关联规则挖掘研究

关联规则挖掘是数据挖掘的一个重要组成部分，频集挖掘是关联规则挖掘的关键步骤，它在很大程度上决定了关联规则挖掘的效率。对于强规则、相关分析、时间序列、频集挖掘也有着十分重要的意义。

史月美 宗春梅 ◎著

兵器工业出版社

关联规则挖掘研究

史月美　宗春梅　著

兵器工业出版社

内容简介

本书从数据挖掘领域引人，介绍了关联规则的经典算法，重点从多种角度阐述作者以及该领域研究者们在关联规则挖掘方面的最新研究成果，包括最新研究思想、最新算法以及最新应用等。本书在各部分都提供了算法所涉及的必要的背景知识，以便读者无需查阅其他资料就可更好地理解。

图书在版编目(CIP)数据

关联规则挖掘研究/史月美，宗春梅著．－北京：兵器工业出版社，2016重印

ISBN 978－7－80248－502－0

Ⅰ．①关… Ⅱ．①史…②宗… Ⅲ．①数据采集—研究 Ⅳ．①TP274

中国版本图书馆 CIP 数据核字(2010)第 056854 号

出版发行：兵器工业出版社
发行电话：010－68962596，68962591
邮　　编：100089
社　　址：北京市海淀区车道沟 10 号
经　　销：各地新华书店
印　　刷：北京毅峰迅捷印刷有限公司
版　　次：201 6年 5月第 1 版第 4 次印刷

责任编辑：周宜今
封面设计：揽胜视觉
责任校对：郭　芳
责任印制：赵春云
开　　本：787×1092　1/16
印　　张：30.25
字　　数：355 千字
定　　价：65.00 元

前　言

随着产生和收集数据的能力不断提高，数据集的规模越来越大。存储数据的几何级数的爆炸性增长激起业界人士对信息新技术和智能化工具的需求，欲将数据转换为知识，以帮助决策者英明决策。

数据挖掘是一个多学科、知识交叉的领域。而关联规则挖掘是数据挖掘的知识模式中的一种主要形式，因其形式与大多数人对知识模式的理解最为相似，使其成为近几年研究的热点。

本书从数据挖掘领域引入，介绍了关联规则的经典算法，重点从多种角度阐述作者以及该领域研究者们在关联规则挖掘方面的最新研究成果，包括最新研究思想、最新算法以及最新应用等。本书在各部分都提供了算法所涉及的必要的背景知识，以便读者无需查阅其他资料就可更好地理解。

全书共8章。第1章主要介绍数据挖掘的基础知识，使读者有一个研究关联规则挖掘的背景知识；第2章介绍关联规则的基本知识，重点介绍关联规则挖掘的经典算法 Apriori 算法及 FP 算法，以及它们的改进算法；第3章介绍粗糙集的基本理论及将粗糙集理论应用于关联规则挖掘的研究成果；第4章简要介绍遗传算法的基本理论及将遗传算法应用于关联规则挖掘的相关算法；第5章介绍蚁群算法的基本原理及将蚁群算法与关联规则挖掘相结合的研究成果；第6章阐述分布式环境下关联规则挖掘的研究进展及其应用；第7章介绍基于时序和极大团

理论的关联规则挖掘相关算法；第 8 章说明关联规则挖掘在相关领域的实际应用。

本书第 1、3、6、8 章由史月美编写，第 2、4、5、7 章由宗春梅编写，由赵青杉副教授认真仔细审查了全书，并提出了许多宝贵意见，在此一并致谢。

本书可作为高等院校计算机科学与技术专业研究人员、技术人员的参考书，也可作为从事信息系统建设人员的参考资料。

由于作者水平有限，本书难免有不妥之处，恳请读者及专家批评指正，或提出建设性意见。敬请通过 sym_2291@sina.com 与作者联系，不胜感激。

作　者

2009 年 11 月

目 录

第 1 章　数据挖掘概述

数据采集和存储技术的进步导致庞大的数据库日益增多，这已经发生在人类耕耘的几乎所有领域，从普通的超市业务数据、信用卡使用记录、电话呼叫清单以及政府的统计数据等领域，到不太普通的天体图像、分子数据库和医疗记录等领域。能否从这些数据中提取出对数据库拥有者有价值的信息呢？人们对这个问题的兴趣在不断增长，而且已经形成了致力于这个任务的一门学科——数据挖掘。

目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏知识的手段，导致了“数据爆炸但信息、知识贫乏”的现象。而我们要想在这些数据和信息的基础上进行迅速准确的决策就必须借助一些新的技术和自动工具，以便将海量的数据转化成有价值的信息和知识。这些都促使数据挖掘的出现并推动其发展。

1.1　数据挖掘的发展

数据挖掘是一个逐渐演变的过程。在电子化数据处理的初期，人们试图通过某些方式来实现自动决策支持，当时的机器学习成为人们关心的焦点，机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，机器通过学习这些

范例总结并生成相应的规则，这些规则具有通用性，使用它们可以解决某一类的问题。随后，随着神经网络技术的形成和发展，使人们的注意力转向知识工程。知识工程不同于机器学习那样给计算机输入范例，让其生成出规则，而是直接为计算机输入已被代码化的规则，计算机通过使用这些规则来解决某些问题。专家系统就是使用这种方法所得到的成果，但它有投资大、效果不甚理想等不足。20 世纪 80 年代人们又在新的神经网络理论的指导下，重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库，从而产生了一个新的术语——数据库中的知识发现（Knowledge Discovery in Database，KDD）。它首次出现在 1989 年 8 月在底特律举行的第十一届国际联合人工智能学术会议上。为了统一认识，在 1996 年出版的总结该领域进展的权威论文集《知识发现与数据进展》中，Fayyd，Piatetsky - Shapiro and Smyth 给出了 KDD 和数据挖掘的最新定义，将二者加以区分。

KDD 的定义为：KDD 是从数据中辨别有效的、新颖的、潜在有用的、最终可理解的模式的过程。数据挖掘的定义为：数据挖掘是 KDD 中通过特定的算法在可接受的计算效率限制内生成特定模式的一个步骤。

由此可见，整个 KDD 过程是一个以知识使用者为中心、人机交互的探索过程。数据挖掘是数据库中知识发现的一个步骤，但又是最重要的一步。因此，往往可以不加区别地使用 KDD 和数据挖掘。一般在研究领域被称作数据库中的知识发现，在工程领域则称为数据挖掘。1989 年举行了第一届专题讨论会后，1991 年、1993 年、1994 年又连续举行了 KDD 专题讨论会。1995 年讨论会开始发展为一年一次的国际学术大会。会议较全面地讨论了数据挖掘与知识发现（Data Mining and Knowledge Discovery，DMKD）的基础理论、新的发现算法、数据

挖掘与数据仓库及 OLAP 的结合、可视化技术、知识表示方法、Web 中的数据挖掘等。迄今为止，对关系数据库和事务数据库进行数据挖掘和知识发现的研究已经取得了一定的进展，最有影响的发现算法有：加拿大 Simon Fraser 大学 J. Han 教授的概念树提升算法[2]、IBM 的 R. Agrawal 的关联算法[3]、澳大利亚的 J. R. Quinlan 教授的分类算法[4]、密西根州立大学 Erick Goodman 的遗传算法等。IBM、GTE、SAS、Microsoft、Silicon Graphics、Integral Solutions、Thinking Machines 等公司，相继开发出一些实用的 KDD 商业系统和原型系统，如市场分析用的 BehaviorScan、Explorer、MDT（Management Discovery Tool），金融投资领域的 Stock Selector、AI（Automated Investor），欺诈预警用的 Falcon、FAIS、Clone detector 等。与国外相比，国内对 DMKD 的研究稍晚，没有形成整体力量。1993 年国家自然科学基金首次支持该领域的研究项目。目前，国内的许多科研单位和高等院校开展了知识发现的基础理论及其应用研究，如清华大学、中科院计算技术研究所等。

1.2　数据挖掘的过程

数据挖掘（Data Mining）就是对观测到的数据集（经常是很庞大的）进行分析，目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据。这些数据可以存放在数据库、数据仓库或其他信息存储中。这是一个年轻的跨学科领域，源于诸如数据库系统、数据仓库、统计学、机器学习、数据可视化、信息检索和高性能计算。其他有贡献的领域包括神经网络、模式识别、空间数据分析、图像数据库、信号处理和许多应用领域，包括商务、经济学和生物信息学等。

数据挖掘也称为术语数据库中的知识发现。数据挖掘过程

一般分为三个阶段:数据准备、数据挖掘、结果评估,如图 1-1 所示。

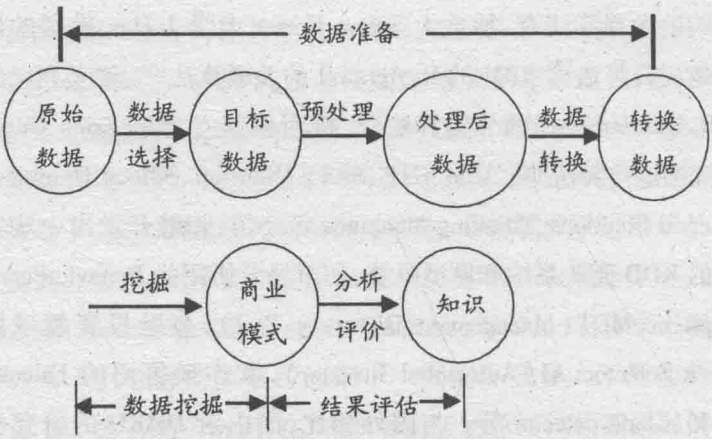

图 1-1 数据挖掘过程

1.3 数据预处理

现实世界中的原始数据经常存在不一致、重复、不完整、含噪声、维度高等问题,怎样预处理原始数据才能使得数据挖掘过程更加简便有效?数据预处理的目的就是为数据挖掘过程提供干净、准确、简洁的数据,提高数据挖掘效率和准确性,是数据挖掘中非常重要的环节。

常用的数据预处理的方法有数据清洗、数据集成、数据变换和数据规约。

1.3.1 数据清洗

数据清洗(Data Cleaning)就是清除数据噪声和与挖掘主题明显无关的数据。通常包括:添补遗漏数据、平滑噪声数据、识