文章编号: 1671-8860(2006)10-0916-04 文献标志码: A

可视化交互空间数据挖掘原型系统设计与实现

贾泽露[1,2] 刘耀林[1] 张 彤[3]

(1 武汉大学资源与环境科学学院, 武汉市珞喻路129号, 430079)
(2 中南大学地质与环境工程学院, 长沙市麓山南路25号, 410083)
(3 圣地亚哥州立大学地理系, 美国圣地亚哥钟塔路5500号, CA 92182-4493)

摘 要: 基于VC++ 6.0和MapObject2.0组件技术设计, 开发了一个可视化交互空间数据挖掘的原型系统VGC(visual geo-classify), 并用实例数据对系统性能和算法、规则有效性进行了验证。结果表明, 该原型系统是一个适用的、可扩展的可视化交互空间数据挖掘工具。

关键词: 空间数据挖掘; 决策树; 贝叶斯网络; 地理可视化; 交互; 空间分类
中图法分类号: P208

空间数据挖掘和地理可视化的目的都是促进人们对于空间数据的科学理解, 并从中进行相关高层次空间知识的构建, 二者都是循序渐进、不断优化的过程, 差别在于它们对人类视觉能力和计算机计算能力的依赖程度不同。因此, 将两者结合起来进行空间数据探索分析是完全可行的。笔者结合国内外研究的成果, 开发了一个可视化交互空间数据挖掘(主要用于分类)的试验系统VGC。

1 VGC系统框架及基本设计思想

1.1 系统总体框架

系统的总体框架如图1所示, 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, 用户接收这些信息后, 通过可视化思考过程决定继续的交互操作。系统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, 用户可以随时捕捉被地图放大了的模型的细微变化, 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的分布规律; 另一方面, 可以比较各模型的优劣。总之, 使用地图作为空间数据探索分析的交互中心, 不论是用户的想法还是计算机运算的结果, 都可以实时地在地图上显示。

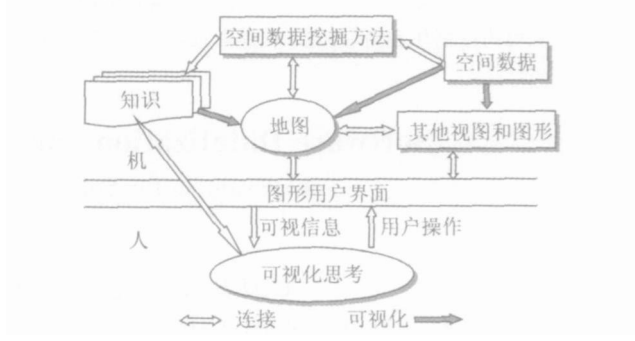

图1 VGC系统的总体框架
Fig. 1 Structure of System

1.2 数据挖掘算法选取

系统采用决策树方法和贝叶斯网络作为数据挖掘方法的基本算法, 其中以决策树算法为主, 贝叶斯网络为辅, 即使用决策树算法对空间数据(主要是矢量数据)进行属性的处理和选择, 通过一定的学习生成规则, 通过这些规则对更为大量的空间数据进行分类; 贝叶斯网络则是描述属性间的相关关系。其总体的目标在于, 通过数据挖掘方法得到分类规则, 并不断加以完善, 尽量提高GIS

收稿日期: 2006-07-25。
项目来源: 国家自然科学基金资助项目(40271088); 国家教育部留学人员回国基金资助项目(152174); 中南大学文理科学研究基金资助项目(0601057)。

空间数据的分类精度。

2　模块和界面设计

2.1　系统模块及工作流程

基于系统总体框架的设计, 系统主要模块如图 2 所示。首先通过一个数据连接和预处理模块从空间数据库中交互地选取参加可视化交互分类分级的数据集、属性和地图, 并对其中的一些数据进行处理(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, 也可以直接从数据集中学习贝叶斯网络, 或者是根据决策树的规则, 把其作为先验知识, 并结合用户对数据的理解来学习贝叶斯网络, 从而得到贝叶斯网络图形。这几种数据挖掘的中间结果可以实时可视化, 并同地图等视图相连, 用户可以从中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两种数据挖掘参数、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, 同样可以从空间数据库中选取测试数据集, 按照分类规则进行测试, 测试的结果可以同训练数据在地图中一起显示, 从而可以比较两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。三个过程之间和不同视图之间可以时时切换, 形成一种循环的渐进的知识发现过程。图 3 描述了系统各部分的联系。

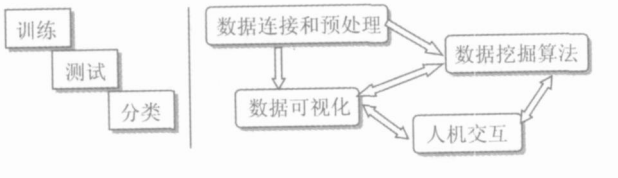

图 2　VGC 系统各模块

Fig. 2　Models of System VGC

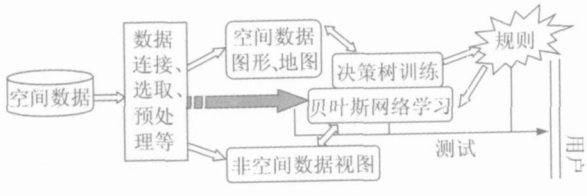

图 3　VGC 系统各模块的连接关系

Fig. 3　Relation of Models of System VGC

2.2　系统主要界面设计

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。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。界面形式的代表截图如图 4 所示。图 4 左边是表单输入的交互形式, 右边是一个典型的菜单形式。程序运行的主界面如图 5 所示, 界面主要由五部分构成, 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。

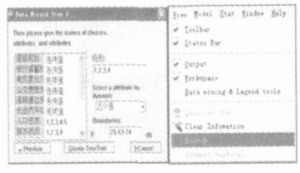

图 4　表单交互、菜单交互

Fig. 4　Alternation of Table and Menu

图 5　系统运行界面

Fig. 5　Main Surface of System

3　系统实现

系统运行和实现环境是 Windows 2000 Professional, 编程语言采用 VC^{++} 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 MapObject2.0 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。贝叶斯网络算法选用贝叶斯学习软件 BNPC(belief network power constructor) 。VGC 采用的策略是将属性数据利用 BNPC 训练得到网络后导入到 VGC 中来。另外, 网络学习的算法比较复杂, 实现困难。VGC 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。

利用 VGC 进行动态分类分级的过程及步骤如下: ① 源数据准备。系统支持 Access2000 的 .MDB文件格式的属性数据和 Shape 格式的空间数据, 属性数据和空间数据分开处理。通过一个数据连接和预处理向导来完成数据导入及为训练程序提供必要的信息。② 初始训练。通过决策树算法进行初始训练, 得到必要的训练信息后, 设定训练参数, 包括树的训练参数和规则的训练参

数两部分。③ 数据分析。得到初步决策树结果后,可以得到剪裁前后的决策树及分类规则。此时,可以根据得到的规则和用户自身的知识作为先验知识,对当前的数据训练学习得到一个贝叶斯网络,然后打开贝叶斯网络视图来对属性之间的关系进行进一步的可视化分析。④ 数据测试。在主控制区中选中测试标签视图。开始测试,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进行比较。文本结果输出在信息输出视图区的测试信息标签视图中显示,地图结果则在地图视图中显示。⑤ 数据分类。测试结果满意后,采用交互分类的方法,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。

由于将地图放在信息交互和传输的中心,因此,进行可视化数据探索分析在大多数情况下是以地图作为探索分析的主要界面。视图连接模型如图 6 所示。

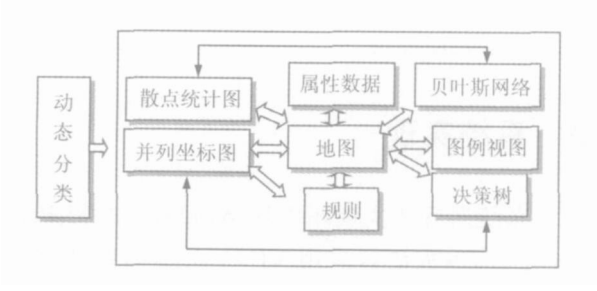

图 6 各视图连接模型

Fig. 6 Model of Views Relation

4 试验与分析

该数据为人工构建,主要是反映土地价格同其他各种影响因素之间的关系,并试图根据训练数据的结果来对未知土地价格进行分类预测,目的只是为了描述 VGC 下可视化交互分类的过程和特点。数据以 shape 格式存储。选取部分对土地价格有影响的因子属性,其中人口密度、用地潜力、地形地质、环境质量、公用设施和基础设施取离散型属性,对外交通、道路通达度、集贸市场和商服中心取连续值属性。这样兼顾了离散和连续值属性的处理,使得数据集更有代表性。训练数据和测试数据比例大约为 7∶3。考虑到动态分类的需要,建立了地价这个连续取值属性,并对其进行动态分类,得到土地价格的各个级别。

按分类过程进行数据的训练和测试工作。试验分类数量和分类界限对决策树训练和测试的影响结果如表 1 所示。试验表明,用户给定的类别个数(对地价进行分级)对分类结果有较大的影响。这说明类别增大时,规则增加很快,但对精度的影响不大。不同分类界限对分类精度的影响也不是很大。为了方便分析,在确定分级边界时,将整个地价分布区间分成四级。表 2 是不同决策树参数下的训练时间、规则个数、剪裁后的树的大小以及分类错误率。

表 1 类别数量对规则分类精度的影响

Tab. 1 Effect of Classification Number to Precision of Classify Rules

分类类别个数	提取的规则数量	分类精度/%
3	4	89
4	7	80
5	7	84
6	11	84

表 2 不同参数配置下的决策树训练情况

Tab. 2 Decision Tree Training Under Different Parameters

参数	执行时间/s	规则大小	剪裁后树大小	错误率
缺省	4.2	8	31	0.200/0.289
成长 5 棵树	32.7	7	23	0.225/0.244
使用信息增益分支方法	3.5	8	31	0.200/0.289
离散属性归并处理	8.5	7	31	0.170/0.290
成长 5 棵树/离散属性归并处理	126.0	6	21	0.230/0.289

试验结果表明,训练多棵决策树可以得到更加简练和精确的决策树(多棵选一),但执行速度却慢很多,所以从交互角度来说,采用单棵决策树训练更为可行。而对离散属性进行归并处理,可以提高分类精度,费时也不多。因此,在出现某些离散属性的属性值过多的情况时可以适当采用。用户在交互操作的任何时候都可以改变这些参数,以达到更好的分类结果。在进行决策树训练的同时,可以根据训练数据生成贝叶斯网络。

5 结 语

对于 VGC 系统,需要对现有的算法努力降低算法复杂度,并在贝叶斯网络结构学习的基础上进行概率参数的学习,这样可以通过贝叶斯网络来验证决策树的分类结果,并直接进行推理预测,做到决策树和贝叶斯网络两种数据挖掘方法的高度结合。同时可以包含更多的数据挖掘方法,并尝试将空间拓扑关系融入到数据挖掘模型之中,真正做到空间数据挖掘。对可视化的界面和内容可以向三维可视化发展,并完善现有的界

面。在地图的表现力方面,加入自动专题制图技术,根据数据的不同特点和用户的不同需要自动选择表示的方法。在可视化和空间数据挖掘方法的结合上,需要扩展可视化在整个空间知识发现中的作用,在数据预处理和选取、数据挖掘、结果解译等各个阶段,都要充分利用可视化的作用,充分体现用户在知识发现过程中的作用。

参 考 文 献

[1] MacEachren A M, Wachowicz M, Edsall R, et al. Constructing Knowledge from Multivariate Spatio-temporal Data: Integrating Geographic Visualization (GVis) with Knowledge Discovery in Databases (KDD)[J]. International Journal of Geographic Information Science, 1999, 13 (4): 311-334

[2] Howard D, MacEachren A M. Interface Design for Geographic Visualization: Tools for Representing Reliability[J]. Cartography and Geographic Information Systems, 1996, 23: 59-77

[3] Quinlan J R. C4.5: Programs for Machine Learning [M]. San Mateo: Morgan Kaufamann, 1993

[4] Cheng Jie. Belief Network PowerConstructor[OL]. http: www. cs. ualberta. ca/~jcheng/ bnpchlp/index. html, 1993

[5] Miller H J, Han J. Geographic Data Mining and Knowledge Discovery: an Overview[M]. London: Taylor and Francis, 2001

[6] Koperski K, Han J. Discovery of Spatial Association Rules in Geographic Information Databases [M]. Berlin: Springer-Verlag, 1995: 47-66

[7] Buja A, Cook D, Swayne D F. Interactive High-Dimensional Data Visualization[J]. Journal of Computational and Graphical Statistics, 1996, 5(1): 78-99

[8] MacEachren A M. An Evolving Cognitive-Semiotic Approach to Geographic Visualization and Knowledge Construction[J]. Information Design Journal, 2001: 26-36

第一作者简介:贾泽露,博士生。主要从事信息系统智能化的研究及其 GIS 理论、方法的研究和教学工作。

E-mail: gis_zljia@163.com

Design and Development of System Based on Visual Interactive Spatial Data Mining

JIA Zelu[1,2]　*LIU Yaolin*[1]　*ZHANG Tong*[3]

(1 School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2 School of Geology and Environment Engineering, Central South University, 25 South Lushan Road, Changsha 410083, China)

(3 Department of Geography, San Diego State University, 5500 Campanile Drive San Diego, CA 92182-4493, USA)

Abstract: A spatial data mining prototype system (visual geo-classify) based on VC++ 6.0 and MapObject2.0 are designed and developed, the basic arithmetic of the spatial data mining of the system is used decision tree and Bayesian networks, and datum classify are used training and learning and the integration of the two to realize. The result indicates that this prototype system is a practical and extensible visual interactive spatial data mining tool.

Key words: spatial data mining; decision tree; Bayesian networks; geo-visualization; interaction; spatial classification

About the first author: JIA Zelu, Ph. D candidate. He focuses on the development of intelligent information system, GIS theories, methods. E-mail: gis_zljia@163.com