地理空间数据挖掘

李连发　王劲峰　等　著

GEOSPATIAL
DATA MINING

科学出版社

地理空间数据挖掘

李连发　王劲峰 等　著

科学出版社
北京

内 容 简 介

本书针对空间数据具有空间关联性即空间变异性的特点，系统总结了空间数据挖掘的理论方法，描述了主要的空间数据分析及挖掘软件，探讨了空间数据挖掘在资源环境调查、自然灾害风险分析等方面的综合运用，展现了其解决实际问题的作用及辅助决策的功能。

本书可供地学、环境及社会科学等领域的研究者对空间数据进行探索性分析、统计建模及预测时参考使用。

图书在版编目(CIP)数据

地理空间数据挖掘/李连发等著. —北京：科学出版社，2014. 6
ISBN 978-7-03-041227-0

Ⅰ. ①地… Ⅱ. ①李… Ⅲ. ①地理信息系统-数据采集 Ⅳ. ①P208

中国版本图书馆 CIP 数据核字(2014)第 126411 号

责任编辑：张艳芬 / 责任校对：鲁 素
责任印制：肖 兴 / 封面设计：蓝 正

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

骏走印刷厂印刷
科学出版社发行 各地新华书店经销
*
2014 年 6 月第 一 版 开本：720×1000 1/16
2014 年 6 月第一次印刷 印张：20 1/4 插页：4
字数：392 000

定价：100.00 元

（如有印装质量问题，我社负责调换）

前　言

空间数据特有的存储格式及空间相关性的特点，使得常用的数据挖掘方法不适合处理空间数据及其时间序列，本书针对空间数据及其时间序列的特点，系统总结了时空数据挖掘的流程，即数据预处理→探索性分析→监督/非监督学习→建模及验证→模型集成，介绍了主要方法原理，提出了建立在时空数据挖掘基础上的辅助决策支持模型。

本书理论方法的介绍深入浅出，实用性强。全书分理论方法、软件工具及综合运用三部分：①理论方法部分探讨采用经典数据挖掘来处理空间数据，在经典数据挖掘方法的基础上融入空间因子，还系统总结了不同于经典方法的空间数据挖掘方法，探讨如何在 GIS 的环境中建立基于时空数据挖掘方法的辅助决策支持模型；②软件工具部分总结了主要的空间数据分析及数据挖掘软件，重点探讨在 GIS 支持下如何实现时空数据挖掘的算法，如何集成多个系统组件形成有机的系统，以达到辅助决策的功能；③综合运用部分则重点介绍了时空数据挖掘在资源环境调查、自然灾害风险分析、软件系统风险识别等方面的运用，展现了其解决实际问题的作用及辅助决策功能。

本书深入探讨了空间数据分析及挖掘的原理及方法，将空间分析、数据挖掘与辅助决策结合起来，探索其在资源环境及风险分析中的应用。本书提出的方法应用到资源环境领域调查及自然灾害风险分析中将有助于提高结果的精度及分析效果。本书的出版将会为推动时空数据挖掘方法论发展及普及，提高 GIS 环境下的海量空间数据挖掘、不确定性分析及辅助智能决策支持功能作出贡献。

同国内外同类书籍比较，本书主要特色是针对空间数据的相关性，将空间分析、数据挖掘及知识推理结合起来，针对地学中许多问题的复杂性及不确定性，建立在 GIS 环境中的建模、预测、不确定性分析的功能，为辅助决策提供强大的支持功能。本书实用性较强，相关的软件工具、编码示例及在抽样调查及风险分析中的应用，可提高读者动手实践的能力，辅助解决相关问题。

本书得到了国家 863 课题、国家自然科学基金及 973 项目的支持，是相关项目成果的总结。在本书的写作过程中，李连发负责主要章节的撰写，谢会云、翟春丽、汪承义及姜成晟参与了本书的部分研究工作，王阳及赵斯思参与第 8 章及第 11 章的撰写，杨勋凤参与第 9 章的撰写并负责全书的整理及校对，最后由李连发和王劲峰统稿。周成虎、朱阿兴、龚建华、吴俊及 Hareton Leung 教授等对本书相关内容给予了指导与帮助，在此一并表示衷心感谢。

限于作者水平，书中难免存在疏漏之处，衷心期望读者不吝批评指正。

目　　录

第三篇　综 合 运 用

引　言

一切事物都与其他事物相关，但是距离近的比远的相关性更强。

——Tobler(1979)

以上为 Tobler 地理学第一定理，它概括了空间数据的主要特点，即空间相关性。由于这种空间相关性的存在，如果把经典的统计学方法不加区别地应用到空间数据分析之中，会使得分析结果产生严重偏差（应用样本独立的前提条件不满足）。因此，在进行空间数据分析及数据挖掘的相关研究时，有必要考虑空间相关性因素，有选择地使用经典统计学中的分析方法与模型（Ripley，1981；Anselin，1992；Cressie，1993；Haining，2003）。

空间数据的时间序列形成了时空数据，本书以时空数据为研究对象，采用空间与非空间（即不考虑相关性）的方法，考察其相关性、变化趋势、模式等。本书将以基于栅格的多维时空数据为基本的数据格式，从数据采集、预处理、知识归纳到建模与预测，形成一套时空数据分析及挖掘的方法，并通过抽样调查分层及动态风险分析的研究案例，说明这套方法的理论研究意义及实际应用价值。

0.1　时空数据的特点

空间数据总是以一定的地学对象作为载体，如资源环境调查中的调查对象（耕地、林业、土地等），又如灾害风险分析中的承灾体等，这些都涉及实际的地学实体，其空间位置关系是最显著的特征，这种特征通过空间相关性得以体现。例如，地震学家研究地震的区域分布，以及地震发生是否存在空间格局及其可预报性；流行病学家分析病例的空间分布规律，及其是否与污染分布有关；警察通过察看盗窃事件发生的空间位置寻找其与社区社会经济特征的空间关联性，据此对未来态势做出估计；遥感专家将遥感图像中的噪声过滤掉以恢复其基本空间格局；地质学家根据空间离散分布的钻孔点集信息推测矿藏储量；地下水文学家用一系列有毒化学浓度样品制作地下水污染地图；零售商根据区域科学家建立的购物模型来估计居民对于其所属零售店的需求，以及是否有新营业的网点等（王劲峰等，2006）。

以空间相关性为核心，可以衍生出其他的基本特点，包括空间变异性及空间位置的拓扑关系等。空间变异性是指空间属性值在空间上的差异，它只与相对位置有关，与具体位置无关，由此可以探索空间属性在空间上的变化规律；而空间拓

扑关系是指相对位置关系，涉及空间推理问题。

专门研究空间特性数据分析的方法形成了一门独特的学科，即空间数据分析，它发源于国外，之后扩展到国内。空间数据分析开始是在地质领域的空间关联性的克里格插值方法[由南非采矿工程师克里格(Krige)于1951年首次提出，并命名为“克里格”法]中被提出，其考虑了空间相互之间的关联性(测点的相互关系和空间分布位置等几何特征)，在实际及理论上得到了较完善的证明，之后，克里格插值方法本身得到了进一步的完善及成功应用，发展出了协同克里格和区域克里格等，在矿产探测及预报等方面取得了很大成功，其他领域的应用也相继开展(Journel et al.，1978)；Whittle于1954年提出了降雨等连续空间关联，从物理机理方面说明了这种关联性是存在的，但从数学角度很难证明(Whittle，1954)；Ignacio与Jose于1974在降雨网络监测步骤中采用这种相关性，提出与克里格不同的基于不同应用对象的空间关联函数(Rodríguez-Iturbe et al.，1974)；同时，英国的Haining(2003)在其专著《空间数据分析》中总结了相关理论(Haining，2003)；而国内专家王劲峰等(2002)又提出了基于离散耕地的Sandwich方法，将空间相关性理论扩展到了离散地物方面。这些发展都使得空间分析日益成为一门独特的学科(Wang et al.，2002)。

本书以空间数据(空间数据的时间序列形成时空数据)为主，采用空间数据分析、数据挖掘及机器学习方法对这些数据进行剖析，提取相关的时空知识，探索空间分布模式，结合具体的领域进行典型案例研究与验证。

0.2 基于栅格的数据分析及发掘流程

本书涉及的空间数据分析将以栅格数据作为主要的数据集格式，预处理、建模及预测都基于栅格格式的数据集展开。栅格数据将会被转化成数据表格的方式输入、输出学习器。采用栅格处理时空数据有以下优点(Li et al.，2008)：

(1) 便于融合多源、异构的时空数据进行分析，包括遥感影像、图片、矢量、栅格数据，通过矢量栅格化、再抽样技术转化成统一的数据集，便于数据的相关性探索分析、分类、聚类及建模与预测等。

(2) 对于多维的栅格时空数据，可以采用影像处理软件(ENVI、ERDAS及ArcGIS Grid模块等)来辅助栅格数据的处理，对数据进行可视化的探索性分析，对分析结果进行表达与再现，突出结果信息。

(3) 便于从微观到宏观的分析，通过栅格单元的微观分析，捕捉和把握区域上的宏观空间分布模式(如空间异构模式的识别)。

(4) 便于以像素单元为粒度的知识归纳与学习，可以将每个像素单元划分成一个训练实例或者测试样本，采用数据挖掘及机器学习方法，从中归纳出一般性

的规则知识集，或者采用贝叶斯网络（Bayesian network，BN）归纳出变量之间的不确定性关系，从而保存到知识库中进行谓词逻辑及不确定性推理，这种知识归纳及推理方法具有重要的理论及应用价值。

当然，转化成统一的栅格数据格式会有信息的丢失，或者引起误差。因此，进行有效性验证也是很必要的。而且，栅格单元的大小决定了学习的最小粒度，因此要结合应用目标及任务小心确定，栅格单元太大会导致精度不够，结论自然不可靠，而栅格单元太小会导致数据量增加，学习过程异常缓慢，以致得不到理想的结果。

图 0.1 描述了基于栅格的多维时空数据的学习过程。

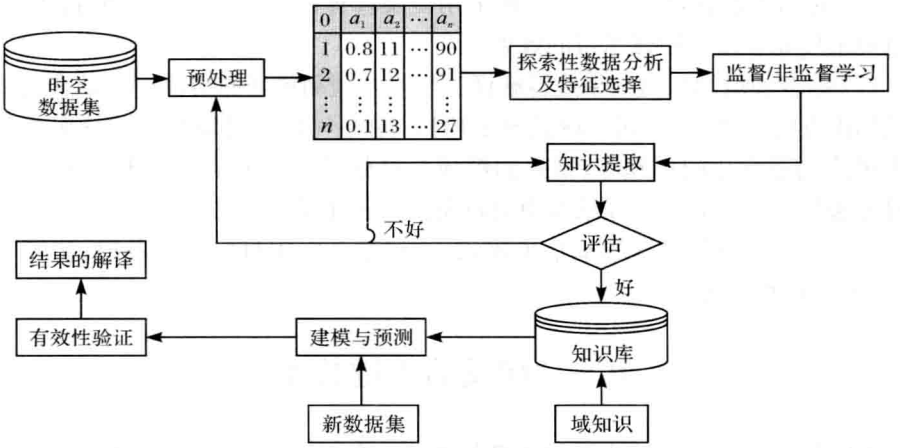

图 0.1　基于栅格的多维时空数据分析及发掘流程

基于栅格的多维时空数据分析及发掘流程主要有以下七个步骤：

（1）时空数据的采集及整理。本步骤根据研究目标选择合适的多个数据来源，包括遥感光谱数据、自然地理条件、社会经济因素、专题数据（如调查的样本数据或者自然灾害数据）等，通过一定的技术手段（如矢量栅格化、最近邻法或者再抽样等）转化成统一格式的多维栅格时空数据集，最终将数据以类似多波段遥感图像的方式存储及表达，并进行一些初步的处理。

（2）数据预处理。预处理包括正则化、去噪（孤立点分析）、插值与缺值数据分析、数据格式转换、模型选择等，预处理的主要目的是清理数据，去除噪声，提高学习及预测的效果。

（3）探索性分析及特征选择。根据需要进行探索性分析，包括可视化的探索性分析（如直方图、散点图、插值三维图及空间相关性分析及空间变异函数）、识别变量之间的关联性及孤立点、去处多重共线性问题等，而特征选择是根据学习的目标选择信息含量丰富的因子变量，探索性分析及特征选择也可以作为预处理的

一个主要步骤。

(4) 监督/非监督学习。根据学习的需要选择监督或非监督学习方法，监督学习需要有可靠的样本作为训练及测试资料，而非监督学习则不需要训练样本，自动地从数据中发现一定的模式(如区域异构性)，学习得到的知识经过实际提炼后可保存到知识库中便于以后调用。

(5) 建模与预测。该步骤主要是采用空间数据挖掘技术及贝叶斯网络进行推理及预测分析，空间数据挖掘技术包括时空关联技术、co-location 技术、空间聚集及空间预测模型，而贝叶斯网络包括网络的学习、调整及不确定性证据推理等方面的技术。

(6) 对结果的有效性验证。一般采用交叉验证的方法来验证结果的精度，同时也可采用元学习法加强学习的效果。

(7) 结合领域知识对结果进行解译。学习和预测得到的结论应当结合具体的领域知识进行解译，如空间区域差异的原因(用在气象上可能是某种气候指数)、灾害风险动态变化的机理等，结果的解译是对分析及发掘结果的认识及深化，可在对方法检验的同时探索领域方面的新模式与新知识。

以上只是分析及发掘的基本步骤，具体到某个应用目标及任务时，步骤会更具体，也会有所不同。

0.3 涉及的关键技术

空间数据分析及挖掘涉及的关键技术有空间统计分析技术、数据挖掘技术、空间数据挖掘技术、基于贝叶斯网络的学习及概率推理技术，下面分别对其进行介绍。

0.3.1 空间统计分析技术

空间统计分析是采用空间统计学分析空间数据的技术，主要特点在于其分析方法中对空间相关性的着重考虑。空间统计分析技术范围比较广泛，本书中涉及的技术包括以下三种。

1. 空间插值技术

空间插值技术主要用于数据预处理。其是一种特殊的插值形式，它特别注意空间关系(如连通性、距离和方向等)，这些空间关系决定了已知点对待估点观测值的影响。目前流行的插值方法主要有最近邻法、距离反比、算术平均、高次曲面、多项式插值、最优插值、克里格插值、样条插值、经验正交函数插值、径向基函数插值等(Franke，1982；Özdamar et al.，1999；Haining，2003)

2. 空间关联性探测

空间关联性探测包括点格局分析、格数据统计和用于相关数据的探索性分析(Cressie,1993;Haining,2003;王劲峰等,2006)。其中,点格局分析是针对不规则分布于感兴趣区域的一系列点集,分析这些对象集在空间的分布特征和相互关系,即空间分布格局,如集聚(clumped)、随机(random)、规则(uniform)分布等;而格数据则是分析具有格网形态的空间事物属性,即代表具有格状统计分析单元的数据,其主要解决的问题是如何衡量空间事物的相关性及其度量,具体包括格数据空间效应的表达,即空间相关性和空间异质性、空间热点区域及可变面域问题分析等。

3. 空间回归预测

空间回归预测包括点回归与格回归等,主要是根据一系列的因子对依赖变量进行预测(Cressie,1993;Haining,2003;王劲峰等,2006)。其中,点回归主要采用Kriging方法,该方法主要利用随机函数对不确定现象进行探索分析,并结合采样点提供的信息对未知点进行估计和模拟(Journel et al.,1978);而格回归则是对具有格网形态的空间事物属性,即代表具有格状统计分析单元的数据进行预测,包括空间自相关回归(spatial autocorrelation,SAR)模型、空间移动平均回归(moving average regressive,MAR)模型和空间条件自回归(conditional autoregressive,CAR)模型(王劲峰等,2006)。

0.3.2　数据挖掘技术

数据挖掘是从大量不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的人们事先不知道的但又潜在有用的信息和知识的过程。数据挖掘是知识发现的关键步骤(Michalski et al.,1998;Mitchell,1997;Tan et al.,2006;Witten et al.,2005)。数据挖掘技术包括以下五个方面。

1. 关联分析

关联规则挖掘是探测两个或两个以上变量取值之间存在的某种规律性,即关联。数据关联是数据库中存在的一类重要的、可被发现的知识。关联分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网,一般用支持度和可信度两个阈值来度量关联规则的相关性,还不断引入兴趣度、相关性等参数,使得所挖掘的规则更符合需求。