基于决策树的土壤质量等级研究

孙微微，胡月明，刘才兴，薛月菊
（华南农业大学 信息学院，广东 广州 510642）

摘要：将广东省土壤资源类型图和各种评价因子的单要素图层进行叠置分析，以土壤资源类型图的图斑作为土壤质量评价单元，从各生成图层的 PAT 文件中提取高程、坡度、土壤有机质含量、土壤质地、土壤 pH、土壤利用类型、地貌类型和土壤类型等评价属性，用决策树方法预测土壤质量等级，并以定量规则方式表达所获取的知识．结果表明，其知识表达易于理解，预测准确率为 96.61%．

关键词：空间数据挖掘；决策树；土壤质量；地理信息系统
中图分类号：TP181　　文献标识码：A　　文章编号：1001－411X（2005）03－0108－03

Soil quality grade evaluation based on decision tree

SUN Wei-wei，HU Yue-ming，LIU Cai-xing，XUE Yue-ju
（College of Information，South China Agric. Univ.，Guangzhou 510642，China）

Abstract：This paper overlayed the coverage of land-resource-type map with single-factor-coverages of each evaluation factor separately．The spots on land-resource-type map were studied as evaluation units for soil quality，attributes were extracted from the polygon attribute table（PAT）of each resulted coverage，such as elevation，slope，soil organic material content，soil texture，soil pH，land uses，landform type and soil type．Decision tree is used to predict soil quality grade，the knowledge are expressed in quantitative rules．The results showed that it was easy to understand the knowledge，and accuracy could be 96.61%．

Key words：spatial data mining；decision tree；soil quality；GIS

土壤质量的评价与监测是评价和重新设计可持续性土地利用系统的基础．传统的土壤质量地学统计方法需要大量的野外采样，给实际应用带来困难．近年来已有一些对于土壤性质的景观建模方法．这些方法需要领域专家的参与，对地理数据的统计不相关假设依赖较强，较难表达土壤性质和环境变量间的非线性关系．另外，这些方法以传统统计学为基础，难以处理字符型数据[1]．数据挖掘中的决策树方法是以实例为基础的归纳学习算法，在确定数据集之后，完全依赖数据本身学习模型．其优点在于不依赖领域知识，易于处理字符型属性数据，表达出的规则易于理解．本研究采用决策树方法，选取了高程、地面坡度、土壤有机质含量、土壤质地、土壤 pH、土壤利用类型、地貌类型和土壤类型等属性进行土壤质量等级预测，取得了较好效果．

1　材料与方法

1.1　数据来源

所研究的空间数据库主要包括 1∶500 000－1 000 000广东省，1∶10 000～50 000中山市、新兴县、高要市、台山市、增城市，1∶10 000东莞赤红壤试验区等各种图形资料，广东省第 2 次土壤普查、土地利用详查、广东省农业区划、广东省土壤资源数据库、广东省土壤资源分类与评价等统计资料．对于图形资料，利用扫描和图像处理等手段输入；对于非图形信息，整理和规范化后通过键盘输入．

1.2　方法

为获取属性数据，采用地理信息系统软件 ArcInfo 的多边形拓扑叠加功能，把广东省地貌类型图、土壤类型图、土地利用现状图进行叠置分析．对叠置过

收稿日期：2004－10－28　作者简介：孙微微（1971－），女，讲师，硕士．通讯作者：胡月明（1964－），男，教授，博士；E-mail：ymhu@scau.edu.cn

程中生成的大量细小多边形进行同质融合, 即将土壤资源类型相同的邻接图斑合并, 生成广东省土壤资源类型图, 将土壤资源类型图的图斑作为土壤质量评价单元[2].

将土壤资源类型图与海拔高度、坡度、土壤质地、土壤 pH、土壤有机质含量和土地利用类型等评价因子的单要素图层分别进行叠加, 得到一系列过渡图层. 在 Foxpro 数据库系统中, 对各过渡图层的 PAT 文件进行操作: 在不打乱原来的土壤资源类型图中的用户码的前提下, 按照过渡图层中多边形面积的大小分别提取各评价因素的属性数据. 以土壤资源类型图中的用户码为公共数据项, 从处理后的各单要素图层 PAT 文件中提取所需要的属性数据项进行合并汇总, 制成一个新表. 最后, 在 ArcInfo 中通过土壤资源类型图层 PAT 文件中用户码的联系, 将各评价因子属性数据的汇总表与土壤资源类型图的 PAT 文件合并或联接. 这样就获得了土壤资源类型图中各评价因子的属性指标值, 并与相应的图形数据连成为一个有机的整体. 共获得16 776条记录.

根据研究区土壤资源和数据源特点, 选取高程、地面坡度、土壤有机质含量、土壤质地、土壤 pH、土壤利用类型、地貌类型和土壤类型等属性作为土壤质量预测属性, 以土壤质量等级为分类属性.

所获取的属性数据大多为连续型数值, 过于细碎. 为了分析的方便和明晰, 将其表示为分级区段值, 如表 1 所示.

表 1　高程、坡度、土壤有机质含量、pH 值分级表

Tab. 1　Grade code for aspect, slope, organic material(OM), soil pH

高程 elevation/m	坡度 slope/(°)	w(土壤有机质 OM)/%	土壤 pH soil pH	分级 grade
＜200	＜3	＞4	6.5~7.5	1
200~500	3~6	3~4	5.5~6.5 或 7.5~8.5	2
500~800	6~15	2~3	8.5~9	3
800~1 000	15~25	1~2	4.5~5.5	4
＞1 000	＞25	0.6~1	＜4.5	5
		＜0.6		6

土壤质地、土壤利用类型大类、土壤利用类型小类、地貌类型、土壤类型等为离散型数据.

土壤质地分为: 轻壤, 砂壤, 中壤, 重壤, 轻粘, 中粘, 砂土. 分别用 1~7 表示.

土壤利用类型大类分为: 水域, 未利用地, 园地, 林地, 耕地, 牧草地, 建设用地.

土壤利用类型小类分为: 基塘, 疏林, 疏林地, 沙地, 针叶林, 热带作物园, 经济林, 工矿区, 裸露地, 竹林, 滩涂, 旱地, 水田, 草山草坡, 阔叶林, 果园, 防护林, 水库, 灌丛草地, 菜地, 盐场, 茶园, 城镇, 混交林.

地貌类型分为: 低山, 中山, 高丘, 台地, 低丘, 平原.

土壤类型分为: 火山灰土, 石灰土, 山地草甸土, 沼泽土, 砖红壤, 紫色土, 潮土, 滨海盐土, 水稻土, 石质土, 红壤, 黄壤, 滨海砂土, 酸性硫酸盐土, 赤红壤, 粗骨土.

分类属性为土壤质量等级, 采用文献[3]的土壤质量模糊变权评价方法对土壤质量评价单元进行质量分级, 按照优质到劣质分为Ⅰ~Ⅴ级和Ⅵ级, 以此质量分级学习决策树及检验预测准确率. Ⅰ~Ⅴ级适用于土壤, Ⅰ级绝大部分为耕地, 另有少量林地、园地和牧草地, 占研究区总面积的 14.62%; Ⅱ级占 24.17%; Ⅲ级占 43.64%; Ⅳ级占 16.63%; Ⅴ级全部为建设用地, 占 0.09%. 另外, 将 105 个水域(水库和滩涂)记录全部划为Ⅵ级, 占 0.84%.

2　结果与分析

采用由广东省土壤资源类型图及各评价因子单要素图层叠置分析获取的属性数据为实验数据集, 用 C4.5 决策树算法[4] 生成和测试决策树.

根结点处各预测属性的增益如下: 土地利用类型小类为 0.448, 土地利用类型大类为 0.433, 土壤 pH 为 0.332, 坡度为 0.224, 高程为 0.193, 土壤质地为 0.182, 土壤类型为 0.173, 地貌类型为 0.125, 土壤有机质含量为 0.077. 可见, 以土地利用类型划分各评价单元的土壤质量等级具有最小随机性, 是确定土壤质量等级的首选属性(根结点处). 而土壤有机质含量的 1~6 级在各级土壤质量子集中分布比较均匀, 所以用土壤有机质含量划分数据集时得到的子集概念纯度最低, 在确定土壤质量等级时不应首先选择该属性. 然而在为确定某个决策树内结点而重新计算剩余属性的增益时, 该属性又可能在剩余属性中增益最高, 因而成为决策树内部某分支的最佳属性.

为检验决策树算法 C4.5 对于土壤质量等级预测的有效性, 对16 776条记录采用十折交叉验证法, 取 90%记录作训练集, 另 10%作测试集, 训练集与测试集无交集. 迭代 10 次, 得到的决策树平均预测准确率为 96.61%. 例如, 其中 1 棵决策树叶结点数为 159, 按照这棵决策树可产生规则 159 条. 产生的规则表明, 首选等级为Ⅱ级, 这与土地实际比例相符.

选取涵盖样本数最多的部分规则如下(其中第 1、2 条规则是几条规则的合并规则): (1) IF 利用类

型小类＝′水库′ or ′滩涂′THEN Ⅵ级;(2) IF 利用类型小类＝′城镇′ or ′盐场′ or ′工矿区′ THEN Ⅴ级;(3) IF 利用类型大类＝′林地′ and pH＝2 and 坡度＝4 THEN Ⅳ级;(4) IF 利用类型大类＝′林地′ and pH＝2 and 坡度＝1 and 土壤质地＝4 and 有机质含量＝4 THEN Ⅲ级;(5) IF 利用类型大类＝′耕地′ and pH＝2 and 坡度＝1 and 土壤质地＝4 and 有机质含量＝4 THEN Ⅱ级;(6) IF 利用类型大类＝′耕地′ and pH＝3 and 坡度＝1 and 土壤质地＝4 and 有机质含量＝3 THEN Ⅰ级;

分析得到的规则发现,土地利用类型、土壤 pH、坡度、土壤质地和土壤有机质含量与土壤质量等级有较强的联系. 优质的Ⅰ、Ⅱ级土壤大多是坡度＜6°、pH 适中的耕地. 广东地区的土壤普遍有机质含量较低,质地较粘重,因此在Ⅰ、Ⅱ级土壤的土壤质地和有机质含量这 2 项指标中也反映出该土壤地域特性. Ⅱ、Ⅲ级土壤本身无本质差别,由于利用类型不同导致不同等级,这是由于模糊变权质量等级评价侧重于土地资源对于农业用地的适宜性. Ⅳ级基本上为坡度较陡的林地、牧草地或未利用地. 表 2 是Ⅰ～Ⅳ质量等级中评价属性的部分特征取值及次序.

表 2 Ⅰ～Ⅳ级中评价属性的特征取值及次序

Tab. 2 Typical attribute value and sequence for grade Ⅰ－Ⅳ

属性 attributes	Ⅰ	Ⅱ	Ⅲ	Ⅳ
高程 elevation	1	1, 2	1, 2	2, 1, 3
坡度 slope	1	1, 2	1, 3, 2	1, 4, 3
土壤有机质 soil organic materials	3, 4	4, 3, 2	3, 4, 2, 5	3, 4, 2, 1
土壤质地 soil texture	4, 5	4, 6, 5	4, 3, 2, 6	4, 2, 3, 5
土壤 pH soil pH	3	2, 3	2, 3	2, 3
土地利用类型 land use type	耕地	耕地, 林地	林地, 牧草地, 耕地	林地, 牧草地, 未利用地
地貌类型 landform	平原, 高丘, 台地	平原, 高丘, 台地	平原, 高丘, 低山	中山, 低山, 高丘
土壤类型 soiltype	水稻土, 赤红壤	水稻土, 赤红壤	赤红壤, 水稻土	红壤, 赤红壤, 水稻土, 黄壤

对同样的数据集,以文献[3]的模糊变权评价方法得到的评价结果为基准,应用常权指数和法[5]得到的评价结果准确率为 96.73%,指数积开方法[5]的准确率为 73.78%,灰关联综合评价法[6]的准确率为 84.29%. 由于模糊变权评价方法是用变权代替常权进行指数和得到评价值,所以常权指数和法准确率较高. 而决策树方法优于指数积开方法和灰关联综合评价法,与常权指数和法非常接近.

3 结语

决策树方法需要预先确定训练集中各评价单元的土壤质量等级,然后完全依赖数据学习得到决策树. 当评价属性集合变更时,传统评价方法需要由专家重新确定属性的权重和计算每一单元的评价值,而决策树方法只需重新学习一次树模型.

本文通过高程、坡度、土壤有机质含量、土壤质地、土壤 pH、土壤利用类型、地貌类型和土壤类型等属性,用决策树方法进行土壤质量等级预测,并用定量规则方式表达所获取的知识,预测准确率可达 96%,利于土壤资源数据的快速更新与质量评价,并为 GIS 数据的综合利用和空间数据挖掘提供了一种思路和方法.

参考文献:

[1] 周 斌,许红卫,王人潮. 基于分类树方法的土壤有机质空间制图研究[J]. 土壤学报,2003,40(6):801－808.

[2] 胡月明,欧阳村香,戴 军,等. 基于GIS 的土壤资源评价单元确定与属性数据获取方法初探[J]. 华南农业大学学报,1999,20(2):65－69.

[3] 胡月明,万洪富,吴志峰,等. 基于 GIS 的土壤质量模糊变权评价[J]. 土壤学报,2001,38(3):266－274.

[4] QUINLAN ROSS. C4.5 Program [EB/OL]. http://www.cse.unsw.edu.au/～quinlan/c4.5r8.tar.gz,2003－07－20.

[5] 胡月明,戴 军,王人潮,等. 基于 GIS 的浙江省红壤资源质量评价[J]. 华南农业大学学报,1999,20(4):80－85.

[6] 胡月明,吴谷丰,江 华,等. 基于 GIS 与灰关联综合评价模型的土壤质量评价[J]. 西北农林科技大学学报,2001,29(4):39－42.

【责任编辑 周志红】