统的统计学分类方法,对数据并不要求正态分布,并且可以处理来自于遥感数据中的噪音和丢失数据(Safavin and Langrebe,1991)

可以反复使用的决策树模型对于填图具有相当重要的意义,方便非专业的技术人员迅速获得较为准确的分类结果,提取隐含知识,是一种有效的研究方法. 本研究的目的是探讨 C5.0 决策树方法和使用决策树方法实现简单的遥感图像分类,进而实现对决策树方法的试验和研究.

1 研究区现状

研究区经纬度范围为 102°~103°E,36°~37°N,位于祁连造山带东段,北邻中朝阿拉善地块,横跨祁连造山带各个单元,南接华南中秦岭地块,显示出地壳结构复杂、构造演化历史悠久的特点. 研究区地面植物覆盖较少,主要农业作物为小麦、青稞、油菜等,高寒草甸分布较广,部分地区有阔叶林;地面主要为黄土和第三纪红层(沉积岩)覆盖,沟底可见岩体、变质岩,遥感图像上可清晰分辨第四纪地貌各级阶地和河流沉积、冲积. 海拔范围大致在 2 000~4 500 m;测区气候干燥,天空少见云,拉脊山海拔高的地区常年积雪,植被因供水充分发育好,地面部分地区(兰州市)烟气污染严重、水土流失严重,并可见滑坡等自然灾害,该区域属于有利于遥感解译区.

2 C5.0 决策树方法

研究中使用的分类方法为 C5.0 决策树理论(Quinlan, 1993),具体的说明见文献,决策树方法如 C5.0 是将输入数据逐步细分,同类数据最终形成集合,同时产生对应的决策树或决策标准的分类方法.

C5.0 决策树的生成是由上而下的,它使用信息获取(information gain)技术来确定划分节点的属性,同时确定最佳阈值. 也就是说对于一个节点,有多个属性可以选择,比如植被指数、湿度等如何确定,使用哪一个为最佳? C5.0 用到了信息理论(Claude Shannon,1940),每个属性将被计算,获得最大信息获取的属性将成为节点划分依据.

Boosting 是 C5.0 使用的另外一项技术,同一模型的决策树并不唯一,对于已经生成的决策树,算法会更注重那些错分的和漏分的数据,在产生新的

决策树的时候,尽可能给予这些数据更多的注意,从而产生一个更准确的决策树,由此不断产生新的决策树,直到达到一定的标准,多数情况下极大地提高了决策树模型的准确性,在某些特定的情况也有准确性降低的例子(Bauer and Kohavi,1999).

3　数据准备和预处理

3.1　地形高程数据准备

数字高程数据来自于 1∶25 万民和幅地质资料,数据格式为 Mapinfo 通用格式 MIF,对该数据首先进行坐标系转换(原始数据为地理坐标系),对数据进行检查,使用 Envi4.0 读入数据,转变为 evf 矢量格式文件,从矢量线中提取数字高程信息,剔除坏点、应用最小距离法生成网格化高程数据.

3.2　遥感 ETM+数据

考虑到价格、数据的即时获取等问题,遥感数据我们选用 ETM+,ETM+数据的处理主要在 Envi4.0 环境下完成.

选用的遥感数据为 1999 年 ETM+数据编号为 131－35－991029,经过地面配准和反射率矫正,使用数字高程数据对 ETM+数据进行光照矫正(Ricchetti,2001),计算 NDVI(标准化植被指数)和 NDWI(标准化水指数)以及缨帽变换(tasseled cap)得到的 Brightness、Greeness、Wetness,使用图层合并工具,合并遥感数据、遥感数据得到的指数、数字地形数据、数字地形特点(坡度、平面),建立测区大小的文件.

参加分类的图层为 1,2,3,4,5,7 波段,6 波段为热红外波段,分类地面分辨率 60 m,故不参加分类,pan 波段地面分辨率 15 m 与其他波段波长有重叠也不参加,NDVI(植被标准化指数)、NDWI(水标准化指数)以及 Brightness、Greeness、Wetness.

4　分类树的建立和结果

目标分类结果确定为:本地生植被(native vegetation,深绿色)、农作物(浅绿色)、黄土(黄色)、红层(红色)、变质岩(黑色)、冰雪(白色)、水体(蓝色)、云层(淡蓝色).

首先使用常规的彩色合成手段,提取目标地物的波谱、高程等信息,不同彩色合成的图像用于区分不同的地物和提取热区(ROI),如西北地区的红层与黄土的区分是很小的,只有从第一和第三波段上才有显著的差别,应用 321(RGB)波段组合可以清晰区分二者之间的色彩差别,从该种图像上直接提取红层和黄土及清洁水的样区,变质岩在该波段组合上也能明显区分;而不同种类的植物从常规的波段组合中很难显示他们的差别,应用比值波段 TM 就可以区分不同的植被种类,使用交互式的方法就可以得到不同植被的热区,使用 IDL 语言编制程序从热区中提取数据,将数据转换为 SPSS 文件格式.

对于已经提取出来的样本,使用 Clementine7.0 数据挖掘工具中的 C5.0 算法,得出决策树,具体的参数使用 Boosting,迭代次数限制为 10(Freund,1996).得到的决策树如图 1 所示,图 1 中结果需要经过了一定的剪枝处理,否则得到的决策树会非常复杂和庞大,并非大的决策树就是理想的,因为很难保证训练样品中不存在噪音,C5.0 采用的方法是 error-based pruning 方法.根据来自 Clementine 的决策树结果可以在 Envi4.0 中建立决策树模型,赋予各个节点属性和取值,运行程序得到最终结果,对结果进行处理,合并小的碎片,赋予各个类别颜色,得到最终的分类图(图 2).

应该说在遥感计算机自动化分类方面决策树已经提供了一个十分有效的方法,图 3 是采用同样的分类标准进行的传统非监督分类的结果同真彩色合成图的比较,图中可以明显发现错分现象严重,地物也较难识别,红层、变质岩没显示出来,植物分区过大,该区域决策树的分类结果如图 4 所示,植被更准确,红层也得到区分,黄土分布更合理.尽管许多科学家采用非监督分类方法进行了遥感数据的分类,在一定程度上也取得了成功,但是成功进行传统的分类(监督和非监督)依赖于许多因素,特别需要遥感解译人员的经验和多次的实践,即便如此也很难同决策树的分类精度相比,而且传统的分类方法也只能使用 ETM+的 5 个波段.

同实际地质图进行比较,决策树方法分类的结果是基本准确的,植被、黄土、红层、积雪分类结果准确,当然也存在一定的漏分、错分现象.变质岩的分类结果不是十分理想,主要原因可能是变质岩的矿物成分种类繁多,因此单纯将变质岩做为一个大类是不合适的,应该划分得更细,具体的研究方法可以参考矿物的光谱特征,同时应当考虑到 ETM+数据的分辨能力.部分地区水体有较大的错分,原因有3

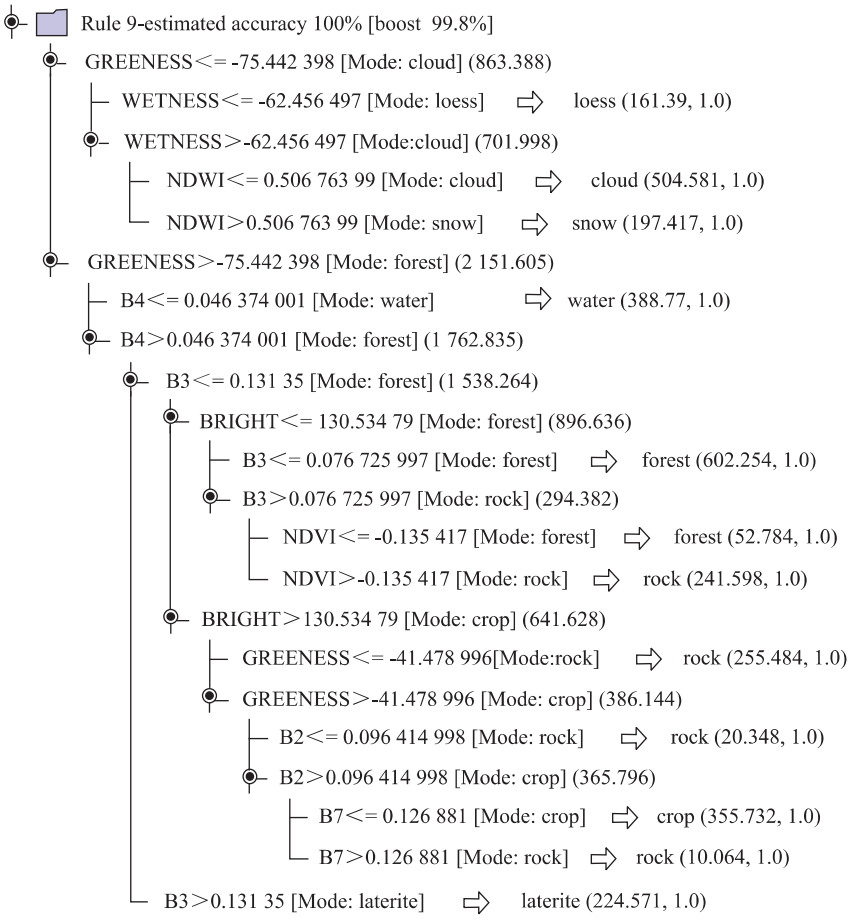

图 1 由样本数据建立的决策树模型

Fig. 1 Model of decision tree extracted from sampling data

点:一是该地区水体污染严重,水中含有大量的黄土、红土;二是部分地区土壤含水高;三是积雪、云层、烟雾和水体的光谱较为接近. ETM+数据的地面分辨率为 30 m,也就是说像元在各波段上的数值是地面地物综合的体现,有部分学者使用线性分离技术分离不同像元,Paul(2001)使用 least mean square methods 方法取得了较好的效果,本研究也试验该种方法,却并没有取得更好的效果,原因可能是地面的复杂性估计不足,仅仅将待分离地物划分为水、黄土、红土等是不够的. 在现有知识加入情况下会提高分类的精度,比如由数字高程所提取的坡度信息可以提升水体的分类精度,水体分布的地区坡度不会太大,数字高程也有利于积雪分类的准确性,积雪分布的地区海拔普遍偏高,并且在一定程度上受到坡向的影响. 地形、光照因素对于分类精度也有相当的影响,可能的条件下可以使用 DEM 数据对 ETM+各波段数据进行矫正,不过注意 DEM 数据地面的地面分辨率不应该小于 ETM+数据的地面分辨率. 值得强调的是 ETM+数据对于植物的研究具有相当的优势(植物在红色波段强吸收,在近红外波段强反射),有利于生态填图,科学家提出多种标准化植被因子,值得注意的是没有证据证明哪一种植被因子更好,具体情况可以在实际工作中调整.

分类精度的提高依赖于多方面的因素:(1)采样,所采样是否具有代表性,比如水体在深水和浅水区的地物光谱有着较大差异,采样的时候建议多点采样;(2)混合像元的问题,任何一点的光谱是方圆 30 m 范围内地物光谱总的反映,Envi 4.0 软件提供了 PPI 方法提纯像元;(3)数据预处理,在条件容许的条件下使用数字高程对遥感数据进行矫正可以提高最终的分类精度.

5 结论

应用决策树分类方法对民和地区进行了遥感自动分类,分类结果证明该方法优于传统的分类方法

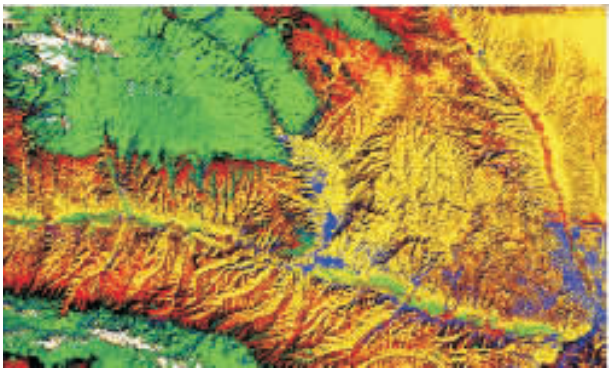

图 2　决策树分类结果

Fig. 2　Diagram of decision tree classification

(监督、非监督分类);由决策树提取了隐藏在大量数据中的规律,建立了区分不同地质体的分类标准;自动分类方法弥补了人工解译的不足,人工解译是完全依赖于解译人员的个人素质的,应用该分类方法为遥感解译提供了新的方法;在提高分类精度方面引入数字高程所提取的信息,为多源数据的应用提供了途径.

将数据挖掘等方法(C5.0 决策树)引入地质调查有利于多学科的综合应用,传统的遥感数据主要是人工解译,尽管目前还无法使用计算机进行自动解译,同时现在进行的这种分类还很简单,但是随着研究的深入相信会有很大的进展. 利用遥感所提供的几类信息:光谱信息、色彩信息、纹理信息,可以进行数字填图的各个方面研究. 美国已经利用航空遥感的光谱信息进行矿物填图(Clark and Swayze, 1995),目前由于价格因素获得这样的数据还有一定困难;同时纹理信息方面的研究也不断深入,特别是面向对象的遥感分类技术已经进入商业市场(如 Ecognition 3.0),充分利用了遥感影像所提供的信息,同时应用计算机的方法弥补人工识别的不足,为数字填图提供了更为高效的手段和强有力的工具.

决策树的方法,从大量的数据中也提取到了知识,比如从区分红层和黄土的属性看,决策树得到的结论是使用波段一、波段三,从光谱上看两者的差别,只有这 2 个波段才能将两者分离,并且给出了分离二者的数值. 在分离地物的研究中,许多研究者提供了不同的区分方法,多是基于经验性的,比如如何区分积雪、冰盖、云层、阴影等,使用决策树就可以从定量化的角度对这些地物进行分析和研究,直接得到区分不同地物最佳波段和取值,从而提供更为有效、快捷的研究工具和方法.

决策树的应用,是遥感自动分类的一项重要进展,有必要在未来的研究中更多应用该项技术,决策树算法也在不断的改进,总的来说决策树算法分为正交化的和斜交化的,C5.0 算法是正交化的,比如 $a_i > \alpha$ 等(a_i 是遥感波段、数字高程等的某项属性,α 是数值),有科学家提出斜交的决策树算法,如 $\sum_{i=1}^{d} w_i a_i + w_{d+1} > 0$,这里的 w_i 是参数,显然斜交算法更好一些,但是有证据表明斜交算法的边界取值是 NP-Hard 的(Heath *et al.*, 1993). 另外一种决策树

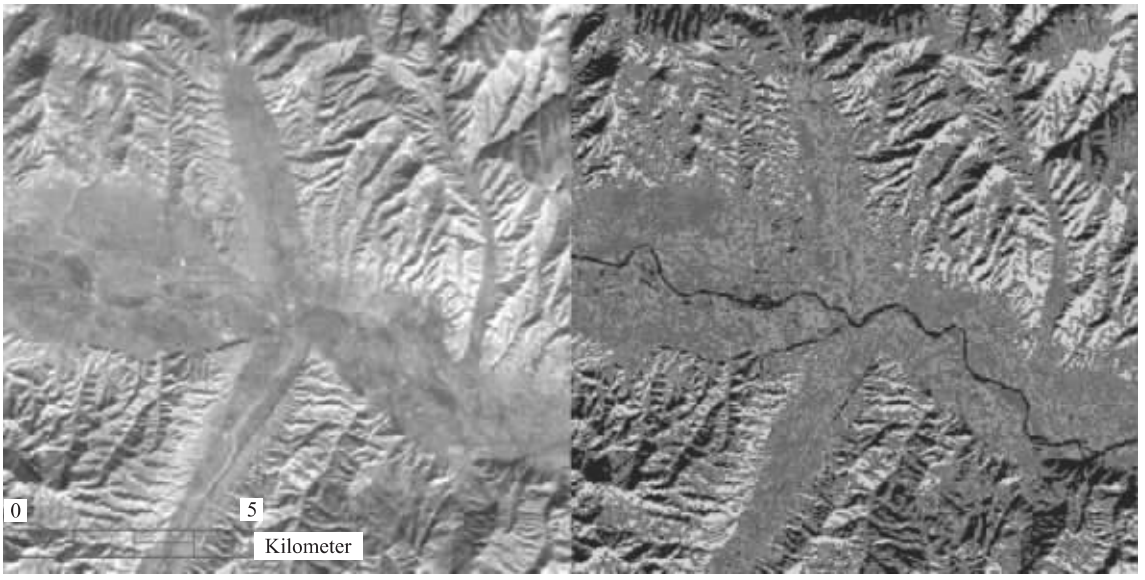

图 3　真彩色合成与非监督(IsoData)分类结果对比(左侧为真彩色合成图)

Fig. 3　Contrast between true color composite and IsoData unsupervised classification images

图 4 决策树分类

Fig. 4 Classification diagram of decision tree

是多变量决策树形式如$\sqrt{\sum_{i=1}^{d}(\delta_i - e_i)^2} \leqslant 0$，$e_i$ 是 a_i 在某节点处的取值. 有研究者提出了折中方案使用 Genetic and Artifical Life Environment(GALE)，据部分研究表明该模型具有更好的适应性和分类精度，在以后的研究中我们会给予更多重视.

致谢：特别感谢中国地质调查局提供的数据，感谢中国地质大学地调院张克信教授、朱云海教授在地质及地质填图技术方面的指导，感谢所有对本研究提供帮助的人们.

References

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36: 105－139.

Clark, R. N., Swayze, G. A., 1995. Mapping minerals, amorphous materials, environmental materials, vegetation, water, ice and snow, and other materials: The USGS tricorder algorithm. Summaries of the Fifth Annual JPL Airborne Earth Science Workshop, January 23－26, R. O. Green, Ed., JPL Publication 95－1, 39－40.

Friedl, M. A., Brodley, C. E., 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing Enviro.*, 61(3): 399－409.

Friedl, M. A., Strahler, C. E., 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Transactions on Geoscience Remote Sensing*, 37(2): 969－977.

Heath, D., Kasif, S., Salzberg, S., 1993. Learning oblique decision trees. In: Proc. 13th Int. Joint Conf. Artificial Intelligence, 1002－1007.

Menenti, M., Lorkeers, A., Vissers, M., 1986. An application of thematic mapper data in Tunisia. *ITC Journal*, (1): 35－42.

Mulders, M. A., Epema, G. F., 1986. The thematic mapper: A new tool for soil mapping in arid areas. *ITC Journal*, (1): 24－29.

Paul, L. R., 2001. Robust pixel unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 39(9): 1978－1983.

Quinlan, J. R., 1993. C4. 5: Programs for machine learning. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Rasoul, S. S., David, L., 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics*, 21(3): 660－674.

Ricchetti, E., 2001. Bisible-infrared and radar imagery fusion for geological application: A new approach using DEM and sun-illumination model. *International Journal of Remote Sensing*, 22(11): 2219－2230.

Safavin, S. R., Langrebe, D., 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics*, 21(3): 660－674.

Swain, P. H., Hauska, H., 1997. The decision tree classifier: Design and potential. *IEEE Trans. Geosci. Remote Sensing*, GE－15: 142－147.

Yoav, F., Robert, E. S., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119－139.