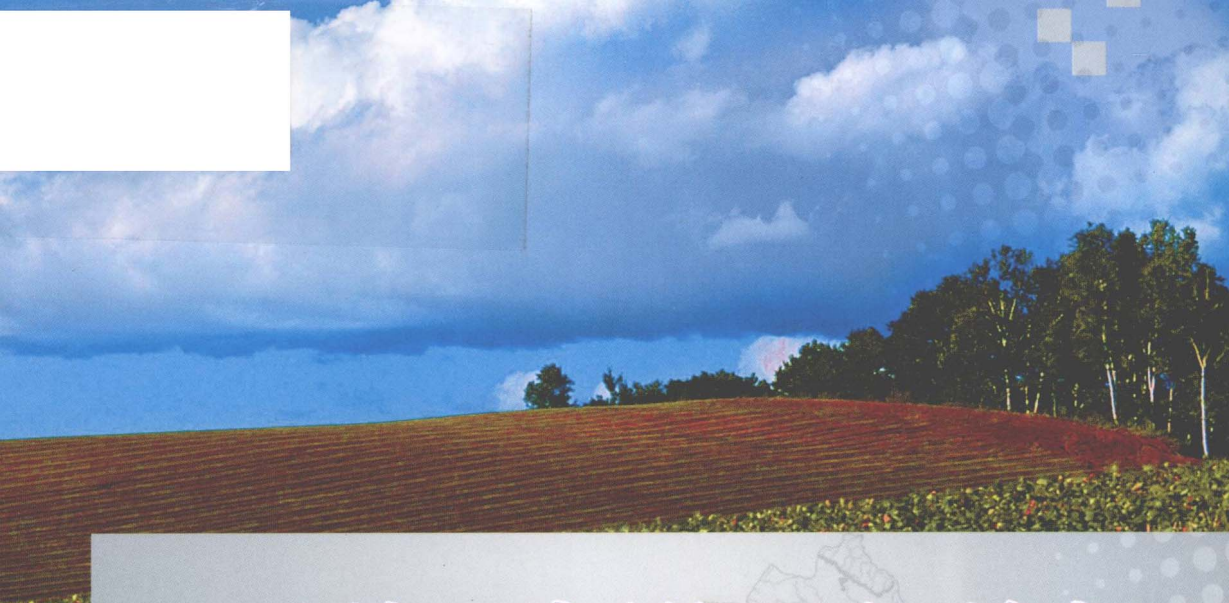

土地资源评价数据挖掘方法与应用

胡月明　张俊平　薛月菊　编著

科学出版社

土地资源评价数据挖掘方法与应用

胡月明　张俊平　薛月菊　编著

科学出版社
北京

内 容 简 介

数据挖掘技术已被广泛应用于资源环境和社会经济调查与数据分析之中。相对于传统的土地资源评价方法，数据挖掘技术的应用具有速度快、准确度高、费用低等特点。本书介绍了数据挖掘的概念和一般程序，结合具体的土地资源评价案例，系统地研究了数据挖掘技术在数据预处理、评价指标权值确定、样本容量计算、土地质量等级评定等环节中的应用。

本书是一部比较全面系统阐述数据挖掘技术在土地资源评价领域中应用的专著，可供土地、GIS、环境、生态等领域的本科生及研究生、相关教学科研人员及技术人员使用，对数据挖掘应用研究的有关学者、高校师生亦有一定的参考价值。

图书在版编目(CIP)数据

土地资源评价数据挖掘方法与应用/胡月明，张俊平，薛月菊编著. —北京：科学出版社，2012

ISBN 978-7-03-035566-9

Ⅰ. 土… Ⅱ. ①胡…②张…③薛… Ⅲ. ①土地资源—资源评价—数据采集 Ⅳ. ①F301-39

中国版本图书馆 CIP 数据核字(2012)第 219096 号

责任编辑：余 丁 张海丽 / 责任校对：宋玲玲

责任印制：张 倩 / 封面设计：耕者设计工作室

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

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2012 年 10 月第 一 版 开本：B5(720×1000)

2012 年 10 月第一次印刷 印张：12 1/2 插页：20

字数：239 000

定价：60.00 元

(如有印装质量问题，我社负责调换)

作者简介

胡月明，男，1964年生，湖南安化人，教授，理学博士，博士生导师。主要从事地理信息系统应用领域的教学与研究工作，现任华南农业大学信息学院学术委员会主任、地理信息工程研究所所长、建设用地再开发国土资源部重点实验室主任、土地利用与整治广东省重点实验室主任、广东测绘地理信息产业技术创新联盟理事长。已主持科研项目100多项，发表学术论文150多篇，出版专著和统编教材16部，获省部级科技成果奖励10项、教学成果奖励1项。

张俊平，男，1976年生，江西南昌人，助理研究员，工学博士。主要从事土地资源评价、土地资源管理与数学建模研究，曾参与国家自然科学基金、国家“十一五”科技支撑计划项目、国土资源部项目共5项，广东省自然科学基金2项，公开发表学术论文43篇。2011年起就职于深圳市房地产评估发展中心，具体从事深圳市土地整备、城市征收地拆迁评估的相关理论、政策与实践方面的研究工作。

薛月菊，女，1969年生，新疆乌苏人，教授，工学博士，博士生导师。主要从事数据挖掘研究及其在农业中的应用、无线传感器网络和RFID射频识别的应用和研究。主持国家级和省部级项目10余项，在国际国内重要学术期刊上发表论文30余篇，其中第一作者论文被SCI检索6篇、被EI检索14篇。荣获军队科技进步一等奖1项，广东省科学技术一等奖1项、二等奖1项，国防科技进步三等奖1项。

前　言

数据挖掘技术是在统计学、人工智能(特别是机器学习)、数据库技术等多种技术的基础上发展起来的多学科交叉新兴技术,它是随着数据的大量积累和人们对信息和知识的迫切需求而产生和发展起来的,并逐渐成为人们关注的热点。数据挖掘强调的是大数据量和算法的可伸缩性,是一门实用性很强的学科,一出现就引起许多领域和部门的重视及应用。根据调研,目前市场上尚没有系统介绍数据挖掘应用于土地科学方面的书籍,仅有一些期刊文章上的零星研究。随着土地科学,特别是土地资源评价工作的逐渐深入开展,各种复杂计算和模型建立问题越来越多,这对土地科研工作者提出了更高的要求。本书可以帮助土地科研工作者快速适应这一形势变化,有利于提高土地资源评价科研工作水平和促进我国土地资源评价的技术水平。

本书以广东省龙川县耕地分等定级数据库为基础数据来源,探讨数据挖掘技术在土地资源评价中的应用。全书主要内容为:简明阐述数据挖掘的概念和一般程序;设计并建立 Geodatabase 地理数据库模型;介绍因子分析、层次分析、模糊层次分析、灰色关联综合分析和 BP 神经网络方法在指标权值制定中的应用;引进抽样技术确定样本容量;应用 K-均值聚类分析、两步聚类分析、模糊综合评判、模糊聚类分析等方法,实现土地质量的无监督学习分类;采用 Fisher 判别分析、Logistic 多元回归分析、决策树、BP 神经网络、径向基概率神经网络、支持向量机等方法,对土地质量进行有监督学习分类;应用关联规则对评价指标间的关联特性进行挖掘,综合描述指标间存在的联系和相互作用。另外,还针对径向基概率神经网络和支持向量机方法编写了相关的计算程序,旨在提高本书的使用价值。

在本书编撰过程中,力图内容全面、层次清晰、文字简洁、通俗易懂,注重基本概念与方法应用,简化操作步骤,并应用 Clementine8.1、MatlabR2007a、SPSS13.0、DPS8.0、ArcGIS9.0 等软件及自编程序实现土地资源的数据挖掘和评价结果的可视化,使读者易于理解、易于操作、易于接受;内容选取和叙述形式不追求"理论的高深和数学推导的严谨",在学术性和实用性发生冲突时,学术性服从实用性。

本书引用和参照了许多相关资料或已有的研究成果,已在书中标明,如有遗

漏之处，恳请谅解！由于作者水平所限，书中不足之处在所难免，欢迎广大读者批评指正。

作　者

2012 年 8 月于广州

目　录

第1章 绪 论

1.1 引 言

我国土地面积辽阔，但却是世界上人均土地资源最少的国家之一。土地资源的稀缺性，要求我们必须“十分珍惜和合理利用每一寸土地，切实保护耕地”，充分挖掘土地资源的生产潜力。随着数据库技术的发展和土地资源评价数据来源的多样化，数据库中隐藏的丰富知识远没有得到充分的挖掘和利用。运用数据挖掘技术进行土地资源评价，对土地的质量与数量进行科学分析，对于土地资源评价方法研究、数据挖掘技术应用领域的拓展以及我国在2020年内守住18亿亩耕地这条红线，具有一定的理论与实践意义。

1.1.1 问题的提出

继第一次土壤普查(1958～1960年)与第二次土壤普查(1979～1990年)之后，中国农业部于“十五”期间组织开展了全国耕地地力调查与质量评价工作(2002～)。国土资源部在1999年开始了新一轮国土资源大调查(1999～2010年)，其中土地资源调查专题包括土地利用动态遥感监测、耕地后备资源调查与评价、农用地分等与定级估价等项目；并开展了新一轮土地利用总体规划修编(2006～2020年)和全国第二次土地调查(2007～2009年)等工作。开展这些工作的一个根本目的就是摸清我国的土地资源现状和未来发展状况，并对土地质量进行综合评价，实现土地质量的科学量化和空间分布的可视化，促进耕地保护和土地节约集约利用。然而，这需要解决好下面两个方面的问题：

(1) 土地资源评价基础数据的预处理。随着数据库技术的发展和人们获取数据手段的多样化，土地资源评价基础数据源可能来自多个数据库，这些数据源的结构和规则可能不同。同时，原始数据中存在大量的不完整的、含噪声的、不一致的、模糊的、随机的数据信息。这些都会影响土地资源评价的效率和结果的准确性，甚至产生一些无效归纳。将这些分散在各部门及地方基层单位的与土地资源有关的调查更新数据与珍贵的历史数据集成和融合，尚没有得到人们足够的重视。

(2) 土地资源评价的复杂性。构建一个土地资源评价模型来描述和刻画“既

要保护土地，又要增加农民收入、实现农民的土地财产权益，同时还要满足经济建设用地对土地需求，并保障土地生态环境的可持续发展”这样一个复杂过程，亟需土地资源评价体系的完善。

土地资源评价是一个多因素的综合评价，在选取评价指标和确定权值方面还存在不少问题。传统的评价方法主要从土地自然生产潜力的角度出发，在土地的自然属性中选取若干个指标因素构建指标体系，并对土地进行适应性评价及分区（Raúl et al.，2007；Buia et al.，2006；郑宇等，2005；张海涛等，2003；张友焱等，2003；邱炳文等，2002；王建国，2001；胡月明等，2001a）。近年来，社会条件、经济发展水平、土地利用对生态环境影响等因素在土地资源评价中受到越来越多的关注，指标选择已由过去偏重于自然属性向较为全面考虑自然与环境、生态、社会、经济等方面逐渐完善。评价方法已从简单的定性描述发展为定量分析为主、定性描述为辅，同时随着 3S 技术、计算机技术和数学模型的应用和研究的深入，评价方法和手段逐渐丰富起来，采用的方法有层次分析、灰色关联度分析、回归分析、聚类分析、判别分析、模糊综合评判、人工神经网络等。然而，无论采用哪种方法，评价指标的确定、指标等级的划分或指标权值的计算都是影响评价结果的重要因素。同时，如何确定合理的样本容量、如何选择适合的土地资源评价方法、如何比较不同评价方法的评价结果等棘手的问题都亟待解决。目前，应用专家经验来决定评价指标体系及指标的数量级别和权值不失为一种较为有效的方法（张凤荣等，2002），然而过多的人为因素难免会影响评价的可信度和准确性。为了降低评价过程中人为因素的影响成分，需要探索一些更加符合实际、准确高效的定量化评价方法，使评价结果更加科学、严谨。因此，挖掘有效的评价方法已成为当今土地资源评价工作的一个重要内容。

随着数据库技术、人工智能、机器学习、数理统计、知识工程、高性能计算、数据可视化、专家系统等技术的发展，数据挖掘技术在 20 世纪 90 年代有了突飞猛进的发展（Han et al.，2006），它以一种全新的概念改变着人们利用数据的方式（毛国君等，2006）。将数据挖掘技术应用于土地资源评价，能够较好地克服评价过程中人为因素过多的缺点，提高评价的效率、可信度和准确性，为开发新的评价方法提供了一个丰富的知识库。

1.1.2 研究意义

土地资源数据库的迅速膨胀，迫切需要新的方法和技术挖掘隐藏其中的丰富的知识，以指导人们更好地利用和保护土地资源。汲取数据挖掘的思想，引进数据挖掘方法对土地资源评价工作具有重要的理论和现实意义，具体表现在以下几个方面：

（1）提高土地资源信息表达的准确性。土地资源评价数据源的多样化，导致数

据杂乱，难以直接使用，即使是同一个数据源的数据，也可能存在重复和不完整的数据信息。因此必须对这些数据进行一定的预处理，保证数据的质量，提高数据挖掘的效率和结果的准确性。通常大量的原始数据记录了土地资源较为全面的属性信息，可能只需要其中一部分的属性就可以获取希望知道的知识，而其他的冗余属性信息的增加会导致无效的数据归纳，可能把数据挖掘结果引向错误的方向。

(2) 促进地理信息系统(geographic information system，GIS)可视化与数据挖掘技术的融合。目前，ArcGIS、MapInfo、MapGIS 等国内外 GIS 软件偏重于空间数据的存储、管理、制图，其分析方法基本上是以空间位置为核心的(韦玉春等，2005)，这为土地资源评价结果的空间位置可视化提供了平台。同时，当前的 GIS 技术对地理的信息表达基本上是静止的(刘湘南等，2005)，而土地信息往往具有一定的空间特征，即存在一定的模糊性、不准确性、不完整性等特征。应用数据挖掘技术，将地块的空间位置作为土地数据的属性来考虑，分析各个要素之间的关系以及这种关系的表现形式，可以弥补 GIS 在空间分析，特别是在属性数据分析方面存在的不足。因此，采用 GIS 和数据挖掘技术相结合的方法对土地资源评价无疑是一种很好的选择。

GIS 本身就是一个优秀的数据挖掘工具，同时也为数据挖掘提供了良好的平台。随着数据挖掘技术的发展，将一些新的数据挖掘技术应用于 GIS，可促使 GIS 在以下几个方面得到较大的发展或突破(蒋良孝等，2003)：

① 使有限数据的 GIS 成为具有无限知识的 GIS。尽管 GIS 中存储了大量的数据，但其容量总是有限的，它是对客观世界的不完全描述。而数据挖掘技术能从这些有限的数据库中发现新的知识，并将这些知识反作用于已有的数据，如此循环的知识发现，使 GIS 不仅是一个信息系统，还成为不断更新的数据源和知识源，将 GIS 静态数据变成了动态的数据和知识。

② 促进 GIS 的数据精练。GIS 数据库中存储了大量的数据，其中有些数据是必需的，也有些数据是冗余的。利用数据挖掘技术，可以寻找出数据间的相互依赖性，得到数据间的层次和层次间的相互关系。因而，数据库中就可只存储那些必需的数据和关系，而不必存储其他的数据。这样不仅节省了存储空间，而且可以提高数据库的管理效率和整个系统的运行速度。

③ 可用于 GIS 的数据更新。现有的 GIS 数据库中存储了描述客观世界的大量数据，而客观世界在人类活动的影响下是时刻变化的，如何将这些变化在 GIS 中进行快速地更新，也是一个十分棘手的问题。应用数据挖掘技术中的空间分析方法可以解决此问题，它通过对不同时域的数据进行比较，得到事物随时间变化的规律，并找到影响此变化的主要因子。这样，在以后的分析中，检查这些主要的因子是否变化，若有变化，则进行数据更新，否则就不予考虑。

④ 使 GIS 成为真正的“智能”空间信息系统。在 GIS 中引入专家系统技术

后，GIS具有了一定的自动性和智能性，但它远不能称为一个真正的“智能”系统，因为它不具备自动学习的能力。应用数据挖掘技术，使得GIS能自动地获取知识而可能成为真正的“智能”系统。

（3）较好地推进数据挖掘技术在土地资源评价中的应用。目前，市场上还很难找到一本较为系统地研究GIS支持下的基于数据挖掘的土地资源评价方面的专著，本书通过对土地资源评价中基础数据的预处理、数据库的建立、指标权值的制定、样本容量的计算、评价方法的应用和模型的建立等问题的研究，较好地实现GIS、数据挖掘、土地科学等学科在土地资源评价中的融合。

（4）丰富土地资源评价的方法，为今后土地资源评价工作提供一些新的思想和技术手段。本书介绍了抽样技术、两步聚类、模糊层次分析、关联规则、决策树、神经网络模型、支持向量机等数据挖掘方法，并实现了它们在土地资源评价中的应用。同时，分析了传统评价方法中可能存在的不足，并对其进行了一定的改进。

1.1.3 研究内容及技术路线

1. 研究内容

采用不同的数据挖掘方法，土地资源评价结果往往有所不同，这需要充分考虑数据挖掘方法的理论基础、评价模型和适宜范围。因此，本书的研究内容主要有以下几个方面：

（1）土地资源评价基础数据预处理。土地资源评价基础数据来源于多个数据库或数据仓库，这些数据源的结构和规则不尽相同，同时，基础数据普遍存在不完整、含噪声、不一致、重复、维度高等问题，如何对数据进行清洗、集成、变换、归约化等处理？

（2）土地资源评价指标权值的制定。指标权值表达了指标对土地质量的影响程度，如何有效地克服常规定量方法（如因子分析和灰色关联度分析）制定的指标权值存在离散性不强、过多人为主观因素判断、权值线性化等问题？

（3）土地资源评价样本容量的确定与样本的自动化抽取。传统评价中样本容量大小无所适从，如何有效地确定样本容量，以及如何从土地资源评价数据库中自动化抽取有代表性的样本记录？

（4）土地资源评价方法的研究。传统的土地质量分类采用无监督学习分类的模式，然而，如何通过对数据库中提取出的样本容量数据（学习样本）进行学习、建模，实现土地资源评价单元（简称土地评价单元，视为测试样本）的有监督学习分类？

2. 技术路线

土地资源评价数据挖掘工作主要包括：资料准备阶段、外业补充调查、数据预

处理、创建数据库、数据挖掘、结果分析与展望6个部分。其中，资料准备阶段的内容是收集并整理与土地资源评价有关的自然条件、社会经济、图件等信息；外业补充调查主要包括野外调查线路设计，调查点设置，调查和核实评价单元的定性描述和定量记录；数据预处理阶段的工作包括坐标变换、数据转换、图形叠加、缓冲区分析、专题地图转绘、统计图表整理等，以消除原始数据中不完整的、含噪声的、不一致的、模糊的、随机的数据信息；创建数据库包括表格、文本文件等属性数据的录入以及图形数据入库，构建属性数据库和空间数据库；数据挖掘阶段应用抽样技术和因子分析、判别分析、两步聚类、多分类Logistic回归模型、层次分析、模糊层次分析、模糊聚类、模糊综合评价、灰关联综合评价模型、关联规则、决策树、BP神经网络(back propagation net-work，BP)、径向基概率神经网络与支持向量机等方法，解决土地资源评价中样本容量的确定问题、土地质量的无监督和有监督学习分类的问题，具体流程如图1.1所示；结果分析与展望部分主要从评价方法的数学理论及其在土地资源评价应用中的优缺点方面出发，对指标权值和评价结果进行归纳和分析，然后对数据挖掘技术在土地资源评价中的应用前景进行展望。

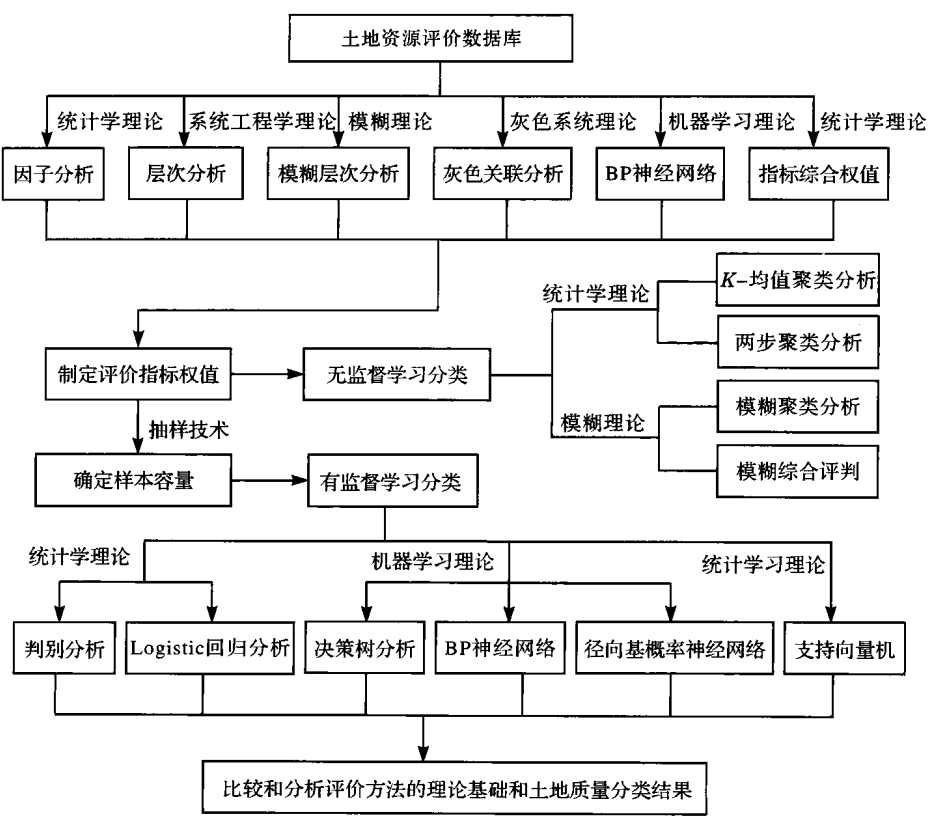

图1.1 土地资源评价数据挖掘流程图

1.2 土地资源评价数据挖掘概述

1.2.1 基本概念

1. 土地资源评价的概念

通常情况下，土地与土地资源、土地(质量)评价与土地资源评价这两组名词是等同起来使用。国内外关于土地(资源)评价的定义主要有：①联合国粮农组织(Food and Agriculture Organization of the United Nations，FAO)对土地资源评价的定义是“当土地用于特定的用途时，对土地特性进行估计的过程”(刘黎明，2007)；②林增杰(1986)认为，土地资源评价是在特定目的之下，对土地生产力高低的鉴定、评定或估价；③傅伯杰(1989)认为，土地资源评价是估价土地生产力和适宜性的过程；④戴旭(1995)认为，土地资源评价是估价土地生产潜力和土地适宜性的过程；⑤倪绍祥(2005)认为，土地资源评价主要是根据土地的自然生产能力或其他方面利用潜力的高低对土地的质量作出评估；⑥刘鑫(2005)认为，土地资源评价是一门揭示土地在各种用途条件下土地质量高低及其在区域内空间分布规律的科学；⑦刘黎明(2007)认为，土地资源评价是指为了一定的目的，在一定的用途条件下，对土地质量的高低或土地生产力的大小进行评定的过程。

综合上述专家的阐述，土地资源评价应具有特定的土地用途条件、评价的时效性、参评指标的全面性、研究结果的指导性等特点。因此，我们结合土地资源评价的定义和特点，对农业用途的土地质量与土地资源评价的理解是：农用土地质量是指在一定的社会历史发展条件下，充分考虑与土地生产有关的自然、生态环境、社会经济等因素，综合评判土地在特定种植条件下的粮食生产能力、农业收益能力和未来用途条件下的潜在收益能力以及土地生产对生态环境的影响程度；在农用土地质量定义的基础上，农用土地资源评价可以简单地概括为根据农用土地的质量状况评定土地等级或价格的过程。过去人们所开展的土壤肥力评价、土壤质量评价、耕地地力调查与质量评价、土壤适宜性评价、土地潜力评价、农用地分等定级与估价等工作均属于农用土地资源评价的范畴。

2. 数据挖掘的概念

数据挖掘(data mining，DM)是一门融合了统计学理论、数据库技术、人工智能、机器学习、模式识别、可视化技术、专家系统等领域技术成果的交叉学科。它借助统计学理论、人工智能等领域的研究成果构建起自己的理论体系；利用数据库技术对数据进行前端处理；应用机器学习的方法从处理后的数据中提取有用的

知识，并对数据背后隐藏的特征和趋势进行分析，挖掘数据的总体特征和发展趋势；运用可视化技术将人的观察力和智能融入系统，以图形形式将信息模式、数据的关联或趋势呈现给用户，用户能够交互地分析数据。

关于数据挖掘的定义，不同学科领域的专家对其理解和定义不完全相同，其中一个比较公认的定义是："数据挖掘是基于数据库平台，对数据进行一定的处理，从大量的、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、人们事先未知的、但又是潜在有用的信息和知识的过程，为科学的分析提供决策支持"(Han et al.，2006)。

1.2.2 数据挖掘的产生

随着信息技术特别是网络技术的飞速发展，人类活动中每一个领域都充斥着大量的统计观测数据。数据出现了爆炸性增长，信息的增长呈现超指数上升。据估计，全世界的信息量不到20个月就增加一倍(贺昌政，2005)，与此形成鲜明对比的是，对决策有价值的知识却非常匮乏。例如，在土地资源评价过程中，需要收集大量的土地资源调查数据和散落在各个部门的与土地质量有关的信息，同时遥感(remote sensing，RS)和全球定位系统(global positioning system，GPS)能够便捷地提供大量的对地观测数据，促使土地资源评价数据库的迅速膨胀。如果仅仅依靠传统的数据检索机制、统计分析方法和GIS技术，难以满足评价工作的需要，尽管现代计算机与数据库技术可以支持存储并快速检索如此巨大规模的数据集，但无论在时间意义上还是空间意义上，人们都困于如何更好地理解和利用这些数据。由此导致越来越严重的"数据灾难"，数据增长与数据分析之间出现了越来越大的距离，传统方法的失效和"数据灾难"是数据挖掘产生的原因，数据和信息之间的鸿沟在迫切地推动着数据挖掘技术的发展和应用。

1.2.3 数据挖掘过程

数据挖掘是一个知识发现的过程，从实际应用的角度讲，数据挖掘过程包括：确定应用目标，了解应用领域相关的先验知识，选择目标数据集，进行数据清理、转换、归约，确定数据挖掘任务并选择挖掘算法，挖掘与搜索有价值的模式，解释、分析、可视化挖掘结果，应用所发现知识等活动。这些活动大致主要包括：目标的确定、数据的准备(如数据的选择、数据的预处理、数据的转换)、数据挖掘、结果分析和知识的同化，具体过程如图1.2所示。

1.2.4 数据挖掘分析模型

数据挖掘利用复杂的分析与建模技术发现隐藏在数据库中的模式和关系，其本质是对数据建立模型，以获得更为简洁的表达，即模式。数据挖掘模型按其功