数据挖掘方法及天体光谱挖掘技术

赵旭俊◎著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

数据挖掘方法及天体光谱挖掘技术

赵旭俊 著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

数据挖掘是一门面向应用的新兴学科分支，涉及人工智能、机器学习、模式识别、统计学、数据库、可视化等多个学科领域，其主要目的是从大量原始数据中提取人们感兴趣的、隐含的、尚未被发现的信息和知识，目前已广泛应用于科学、工程、商业、医学等领域。

本书适合从事天文学研究、数据挖掘及知识发现和人工智能等领域的技术人员阅读，也可以作为高等院校天文学、计算机科学与技术等学科的高年级本科生及研究生的学习参考书。

图书在版编目（CIP）数据

数据挖掘方法及天体光谱挖掘技术/赵旭俊著. —北京：电子工业出版社，2013.6
ISBN 978-7-121-20532-3

Ⅰ. ①数… Ⅱ. ①赵… Ⅲ. ①数据采集－研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2013）第 111696 号

策划编辑：赵　娜
责任编辑：谭丽莎
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1 000 1/16 印张：11.75　字数：240 千字
印　　次：2013 年 6 月第 1 次印刷
定　　价：39.80 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

随着 LAMOST 望远镜的正式投入使用，获取的光谱急剧增多。据统计，每晚将有 2 万～4 万条光谱需要进行自动分类识别及参数测量，如何快速、准确地处理海量天体光谱成为瓶颈。数据挖掘是一门面向应用的新兴学科分支，涉及人工智能、机器学习、模式识别、统计学、数据库、可视化等多个学科领域，其主要目的是从大量原始数据中提取人们感兴趣的、隐含的、尚未被发现的信息和知识，目前已广泛应用于科学、工程、商业、医学等领域。因此，采用数据挖掘作为天体光谱数据的分析方法是可行的、有价值的。

本书是作者近年来科研成果的总结。全书共 5 章，在绪论之后，全书可以分为以下 3 部分。

（1）关联规则挖掘方法及应用，包括第 2 章。这一部分提出了基于准频繁项目集的关联规则挖掘、基于背景知识的关联规则挖掘、约束 FP-tree 及其构造方法、基于信息熵的加权频繁模式树构造共 4 个算法，用来解决关联规则挖掘中效率较低、扫描数据库次数较多、背景知识无法直接给出等问题。同时，将这几个算法用在天体光谱的数据处理中，实现了天体光谱属性之间的相关性分析，为探索新的天体规律提供了技术支持。

（2）离群数据挖掘方法及应用，包括第 3 章。这一部分提出了基于距离支持度的离群数据挖掘、基于分阶段模糊聚类的离群数据挖掘、基于信息熵的离群数据挖掘、基于特征属性子空间的离群数据挖掘共 4 种算法，从而提高了离群挖掘的效率及准确率，同时实现了天体光谱数据的离群挖掘，为发现未知天体提供了一种新的方法。

（3）天体光谱数据的其他挖掘方法及天体光谱数据挖掘原型系统，包括第 4 章和第 5 章。这一部分介绍了天体光谱数据的正、负项目集挖掘、基于约束概念格的恒星光谱分类规则提取、恒星光谱的分类规则后处理等方法，之后给出了几个天体光谱数据挖掘原型系统，介绍了系统的功能模块、体系结构，以及系统运行的相关界面。

本书的完成得到了太原科技大学人工智能实验室、计算机科学与技术学院各位同人的大力支持，尤其是张继福教授、蔡江辉博士为本书提出了很多很好的建议，在此一并致以诚挚的谢意。

本书所涉及的部分研究工作得到了山西省青年科学基金项目(项目编号：2012021015-4）和山西省高校高新技术产业化项目（项目编号：20121011）的资助，在此谨向山西省自然科学基金委员会和山西省教育厅表示深深的感谢并致以敬意。

由于作者的水平有限，书中难免有不妥之处，恳请各位专家和广大读者给予批评指正。

编　者

2013 年 5 月

目　录

第1章 绪 论

随着数据库和计算机网络的广泛应用，数据处理领域面临两方面的难题。一方面是数据雪崩：现实世界中产生的数据量呈指数级增长，人们所拥有的信息量急剧增大，超大规模的数据集与日俱增，待处理的海量数据层出不穷，信息量远远超过了人脑掌握、消化的能力，这就是数据雪崩。另一方面，先进的观测技术和现代监测仪器的推广和应用使我们的监测范围更加广泛，随着数据维度的增加，许多数据分析变得非常困难，特别是随着维度的增加，数据在它所占据的空间中越来越稀疏。对于分类，这可能意味着没有足够的数据对象来创建模型，将所有可能的对象可靠地指派到一个类；对于聚类，点之间的密度和距离的定义（对聚类而言是至关重要的）失去了意义，这就是“维灾难”。

如此庞大的信息量已经远远超过了人脑可以驾驭的范围，传统的人工处理方法已经无法处理和利用如此大规模的海量、高维数据，更无法快速、准确地从中获取有用知识，传统的数据库技术和数据处理手段也已经不能满足要求。由于人们迫切需要将这些数据转换成有用的信息和知识，所以如何从海量、高维数据中快速提取有用信息已成为亟待解决的问题之一。正是基于这样的需求，数据挖掘技术受到了广泛关注，并得以快速发展。

1.1 数据挖掘

数据挖掘（Data Mining）是一个从大量的数据中发现潜在知识的过程，是半自动或自动地从海量数据中发现模式、相关性、变化、反常规律性的过程。根据挖掘任务划分，数据挖掘主要发现五类知识：广义型知识（Generalization）——根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、微观或宏观的知识；分类型知识（Classification）——反映同类事物共同性质的特征知识和不同事物之间差异型的特征知识，用于描述数据的汇聚模式或根据对象的属性区分其所属类别；关联型知识（Association）——反映一个事件和其他事件之间依赖或关联的知识，又称

为依赖（Dependency）关系，这类知识可用于数据库的归一化、查询优化等；预测型知识（Prediction）——通过时间序列型数据，由历史的和当前的数据去预测未来的情况，它实际上是一种以时间为关键属性的关联知识；偏差型知识（Deviation）——离群数据（孤立点）的挖掘（Outliers Mining），通过分析标准类外的特例、数据聚类外的异常值、实际观测值和系统预测值间的显著差别，来对差异和极端特例进行描述。

1.1.1 产生和定义

随着数据库和计算机网络的广泛应用，人们所拥有的数据量急剧增大，海量数据层出不穷。先进的现代科学观测仪器的使用造成每天都要产生巨量的数据，如我国建成的 LAMOST 望远镜，每晚将有 2 万～4 万条光谱需要进行自动分类识别及参数测量。显然，大量信息在给人们带来方便的同时也带来了一系列问题，如信息量过大，超过了人们掌握、消化的能力；一些信息的真伪难辨，从而给信息的正确运用带来了困难；信息组织形式的不一致性导致难以对信息进行有效统一处理等，这种变化使得传统的数据库技术和数据处理手段已经不能满足要求。如何在海量数据中获取有价值的信息和知识成了信息系统的核心问题之一。数据挖掘正是为了解决这一问题，并针对大规模数据的分析处理而出现的。

数据挖掘就是从大量原始数据中提取人们感兴趣的、隐含的、尚未被发现的、有用的信息和知识，使它们可以有利于为专家进行决策提供技术支持，其提取的知识可以表示为概念、规则、规律、模式等形式。数据挖掘是当今数据库和人工智能相互结合的最前沿和极富应用前景的研究领域，已引起了国内外众多学者和业界的高度重视，他们已对数据挖掘的方法论、理论和工具开展了广泛深入的研究工作。由于数据挖掘获取的信息和知识可以广泛地应用于生物医学和 DNA 分析、银行与金融机构、零售业、电信业、商务管理、市场分析和企业决策管理等领域，所以数据挖掘技术引起了信息产业界的极大关注。目前，国内外学者已研究和开发出了一些数据挖掘系统，比较有代表性的通用数据挖掘系统有 IBM 公司的 Almaden 研究中心开发的 Quest、加拿大 Simon Fraser 大学开发的 DBMiner、SGI 公司和美国 Standford 大学联合开发的 MineSet、南京大学开发的 Knight 原型工具等。一个典型的数据挖掘系统可以由以下几个主要成分组成（如图 1-1 所示）。

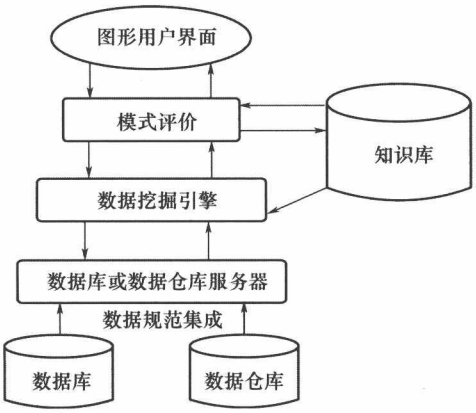

图 1-1　数据挖掘系统的组成图

数据库、数据仓库或其他信息库：这是一个或一组数据库、数据仓库、电子表格或其他类型的信息库。可以在数据库上进行数据的清理和集成。

数据库或数据仓库服务器：根据用户的数据挖掘请求，数据库或数据仓库服务器负责提取相关数据。

知识库：这是领域知识，用于指导搜索或评估结果模式的兴趣度。

数据挖掘引擎：这是数据挖掘系统的基本部分，由一组功能模块组成，用于特征化、关联、分类、聚类分析，以及演变和偏差分析。

模式评价：通常，此成分使用兴趣度度量，并与数据挖掘引擎模块交互，以便将搜索聚焦在有趣的模式上。

图形用户界面：本模块在用户和数据挖掘系统之间进行通信，允许用户与系统交互，制定数据挖掘查询任务，提供信息，帮助搜索聚焦，根据数据挖掘的中间结果进行探索式数据挖掘。

1.1.2　挖掘的过程

数据挖掘的一般过程可用图 1-2 进行描述。

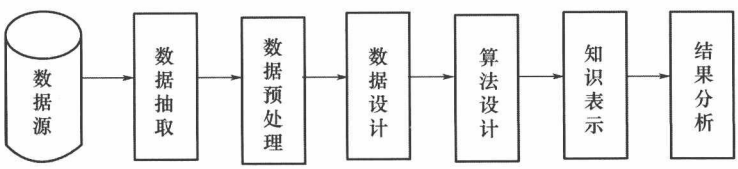

图 1-2　数据挖掘的一般过程

1. 数据抽取

大多数时候，与数据挖掘任务有关的数据是存储在应用数据库中的，这些数据库往往是为应用的目的而建立的，通常不能直接运行数据挖掘算法，而需要进行必要的抽取和格式的整理工作。

2. 数据预处理

数据预处理是指处理掉一些噪声数据（冗余的、不一致的）或添补一些丢失的数据，以便使被挖掘的数据保持完整和干净。

3. 数据设计

数据设计是指去掉一些无关的属性或对数据量过大的数据库进行抽样等。

4. 算法设计

算法设计主要是指针对特定的挖掘任务，设计挖掘方法模型与高效的算法和相应的数据结构。

5. 知识表示

知识表示是指从数据库或数据仓库中获取特定的知识类型，如分类、关联规则、聚类和序列模式等。

6. 结果分析

结果分析是指由领域专家（Domain Expert）分析结果的可靠性、合理性及可用性，有时还需要对结果进行可视化处理。

从图 1-2 可以看出，数据挖掘的核心步骤是算法设计，一个好的数据挖掘模型、一个好的算法（速度快、伸缩性好、结果容易使用且符合用户的特定需求）是影响数据挖掘效率的最重要的因素。

1.1.3 挖掘的任务

数据挖掘的任务是从大量数据中发现有趣模式。模式是用语言 L 来表示的一个表达式 E，它可用来描述数据集 DS 中数据的特性。E 所描述的数据是

DS 的一个子集。在实际应用中，数据挖掘模式可分为分类和回归模式、关联规则模式、聚类模式、孤立点模式、时间序列模式等。

分类（Classification）是找出描述并区分数据类或概念的模型（或函数）的过程。数据分类通常可以分为以下两步。

第一步，建立一个模型，描述预定的数据类集或概念集。可以通过分析由属性描述的数据库元组来构造模型，其中数据库元组是指被分析的样本、实例和对象。

第二步，使用模型进行分类。

常用的数据分类方法有判定树（Decision Tree）、贝叶斯分类（Bayesian）、神经网络（Neural Network）、K-最近邻分类、基于案例的推理（Case-based Reasoning，CBR）、概念格方法、粗糙集方法和模糊集方法。对于分类来说，其目的主要是提高分类的准确率与效率。

关联规则模式由 Agrawal、Imielinski、Swami 于 1993 年提出，它是描述在一个事务（项目）中物品（交易）同时出现的规律的知识模式，即通过量化的数字描述物品甲的出现对物品乙的出现的影响程度。关联规则模式的提取通常可以分为两步：找出满足用户的最小支持度的频繁项目集；提取出满足用户的最小置信度的关联规则。从大量商务事务记录中发现有趣的关联联系，可以帮助许多商务决策的制定，如市场规划、广告规划、分类设计、交叉购物和贱卖分析等。又如，在现今中国贷款购买住房和汽车的顾客中，调查发现 70%的人的年龄在 35～45 岁之间，这样银行就可以通过分析这些客户的特点来调整一些相应的政策，以便将贷款发放给这类客户群体。自从 Agrawal 等提出从大型数据库中挖掘关联规则以来，关联规则的挖掘已广泛地应用在电子通信行业、信用卡公司、股票交易所、银行和超级市场等场合。目前，国内外研究者正在从多种角度、多种渠道研究基于各种数据模型的关联规则的提取。

聚类（Clustering）分析是指根据对象属性标识对象集的类（组、簇）的过程。将对象按某种聚类准则聚类后，可以使对象组内的相异性最小、组间的相异性最大。例如，在保险业上，聚类能帮助保险公司分析投保人群的特征，以加大对这些客户群体的投保率。

孤立点（Outlier）分析是指挖掘出与数据的一般行为或模型不一致的数据对象的过程。例如，孤立点分析通过监测一个给定账号与正常的付费的比较，以付款数额特别大来发现信用卡的欺骗使用。

时间序列（Timer Serial）分析是指把数据之间的关联性与时间联系起来，

根据数据随时间变化的趋势预测未来的相关数值。

1.1.4 挖掘的分类

数据挖掘可以从很多不同的角度进行分类。

（1）根据挖掘的数据库类型，数据挖掘可分为关系数据库挖掘、空间数据库挖掘、时间数据库挖掘、文本数据库挖掘和多媒体数据库挖掘。

（2）根据发现知识的种类不同，数据挖掘可分为分类规则挖掘、聚类规则挖掘、关联规则挖掘和序列模式挖掘。

（3）根据挖掘使用技术的不同，数据挖掘可分为决策树、贝叶斯网络（Bayesian Networks）、模糊集、粗糙集、遗传算法和概念格。

1.1.5 面临的主要问题

目前，数据挖掘面临的主要问题有三大类：挖掘方法和用户交互的问题、性能问题和存储数据的数据库类型具有多样性的问题。

1．挖掘方法和用户交互的问题

这一问题反映了所挖掘的知识类型、在多粒度上挖掘知识的能力、领域知识的使用、特定的挖掘和知识显示。由于不同的用户可能对不同类型的知识感兴趣，所以数据挖掘系统应当覆盖范围很广的数据分析和知识发现任务，并且用户可以和数据挖掘系统交互，以不同的粒度和从不同的角度观察数据和发现模式；发现的知识应易于理解，能够直接被人们使用。

2．性能问题

这一问题包括数据挖掘算法的有效性、可伸缩性和并行处理。为了有效地从数据库中提取信息，数据挖掘算法必须是有效的和可伸缩的，即对于大型数据库来说，数据挖掘算法的运行时间必须是可预计的和可接受的，这是促使开发并行和分布式数据挖掘算法的因素。此外，当数据库更新时，不必重新挖掘全部数据，只要进行知识更新，修正和加强已经发现的知识即可。

3．存储数据的数据库类型具有多样性的问题

这一问题包括关系的、复杂的数据库处理和在异种数据库之间挖掘信息。

目前，有些数据库可能包含复杂的数据对象、超文本和多媒体数据、空间数据、时间数据或事务数据，由于数据类型的多样性和数据挖掘的目标不同，所以指望一个系统挖掘所有类型的数据是不现实的。从具有不同数据语义的，结构化的、半结构化的和非结构化的数据源来发现知识，对数据挖掘提出了巨大挑战。

以上问题是数据挖掘技术未来发展的主要需求和挑战。在近年来的数据挖掘的研究和开发中，一些挑战业已受到一定程度的关注，并考虑到了各种需求，而另外一些仍处于研究阶段。

1.1.6 主要应用

数据挖掘的研究方兴未艾，具有非常广阔的前景。面向对象数据库、分布式数据库、文本数据库等的数据挖掘；贝叶斯网的兴起；面向多策略和合作的发现系统；结合多媒体技术的应用等都是新的研究方向。数据挖掘原型系统和商业软件已开始在多个方面得到应用。

（1）客户分析：在银行信用卡和保险业中，确定有良好信誉和无不良倾向的客户是经营成功与否的关键。数据挖掘可以从以往的交易记录中“总结”出客户这些方面的信息。

（2）客户关系管理：数据挖掘可以识别产品使用模式或协助了解客户行为，从而可以改进通道管理（Channel Management）。例如，适时销售（Right Time Marketing）就是基于可由数据挖掘发现的顾客生活周期模型来实施的一种商业策略。

（3）零售业：数据挖掘对顾客购物篮数据（Basket Data）的分析可以协助货架布置、确定促销活动时间、促销商品组合及了解畅销和滞销商品的状况。

（4）产品质量保证：通过对历史数据的分析，数据挖掘可以发现某些不正常的数据分布，暴露制造和装配操作过程中出现的问题。

（5）WEB 站点的数据挖掘：电子商务网站每天都可能有上百万次的在线交易，生成大量的记录文件和登记表，可以对这些数据进行分析和挖掘，充分了解客户的喜好、购买模式，甚至是客户的一时冲动，以设计出满足不同客户群体需求的个性化网站，甚至从数据中推测客户的背景信息，进而增加其竞争力。

另外，在各个企事业部门，数据挖掘在假伪检测、风险评估、失误回避、

资源分配、市场销售预测和广告投资等方面都可以发挥作用。在国外，数据挖掘已应用于银行金融、零售批发、制造、保险、公共设施、行政、教育、通信、运输等多个行业部门，并且已经出现了许多数据挖掘和知识发现系统。例如，Quest是由IBM Almaden研究中心开发的数据挖掘系统，它可以从大型数据库中发现关联规则、分类规则、时间序列模式等；DBMiner是加拿大Jiawei Han教授领导的小组开发的一个数据挖掘系统；SKICAT系统是由U.M.Fayyad等人开发的知识发现系统，它将图像处理、数据分类、数据库管理等功能集成在一起，能够自动地对数字图像进行搜索和分类；KEFIR（Key Finding Reporter）是由GTE实验室开发的一个知识发现系统等。

1.2 关联规则挖掘

关联规则挖掘是数据挖掘的一个重要研究方向，它描述了交易数据集DB中两组不同对象（对象指交易中的内容，又称为交易项目）之间存在的某种关联关系。在关联规则挖掘过程中，需要多次对交易数据集进行扫描并与候选频繁项目集进行匹配和计数。由于面对巨量交易数据集，这一匹配和计算过程需要花费大量时间，所以效率是设计挖掘算法的关键。

关联规则是由Agrawal等人首先提出的一个重要KDD研究课题，它反映了大量数据中项目集之间有趣的关联或相关联系。目前，关联规则挖掘算法中较有影响的挖掘算法为R.Agrawal在1993年提出的Apriori算法。

面对巨量数据，效率是数据挖掘的关键问题所在，因此关联规则的研究也以效率为中心展开了多角度、多层面的讨论与研究。目前，针对关联规则的研究主要有以下几个方面。

（1）从研究对象上讲，有布尔型数据挖掘、数值型数据挖掘和模糊数据挖掘。

（2）从研究技术上讲，有概念格、云模型、粗糙集等技术，或将这些技术相结合以利于提高挖掘效率。

（3）从研究内容上讲，有挖掘算法、更新算法、分布式挖掘、抽样挖掘等几个方面。

挖掘算法以支持度、置信度框架为考虑因素，称为支持度-置信度框架，其中以R.Agrawal提出的Apriori算法较有影响。在此基础上又有很多Apriori算法的变种，如AprioriTid算法、AprioriHybrid算法；更新算法基于交易数据、