上海市科学技术协会
“晨光计划”资助出版

数据资源的聚类预处理

夏骄雄　著

上海科学普及出版社

数据资源的聚类预处理

Clustering Preprocessing of Data Resource

夏骄雄 著

上海科学普及出版社

图书在版编目(CIP)数据

数据资源的聚类预处理/夏骄雄著. —上海：上海科学普及出版社，2011.11

ISBN 978-7-5427-5054-9

Ⅰ. ①数… Ⅱ. ①夏… Ⅲ. ①数据处理—聚类分析 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2011)第 194986 号

责任编辑 陈爱梅

数据资源的聚类预处理

夏骄雄 著

上海科学普及出版社出版发行

(上海中山北路 832 号 邮政编码 200070)

http://www.pspsh.com

各地新华书店经销 上海叶大印务发展有限公司

开本 787×1092 1/16 印张 10 字数 238 000

2011 年 11 月第 1 版 2011 年 11 月第 1 次印刷

ISBN 978-7-5427-5054-9 定价：25.00 元

本书如有缺页、错装或坏损等严重质量问题

请向出版社联系调换

上海科技发展基金会(www. sstdf. org)的宗旨是促进科学技术的繁荣和发展，促进科学技术的普及和推广，促进科技人才的成长和提高，为推动科技进步，提高广大人民群众的科学文化水平作贡献。本书受“上海科技发展基金会”资助出版。

“上海市科协资助青年学者出版科技著作晨光计划”出版说明

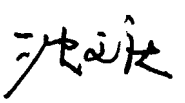
“上海市科协资助青年学者出版科技著作晨光计划”由上海市科协和上海科技发展基金会主办，上海科学普及出版社协办。该计划定向资助40周岁以下的上海青年学者出版首部个人原创性科技著作，旨在支持和鼓励学有所成的上海青年学者著书立说，加快培养青年科技人才的成长，切实推动“科教兴市”战略的实施。该计划每年资助不超过5人，每人资助1 500册以内的出版费用。申请资助的作者需要通过其所在学会（协会、研究会）向上海市科协学术部推荐，申请表下载网址：www. sast. stn. sh. cn。

总　序

尊重知识、尊重人才，在积极发现、培养、使用、凝聚优秀科技人才的同时，大力促进创新人才特别是年轻人才脱颖而出，是推动科技进步和创新的重要任务，也是上海市科学技术协会及其所属科技团体的重要职责。上海市科协联合上海科技发展基金会、上海科学普及出版社，新推出的“上海市科协资助青年学者出版科技著作晨光计划”，是履行这一职责的重要体现。

上海是我国科技人才集聚和青年科技人才涌现的地区之一。上海的青年科技工作者，长期以来为贯彻实施“科教兴国”战略，推动科技进步和创新，在科研和教学实践中默默耕耘，逐渐形成了一些新的工作成果，推出了不少新的学术思想，然而，这些优秀青年想要为自己的创新成果或创新思想著书立说，却受到资金、渠道等多种因素的困扰。“上海市科协资助青年学者出版科技著作晨光计划”，就是为这些优秀青年科技人才而设立的，就是要雪中送炭，支持和鼓励学有所成、干有所长的上海青年科技人才著书立说，从而促进青年科技人才的成长，繁荣学术交流，加快科学技术新思想、新方法和新知识的传播。

众人拾柴火焰高，科学事业的繁荣要依靠社会各界的关心和支持，尤其需要科技团体发挥独特作用。纵览目前国内的资助出版项目有很多，但“上海市科协资助青年学者出版科技著作晨光计划”在资助青年学者出版首部个人原创性科技著作上具有鲜明的特色。我衷心希望这项计划的实施，能对上海青年科技人才的成长有所帮助，能向世界展示上海青年科技人才的新面貌。

上海市科学技术协会主席　 

自　序

水呵水，到处都是水，船上的甲板却在干涸；
水呵水，到处都是水，却没有一滴能解我的焦渴。

数据呵数据，到处都是数据，各类用户却在迷茫；
数据呵数据，到处都是数据，却没有任何提示能帮我决策。

美国前副总统 Al Gore 在 1998 年 1 月 31 日所做的演讲《数字地球：21 世纪认识我们的星球》[Gore1998] 中指出：一场新的技术革新浪潮正允许我们能够获取、储存、处理并显示有关地球的空前浩瀚的数据以及广泛而又多样的环境和文化数据信息，而充分利用这些浩瀚数据的困难之处在于把这些数据变得有意义——即把原始数据变成可理解的信息。今天，我们经常发现我们拥有很多数据，却不知如何处置。现在，我们贪婪地渴求知识，而大量的资料却闲置一边，无人问津。

没有物质，就什么都不存在；没有能源，就什么都不会发生；没有信息，就什么都没有意义[Oet1965]。作为三大资源之一的信息，对于我们的生活越来越具有深远的影响。面对如此丰富、繁杂的数据，如何才能从中提取有价值的信息和知识，由此诞生了一个新的研究方向：基于数据库的知识发现 KDD(Knowledge Discovery in Database)以及相关的数据挖掘 DM(Data Mining)理论和技术。

数据资源(Data Resource)作为信息领域基本的研究对象，是从资源的角度对数据及其本身所存在的状态给予的重新认识与高度概括。综合利用各类有效的 KDD 和 DM 技术来提高数据资源本身的质量、增强数据对象的利用效率成为数据资源有效开发利用的主要研究方向。数据资源的预处理作为 KDD 和 DM 过程的重要环节，聚类分析作为 KDD 和 DM 领域成熟的技术，这两者相结合的研究具有重要的探讨意义和应用价值。

本书是本人在 2000 年至 2007 年期间，将聚类分析技术引入数据资源预处理阶段的过程中，分别借鉴运筹学、数理统计学、哲学本体论、数字图像处理学、分子动力学、物理学等领域的具体理论与方法所取得的一些潜心研究的成果。

全书共分八章。

第一章说明了本书研究领域的背景和现状，并就本人从事这方面研究所做的理论与实践方面的准备工作进行了阐述。

第二章着重就研究手段的理论背景和实践背景进行全面概述。

借鉴分裂型层次化聚类方式，本人在第三章中对分别从平面、立面、空间等三个层次综合构建基于层次分析法的数据库聚类预处理方法进行了描述，突出运用层次化思维来迭代评估目标，剔除相异度高的数据对象集合，达到聚类清理数据对象集合的目的，减少定性问

题定量化后误差的影响。

按照相关性最小原则，本人在第四章提出数据库主成份提取的聚类预处理方法进行高维数据系统的降维处理，获取数据对象变异最大方向的投影作为特定数据对象集合中的各个主成份，实现分层次的主成份聚类提取；同时此方法也验证了主成份对于原有信息全面覆盖的特性，同步解决了综合变量覆盖和降维问题，降低了数据对象集合的相异度和维度，实现了数据对象集合的聚类归约。

第五章中，本人利用数据对象的物理存储属性本身所具有的“0、1”特性，针对同体不同源数据对象提出同体不同源数据对象聚类数化算法，将数据资源中所有数据对象都通过数据对象预处理的过程转换成数字状态，然后利用数化后数据对象的数字状态作为聚合归类的依据，在不考虑数据对象其他属性的情况下，提高同体不同源数据对象的凝聚程度，达到降低比较次数、减少总体执行时间的目的，实现数据对象的聚类集成。

为了贯彻“复杂问题求解”的思想，本人在第六章提出了基于本体核与直方图的聚类预处理方法。在对数据对象进行聚类预处理时，首先得到弱量本体核的客体数据频数，然后根据用户明确的需求信息，获得所有需要的弱量本体核，并将其结合成强量本体核，最后通过“直方图”的构建与分析，明确数据对象的相关类属。

借鉴“能量”与“碰撞”的基本理念，本人以数据资源预处理得到的数据对象类或簇作为主要研究对象，构建了基于能量的“有效”动态阈值，在第七章实现了基于能量碰撞的聚类优化策略；对已经具备聚类初步特征的数据空间进行用户主题需求的能量驱动，把聚类内部的数据对象与孤立点数据对象放在统一的认识平台中加以统筹处理，保证了数据对象的聚类优化。

第八章对本人在2000年至2007年期间的研究工作进行了总结，并对后续的研究工作进行了展望。

作为理论成果的应用研究，本书选择了高校教育评估体系作为应用研究对象，将聚类分析技术引入高校数据资源的预处理环节，给出了应用实例，为有效利用现有数据资源，理性分析高校各方面工作的成效，深入探索学生培养的模式提供了有效的分析方法。

聚类分析技术的发展日新月异，数据资源方面的理论与实践探讨也日趋丰富。由于本人的研究所涉及的学科广泛、内容丰富，而本人的水平有限，谬误与疏漏在所难免，恳请大家批评指正。

夏骄雄

2010年6月20日于小心斋

目　录

第一章 绪 论

方以类聚，物以群分，吉凶生矣。在天成象，在地成形，变化见矣。

——《周易·系辞上》

水、火、金、木、土、谷，惟修；正德、利用、厚生，惟和；九功惟叙，九叙惟歌。

——《尚书·虞书·大禹谟》

Water, Water, Every where, And all the boards did shrink;

Water, Water, Every where, Nor any drop to drink.

19世纪英国湖畔诗人Samuel Taylor Coleridge在*The Rime of the Ancient Mariner*中对于"守着大海无水喝"的尴尬曾发出这样的叹息[Col1999]，当今的信息社会何尝不是面对同样的困境——到处都是累积的数据，到处都是对信息的渴望，这种期盼永难抑制：

数据呵数据，到处都是数据，各类用户却在迷茫；

数据呵数据，到处都是数据，却没有任何提示能帮我决策。

1.1 研究背景和现状

信息时代，多种媒介所携带的大量信息呈现出纷繁复杂的各种特性。随着计算机技术、通信技术、网络技术的飞速发展以及Internet应用技术的日益普及，以电子形式表达信息的各类数据日积月累形成"数据海洋"，其总量显现出爆炸性增长的势态，充分验证了1998年图灵奖获得者、数据库技术和"事务处理"专家Jim Gray所提出的"新摩尔定理"——从现在起，每18个月新增的存储量等于有史以来存储量之和[吴鹤龄2000]！

目前，随着数据采集和存储技术的进步，各类数据库系统高效率地实现着数据的录入、查询、统计等功能，忠实地完成着数据记录者的任务。但是，通过这些数据库系统中的数据所获得的信息量仅占整个数据库系统信息量的一小部分，因为用来对这些数据进行分析处理的工具很少，而且存在局限性[BL2000]。在信息时代，大量信息在给人们带来方便的同时，也带来了一系列问题，包括：信息量过大，超过了人们掌握、消化的能力；一些信息真伪难辨，给信息的正确运用带来困难；网络上的信息安全难以保障；信息组织形式的不一致性，增加了对信息进行有效统一处理的难度等。另一方面，人们意识到隐藏在这些数据之后的更深层次、更重要的信息能够描述数据的整体特征，能够预测发展趋势，这些信息在决策生成的过程中具有重要的参考价值[赵安郎1992]。面对海量规模和数量的数据库、大量繁杂的数据，如何才能从中提取有价值的知识，进一步提高数据的利用率，由此诞生了一个新的研究学科：

基于数据库的知识发现 KDD(Knowledge Discovery in Database)以及相关的数据挖掘 DM (Data Mining)理论和技术的研究[HMS2001]。

知识发现和数据挖掘领域的研究工作是适应市场竞争需要的,它将为决策者提供重要的、前所未料的信息或知识,从而产生不可估量的效益[Ago2000]。目前,关于 KDD 和 DM 的研究工作已经涉及众多领域,包括过程控制、信息管理、市场营销、风险投资、金融财务、医药卫生等领域[Bao2005]。同时,作为先进的分析工具和技术研究主体,数据及数据存在的状态也正成为 KDD 和 DM 研究的一个方向,继而成为数据库领域和人工智能领域交叉研究的一个热点。

数据资源(Data Resource)作为信息领域基本的研究对象,是从资源的角度对数据及其本身所存在的状态给予的重新认识与高度概括[孙九林2003]。任何有效利用数据资源的手段和方法都必须建立在实效分析和全面认识数据资源的基础之上[KLB2002]。

2000 年,时任中国科学院院长的路甬祥院士在国际数字地球研讨会议开幕式大会的报告中指出:现实状态中各类数据资源的产生与积累都得益于计算机应用技术的普及、数据库技术的迅速发展,以及数据库管理系统应用领域的不断扩展[路甬祥2000]。同时,大容量、高速度、低价格的存储设备大量出现,也对各类数据资源的存在形态给予强大的物理支撑。

作为数据资源综合处理的一种媒介和手段,数据库及其相应技术的发展对于数据资源的有效应用提供了具有划时代和里程碑意义的支持[施伯乐1999]。无论是早期的文件系统阶段,还是现在的数据库系统阶段,数据资源都能够以统一管理的数据库形式进行统筹描述,并为数据资源的共享应用提供了具体的操作平台[萨师煊1991]。根据数据资源存在状态的结构化程度,数据资源主要分为结构化(Structured)、半结构化(Semi-Structured)和非结构化(Non-Structured)三大类[刘宏2003]。

结构化数据资源,是以统一的结构方式表示特定数据对象集合的各个层次信息,广泛存在于局部的网络环境中。它具有良好定义的(Well-defined)语法和明确定义的语义,能够以数据对象本身、数据对象集合的子集以及数据对象集合本身等不同层次的特性,统筹反映客观事物及其状态、特点,同时各层次又都具有自身特色的反映能力[李庆忠2006]。这些特性折射在数据库层面,就是指具有良好的数据库结构定义和明确的数据库内容信息,数据库本身与周边数据库之间的联系通过结构化的模式加以体现,以便突出数据库本身所具有的针对性、方向性等功能。

半结构化数据资源,是以统一的结构方式表示特定数据对象集合本身及其子集的层次信息,而对数据对象本身的语法与语义并没有给予明确的定义。通常,半结构化数据资源是围绕数据对象集合本身综合功能的特点,在不同领域、不同层次的数据对象集合中间以一定组织形式进行反映。它能够整体反映客观事物的状态及其特点,但对于数据对象本身并不具有约束力。因此,半结构化数据资源具备管理领域与适应领域宽泛的特点,但对于细节信息的描述十分缺乏。

非结构化数据资源广泛存在于网络环境中间,仅对特定数据对象集合本身进行信息描述,没有严格意义上的结构化定义模式。数据对象的存在只遵循其自身特点,并不提供对数据对象集合独立的语义和语法支持。而且,对于数据资源本身的有效利用只能停留在对用户需求的被动接受层面,缺乏主动、积极的管理应对模式,并不利于数据资源整体环境中的综合开发和利用。

尽管数据资源的结构化状态具有不同的形式，但是在现实状态中，每一个数据库所具有的结构模型确实能够促使数据资源在特定平台上具有结构化特性。而且，充分利用数据库本身的结构化特性，对于有效进行数据资源的管理是十分高效的[黄鼎成2003][董诚2006]。当然，数据库对象的结构化特性对于数据资源的形态支撑作用仍然是不充分的，尤其是在 Internet 上所提供的万维网(World Wide Web)服务，其拥有的大量、动态、互连的数据资源在整个服务体系中仍然显示出非结构化或者半结构化的特性[辜寄蓉2003]。因此，合理应用非结构化或者半结构化特性，综合利用各类有效的 KDD 和 DM 成熟技术来提高数据资源本身的质量、增强数据对象的利用效率，已经成为数据资源有效开发利用的主要研究方向。

作为提高数据资源本身质量的主要技术手段，数据资源的预处理是 KDD 和 DM 过程中的重要环节，其相关研究具有重大的现实意义[KLB 2002]。它以适应数据资源应用任务为目标，以数据资源背景领域知识为指导，运用新颖的模型重新组织原有数据资源的结构与内容，弱化与应用任务目标不相符部分的作用，为 KDD 和 DM 过程核心算法提供整合并具有针对性的数据资源，从而减少相关算法的数据对象处理数量，提高数据资源利用效率，提高 KDD 过程知识发现的起点和 DM 过程的准确度。本书所体现的研究工作着重这一研究内容，正是对其理论研究的深入思考。

参考文献[Luan 2002]指出：高等教育领域多年来教学和管理工作所积累的大量数据对象，目前尚未得到有效利用，是一类运用 KDD 和 DM 过程能够大规模开发的“资源宝藏”。中国高等教育领域经过近 20 年的教育信息化建设和管理网络化建设，相关的应用系统业务趋于成熟，相关的操作管理模式渐成体系，特别是财务、人事、设备、教学、图书管理等部门都基本实现了本部门数据对象的单一封闭式管理，从而在同一高校范围内形成种类繁多、形态各异的庞大数据资源[董红斌1999]。

鉴于社会对高等院校发展的总体需求和目前高校各类数据系统信息管理的现状，利用这些现有数据资源理性分析高校各方面工作的成效以及学生培养过程中的得失就变得十分重要，这将对高等院校教学、管理、育人等多个方面的决策支持起到广泛的辅助作用[刘星晔2005]。当前，国内运用 KDD 和 DM 技术针对高校教育数据资源的研究(包括应用能力培养[刘贤龙1998b]、保持工作[黄发良2004]、教务管理[黄万华2004]、教学系统[魏萍萍2003]等方面)尚处于起步阶段。本书选择高校教育评估体系作为应用领域，正是对高校数据资源涉及的实际问题的具体思考。

1.2 主要研究内容和结构

本书的理论研究工作起步于本人曾经申请的上海市高等学校科学技术青年基金项目《基于地震学理论的数据仓库决策问题研究》(项目编号 01QN59)。由于地质构造学中地质能量的分布分析是预测地震重要的决策依据和数据来源[冯德益1996]，因此最初的理论研究工作从了解数据资源分布的构造为起点，探讨运用能量描述的手段来说明数据资源的分布情况[徐俊2006]。

随着上海市高等学校科学技术发展基金项目《基于相图理论和类能分布的数据仓库决策问题研究》(项目编号 04AB29)的开展，研究工作进一步向三个层面展开，并取得一些成

绩。其一，在现有成熟的KDD和DM领域技术中选择能够作为数据资源预处理的主要应用技术，并在模式分析[李强2004]、特征选择[施佳2007]、并行序贯模式[金沈杰2004]、聚合整理、算法组件[李强2005]、协同分解[ZJC+2005]等方面开展初步研讨；其二，从决策支持系统和管理信息系统的理念[夏骄雄2000a]出发，以数据信息存取环节作为突破口[夏骄雄2000b]，在各类环境中（包括网络环境[夏骄雄2003b]）研讨数据资源的基本结构特性及其状态[夏骄雄2001]，并在以往需求驱动研究[夏骄雄1998]的基础上，探讨数据资源中数据对象基于用户主题需求驱动的整合问题；其三，在原有数据资源分布构造的基础上，充分研讨各类数据资源中的数据对象在能量触发机制条件下对于决策的支撑程度[夏骄雄2006b]，优化和精简数据资源预处理过程的范围与细节，提高与用户主题需求的匹配程度[夏骄雄2003d]。

在这些研究项目的支持下，在总结现有数据资源整理阶段的相关理论和数据资源预处理相关技术的基础之上，以统计方法中的聚类分析作为基本研究手段，分别借鉴运筹学中层次分析方法和统计学中主成份分析方法的模式对数据资源中数据对象集合进行预处理，利用数据对象的物理存储特性，以及将数据对象之间存在的语义共享信息转化为概率分布的可视化模型，对数据资源中的数据对象进行预处理，并通过“能量”与“碰撞”理念的引入达到优化预处理结果的效果。

本书的实践研究工作来源于本人以往教学实验中数据信息的管理和现实工作中积累数据的处理这两个方面。一方面，教学实验环节利用计算机来仿真，并对实验数据进行积累和管理，以便提高仿真的真实性，这是计算机辅助教育领域教学实验仿真环节最重要的目标之一。因此，通过探讨一些理论细节[夏骄雄2002]，支持教学实验仿真过程[夏骄雄2000c]，提高积累数据的综合利用率[夏骄雄2003c]，也是本书重要的应用研究内容之一。另一方面，面对现实工作（尤其是共青团工作）中大量积累的数据只能作为文字形态保存，没有体现信息时代信息共享的实效，也没有给予现有工作任何有效启示的状况，利用2002年共青团上海市委员会重点调研课题《高校共青团工作信息化建设》的项目平台[夏骄雄2003a]，通过对信息化平台建设的重点研究[夏骄雄2006c]，取得一定的实践经验积累。

基于2004年上海大学《信息服务统一平台》建设项目中《学生信息统一整合平台》子项目的支持，以及2004年上海大学学生工作系统科研项目《学生个体对象的数据聚类研究及其属性聚类研究》（项目编号0405d14）的支持，本书实践研究部分充分利用本人多年来在上海大学教学、管理、服务等相关工作领域所积累的数据资源进行数据整理，对上海大学部分学生状态的描述及其最终结果与上海大学相关教学、教育领域的关联进行了客观的验证，从中得到一些有效开展上海大学学生工作的重要启示，为上海大学“全员育人”工作的开展和相关措施的决策提供了有效的信息支撑和知识帮助。

因此，综观本人所涉及的理论研究工作和实践研究工作的开展情况，本书的主要结构及其相关研究内容包括：

1. 数据资源聚类预处理问题概述

本书第二章主要论述数据资源存在的状态和聚类分析的技术要点、发展趋势，并在概要描述数据资源预处理主要内容的基础上，明确了本书所呈现的数据资源聚类预处理的基本内容。同时，结合本人的研究内容，详细描述了应用实践的主要背景、具体内容及使用策略。

2. 基于层次分析法的数据库聚类预处理方法

针对数据资源中数据对象集合的预处理，本书第三章借鉴运筹学理论的相关表述，引入

“层次分析法 AHP(Analytic Hierarchy Process)”的基本理念,将复杂的决策系统层次化,通过逐层比较各种关联因素的重要性,构建独立的、针对数据资源中每个数据库进行聚类评估的过程模型,并基于数据库聚类预处理的三个定义,提出基于层次分析法的数据库聚类预处理方法 DCP-AHP(Database Cluster Preprocessing on Analytic Hierarchy Process)。同时,通过“培养优秀本科生攻读硕士学位”这一主题信息,对选样的学生基本信息数据资源进行基于层次分析法的数据库聚类预处理,取得较高准确率的实验结果。

3. 数据库主成份提取的聚类预处理方法

围绕数据资源中数据对象集合的预处理,本书第四章借鉴数理统计的相关理论,引入“主成份分析方法 PCA(Principal Component Analysis)”中将多元统计分析的多变量状态简化为较少综合变量描述状态的基本理念,以数据库本身作为主要研究对象,构建由低维的有效特征成份实现最大程度信息覆盖的模型。同时,结合“学习质量”的主题信息,对数据资源信息运用数据库主成份提取的聚类预处理方法 DCP-PCE(Database Cluster Preprocessing on Principal Component Extraction)进行实验,实验结果验证了方法的可行性和在空间、时间上的节约性。

4. 同体不同源数据对象的聚类预处理数化算法

针对数据资源中的数据对象,本书第五章围绕这些数据对象的特点,针对同一实体数据对象具有不同现实表述的情况,充分应用数据对象存储的物理属性状态,引入同体不同源数据对象聚类数化算法 NC-SEDS(Numerical Cluster on Same Entity from Different Sources),对这些聚类预处理的主体数据对象中重复型、意义相近型数据对象进行聚类提取与质量分析。同时,通过与经典笛卡儿算法、增强型笛卡儿算法和优先队列算法进行实验比较,验证了数化过程在执行时间及预处理结果质量上的优势。

5. 基于本体核与直方图的聚类预处理方法

本书第六章针对数据对象在数据资源空间中存在的形式及其动态分布的特点,引入在语义和知识层次上描述共享和重用知识的“本体”概念,对数据对象之间的相互关系进行问题求解;并通过统计学的方法确定数据对象的概率分布,借鉴“灰度级直方图”的可视化分形理念反映数据对象存在状态的非确定性和密度几率,使用控制聚类合理性函数确定数据对象预处理聚类分布的基本划分,从而获取数据对象的最终聚类归属。实验结果表明,基于本体核与直方图的聚类预处理方法 CPOKH(Cluster Preprocessing on Ontic Kernel and Histogram)保障了数据资源中数据对象的预处理能力,并在实际应用中具有较大优势。

6. 基于能量碰撞的聚类优化策略

本书第七章借鉴物理学与分子动力学的相关理论,引入“能量”与“碰撞”的基本理念,以数据资源预处理得到的数据对象类或簇作为主要研究对象,构建基于能量的“有效”动态阈值,实现基于能量碰撞的聚类优化 COEC(Clustering Optimization by Energy Collision)的基本过程模型,为构建高质量的数据对象聚类划分、有效精简数据对象聚类结果的数量提供了准确的优化方式。实验结果表明,基于能量碰撞的聚类优化过程得到的结果相比纯粹的最大密度相连区域获取算法 DBSCAN(Density-Based Spatial Clustering of Applications with Noise)得到的结果更加符合实际情况。

本书第八章总结了本人的主要研究工作及研究工作的基本成果,并对后续工作(包括已经开展的后续工作)进行了展望。

第二章　数据资源聚类预处理问题概述

君子之道，譬如行远，必自迩；譬如登高，必自卑。

——《礼记·中庸》

不知敌之情者，不仁之至也。非民之将也，非主之佐也，非胜之主也。

——《孙子兵法·用间》

面对数量如此巨大而又不断累积的数据，如果没有行之有效的管理手段和信息获取途径，极易导致“数据爆炸但信息贫乏”这一悖论现象的发生[Gra2005]。因此，探索有效利用这些数据、开发应用隐含的信息和知识，成为信息领域发展的永恒主题。

2.1　KDD 与数据资源

分析数据资源存在的形态，现实状态中各类数据资源随着网络技术的快速发展，以及数据库网络应用领域的不断扩大，仍然处于蓬勃发展的阶段[黄建军1999]。同时，随着人工智能技术介入数据资源管理的各个阶段，基于数据库的知识发现 KDD(Knowledge Discovery in Database)对于数据资源的强大支撑作用得以显现。

根据参考文献[FU1996]，Usama M. Fayyad 给出的 KDD 定义为：基于数据库的知识发现是指从大量数据中提取有效的(Identifying Valid)、新颖的(Novel)、潜在有用的(Potentially Useful)、最终可被理解的模式(Ultimately Understandable Patterns)这一非平凡(Nontrivial)过程。

数据：指一个有关事实 F 的集合，用以描述事实的基本项目和信息。

模式：语言 L 中的表达式 E 所描述的数据是集合 F 的一个子集 F_E。F_E 表明集合 F 中的数据具有特性 E。作为一个模式，描述特性 E 比枚举数据子集 F_E 简单。

非平凡过程：KDD 是由多个步骤构成的处理过程，包括数据预处理、模式提取、知识评估及过程优化等。所谓非平凡是指具有一定程度的智能性和自动性，而绝不仅仅是简单的数值统计和数学计算。

有效性(可信性)：从数据中发现的模式必须有一定的可信度，通过函数 C 将语言 L 中的表达式映射到度量空间 M_C 上，c 表示模式 E 的可信度，$c=C(E,F)$。其中，$E\in L$，模式 E 所描述的数据集合 $F_E\subseteq F$。

新颖性：提取出的模式必须是新颖的。模式是否新颖可以通过两个途径来衡量：一是通过比较当前得到的数据和以前的数据或期望得到的数据来判断；二是通过对比发现的模

式与已有模式的关系来判断。通常用一个函数来表示模式的新颖程度 $N(E,F)$，该函数的返回值是逻辑值或是对模式 E 新颖程度的一个判断数值。

潜在作用：指提取出来的模式将来会被实际运用的几率和程度，通过函数 U 把语言 L 中的表达式映射到度量空间 M_U 上，u 表示模式 E 的有作用程度，$u=U(E,F)$。

可理解性：发现的模式应该能够被用户理解，尤其是在简洁性方面，这样才能帮助用户更好地了解和使用数据集合中的信息。一个模式能否被用户理解并不是一件容易的事，需要对其简单程度 S 进行度量。用 s 表示模式 E 的简单度(可理解度)，$s=S(E,F)$。

上述度量函数从不同角度进行模式评价，往往需要采用权值来进行综合评判。若使用函数来获取模式 E 的权值 $i=I(E,F,C,N,U,S)$，则从 KDD 角度给知识的定位为：一个模式 E 对用户设定的阈值 I，如果 $I(E,F,C,N,U,S)>I$，则模式 $E\in L$ 称为知识[FPS1996a]。

基于数据库的知识发现是一个反复迭代的人机交互处理过程。该过程需要经历多个步骤(如图 2.1 所示)，并且许多决策信息需要由用户提供。从宏观上分析，基于数据库的知识发现过程主要由三个部分组成，即数据整理(Data Arrangement)阶段、数据挖掘(Data Mining)阶段和结果的解释评估(Results Interpretation and Evaluation)阶段[FPS1996b]。

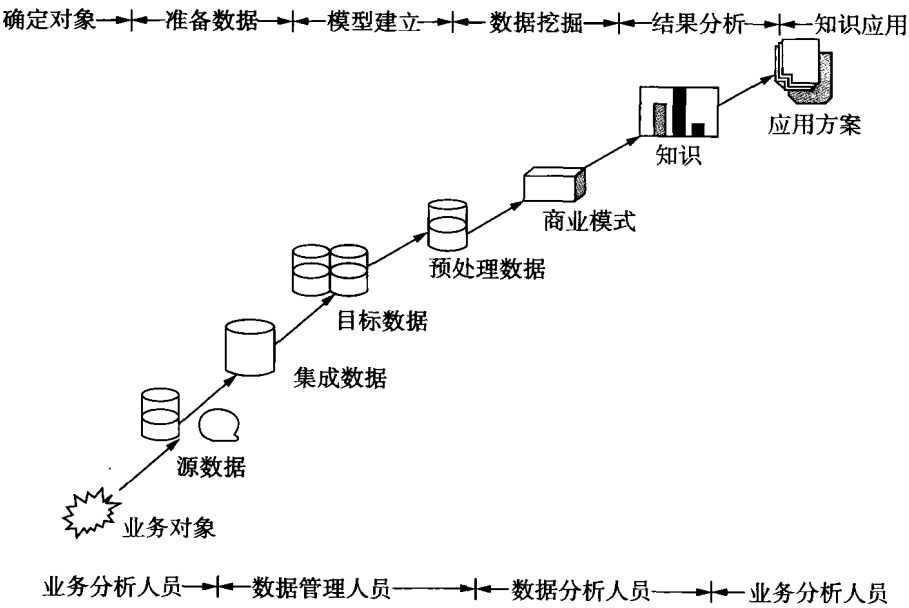


图 2.1　KDD 过程示意图

Fig. 2.1　Process of Knowledge Discovery in Database

基于数据库的知识发现过程中，数据整理阶段和数据挖掘阶段是两个重要的环节[李雄飞2003]。由于高质量的决策必须依赖于高质量的信息支持，因此数据整理阶段成为数据挖掘阶段的必要前提和重要基础。数据整理阶段对于数据资源的质量保证，能够大幅度提高数据挖掘阶段的精度和性能。数据整理阶段主要面对数据资源中的具体对象，进行包括数据准备(Data Preparative)、数据选取(Data Selection)、数据处理(Data Processing)、数据变换(Data Transformation)在内的四项主要任务[FPS1996c]。

数据准备的任务在于强调对数据资源应用领域情况的全面了解和掌握，以及对应用所