

空间数据挖掘技术在土地定级估价中应用

贾泽露^{1,2}, 刘耀林², 张彤³

(1. 中南大学 地质与环境工程学院 湖南 长沙 410083; 2. 武汉大学 资源与环境科学学院, 湖北 武汉 430079; 3. 美国圣地亚哥州立大学 地理系, 圣地亚哥钟塔路 5500 号, CA92182 4493)

[摘要] 介绍了空间数据挖掘技术和决策树算法。通过对其研究, 将可视化空间数据挖掘技术应用于土地定级估价, 并介绍了基于 Visual C++ 6.0 和 ESRI 公司的 MapObject 2.0 组件技术设计和开发了一个可视化交互空间数据挖掘土地定级估价原型系统。系统采用决策树方法作为数据挖掘方法的基本算法, 采用训练与学习相结合实现土地定级估价。阐述了基于决策树空间数据挖掘土地定级估价的系统模型, 系统总体框架、主要模块、系统界面及系统实现定级估价的工作流程。该方法是对土地定级估价方法的一种新的探索, 是对土地信息系统开发的一种新的尝试, 也是土地信息系统智能化发展的一个方向。

[关键词] 空间数据挖掘; 决策树; 土地定级; 视图; 数据分类

[中图分类号] P208 [文献标识码] A [文章编号] 1672-6561(2005)03-0072-06

[作者简介] 贾泽露(1977-), 男, 湖北巴东人, 武汉大学博士研究生, 从事土地定级估价、土地规划及信息系统智能化研究。

空间数据挖掘是在大量空间数据中进行知识发现的技术, 是一个非常有发展前景的研究领域^[1]。决策树是数据挖掘中解决分类问题的主要方法之一, 在空间分类领域也有很多应用。

目前, 随着房地产和土地交易的日益活跃, 土地定级估价方面的工作日趋繁重。土地定级是一项综合性强的研究工作, 结果的评定不仅与许多自然因素有关, 还涉及经济、社会等多种因素, 工作中涉及大量数据和图件的分析与处理, 其计算量很大。对于土地信息, 各种资料可能会出现不完整的情况, 在处理缺失和错误数据方面, 传统的土地定级方法不能很好地解决问题。决策树算法擅长的数据挖掘任务就是混合数据的分类预测问题, 并且对于上述问题都有相应的处理策略, 而土地定级问题最终也是一个对数据处理以后的混合数据的分类预测问题。

综合以上原因, 可以考虑使用基于决策树的空间数据挖掘技术进行土地定级工作。

1 空间数据挖掘技术与土地定级估价

1.1 空间数据挖掘

空间数据挖掘(Spatial Data Mining, 简称 SDM), 也即空间知识发现(Knowledge Discovery in Spatial Databases, 简称 KDSD), 是指从空间数据库中抽取隐含的知识、空间关系或非显式地存储在空间数据库中的有意义的特征或模式^[2]。它是从数据挖掘技术中分支和发展起来的一项旨在使用最新计算机技术从大型空间数据库中发现未知的各种空间规律、关系、趋势等有助于人们进行更好的科学决策的各种空间知识。它可用于理解空间数据、发现空间联系、发现空间数据与非空间数据之间的关系, 构造空间知识库, 重组空间数据库, 优化空间查询和获取简明的总体特征等。SDM 通常认为是地理知识发现(geographic knowledge discovery, 简称 GKD)的一个阶段, 由于处理的对象是特殊的空间数据, 因此同一般的数据挖掘技术相比, 需要考虑到空间数据的各种特性。

(1) 地理空间数据不仅仅是在现有多维数据中简单的加入空间三维坐标, 同其他属性不同, 空间

[收稿日期] 2005 03 15

[基金项目] 教育部留学人员回国基金资助项目(152174); 国家自然科学基金项目(40271088)

各维数据是相互关联且不可以分开处理的。

(2) 由于时空模型的高度复杂性, 空间知识的提取更加困难。

(3) 由于一般用户对数据挖掘方法并不熟悉, 使用中往往不能正确选取合适的方法, 对于结果的解释也可能有所偏差。

对于以上问题, 可视化技术可以部分地加以改进和解决。因此, 空间数据挖掘技术通常与可视化技术结合使用, 以增强其解决问题的能力。

1.2 决策树算法

决策树是一种树状结构, 一般是对一组训练数据训练之后得到的结果, 其内结点作为按某一属性对数据集合的测试, 按照各数据记录对该属性值的不同分为多个分支, 最后的叶结点表示最终类别或者类别的分布。决策树方法的起源可以追溯到 20 世纪 60 年代 Hunt 等人在研究人类概念建模时建立的学习系统 CLS (Concept Learning System)^[3]。Quinlan 在 20 世纪 70 年代末提出了著名的 ID3 算法^[4], 是最早的决策树算法之一, 同时 Breiman 和 Friedman 等也研发出了类似于 ID3 的基于统计方法的 CART 算法^[5]。作为一种数据挖掘的基本算法, 决策树方法研究历经几十年, 已经产生了不少成熟有效的算法, 其算法稳健性好且计算复杂度不高、处理复杂的多维属性效果比较好、检验方法比较完善而可信、算法和结果易于理解、可以提供规则、结果便于可视化, 有利于用户进一步分析。由于决策树方法有其本身特有的优点, 决策树分类方法的应用领域比较广泛, 很早就应用于地理信息科学之中。

1.3 决策树的应用与土地定级估价

1.3.1 决策树用于可视化空间数据挖掘

决策树图形容易可视化, 同时用户对其原理和形式也容易理解, 所以空间数据挖掘中通过将决策树可视化, 并与地图动态地连接, 可以更生动有效地表达空间数据结构, 有利于用户深入进行空间数据分析, 提高人们的分析决策能力。Andriendo 等人将决策树算法 C4.5 和可视化交互图形结合起来动态地对空间数据进行分类^[6]。首先使用 Descartes 系统对空间目标分类分级, 然后将数据导入 Kepler 数据挖掘工具, 利用其中的 C4.5 算法生成决策并以图形的形式显示。通过决策树图形、规则与地图动态连接即树中的各结点、得到的规则与地图上的空间目标分类结果相关联。人们可以通过这种动

态交互式操作, 来分析空间目标的分类结果, 从而更加深入地了解分类分级问题的内在结构和意义。

1.3.2 改进决策树用于空间目标分类

Fayyad 等是较早使用决策树进行空间目标分类的研究人员, 他们使用决策树方法对星云图像分类^[7], 其缺点在于不能应用于矢量数据^[8]。Ester 等人在 ID3 算法^[5,7]的基础上使用邻域图来引入空间目标之间的关系进行分类^[9]。Koperski 等提出了一种基于决策树的两步分类方法^[8], 他们认为, 对 Ester 等人的方法没有考虑目标邻域的非空间信息聚合方面, 也未进行属性的相关分析, 因此分类效果会受到影响, 他们的方法不仅考虑了这些问题还引入了概念层次的概念以得到更简单的决策树, 运算速度也加快了。

1.3.3 决策树与土地定级估价

土地定级工作通常就是将影响土地质量的各类自然、社会和经济因素进行综合分析, 并对各类影响因素因子进行定性、定量化处理, 然后对各类因子影响分值进行叠加, 计算出各类因子对某一地块的综合影响分值。最后根据定量化计算指标, 参照一定的分类标准对定量化计算的分值影响数据进行分类分级, 以对土地质量的优劣和收益的高低进行评价, 并最终使评价结果等级化、具体化。土地估价工作则是在土地定级工作的基础上, 参照标准宗地的价格对于未知宗地价格进行的一个预测分类。土地定级估价的实质是对各类影响土地质量和收益的因素进行量化后的混合数据的一个动态分类分级问题, 其结果就是对于各类综合因素分值的分类分级结果。土地定级与估价的工作原理与决策树用于空间目标分类和预测分类的特点正好吻合, 因此, 利用空间数据挖掘的决策树算法进行土地定级估价工作在理论上是可行的, 并且与传统方法相比, 具有更强的理解空间数据、发现空间联系、发现空间数据与非空间数据之间关系的能力以及处理缺失和错误数据的独特优势。

2 一个基于决策树的土地估价模型

2.1 系统基本框架

系统的总体框架设计如图 1。系统通过图形用户界面进行人机交互, 计算机将空间源数据、数据挖掘中间结果和发现的高层次知识都以地图为主的各种图形表达出来, 用户接收这些信息之后通过

可视化思考过程决定继续的交互操作。这样的交互操作可能会改变当前的地图以及其他视图的表现形式,使得空间数据的内在规律变得更加明显,同时结合空间数据挖掘方法得到的一些知识,用户可以对所得的中间结果进行推理、验证、提炼直到满意为止。地图放在人机交互的中心位置可以充分利用地图作为空间信息存储、传输的中心作用,同时通过交互操作将地图动态化,使其进一步成为信息处理和认知的中心。其他视图(统计视图和图例等),也可以进行动态交互操作,即同样服务于可视化交互分析空间数据分析的目的,它们同地图一起从多角度表现当前数据的各个侧面,这样互相连接的各种图形就提供了较为完整地空间数据观察模式。地图还可以表达初步挖掘出来的各种知识,用户可以实时地了解数据挖掘的进展并及时做出调整,使得最后的结果满足自己特定的需要。将数据挖掘方法得到的模型根据其特点可视化,并同地图相连,这样用户可以随时捕捉模型被地图所放大的细微变化,一方面可以从不同模型的不同角度和具体细节来发现空间数据分布规律,另一方面可以比较各模型的优劣和不同。从而使划分级别的结果更加准确。

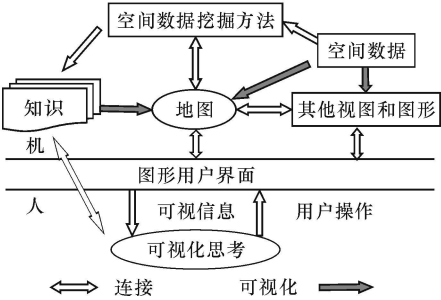


图 1 土地定级估价实验系统的总体框架

Fig. 1 Collectivity structure of land grading and evaluating

2.2 系统模块设计

系统模块设计如图 2。主要模块包括:数据连接和预处理、决策树数据挖掘算法和可视化数据处理。数据连接和预处理模块主要完成因素的选取分析、因素作用分值的量化处理、定级估价基础资料的预处理和宗地样本抽取等;决策树数据挖掘算法是在数据连接和预处理的基础上,对处理的中间结果数据通过训练生成基本决策树,并对决策树不断修剪改良,在此基础上,通过标准宗地库信息并结合规则库中的规则进行初步的定级和估价,对于定级估价的结果通过用户的先验知识及传统定级

估价方法的比较进一步进行调整。将最终认为满意的定级估价结果属性数据与 GIS 图形数据库相连接通过可视化技术输出定级估价成果图。

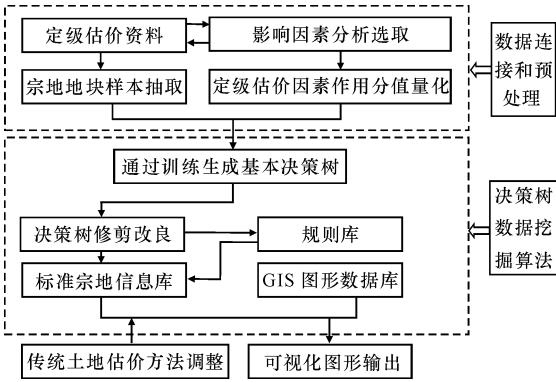


图 2 基于决策树的土地定级估价系统模块

Fig. 2 Models of system of land grading and evaluating based decision tree

2.3 系统界面设计

这里指的界面并不局限于程序的外观界面,而是包括了所有用户同计算机系统交互的工具,Howard 等总结了 5 种交互手段,即:命令行、自然语言、表单输入、菜单和直接交互等。根据当前计算机软件的主要界面特征,系统提供了表单输入、菜单和直接交互等几种形式。界面形式如图 3(定级因子预处理图)。图 3 左边是表单输入的交互形式,为定级估价训练提供必要的属性信息。这种形式可以让用户更为精确地将已有信息和自己的想法告诉计算机,同时计算机也以一定形式在表单中通知用户当前系统的各种状态。它在进行数据探索分析的开始时期为用户提供一个明确、清晰和快速的信息交互管道。图 3 右边是一个典型的菜单形式,点选某项菜单得到某种功能,其好处是不容易出错,明了直观。此外,此处所指的菜单交互不仅仅是普通意义上的下拉菜单或者快捷菜单,也包

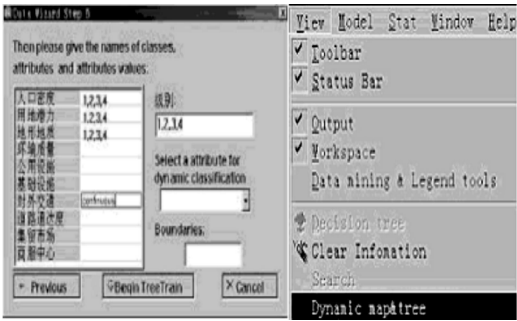


图 3 表单交互、菜单交互

Fig. 3 Alternation of table and menu

括系统提供的比较固定的控制和显示工具, 如工具条、树型菜单等, 它们往往提供了最常用的功能, 同时也是最基本的空间数据分析工具。

系统运行的主界面如图 4, 界面主要由 5 部分构成, 包括菜单区、主控制区、地图视图、辅助视图和信息输出等。另外还有数据视图和统计视图等几个额外的视图。菜单区指的是系统提供的包括下拉菜单和工具条在内的基本空间数据分析工具, 主要有对模型及其参数的选取控制、地图和其他各种视图控制观察工具等。地图视图放在整个界面的中间, 突出体现地图作为信息传输中心的地位。体现了以地图为中心, 其他视图和信息窗口辅助进行交互空间数据挖掘的思想。各个部分可以浮动, 关闭和并列比较, 更加方便了用户的可视化思考。

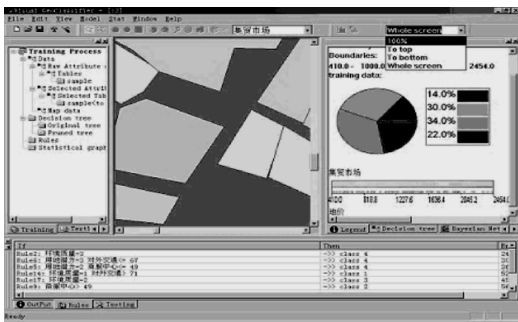


图 4 系统运行界面

Fig. 4 Main surface of the system

2. 4 系统实现环境

系统运行和实现环境是 Windows 2000 professional 版, 编程语言环境是 Visual C++ 6. 0, 程序调试和运行在微机单机上进行。空间数据的属性部分使用 ADO 数据库访问技术来连接处理, 图形部分使用 ESRI 公司的 MapObject 2. 0 组件管理。属性部分和图形部分分开处理可以加快数据连接和访问的速度, 同时也可以避免同时访问同一数据的错误发生。数据挖掘的算法, 选用 C4. 5 决策树算法, 这是目前机器学习和数据挖掘领域使用的最为广泛的算法之一。系统中除地图视图使用 MO 引擎外, 其他各视图的可视化交互形式都在 VC++ 中直接从底层实现。

3 系统工作流程

首先, 通过一个数据连接和预处理模块来从空间数据库中交互选取定级估价的数据集、属性和地图(因素因子图、定级单元图等), 并对其中一些数

据进行处理, 包括对缺失数据和不确定数据的处理、数据压缩、转换等。然后, 将处理好的数据在地图和其他各种视图中显示, 同时导入到决策树训练模块进行决策树的学习过程, 其训练结果包括剪裁前后的决策树以及一些分类规则。这些中间结果可以实时可视化并同地图等视图相连, 用户可以从进一步分析训练数据集的内在特征并考虑对训练数据集的修改或者对数据挖掘参数、先验知识、处理结果进行修改来提高整个学习训练过程的效果。当用户认为当前的分类规则满足一定的需要后, 可以同样从空间数据库中选取测试数据集, 按照分类规则进行测试, 测试结果可以同训练数据在地图中一起显示, 从而可以比较 2 个数据集的空间分布和分类结果。完成测试过程后, 用户就可以对其他未知数据进行交互分类。3 个过程之间和不同视图之间可以随时切换, 形成一种循环渐进的知识发现过程。系统实现定级估价过程, 实际上是通过决策树算法对数据进行训练、分析、测试, 最后达到一个符合要求的数据分类, 分类的结果就是满足要求的各土地级别及其对应的各级别的土地价格。因此, 对土地进行定级估价, 在此实际上简化为一个普通的分类分级问题。系统实现定级估价的大致经历步骤过程是:

3. 1 数据准备

导入定级估价所需的属性数据(Access 2000 的. MDB 文件格式) 和空间数据(Shape 格式), 属性数据和空间数据分开处理。为训练程序提供必要的信息, 如参加训练的属性名称、属性是否连续值、离散属性的可能数值、需要预测的类别属性及类别信息等等, 这些通过一个数据连接和预处理向导来完成。图 5 就是数据向导中的一步, 即选定数据库中的某个表及参加训练的属性(定级因子) 名称。类别属性为连续值时, 需要让用户指定分类分级的界限。

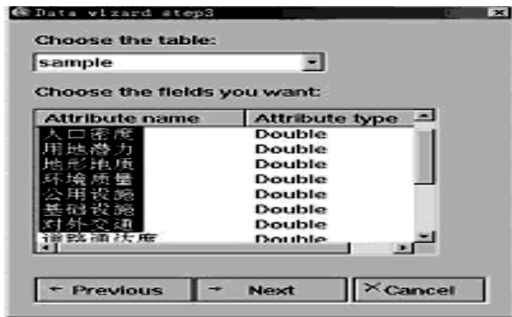


图 5 数据向导

Fig. 5 Guide surface of data

3.2 初始训练

通过决策树算法进行初始训练, 得到必要地训练信息后, 设定训练参数。如图 6, 包括树的训练参数和规则的训练参数。指定参数后开始决策树训练, 当前决策树是对训练数据的拟合, 同时按照剪裁参数对原始树进行剪裁 (CF 值越大则剪裁越少) 得到剪裁后的树。



图 6 决策树定级训练参数设定

Fig. 6 Parameter of grading based decision tree training

3.3 数据分析

得到初步的决策树结果后, 可以得到剪裁前后的决策树、分类规则。同时指定各类别 (因子级别) 的颜色, 同类别 (同一因子同一级别) 的颜色在可视化环境下是统一的, 便于各连接视图比较分析。在继续进行分类过程之前, 对目前得到的决策树进行可视化分析。此时, 可以对参加训练的属性包括训练参数进行调整, 以得到更加满意的分类规则。

3.4 数据测试

在主控制区中选中测试 (testing) 标签视图, 开始测试过程。测试数据直接按照分类规则进行分类, 然后同已知类别结果比较。文本结果输出在信息输出视图区的测试信息标签视图 (testing) 中, 地图结果则显示在地图视图中。此阶段可以使用各种可视化交互技术对测试数据和测试结果进行分析, 也可以返回到训练阶段重新开始训练或者对训练数据进行操作。

3.5 数据分类

测试结果满意后, 就可以进行数据的预测分类 (初步定级)。采用交互分类的方法, 即将待分类数据导入地图, 使用鼠标在地图上点选需要预测的空间目标来实时得到分类的结果。当结果不满意的时候, 修改数据参数并进一步训练测试直到结果满意为止, 将其结果可视化输出即为所求级别及级别对应的土地价格。

以上 3.2~3.5 过程可以根据需要循环操作,

直到定级估价结果使用户满意为止。图 7 即为模拟实验数据进行初步分类定级的结果显示。

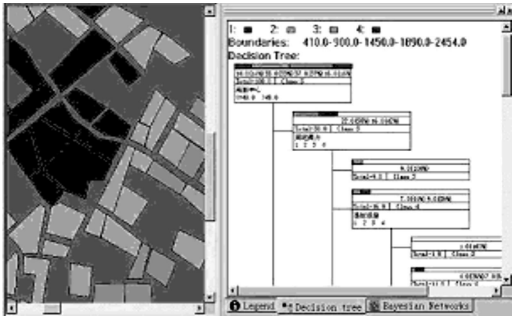


图 7 初步定级结果及决策树分析显示

Fig. 7 Display of the analysis based decision tree and the primary result of land grading

4 结语

可视化交互空间数据挖掘技术应用于土地定级估价领域与传统的土地定级估价相比有如下优点:

(1) 整个定级估价过程是一个循环动态、以用户为中心、知识不断得到提炼深化的过程, 用户可以在可视化交互环境下, 在各个过程中和各视图之间切换观察并进行交互操作, 还可以具体到单个空间目标的操作, 做到了从宏观到微观, 纵向与横向的结合, 增强了辅助决策和空间分析能力。

(2) 地图作为系统中所有视图连接的中心, 用户可以参考地图的显示情况进行可视化空间思考。对视图内部的数据采用聚焦技术加以分析。分析时可以将内容局限于某一小部分, 如分析某类别的数据时不显示或者低调显示其他数据, 便于排除干扰, 集中探索分析有用和感兴趣的内容, 同时可以将数据的细微差别以地图的形式加以放大分析, 增强了数据分析的效率。

(3) 结合人敏锐的观察能力和可能的用户专业知识, 使定级估价过程成为一个互动、可视化、易于理解的重复过程, 而不是完全自动的暗箱操作。

(4) 对于土地信息各种资料不完整的情况下, 增强了处理缺失和错误数据方面的能力。

综上所述, 将可视化交互空间数据挖掘技术应用于土地定级估价, 是对土地定级估价方法的一种新的探索, 是土地信息系统开发的一种新的尝试, 也是未来土地信息系统智能化发展的一个方向, 是进行科学的土地定级估价及其成果进行科学管

理的一条有效途径。此技术不仅具有空间数据提取、表达、分析和可视化的处理能力,也具有辅助空间决策、知识表示和推理的能力,从而有助于解决土地定级估价中的一些半结构和非结构的量化问题,这将对土地定级估价的科学性、合理性及智能化程度做出贡献。今后可以将更多的空间数据挖掘算法如贝叶斯网络等相结合加入到系统中,进一步增强系统综合解决问题的能力,还有很多问题有待于科学工作者更加深入的研究。

[参 考 文 献]

[1] 吴金华,祝国瑞.空间数据仓库的认知过程[J].地球科学与环境学报,2004,26(4):67~71.
[2] Koperski K, han J. Discovery of spatial association rules in geographic information databases[A]. In: Advances in Spatial Databases[C]. Springer Verlag, Berlin, 1995. 47~66.
[3] Hunt E B, Martin J, Stone P J, et al. Experiments in induction[M]. New York: Academic Press, 1966.
[4] Quinlan J R. Induction of decision trees[J]. Machine Learn-

ing, 1986, (1): 81~106.
[5] Breiman L, Friedman J H, Olshen R A. Classification and regression trees[M]. Belmont: Wadsworth International, 1984. 152~158.
[6] Andrienko G L, Andrienko N V. Data mining with C4.5 and interactive cartographic visualization[A]. In: Paton N W, Griffiths T. User Interfaces to Data Intensive Systems[C]. Los Alamitos CA: IEEE Computer Society, 1999. 162~165.
[7] Fayyad U M, Djorgovski S G, Weir N. Automating the analysis and cataloging of sky surveys[A]. In: Fayyad U M, Piattetsky Shapino G, Smyth P, Ullthurusamy R. Advances in Knowledge Discovery and Data Mining[C]. Menlo Park, CA: AAAI / MIT Press, 1996. 88~97.
[8] Koperski K, Han J, Stefanovic N. An efficient two step method for classification of spatial data[A]. In: Proc. International Symposium on Spatial Data Handling SDH' 98[C]. Vancouver, BC, Canada, 1998. 45~54.
[9] Ester M, Kriegel H P, Sander J. Spatial data mining: a database approach[A]. In: Proc 5th Int Symp on Large Spatial Databases (SSD' 97), Lecture Notes in Computer Science [C]. Berlin, Germany, 1997. 47~68.

Land grading and evaluating using spatial data mining

JIA Ze lu^{1,2}, LIU Yao lin², ZHANG Tong³

(1. School of Geology and Environment Engineering, Central South University, Changsha 410083, China;
2. School of Resource and Environment Science, Wuhan University, Wuhan 430079, China; 3. Department of
Geography San Diego State University, 5500 Campanile Drive San Diego, CA 92182-4493, USA)

Abstract A brief introduction to spatial data mining and decision tree is proposed. Thereupon, by researching on the arithmetic of decision tree, this paper applies the visual spatial data mining technique into the field of land grading and evaluating. And then based on visual C++ 6.0 and MapObject 2.0, a spatial data mining prototype system for land grading and evaluating is designed and developed, in which the decision tree is used as the basic arithmetic of the spatial data mining of the system. For land grading and evaluating, training and learning method is adopted and the integration of them implemented. Furthermore, a model of land grading and evaluating based on decision tree is addressed, especially the system framework, the main modules, the interface and the rough workflow of the system. The approach used in the paper is a new exploration for the methodology of land grading and evaluating, a new attempt for implementation of land information system(LIS), a developing direction for the intellectualized LIS as well.

Key words spatial data mining; decision tree; land grading and evaluating; view; classifying data

[英文审定: 马智民]