TIAN Jing, AI Tinghua, DING Shaojun. Grid Pattern Recognition in Road Networks Based on C4.5 Algorithm[J]. Acta Geodaetica et Cartographica Sinica, 2012, 41(1): 121-126. (田晶, 艾廷华, 丁绍军. 基于 C4.5 算法的道路网网格模式识别[J]. 测绘学报, 2012, 41(1): 121-126.)

基于 C4.5 算法的道路网网格模式识别

田 晶, 艾廷华, 丁绍军

武汉大学 资源与环境科学学院 地理信息系统教育部重点实验室, 湖北 武汉 430079

Grid Pattern Recognition in Road Networks Based on C4.5 Algorithm

TIAN Jing, AI Tinghua, DING Shaojun

Key Laboratory of Geographic Information System, School of Resources and Environment Science, Wuhan University, Wuhan 430079, China

Abstract: A method for grid pattern recognition based on C4.5 algorithm is proposed. Meshes in road networks can be classified as belonging to grid and not belonging to grid according to their context. Firstly, shape measures and relation measures are defined to characterize meshes in road networks. Secondly, two classifiers are trained using C4.5 algorithm based on five measures data and three measures data. A 10-fold cross validation process is applied in order to obtain a sounder result. Finally, the performance of the classifiers is evaluated by means of the Kappa index and the overall correct rate. The Kappa classification accuracy for five dimensions data and three dimensions data is 0.63 and 0.66. The overall correct rate is 81.7% and 82.9% for each. The confidence interval of 90% confidence is [0.785, 0.846] and [0.797, 0.857] respectively. The classifiers are tested by a new data set and the results show that the classifiers are valid in grid pattern recognition.

Key words: road network; grid pattern; pattern recognition; C4.5 algorithm

摘 要: 提出一种基于 C4.5 算法的网格模式识别方法。该方法以道路网中的网眼为基本单元, 根据上下文关系将其标识为属于网格模式和不属于网格模式两类。首先采用形状参量和关系参量描述网眼, 然后, 基于决策树 C4.5 算法分别对 5 参量描述和 3 参量描述数据构造分类器, 运用 10 折交叉验证获得具有说服力的结果, 其 Kappa 值分别为 0.63 和 0.66, 正确率分别为 81.7%和 82.9%, 置信度 90%的置信区间分别为[0.785, 0.846]和[0.797, 0.857]。在新数据上进行了识别效果的验证, 结果表明该分类器可用于网格模式的识别。

关键词: 道路网; 网格模式; 模式识别; C4.5 算法

中图分类号: P208 **文献标识码**: A **文章编号**: 1001-1595(2012)01-0121-06

基金项目: 中国博士后科学基金 (20100480863); 国家 863 计划(2009AA121404); 武汉大学自主科研资助项目(111156)

1 引 言

道路是 GIS 中的核心要素类型, 它构成了城市的结构框架, 是城市意象的主要组织元素。道路网的模式反映了道路的分布特点, 蕴涵着特定历史时期的政治、经济和文化特征。地图综合的智能化研究是地图制图学与地理信息工程学科的发展趋势[1], 空间模式的识别是地图综合朝着智能化方向发展的关键问题[2]。道路选取是一种地图综合方法, 随着地图比例尺的缩小, 道路网密度加大, 应对道路进行选取, 保留重要的道路, 舍去次要的道路。道路选取应保持道路网的结构特征, 文献[3]强调应将道路网的模式作为道路选取算法的参数。文献[4]基于网眼密度进行道路选取也是为了在密度这个层面保持道路网选取前后的分布模式。

网格模式是道路网中的典型模式。对于网格模式的识别, 已提出了两种方法。文献[5—6]以连通度为 4 的道路交叉点为起点, 搜索包含该节点的网眼, 然后通过其中一个网眼, 寻找符合质心排列一致的邻近网眼, 综合考虑网眼间的相似性, 完成网格的识别, 该方法简单高效, 但对于连通度不为 4 的复杂道路交叉点在处理上存在困难。文献[7]将相邻网眼排列一致性、相邻网眼形状相似性、网眼自身形状指标集成为一个参数, 根据该参数搜索邻近网眼, 完成网格模式的识别, 参数的阈值可以人为设定, 以达到自适应的效果, 该方法克服了文献[5]中的方法不能从复杂节点识别网格

的局限。上述两种方法均是根据网眼的属性搜索相邻网眼,其基本思想是利用网眼属性导出一个网眼与其相邻网眼是否构成网格的程度参数。

笔者换一种思路,将网格模式的识别看做是考虑上下文关系的分类任务。其理由是:网格模式由一系列形状和尺寸相似并满足特定的排列方式的网眼多边形构成,构成网格模式的网眼多边形与其他多边形具有不同的特征,所以可以将多边形分为构成网格模式和不构成网格模式两类。值得强调的是,群目标的模式识别与单个目标的识别有所区别,它更关心整体的分布态势。例如,在网格模式中,有一两个"坏"的网眼多边形,在人的肉眼判断上不会对网格模式构成影响,而那一两个"坏"的网眼将在这种上下文关系下被识别为属于网格模式。

运用分类思想进行网格模式的识别需要回答以下3个基本问题:① 运用哪些属性参量描述道路网中的多边形网眼,这些属性参量对于分类的作用如何;② 运用何种算法去构造分类器,其适宜性表现在哪些方面;③ 分类器的性能如何,能否将其应用于新的数据。

本文提出一种基于C4.5算法的网格模式识别方法,对上述3个基本问题进行回答。与已有网格模式识别方法的显著区别在于它是一种基于分类思想的方法,与传统的模式识别和数据挖掘学科联系紧密,强调将模式识别与数据挖掘的理论与方法融入到空间问题的解决中。

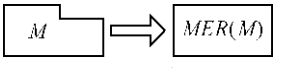
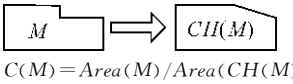
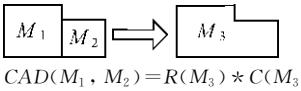
2 网眼多边形的属性参量定义与初步分析

2.1 参量定义

网眼多边形属性参量主要从自身的形状、与周围网眼的排列关系以及相似关系方面定义。下面介绍参量的定义与计算方法,如表1。下述参量与尺度无关,参量值的大小具有绝对性。

表1 网眼参量

Tab.1 Mesh's measure

参量	计算方法	说明
矩形度(rectangularity,R)	描述多边形呈矩形的程度。其计算方法主要有4种,根据文献[8]的建议,选择网眼面积与其最小外接矩形(MER)面积的比值。 取值范围(0,1]	 $R(M)=Area(M)/Area(MER(M))$ $Area()$是面积计算函数,$MER()$返回最小外接矩形。
凹凸度(convexity,C)	描述多边形的凹凸程度,其计算方法详见文献[9],这里选择常用的计算方法,即网眼面积与其凸壳(CH)面积的比值。 取值范围(0,1]	 $C(M)=Area(M)/Area(CH(M))$ $Area()$是面积计算函数,$CH()$返回凸壳。
排列一致度(consistent arrangement degree,CAD)	描述两个网眼是否排列得像网格的程度,其计算方法为:将两个相邻网眼合并成新的多边形,计算该多边形的矩形度与凹凸度,将计算得到的矩形度与凹凸度相乘,然后乘以小网眼与大网眼的面积比值。 取值范围[0,1]	 $CAD(M_1, M_2)=R(M_3)*C(M_3)*\min Area((M_1, M_2))/\max Area((M_1, M_2))$ $\min Area()$是最小面积计算函数,$\max Area()$是最大面积计算函数。
周围排列度最大的网眼的矩形度(rectangularity of mesh having largest CAD, RMHLCAD)	与该网眼排列一致度最大的网眼的矩形度。 取值范围(0,1]	$RMHLCAD=R(ID(\max(CAD(M, M'))))$ M'是与M相邻的网眼多边形,相邻的意思指有公共边,$\max()$是计算最大值函数,$ID()$是返回多边形标识函数。
周围排列度最大的网眼的凹凸度(convexity of mesh having largest CAD,CMHLCAD)	与该网眼排列一致度最大的网眼的凹凸度。 取值范围(0,1]	$CMHLCAD=C(ID(\max(CAD(M, M'))))$ M'是与M相邻的网眼多边形,相邻的意思指有公共边,$\max()$是计算最大值函数,$ID()$是返回多边形标识函数。

2.2　参量的初步分析

箱须图(box-whisker plot)是用于描述数据分布的统计图形，它表示参量的最小值、第 1 四分位数、中位数、第 3 四分位数和最大值，利用它可以从视觉角度观察参量值的分布情况。整个箱图中最上方和最下方的线段分别表示数据的最大值和最小值，对于箱来说，其上下两端的线段分别表示第 3 四分位数和第 1 四分位数，中间的粗线段表示数据的中位数，“o”标出温和的异常值，“＊”标出极端的异常值。箱须图常用于观察数据的分布、识别数据中的异常值，但文献[10]应用箱须图初步估计每个参量对类的可分性。图 1 是由武汉市部分道路网构成的 427 个网眼生成的关于上述 5 个参量的箱须图。该图由 SPSS 软件生成、版本号 16.0。每个箱须图左端是参量名称，下端是是否属于网格的标识(IsGrid)，yes 代表属于，no 代表不属于。从中可以观察出凹凸度(C)和周围排列度最大的网眼的凹凸度($CMHLCAD$)的一维参量类可分性较差，排列一致度(CAD)的类可分性较好。受文献[10]的启发，后文的试验中将分别采用 5 参量(R、C、CAD、$RMHLCAD$、$CRMHLCAD$)和 3 参量(R、CAD、$RMHLCAD$)进行试验。

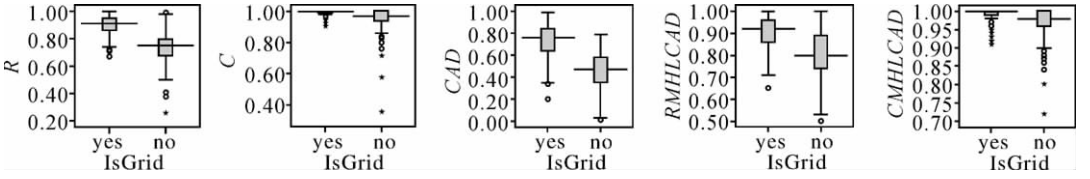


图 1　由 427 个网眼生成的 5 个参量的箱须图

Fig. 1　Box-whisker plot for five measures calculated from 427 meshes

3　基于 C4.5 算法的网格模式识别方法

3.1　决策树与 C4.5 算法

决策树学习是应用最广的归纳推理算法之一，它是一种逼近离散值函数的方法，对噪声数据有很好的抗差性且能够学习析取表达式。决策树通过把实例从根节点排列到某个叶子节点来分类实例，叶子节点即为实例所属的分类。树上的每一个结点说明了对实例的某个属性的测试，并且该节点的每一个后继分支对应于该属性的一个可能值。分类实例的方法是从这棵树的根节点开始，测试这个节点所指定的属性，然后按照给定实例的该属性值对应的树枝向下移动。然后这个过程在以新节点为根的子树上重复。构造过程是从“哪一个属性将在根节点被测试？”这个问题开始的。分类能力最好的属性被选作树的根节点的测试，然后为根节点属性的每个可能值产生一个分支，并把训练样例排列到适当的分支之下。然后重复整个过程，用每个分支节点关联到训练样例来选取在该点被测试的最佳属性。这形成了贪心搜索，也就是算法从不回溯重新考虑以前的选择[11]。C4.5 算法是著名的决策树学习算法，详细的介绍参见文献[12]。

判别式、神经网络和决策树及它们对应的算法都可用来构造分类器，但本文选用决策树 C4.5 算法作为构造分类器的方法，原因在于：① 它具有里程碑意义，是目前为止在实践中应用最为广泛的机器学习工具[13]，在地理信息科学中，有很多研究者采用决策树 ID3 算法、C4.5 算法及其改进版本 C5.0 进行问题的解答[14-16]；② 算法输出的决策树可以很自然地表示成规则的形式，易于理解。

3.2　网格模式识别方法

在介绍了参量描述和分类器构造算法后，描述基于知识的网格模式的识别方法，方法的基本步骤如下：

(1) 数据预处理。受文献[17]研究的启发，重复删除道路网图中的桥和孤立点。重复删除的意思是删除一个桥后，重新检查剩余的道路网是否还存在桥，如果是的话，继续删除，直到不存在桥和孤立点，对处理完的道路网构建多边形拓扑结构。对于已有线面拓扑关系的数据可以跳过这一步。

(2) 对每个网眼多边形，计算 2.1 节定义的参量。

(3) 将训练数据标识为属于网格模式与不属于网格模式两类，运用 C4.5 算法构造分类器。

(4) 评估分类器性能。

(5) 将待识别的道路网数据导入,用该分类器进行模式识别。

4 试验与分析

4.1 试 验

(1) 训练数据:训练数据是武汉市部分道路网构成的 427 个网眼。

(2) 试验环境与试验平台:试验在 Window XP 操作环境下进行,数据处理功能和参量计算用 Visual C++ 6.0 扩展 DoMap 平台开发完成。运用开源数据挖掘软件 Weka 进行分类器的构造,该软件的介绍与操作详见文献[10]。

(3) 试验结果评估:在数据量有限的情况下为获得好的结果,运用 10 折交叉验证方法。根据正确率(正确分类的样本的比例)和 Kappa 统计值评价分类精度。

试验 1:对 5 个参量(R、C、CAD、$RMHLCAD$、$CRMHLCAD$)数据进行试验,导出的决策树如图 2,混淆矩阵如表 2。其 Kappa 统计值为 0.63,正确率为 81.7%,置信度 90%的置信区间为[0.785, 0.846]。

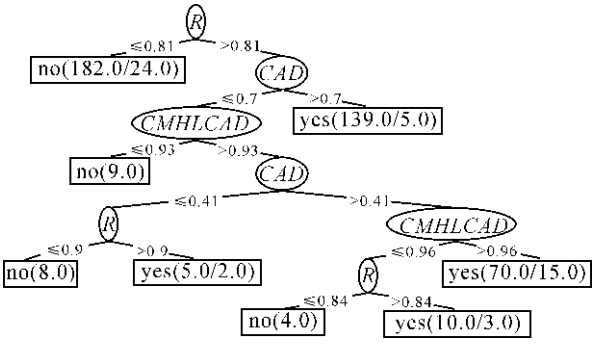


图 2 由 5 参量描述的训练数据生成的决策树

Fig. 2 Decision tree inducted by five measures data

表 2 5 参量描述数据的混淆矩阵

Tab. 2 Confusion matrix for five measures data

	yes	no	total
yes	189	34	223
no	44	160	204
total	233	194	427

试验 2:根据 2.2 节的参量分析,对 3 参量(R、CAD、$RMHLCAD$)数据进行训练。导出的决策树如图 3,混淆矩阵如表 3。其 Kappa 统计值为 0.66,正确率为 82.9%,置信度 90%的置信区间 [0.797, 0.857]。

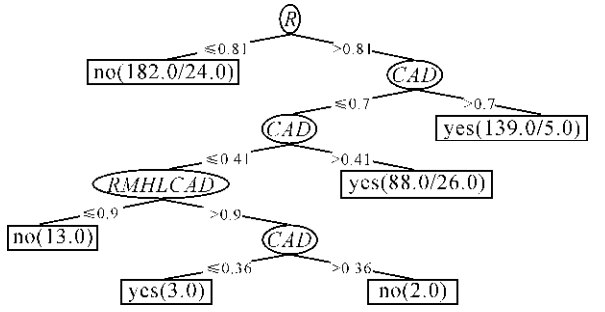


图 3 由 3 参量描述的训练数据生成的决策树

Fig. 3 Decision tree inducted by three measures data

表 3 3 参量描述数据的混淆矩阵

Tab. 3 Confusion matrix for three measures data

	yes	no	total
yes	192	31	223
no	42	162	204
total	234	193	427

4.2 分 析

(1) 从构造的分类器的性能来说,在 10 折交叉验证方法下,5 参量描述和 3 参量描述的正确率均大于 80%,虽然 3 参量描述的正确率和 Kappa 统计值略高于 5 参量描述,但不能说明可用 3 参量描述代替 5 参量描述。

(2) 遍历决策树中由根节点到叶节点的路径,经过提炼(保留叶节点是 yes 的规则,合并某些规则),得到的规则如表 4。由决策树导出的规则易于理解,实现了知识的显示表达。例如,if(R>0.81) and (CAD>0.7) then yes,表达了如果网眼自身矩形度和排列度较大,则该网眼属于网格模式。

表 4 由决策树导出的规则

Tab. 4 Rules derived from decision tree

5 参量描述的数据	3 参量描述的数据
① if (R>0.81) and (CAD>0.7) then yes ② if (R>0.9) and (CAD≤0.41) and ($CMHLCAD$>0.93) then yes ③ if (R>0.81) and (0.41<CAD≤0.7) and ($CMHLCAD$>0.96) then yes ④ if (R>0.84) and (0.41<CAD≤0.7) and (0.93<$CMHLCAD$≤0.96) then yes	① if (R>0.81) and (CAD≥0.41) then yes ② if (R>0.81) and ($CMHLCAD$>0.9) and (CAD≤0.36) then yes

对于在决策树中参量的取值是由 C4.5 算法

决定的。例如图 2 中，决策树第一个分支在 R 上为什么要取 0.81，这是因为，对于连续值，C4.5 算法处理过程如下[18]：根据参量的值，对数据集排序；用不用的阈值将数据集动态的进行划分；当输出改变时，确定一个阈值；取两个实际值中的中点作为一个阈值；取两个划分，所有样本都在这两个划分中；得到所有可能的阈值以及增益比率；每一个参量会变为两个取值。下面举一个例子：参量 A 具有连续值，则在训练数样本中可以按照升序方式排列 A 的值，如果 A 共有 n 种取值，则对每个取值将所有的数据进行划分，针对每个划分计算信息增益比率，选择最大的划分来对相应的参量进行离散化。

(3) 在新的数据上进行方法验证试验，数据是深圳市道路网的一部分，包含 399 个网眼，其识别结果如图 4。其中，灰色填充——仅由 3 参量描述的训练数据生成的决策树所导出的规则的识别结果；灰色＋晕线——由 3 参量描述的训练数据生成的决策树所导出的规则和由 5 参量描述的训练数据生成的决策树所导出的规则均识别为网格模式的结果；空白＋晕线——仅由 5 参量描述的训练数据生成的决策树所导出的规则的识别结果；空白——由本方法判定不属于网格模式的网眼。由上述两组规则分别进行识别的结果与肉眼识别结果相似，部分地实现了对某些“坏”网眼的正确识别。

值得一提的有两点：第一，训练数据与测试数据来自不同的城市道路网，不存在用训练数据进行测试所产生的重新带入误差(re-substitution error)；第二，不是将同一城市的数据分为训练数据和测试数据，而是将一个城市的道路网作为训练数据，另一个城市的道路网作为测试数据，使得结果更具说服力，总体上表明本文提出的方法可行。

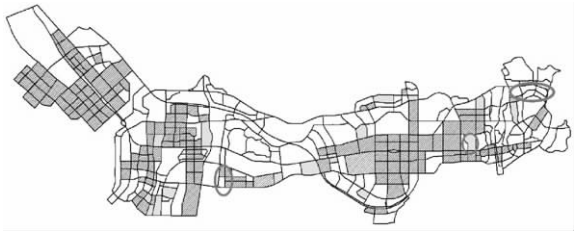


图 4　深圳数据的验证

Fig. 4　Validation by Shenzhen data

(4) 对图 4 的识别结果进一步分析，发现其中有些地方不尽如人意，如图 5 所举的 3 个例子(对应于图 4 的圈出部分)。究其原因，主要有 3 点：第一，构面对识别结果的影响，如图 5(a)，这是基于网眼的网格模式识别方法的通病，其改进有待于 Gestalt 原则的应用；第二，网眼参量的影响，如图 5(b)，矩形度的计算方式使得它对于网眼上小突出很敏感；第三，人为因素的影响，对于较为规则的情况没有异议，但对于一些退化较为严重的情况，“坏”的网眼是否属于网格模式那就是仁者见仁，智者见智了，从而导致训练样本的类标识不同，直接影响由此得到的分类器的性能，从而影响识别结果，如图 5(c)所示，如果该图中的网眼被标识为属于网格模式，那么根据监督学习的特点，遇到类似情况，它们会将其判断为属于网格模式。决策树的泛化功能取决于样本的选择和参量的描述。

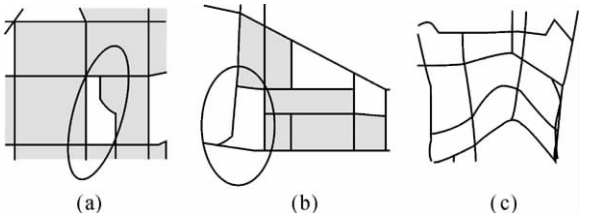


图 5　异常情况

Fig. 5　Exceptional situation

5　结　论

本文提出一种基于 C4.5 算法的网格模式识别方法。该方法以道路网中的网眼多边形为基本单元，将其分为属于网格模式和不属于网格模式两类。采用矩形度、凹凸度、排列一致度、周围排列度最大的网眼的矩形度、周围排列度最大的网眼的凹凸度来描述网眼多边形，然后，基于决策树 C4.5 算法分别对 5 参量描述和 3 参量描述数据构造分类器。试验结果表明该方法有效，能够运用于新的数据进行网格模式识别。

进一步的研究工作将在以下几个方面展开：① 组织不同年龄、不同专业、不同文化背景的人员进行认知试验，对网眼是否属于网格模式进行分析；② 网眼多边形属性参量的进一步分析与甄选，研究其他一些形状参量，如紧凑度、延展度等对识别产生的作用；③ 判别式、神经网络等其他分类器构造方法与决策树方法在分类性能上的比较。

参考文献：

[1] WANG Jiayao. Development Trends of Cartography and Geographic Information Engineering [J]. Acta Geodaetica et Cartographica Sinica, 2010, 39 (2): 115-119. (**王家耀. 地图制图学与地理信息工程学科发展趋势**[J]. **测绘学报**, 2010, 39(2): 115-119.)

[2] MACKANESS W, EDWARDS G. The Importance of Modeling Pattern and Structures in Automated Map Generalization[C]// Proceedings of the Joint ISPRS/ICA Workshop on Multi-scale Representations of Spatial Data. Ottawa:[s. n.], 2002.

[3] ZHANG Qingnian. Modeling Structure and Patterns in Road Network Generalization [C]// Proceedings of ICA Workshop on Generalization and Multiple Representation. Leicester:[s. n.], 2004.

[4] HU Yungang, CHEN Jun, LI Zhilin, et al. Selective Omission of Road Features Based on Mesh Density for Digital Map Generalization [J]. Acta Geodaetica et Cartographica Sinica, 2007, 36(3): 351-357. (**胡云岗, 陈军, 李志林, 等. 基于网眼密度的道路选取方法**[J]. **测绘学报**, 2007, 36(3): 351-357.)

[5] HEINZLE F, ANDERS K H, SESTER M. Graph Based Approaches for Recognition of Patterns and Implicit Information in Road Networks[C]// Proceedings of the 22nd International Cartographic Conference. La Coruna: [s. n.], 2005.

[6] HEINZLE F, ANDERS K H. Characterising Space via Pattern Recognition Techniques: Identifying Patterns in Road Networks[C]// Generalisation of Geographic Information: Cartographic Modelling and Applications. [S. l.]: Elsevier Ltd, 2007: 233-253.

[7] YANG Bisheng, LUAN Xuechen, LI Qingquan. An Adaptive Method for Identifying the Spatial Patterns in Road Networks [J]. Computers, Environment and Urban Systems, 2010, 34(1): 40-48.

[8] ROSIN P L. Measuring Rectangularity [J]. Machine Vision and Application, 1999, 11(4): 191-196.

[9] ZUNIC J, ROSIN P L. A New Convexity Measure for Polygons [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(7): 923-934.

[10] STEINIGER S, LANGE T, BURGHARDT D, et al. An Approach for the Classification of Urban Building Structures Based on Discriminant Analysis Techniques [J]. Transactions in GIS, 2008, 12 (1): 31-59.

[11] MITCHELL T M. Machine Learning [M]. ZENG Huajun, ZHANG Yinkui, translation. Beijing: Mechanical Industry Press, 2003. (MITCHELL T M. **机器学习**[M]. **曾华军, 张银奎, 译. 北京: 机械工业出版**, 2003.)

[12] QUINLAN J R. C4.5: Programs for Machine Learning [M]. San Francisco: Morgan Kaufmann Publishers Inc, 1993.

[13] WITTEN I H, FRANK E. Data Mining: Practical Machine Learning Tools and Techniques [M]. DONG Lin, QIU Quan, YU Xiaofeng, translation. 2nd ed. Beijing: Mechanical Industry Press, 2005. (WITTEN I H, FRANK E. **数据挖掘: 实用机器学习技术**[M]. **董琳, 邱泉, 于晓峰, 译. 第2版. 北京: 机械工业出版社**, 2005.)

[14] SESTER M. Knowledge Acquisition for the Automatic Interpretation of Spatial Data [J]. International Journal of Geographical Information Science, 2000, 14(1): 1-24.

[15] QI F, ZHU A X. Knowledge Discovery from Soil Maps Using Inductive Learning [J]. International Journal of Geographical Information Science, 2003, 17(8): 771-795.

[16] MUSTIERE S. Cartographic Generalization of Roads in a Local and Adaptive Approach: a Knowledge Acquisition Problem [J]. International Journal of Geographical Information Science, 2005, 19(8-9): 937-955.

[17] XIE F, LEVINSON D. Measuring the Structure of Road Networks [J]. Geographical Analysis, 2007, 39 (3): 336-356.

[18] MAO Guojun, DUAN Lijuan, WANG Shi, et al. Principles and Algorithms of Data Mining [M]. Beijing: Tsinghua University Press, 2005:123. (**毛国君, 段立娟, 王实, 等. 数据挖掘原理与算法**[M]. **北京: 清华大学出版社**, 2005:123.)

(**责任编辑:雷秀丽**)

收稿日期: 2010-12-13
修回日期: 2011-02-23

第一作者简介: 田晶 (1982—), **男, 博士后, 讲师, 主要从事地图自动综合和模式识别的研究。**
First author: TIAN Jing (1982—), male, postdoctoral fellow, lecturer, majors in automated map generalization and pattern recognition.
E-mail: yutaka-2010@163.com