多源遥感数据挖掘系统技术框架

宫辉力[1)] 赵文吉[1)] 李 京[2)]

[1)](首都师范大学资源环境与 GIS 北京市重点实验室，北京 100037) [2)](北京师范大学资源信息工程中心，北京 100875)

摘 要 陆地资源卫星源源不断地把遥感数据传输至地面，卫星地面接收站积累了海量的卫星遥感数据。遗憾的是，由于缺乏针对遥感数据的有效的数据挖掘和知识发现技术，致使遥感数据中的绝大部分信息没有得到充分的利用。对传统的数据挖掘和知识发现技术进行技术革新和改造，研究针对多源遥感图像的数据挖掘和知识发现技术，不仅可以提高遥感解译的自动化和智能化水平，而且可最大限度地开发和利用遥感信息。为了能充分利用遥感数据，在传统数据挖掘和知识发现技术的基础上，首先探讨了遥感数据挖掘和知识发现的技术流程，然后设计了多源遥感图像数据挖掘系统框架，最后提出了多源遥感图像数据挖掘系统的原型，从而为进一步开发和研制多源遥感数据挖掘系统奠定了技术基础。

关键词 遥感图像 数据挖掘 知识发现 遥感信息

中图法分类号：P208 TP18 **文献标识码**：A **文章编号**：1006-8961(2005)05-0620-04

The Technological Framework of Data Mining from the Polygenetic Remotely Sensed Data

GONG Hui-li[1)], ZHAO Wen-ji[1)], LI Jing[2)]

[1)](*The Key Laboratory of Resource Environment and GIS of Beijing, Capital Normal University, Beijing* 100037)

[2)](*Engineering Center of Resource Information, Beijing Normal University, Beijing* 100875)

Abstract The Land Resource Satellites continuously send the remotely sensed data to the earth, so large quantity of data has been acquire by the satellite receiving ground station. Unfortunately, because of lacking the effective techniques for data mining and knowledge discovering from the polygenetic remotely sensed data, the majority of the useful information existed in the remotely sensed data not has been fully exploited. Researching the data mining and knowledge discovering technologies, especially suitable to the polygenetic remotely sensed data, through innovating and improving the traditional data mining and knowledge discovering technologies, can promote the level of automatically and intelligently interpreting the polygenetic remotely sensed data, and exploit the useful information existed in the remotely sensed data as much as possible. Basing on the traditional data mining and knowledge discovering technologies, the authors had investigated the technology flowing chart of the remotely sensed data mining and knowledge discovering procedure, and then designed the technological framework of the remotely sensed data mining and knowledge discovering system, finally proposed the prototype of the system. So the technological foundation for further developing the polygenetic remotely sensed data mining system had been established.

Keywords remotely sensed image, data mining, knowledge discovery, remotely sensed information

1 引 言

随着卫星遥感技术的飞速发展，全天候、多光谱、多时相、多分辨率和多传感器的遥感卫星对地观测数据被不断地输送到地面，促进了遥感资料在各个领域的应用研究。近 20 年来，我国的遥感技术应用研究主要是在地学思维引导下进行的，而遥感应

基金项目：国家自然科学基金项目(70073045)；国家重点基础研究发展规划“973”项目(G19990346-06)；国家高技术研究发展计划“863”项目(2002AA134074)；国家高技术研究发展计划“863”项目(2003A135010)

收稿日期：2004-08-30；**改回日期**：2004-12-07

第一作者简介：宫辉力(1956～)，男，教授，博士生导师。首都师范大学资源环境与旅游学院院长，1996 年获长春科技大学水资源管理博士学位。目前主要从事空间信息技术在地下水资源、环境与旅游领域的应用研究。E-mail：gonghl@263.net

用水平则远远滞后于空间遥感技术的发展。这种滞后突出表现在:日益积累的多源卫星遥感数据中蕴涵的有用信息没有得到充分的挖掘和利用,其主要原因是遥感数据的接收、处理、信息提取都依赖于国外少数几个大型软件(如 PCI、ERDAS 等),而国内尚未开发出成型的商用遥感信息处理软件。由于国外软件的高昂价格和使用上的局限性,使国内 80% 以上的遥感信息用户(尤其是地方生产、科研单位)始终停留在对原始图像的目视解译上,从而造成遥感信息分析与提取技术的滞后,使遥感数据中隐藏着的丰富知识远远没有得到充分有效的利用,这就造成了遥感信息资源的巨大浪费及其应用价值的降低,并最终导致了我国目前的这种大量遥感数据积累而有用信息却相对匮乏的局面,因此,多源遥感信息的提取能力与效率问题已经构成了充分利用遥感信息的瓶颈问题[1,2]。

数据挖掘(datamining, DM)与知识发现(knowledge discovery from databases, KDD)技术出现于 20 世纪 80 年代末期,是人工智能、机器学习与数据库技术相结合的产物,其主要用于从商业数据库和数据仓库中,即通过提取有用知识来支持高层管理部门的决策[3]。但由于它不是简单地从数据库管理系统检索和查询信息,而是从数据库中发现隐含的、先前不知道的潜在有用信息,因此是一种从数据中鉴别有效模式的非平凡过程,且该模式是新的、可能有用的和最终可理解的。数据挖掘的目的是把大量的原始数据转换成有价值的知识,因为借助于数据挖掘与知识发现的理论和技术,将有助于解决多源遥感数据急剧增长而人们又缺乏有效技术手段从遥感数据中提取有用信息的问题。

数据挖掘是知识发现(KDD)的重要环节,其属于机器学习的范畴,也是数据库与人工智能技术相结合的产物,而且知识发现过程是多个步骤相互连接、反复进行的人机交互过程,其具体包括:

(1) 学习某个应用领域,包括应用中的预先知识和目标;

(2) 建立目标数据集,在选择的一个数据集或多数据的子集上聚焦;

(3) 数据清理和预处理,去除噪声或无关数据,去除空白数据域,考虑时间顺序和数据变化等等;

(4) 数据换算和投影,找到数据的特征表示,用维变换或转换方法减少有效变量的数目或找到数据的不变式;

(5) 选定数据挖掘功能,决定数据挖掘的目的;

(6) 选定数据挖掘算法,用 KDD 过程中的准则,选择某个特定数据挖掘算法(如汇总、分类、回归、聚类等),用于搜索数据中的模式,且该算法可以是近似的;

(7) 数据挖掘。搜索或产生一个特定的感兴趣的模式或数据集;

(8) 解释。解释某个发现的模式,去除多余不切题意的模式,并将其转换成某个有用的模式,以便于用户应用;

(9) 发现知识。把这些知识结合到运行系统中,以获得这些知识的作用或证明这些知识;用预先可信的知识检查、解决知识中可能的矛盾。

可见,知识发现和数据挖掘的目的是把大量的原始数据转换成有价值的知识。Piatet 提出了数据、信息和知识之间的金字塔关系[4]。这有别于数据库管理系统检索和查询出的信息,通过数据开采发现的是隐含的、精炼的和高层次的知识。陈文伟归纳总结了 10 种数据挖掘和知识发现的方法和技术:(1)决策树方法;(2)神经网络方法;(3)覆盖正例排斥反例方法;(4)粗糙集(rough set)方法;(5)概念树方法;(6)遗传算法;(7)公式发现法;(8)统计分析方法;(9)模糊集合论方法;(10)可视化技术[5]。

数据挖掘和知识发现是目前国内外研究的热点问题,它既是人工智能学者的研究热点,也是数据库专家的探索对象,其研究工作涵盖了医学、机器学习、人工智能、数学、市场营销等诸多领域,并已获得了多门类的有用知识。迄今为止,不仅国内开展这方面研究的专家还不多,而且把 KDD 和 DM 技术应用于卫星遥感的信息处理,更是一项崭新的课题。

2　多源遥感数据挖掘技术背景

数据挖掘是一种从大型数据库或数据仓库中提取隐藏的预测性信息的技术,它能挖掘出数据间潜在的模式(pattern),并能找出有价值的信息和知识(knowledge),用于指导实际应用或辅助科学研究。其中,模式是利用挖掘算法得到的结果,是对一种可能性分布的简单描述;知识或信息则是通过对模式进行处理而得到的易于理解和应用的结果。

通常数据挖掘分为预测型(predictive)和信息

型(informative)两种类型模式,也称为监督型(supervised)和非监督型(unsupervised)模式。挖掘过程可分为证明驱动(verification-driven)和发现驱动(discovery-driven)两种类型。预测型的模式是通过输入集合值来计算某一属性,或某几种属性的值,预测型的模式可用来解决一个指定的问题,如从数据库中的一些属性来预测另外一个或多个属性值。它的重要特征是利用已知的属性值去合理地猜测一个未知的属性值。信息型的模式是用于预测将来要发生的事情。信息型模式不解决某一个指定问题,而是提供给某领域的专家以前可能不知道的有兴趣的模式。信息型模式比预测型模式难评估,因为它们的价值在于其能提供给某领域专家的一些建议和这些建议的有效性。数据挖掘工具是通过预测未来趋势及行为,为应用实践做出前摄的(procative)、基于知识的决策。在典型的决策支持系统中,数据挖掘可自动提供对未来情况的分析结果,这远远超过传统工具所提供的历史情况分析。数据挖掘技术利用并发展了数据存储技术和实时数据导航技术,它由以下 4 个成熟技术支持:(1)大规模数据采集;(2)功能强大的并行处理机;(3)数据挖掘算法;(4)数据库技术。

由此可见,数据挖掘是一个集多种领域知识为一体的综合技术,它涉及了统计学、机器学习、人工智能、不确定性理论、数据库、知识获取、模式识别、信息抽取、可视化、分布式多媒体环境的智能代理、数字图书馆(digital libraries)和管理信息系统等诸多技术领域。

3 多源遥感数据挖掘技术流程

多源遥感数据挖掘过程有数据准备、挖掘算法和结果表达等几个阶段。数据挖掘技术流程可分为以下几个步骤(图 1):

(1) 理解和定义问题;
(2) 多源遥感数据的搜集和抽取;
(3) 遥感数据净化;
(4) 遥感数据挖掘引擎;
(5) 遥感数据算法引擎;
(6) 运行数据挖掘算法;
(7) 评估结果;
(8) 重新精化数据和问题;
(9) 结果应用。

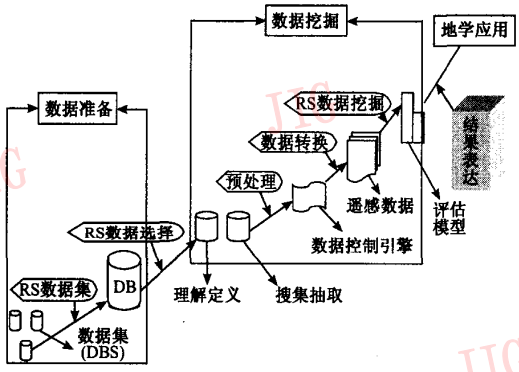

图 1 多源遥感数据挖掘技术流程图

Fig. 1 The flow chart of multisource remote sensing datamining technique

4 遥感数据挖掘系统原型框架设计

4.1 技术框架

遥感数据库作为数据库的一种,若对蕴涵其中的信息进行处理与识别,则自然可以借鉴一般意义上的 DM 和 KDD 技术,但是作为一类特殊的数据库——图像数据库,有着区别于一般关系数据库和事务数据库的信息内容,且数据库中隐含着丰富的时间、光谱和空间信息,因而,就这类数据库中的知识发现而言,数据挖掘与知识发现的过程和方法也应具有特殊性。

针对卫星遥感数据的特殊性,笔者在 DM 和 KDD 技术流程基础上,提出了面向地学应用的卫星遥感数据挖掘和知识发现的系统框架(图 2)。

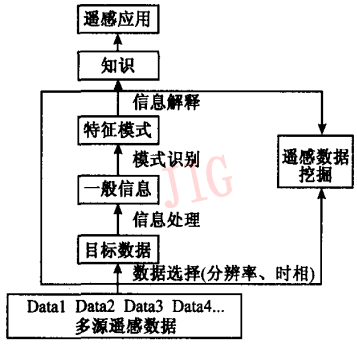

图 2 遥感数据挖掘与知识发现系统框架

Fig. 2 Systems frame work of remote sensing datamining and knowledge discovery from database

在此框架中,数据挖掘占据了极为重要的地位,它包括遥感数据的时相选择、应用预处理、特征分

析、信息识别与知识解释等。但由于现实生活中,许多遥感信息应用者忽略了该过程的特殊作用,直接把原始遥感图像的解释结果作为应用的基础(虽然在解译过程中也加入了人的知识),因而获得的知识往往是肤浅的、表面化的和不精确的。笔者认为,只有充分考虑原始数据的波谱、空间和时间特征,遥感数据挖掘过程才能更好地实现针对遥感应用的有价值的、较精确的和高水平的知识发现。

4.2 总体结构

在遥感数据挖掘和知识发现技术框架基础上,参照 Christophe 和 SIlapro 等人描述的 KDD 参考模型[8],笔者提出了遥感数据挖掘原型系统。该系统可由用户通过交互界面来实现各种功能。该系统由图像数据接口、数据可视化、数据预处理、数据挖掘控制、原型编辑、知识编辑、原型库、知识库、结果显示等几部分构成(如图 3 所示),其中主要包括:

(1) 多源遥感图像预处理　包括数据转换与统一、数据分类、镶嵌和纠正。

(2) 数据挖掘工具　包括多源遥感数据融合、多光谱分析(光谱特征的建立和修改、散点图生成、报告输出及混淆矩阵、分析与增强工具)、图像分割识别(基于知识的专家分类、神经网络分类器、3 维可视化、地形分析、矢量数据的数字化与编辑、矢量-栅格数据的相互转化)。

(3) 结果表达与存储　包括图像管理、可视化表达、语言表达。

(4) 多源遥感数据接口设计　接受不同格式的图像数据(非标准数据格式、无标签数据、其他软件数据格式,如 PAMAP, ERDAS, ARC/INFO, TIFF 等),接受不同矢量格式数据,自定义格式。

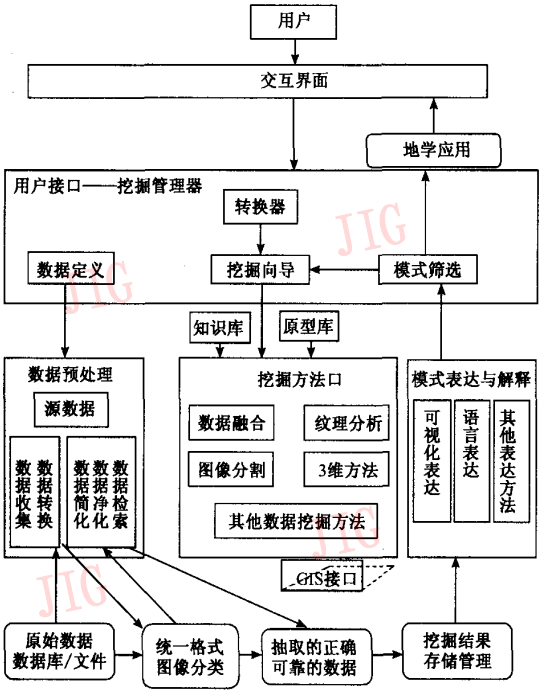

图 3　遥感数据挖掘系统原型

Fig. 3　The pretotype of remote sensing datamining system

5 结　论

多源遥感图像数据挖掘是比较新的研究课题,本文只是初步探讨了遥感数据挖掘与知识发现的技术流程,而系统原型的实现,还有一些关键技术需要进一步讨论。

参考文献(References)

1 (美) Kenneth R. Castleman 著. Digital image processing [M]. Beijing: Electronic and Industrial Press, 2000. [朱志刚,林学闻,石定机等译. 数字图象处理[M]. 北京:电子工业出版社,2000.]

2 Liu Yu-jie, Yang Zhong-dong, *et al*. MODIS remote sensing information processing theory and algorithm [M]. Beijing: Science Press, 2001. [刘玉洁,杨忠东等. MODIS 遥感信息处理原理与算法[M]. 北京:科学出版社,2001.]

3 Jiawei Han, Micheline Kambr. Data mining [M]. Xian: Xian Jiaotong University press. 2001. [Jiawei Han, Micheline Kambr. 候迪,宋擒豹译. 数据挖掘[M]. 西安:西安交通大学出版社,2001.]

4 Flury B. A first course in multivariate statistics (Section 7.5: Simple Logistic Regression) [M]. New York: Springer-Verlag, 1997.

5 Cheng Wen-wei. Development of decision support system [M]. Beijing: Tsinghua University Press, 2000. [陈文伟. 决策支持系统及其开发[M]. 北京:清华大学出版社,2000.]

6 Griffith D. Statistical and mathematical sources of regional science theory: Map pattern analysis as an example [J]. Papers in Regional Science, 1999, **78**(1):21 ~ 45.