倪志伟 倪丽萍
刘慧婷 贾瑞玉 著

动态数据挖掘

科学出版社
www.sciencep.com

动态数据挖掘

倪志伟　倪丽萍　刘慧婷　贾瑞玉　著

科学出版社
北京

内 容 简 介

动态数据挖掘是针对动态数据库和实时数据库进行知识提取的数据挖掘技术。随着信息技术的进一步发展，对知识新颖性的需求越来越强，采用传统的静态数据挖掘技术来分析不断产生的信息无法满足现实应用的要求，对实际应用数据源在其运行的同时进行动态数据挖掘得到相关知识显得日益重要。

本书是关于动态数据挖掘相关技术及其应用的著作，涉及数据流挖掘、分形数据挖掘、联机分析挖掘、经验模态分解和联系发现技术等。本书内容新颖，融入了近年来在学术界和工程界普遍关注的诸多热门课题，是作者及其课题组几年来完成国家级科研项目的成果结晶。

本书可作为管理科学与工程、计算机应用技术等学科高年级的本科生和研究生用书，也可供相关研究人员参考。

图书在版编目(CIP)数据

动态数据挖掘/倪志伟等著. —北京：科学出版社，2010

ISBN 978-7-03-028347-4

Ⅰ. 动… Ⅱ. 倪… Ⅲ. 数据挖掘 Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2010)第 138774 号

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

*

2010 年 8 月第 一 版 开本：B5(720×1000)

2010 年 8 月第一次印刷 印张：16 1/2

印数：1—3 500 字数：321 000

定价：55.00 元

(如有印装质量问题，我社负责调换)

前　言

随着计算机技术、网络技术和通信技术的发展，各行各业获取数据的能力得到了极大的提高，数据呈现出海量、高维、动态的特点。传统的数据挖掘技术很难有效地对这些数据进行分析，这成为商务智能、决策支持以及知识管理等系统中的一个主要瓶颈，影响了管理决策的效果。为解决这一瓶颈，动态数据挖掘技术应运而生，成为当前研究的热点问题。

动态数据挖掘是集过去、现在与未来于一体的动态过程，其动态性体现在数据的采集、处理等环节。动态数据挖掘通常以动态数据库和实时数据库为数据的主要来源。数据处理是动态数据挖掘的核心部分，为了适应数据的动态变化特点，更好地挖掘动态数据中隐含的事先未知的有用信息和知识，必须改进传统的或设计新颖的数据挖掘技术。

本书介绍了近年来应用于动态数据挖掘领域中的几种较为新颖的挖掘技术，包括数据流挖掘、分形数据挖掘、联机分析挖掘、经验模态分解以及联系发现技术，研究了它们的原理、特点、性能及应用情况。全书共分为六章：第一章概述传统的数据挖掘技术和动态数据挖掘技术，重点介绍动态数据挖掘技术产生的背景、发展概况和未来研究方向；第二章探讨数据流挖掘技术及其应用、数据流挖掘算法以及数据流管理系统；第三章介绍分形数据挖掘技术，基于分形维数的约简、聚类、分类及其改进算法；第四章研究联机分析挖掘中的三个重要技术组成，即数据立方体的构建和优化技术、联机分析处理查询技术及联机分析挖掘调度机制和挖掘算法；第五章介绍经验模态分解技术的基本原理、算法思想以及与其他算法结合的应用情况；第六章介绍联系发现技术的发展现状、应用情况和发展趋势。各章内容相对独立又相互联系，较为系统地阐述了动态数据挖掘技术的研究现状。

本书是合肥工业大学管理学院智能管理研究所全体研究人员近年来对动态数据挖掘技术研究与应用中的一些研究和成果的系统总结，得到了国家高技术研究发展计划(863 计划)“面向制造业售后服务的商务智能关键技术研究”和国家自然科学基金“商务智能中的动态数据挖掘与分形技术的研究”两个项目的资助。

在撰写书稿的过程中得到了王超、高雅卓、胡汤磊、伍章俊的大力支持和帮助。倪志伟教授负责全书策划和大纲的制订，并负责全书的统纂和修改。安徽大学刘慧婷、贾瑞玉和合肥工业大学倪丽萍为本书的校对和排版做了大量的工作。同时智能管理研究所的全体研究人员，特别是忻凌、郑盈盈、杨葛钟啸、梁敏君、吴姗、罗义钦、郭峻峰、姜苗、戴奇波、周之强、查春生、赵裕啸等研究生在书稿的撰写过程中

给予了无私的帮助，在此向他们表示衷心的感谢。特别感谢我国著名学者、合肥工业大学博士生导师杨善林教授对本书出版给予的热情鼓励与大力支持。

此外，在本书的撰写过程中，参考了国内外的相关研究成果，在此感谢涉及的所有专家和研究人员。

由于作者水平有限，不妥之处在所难免，恳请同行与读者批评和指正。

作　者

2010 年 5 月于合肥

目　录

前言

第一章　绪论 …… 1

1.1　引言 …… 1

1.2　数据挖掘概述 …… 2

1.2.1　数据挖掘的基本概念 …… 2

1.2.2　数据挖掘技术 …… 10

1.3　动态数据挖掘 …… 14

1.3.1　动态数据挖掘的产生 …… 14

1.3.2　动态数据挖掘技术概述 …… 20

参考文献 …… 24

第二章　数据流挖掘技术 …… 28

2.1　概述 …… 28

2.2　数据流挖掘技术 …… 34

2.2.1　窗口技术 …… 34

2.2.2　动态抽样技术 …… 37

2.2.3　概要数据结构 …… 39

2.2.4　更新策略 …… 41

2.3　数据流挖掘算法 …… 43

2.3.1　数据流聚类算法 …… 44

2.3.2　数据流分类算法 …… 54

2.3.3　数据流频繁项集挖掘算法 …… 59

2.3.4　多数据流挖掘算法 …… 69

2.4　数据流挖掘技术的应用 …… 74

2.4.1　数据流管理系统 …… 74

2.4.2　案例推理在数据流管理中的应用 …… 78

参考文献 …… 80

第三章　分形数据挖掘技术 …… 84

3.1　概述 …… 84

3.2　数据集的分形维数 …… 86

3.2.1　数据集分形维数的含义 …… 86

3.2.2　数据集分形维数的计算方法 …… 89
3.3　基于分形维数的约简技术 …… 96
3.3.1　分形属性选择及其改进算法 …… 96
3.3.2　基于分形维数的案例库维护算法 …… 103
3.4　分形聚类算法 …… 106
3.4.1　基于网格和分形维数的聚类算法 …… 107
3.4.2　基于分形维数的数据流聚类算法 …… 110
3.4.3　基于多重分形的聚类层次优化算法 …… 114
3.5　分形分类与预测技术 …… 115
3.5.1　分形分类技术 …… 115
3.5.2　分形预测技术 …… 117
3.6　分形数据挖掘技术的应用 …… 121
3.6.1　金融数据分析 …… 121
3.6.2　网络入侵检测 …… 128
参考文献 …… 129
第四章　联机分析挖掘 …… 132
4.1　概述 …… 132
4.2　数据立方体 …… 134
4.2.1　数据立方体简介 …… 134
4.2.2　数据立方体优化方法 …… 136
4.2.3　数据立方体物化方法研究 …… 138
4.3　联机分析处理 …… 144
4.3.1　OLAP 概念及分类 …… 144
4.3.2　支持 OLAP 查询的索引技术研究 …… 147
4.3.3　OLAP 动态查询方法 …… 156
4.4　联机分析挖掘 …… 158
4.4.1　联机分析挖掘简介 …… 158
4.4.2　联机分析挖掘体系结构 …… 160
4.4.3　OLAP 与数据挖掘技术的结合方法 …… 163
参考文献 …… 172
第五章　经验模态分解技术 …… 175
5.1　概述 …… 175
5.1.1　经验模态分解基本理论 …… 175
5.1.2　经验模态分解研究现状 …… 177
5.2　基于经验模态分解的序列趋势的提取 …… 179

5.2.1 引言 …… 179
5.2.2 基于 EMD 方法的序列趋势的提取 …… 179
5.3 基于经验模态分解的时间序列匹配算法 …… 185
5.3.1 引言 …… 185
5.3.2 基于交叉覆盖算法的序列匹配算法 …… 186
5.3.3 基于经验模态分解和覆盖算法的序列匹配算法 …… 191
5.4 基于经验模态分解的聚类算法 …… 194
5.4.1 引言 …… 194
5.4.2 基于经验模态分解的数据降维技术 …… 195
5.4.3 基于经验模态分解和 K-means 聚类算法 …… 198
5.5 基于经验模态分解的流数据挖掘技术 …… 204
5.5.1 引言 …… 204
5.5.2 基于经验模态分解的数据流概要生成技术 …… 205
5.6 经验模态分解动态数据挖掘技术的应用 …… 208
5.6.1 引言 …… 208
5.6.2 基于经验模态分解和交叉覆盖算法的个人信用的评估 …… 209
5.6.3 基于经验模态分解和 K-means 算法的客户行为聚类 …… 213
参考文献 …… 217
第六章 联系发现技术 …… 222
6.1 概述 …… 222
6.2 基于图挖掘的联系发现 …… 223
6.2.1 图挖掘的相关概念和定义 …… 223
6.2.2 基于图论的无监督的联系发现算法 …… 231
6.3 基于一阶谓词逻辑的联系发现 …… 235
6.3.1 一阶谓词逻辑的相关概念和定义 …… 235
6.3.2 基于 ILP 的联系发现算法 …… 238
6.4 基于联系发现的结合型数据挖掘方法 …… 240
6.4.1 基于相关分析和联系发现的结合 …… 240
6.4.2 图熵和联系发现的结合 …… 244
6.4.3 概率统计方法和联系发现的结合 …… 247
6.5 联系发现技术的现实应用 …… 250
6.5.1 联系发现在反恐中的运用 …… 250
6.5.2 联系发现在金融反洗钱中的运用 …… 251
参考文献 …… 254

第一章 绪 论

数据挖掘(data mining,DM)是一个多学科交叉的领域,兴起于20世纪80年代末,取得了重大进展。随着信息技术的进一步发展,对知识的新颖性要求越来越高,传统的数据挖掘不能满足对动态环境或动态数据源的数据分析要求,动态数据挖掘(dynamic data mining,DDM)的研究显得日益必要。动态数据挖掘是相对于传统数据挖掘的数据处理过程而言的。传统的数据挖掘只是针对固定的数据集进行,而动态数据挖掘中,为了找出新颖的、有价值的、感兴趣的知识,在数据处理过程中要求能动态处理各种实时数据。本章首先介绍数据挖掘的基本概念和传统数据挖掘的常用技术,然后阐述动态数据挖掘产生的背景和发展现状,最后对一些重要的动态数据挖掘技术进行概述。

1.1 引 言

随着Internet的广泛应用,计算机技术、网络技术和通信技术正改变着整个人类的生活方式和社会的发展。各个行业都开始采用计算机及相应的信息技术进行管理,生成、收集、存储及处理数据的能力大大提高,数据量与日俱增。尽管目前的数据库系统能够高效率地实现数据的录入、查询、统计等功能,但由于数据量过于庞大以及数据库系统中分析方法的严重缺乏,使得它无法直接发现数据中隐藏的相互联系,更无法根据当前的数据去预测未来的发展趋势。因此,出现了所谓"数据多,知识少"的现象,造成了严重的资源浪费。

如何理解已有的历史数据并预测未来的行为;如何从这些海量数据中发现更加有用的信息,变被动的数据为主动有价值的知识;如何快速、准确地获得有价值的信息,为政府和企业提供决策依据,获取更大的经济效益和更好的社会效益,都迫使人们去寻找新的、更为有效的数据分析手段对各种"数据矿藏"进行有效的挖掘以便发挥其应用潜能。数据挖掘正是在这样的需求背景下应运而生的。

1989年8月,在美国人工智能协会举办的专题研讨会上,首次提出基于数据库的知识发现技术(knowledge discovery in database,KDD)。该技术涉及机器学习、模式识别、统计学、智能数据库、知识获取、专家系统、数据可视化和高性能计算等领域。数据挖掘是从数据库中抽取隐含的、未知的、具有潜在使用价值信息的过程。由于数据挖掘是KDD过程中最为关键的步骤,在实际应用中对数据挖掘和KDD这两个术语的运用往往不加区别[1,2]。

目前，数据挖掘的理论基础主要有数学理论、机器学习理论、数据库理论和可视化理论。数学理论包括统计学理论、支持向量机理论、模糊集理论、粗糙集理论、概率论、贝叶斯概率和贝叶斯学习理论等。机器学习理论包括归纳学习、决策树、类比学习与基于案例的学习和计算智能等。数据库理论包括关系数据库理论、事务理论、逻辑与数据库、面向对象数据库理论等。可视化涉及计算机图形学、图像处理、计算机辅助设计、计算机视觉及人机交互等多个领域。

随着数据挖掘研究的不断深入，数据挖掘技术已逐渐成熟。数据挖掘技术的三大支柱是数据库技术、人工智能技术、概率与数理统计。在数据库技术已充分发展的基础上，数据挖掘利用了人工智能和统计分析的应用程序，把这些高深复杂的技术封装起来，来完成复杂的知识挖掘功能。

数据挖掘的应用也越来越广泛。例如，数据挖掘可应用在金融数据分析、商业零售数据分析、电信和网络数据分析、生物医学和 DNA 数据分析、天文和海洋地理探测数据分析等方面[3]。

总之，目前许多学科和行业都投入到数据挖掘的理论和应用研究中，并取得了较好的理论结果和数据挖掘产品。

然而，数据挖掘发展的同时也面临着挑战。传统的数据挖掘是从静态的数据库中发现知识[4~9]，在实际应用中数据库往往是动态变化的，新的数据积累可能导致之前采用的挖掘算法所发现的知识失效，因此发现的知识或模式也需要动态维护，即时更新。如果每次新数据加入数据库后都需要重新运行数据挖掘算法，那么随着数据集的更新，规则集也要做相应的更新，重建整个知识库耗费的时间代价是巨大的[10]。因此，需要研究新的动态数据挖掘算法来应对以增量形式获得的新数据。

1.2 数据挖掘概述

1.2.1 数据挖掘的基本概念

1. 数据挖掘的定义

从商业角度看，数据挖掘就是按企业的既定业务目标，对大量的企业数据进行探索和分析，以揭示隐藏的、未知的规律性并将其模式化，从而支持商业决策活动。数据挖掘技术只有面向特定的商业领域才有应用价值。数据挖掘并不是要求发现放之四海而皆准的真理，所有发现的知识都是相对的，并且只有对特定的商业行为才有指导意义。

从技术角度看，数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含在其中的、人们不知道的，但又是潜在有用的信息和知识的过

程。原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至是分布在网络上的异构数据。发现知识的方法可以是数学的，也可以是非数学的，可以是演绎的，也可以是归纳的。发现的知识可以用于信息管理、查询优化、决策支持和过程控制等。因此，数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询，提升到从数据库中挖掘知识，提供决策支持[2]。

根据定义可以得到数据挖掘的特点：

① 数据量常常是巨大的，因此如何高效地存取数据，如何根据具体应用领域找出数据关系即高效率算法，以及是使用全部数据还是随机或有目的地选择出数据子集，都成为数据挖掘工作者要考虑的问题。

② 在一些应用中（如商业投资、证券交易等），由于数据变化迅速，因此要求数据挖掘能快速作出反应以及时提供决策支持。

③ 数据挖掘既要发现潜在的规则，也要管理和维护规则。规则的改变随着新数据的不断更新而更新。

④ 数据挖掘不是一个简单算法，而是一个较为复杂的系统，它需要业务理解、数据准备、建模、评估等一系列步骤，是个循环往复不断完善的系统工程。

⑤ 数据挖掘中的规则不必适用全部的数据，也不可能挖掘出普遍适用的规则。

2. KDD 与数据挖掘的关系

在 1996 年出版的总结该领域进展的权威论文集中，Fayyad 等重新给出 KDD 和数据挖掘的定义，并将二者加以区分：KDD 是从数据中辨别有效的、新颖的、潜在有用的、最终可理解模式的过程；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[11]。换句话说，KDD 是一个包括数据清理、数据集成、数据选择、数据变换、数据挖掘、模式评价等步骤，最终得到知识的全过程，而数据挖掘只是其中的一个关键步骤[12]。

可以把 KDD 看做是一些基本功能构件的系统化协同工作系统，而数据挖掘则是这个系统中的一个关键部分。将数据挖掘作为 KDD 的一个重要步骤看待，可以使我们更容易聚焦研究重点，有效解决问题[13]。

事实上，在现今的文献中，这两个术语经常被不加区分地使用着。

3. 数据挖掘的理论基础

数据挖掘方法可以是基于数学理论的，也可以是非数学的。从研究的历史来看，它们可能是人工智能、数理统计、计算机科学等学科诸多学者和工程技术人员

在数据挖掘的研究过程中所创立的理论体系。下面列举一些重要的理论框架，可以帮助我们准确地理解数据挖掘的概念与技术特点[1,14,15]：

（1）基于概率和统计理论

统计学作为一门古老的学科，在数据挖掘中得到了广泛的应用，已经成为支撑知识发现和数据挖掘技术的重要理论基础之一。在这种理论框架下，数据挖掘技术被看做是从大量源数据集中发现随机变量的概率分布情况的过程[16]。例如，贝叶斯网络模型，它在分类和聚类的研究与应用中取得了很好的成果。

（2）数据归约

按照这一理论，数据挖掘就是减少数据的描述。在大型数据库里，数据归约能换来对查询的快速近似应答。数据归约技术主要包括奇异值分解、小波、回归、对数线性模型、直方图、簇、取样和索引树构造等。

（3）规则发现架构

在这种理论架构中，将三类数据挖掘目标，即分类、关联及序列作为一个统一的规则发现问题来处理[17]，给出了统一的挖掘模型和规则发现过程中的几个基本运算，基本解决了数据挖掘问题如何映射到模型和通过基本运算发现规则的问题。

（4）模式发现架构

基于这种理论框架，数据挖掘技术被认为是从源数据集中发现知识模式的过程[18]。按照这种架构，可以针对不同的知识模式的发现过程进行研究，例如，在关联规则、序列模式、分类、聚类、决策树归纳等模式发现的技术与方法上，取得了一些成果。这是目前较为常用的数据挖掘研究与系统开发架构。

（5）机器学习理论

机器学习技术是数据挖掘的核心之一。数据挖掘的关键技术是模式识别和关系识别的算法，其中许多算法正是来源于机器学习研究领域。机器学习是根据生理学、认知科学等对人类学习机理的了解，建立人类学习过程的计算模型或认知模型，发展各种学习理论和学习方法，研究通用的学习算法并进行理论上的分析，建立面向任务的具有特定应用的学习系统[19]。常用的机器学习方法有决策树、案例推理、贝叶斯信念网络、科学发现、遗传算法等。

（6）数据库理论

数据库技术是20世纪60年代初开始发展起来的一项数据管理自动化的综合性新技术，是数据管理最有效的手段。数据库的应用领域相当广泛，从一般事务处理到各种专门化数据的存储与管理。它的出现极大地促进了计算机应用的发展，数据库技术已经成为现代计算机信息系统和计算机应用系统的基础和核心[20]。

数据库技术与网络通信技术、人工智能技术、并行计算技术等互相渗透，互相结合，涌现出多种新型数据库系统。例如，数据库技术与分布式处理技术相结合，形成了分布式数据库系统；与多媒体技术相结合，形成了多媒体数据库系统；与人

工智能技术相结合，形成了知识库系统和主动数据库系统。数据库是数据挖掘的基础。

(7) 可视化理论

可视化技术能够实现对信息数据的分析和提取，以图形、图像、虚拟现实等易为人们所辨识的方式展现原始数据间的复杂关系、潜在信息以及发展趋势，从而丰富科学发现的过程[21]。可视化技术能够准确地表达数据挖掘的过程、挖掘结果，使用户深入地理解问题并选择更适当的数据挖掘算法，达到深入剖析数据的目的。

(8) 云模型和数据场理论

李德毅院士在传统的概率统计理论和模糊理论的基础上提出了定性定量不确定性转换模型——云模型，实现了定性概念和定量值之间的不确定性转换[22]。云是用语言值表示某个定性概念与其定量表示之间的不确定性转换模型，主要反映客观世界或人类知识中概念的模糊性和随机性，并且把二者集成在一起，构成定性和定量之间的映射。云模型是云的具体实现方法，云的具体实现方法可以有多种，这些方法构成了不同类型的云，如对称云模型、半云模型、组合云模型。云模型具有期望(expected value，Ex)、熵(entropy，En)、超熵(hyper entropy，He)这三个数字特征，云的数字特征是描述云模型、产生虚拟云、实现云计算、完成云变换的数值基础。数据挖掘过程中有很多的不确定性，发现的知识本身也有不确定性[22]。因此，在数据挖掘中采用云模型这种分析不确定信息的理论是必要的。

物理学中研究的问题往往涉及分布在一定空间区域中的物理量，通常称之为场，如引力场、电场等。文献[22]受到物理学中场论的启发，将以上物理学中场的概念推广到数域空间，认为数域空间中每个数据对象都相当于一个点电荷或质点，其周围存在一个作用场，位于场中的任何其他数据对象都将受到影响，因此整个数域空间就构成一个数据场。数据场理论的引入是数据挖掘领域中的一个突破，因为传统的数据挖掘算法只考虑对象之间一对一的映射关系，忽视了一对多或者多对一关系。而数据场理论克服了这样的问题，因为它把空间中某点的状态看成是其他对象共同作用的结果[23]。

(9) 双库协同与双基融合机制

双库协同与双基融合机制由文献[24]提出，双库协同机制即采用中断型协调器和启发型协调器，分别对所获得的假设规则进行处理和利用关联强度激发数据聚焦进行数据发掘。双库协同机制基本解决了在数据库基础上进行数据发掘过程中，对领域固有基础知识库的实时维护，同时在一定程度上解决了认知自主性的问题。即利用启发型协调器，实现了计算机自动发现知识短缺，系统自身根据知识短缺产生创见意向，形成定向发掘；对挖掘出来的知识通过中断型协调器，对知识库进行实时管理与维护。

双基融合机制是指构造基础数据库与知识库的内在联系的通道，从而用数据

库与 KDD 去制约、驱动 KDK (knowledge discovery in knowledge base)的挖掘过程,改变 KDK 固有的运行机制,在结构与功能上形成一个开放优化的扩体。

(10) 基于多 Agent 技术的数据挖掘构架

多 Agent 技术(multi-agent system,MAS)是人工智能技术的一次质的飞跃。多 Agent 系统是指由多个 Agent 组成的集合[25],这些 Agent 成员之间相互协调、相互服务,共同完成一个复杂的任务。MAS 系统中各个 Agent 是自主的,可以有共同的目标,也可以有各自不同的目标。Agent 间协作形式多种多样,MAS 系统需要协调这些自制的 Agent 行为。由于各 Agent 空间上的分布性、时间上的并行性和逻辑上的依赖性使得 MAS 系统的问题求解过程较为复杂[26]。

数据挖掘的基本问题在于数据的数量、维数、数据抽样方式以及数据挖掘结果的不确定性等,而把多 Agent 技术引入数据挖掘系统能在一定程度上解决此类问题[27]。

把多 Agent 技术引入到数据挖掘中,用 Agent 来描述数据挖掘过程的各个部分,整个知识发现的过程是一个 MAS 系统,利用 Agent 本身具有的知识、目标及推理、决策、规划、控制等能力,自主性、社会性、反应性、能动性等特性,可以实现整个数据挖掘过程的智能化[28]。

(11) 案例推理系统上的数据挖掘框架

虽然案例推理(CBR)系统比传统专家系统较为容易获取知识,但案例推理系统也存在一定程度的知识获取瓶颈问题,如案例、修正知识、相似性评估等知识的获取问题,无疑案例推理中的知识获取瓶颈问题也影响了案例推理系统的性能。通过在案例推理中引入数据挖掘技术,针对案例推理系统中的多种数据源,在其上进行数据挖掘,以提高知识获取的自动化程度,提高系统的性能,加快智能系统的开发周期[29]。

在案例推理系统中,数据挖掘要进行辅助案例推理方面的工作主要有两点,即自动知识发现和范例知识维护。首先是建造相关知识结构,这涉及范例工程的初始状态与案例推理系统运行时的知识获取,其次是现有范例知识结构的优化。具体地,包括如下内容:从数据库中建立范例一级知识结构,从数据库中建立范例库的组织结构,从数据库中更新范例知识的结构,从案例库中挖掘修正知识,从案例库中进行案例维护的数据挖掘[29]。

4. 数据挖掘过程

数据挖掘是一个完整的过程,该过程从大型数据库或数据仓库中挖掘先前未知的、有效的、可实用的信息,并使用这些信息做出决策或丰富知识。在实施数据挖掘之前,先决定采取什么样的步骤,每一步都做什么,确定目标和实施方案。一般的,数据挖掘在任何一个问题上的应用,都可以大致分为四个阶段[30,31]:

(1) 确定业务对象阶段

清晰地定义出业务对象,认清数据挖掘的目的是数据挖掘的首要任务。虽然挖掘的最后结果是不可预测的,但是要探索的问题应该是事先明确的。

(2) 数据准备阶段

数据挖掘处理的对象数据是原始数据,不适合在这些数据上进行知识挖掘,需要进行相应的处理,如数据的选择、净化(消除噪声、冗余数据)等,使其生成过程数据。然后进行转换,包括离散值数据与连续值数据之间的相互转换、数据值的分组分类、数据项之间的计算组合等,为后面的数据挖掘准备好正确的数据。

(3) 数据挖掘阶段

这是整个数据挖掘过程中最重要的一步,即使用适当的数据挖掘算法对前面处理过的数据进行分析,进而得到可能的模式或模型。根据不同数据的特点以及用户不同的需求,对同样的任务,可以选用不同的算法。

(4) 解释与评估阶段

数据挖掘将获得的信息通过用户可以理解和观察的方式反馈给用户。分析人员在使用数据挖掘结果之前,希望能够对挖掘的结果进行评估,以保证数据挖掘结果在实际应用中的成功率。在对挖掘结果进行评价时,需要考虑这样几个方面的问题:此结果要优于用不同的数据集在模型上的操作结果;模型的结果要比其他模型的结果更加准确;由于模型是以样本数据为基础建立的,因此实际结果往往要比建模时的结果差。

整个挖掘过程是一个不断反馈的过程。例如,用户在挖掘途中发现选择的数据不太好,或使用的挖掘技术产生不了期望的结果。这时,用户就需要重复先前的过程,甚至从头重新开始。严格地说,上述整个过程称为数据库上的知识发现,而仅把第三阶段称为数据挖掘,认为它是知识发现过程的一个步骤[32]。数据挖掘的四个主要阶段及各阶段的相关步骤如图 1.1 所示。

5. 数据挖掘系统的分类

数据挖掘源于多个学科,因此对数据挖掘研究就产生了大量的、各种不同类型的数据挖掘系统。这样,就需要对数据挖掘系统给出一个清楚的分类。这种分类可以帮助用户区分数据挖掘系统,确定最适合其需要的数据挖掘系统。根据不同的标准,数据挖掘系统可以分类如下[33]:

(1) 根据挖掘的数据库类型分类

由于数据库系统本身可以根据不同的标准分类,因此数据挖掘系统可以依此相应分类。如果根据数据模型分类,可以分为关系的、事务的、面向对象的、数据仓库的数据挖掘系统。如果根据所处理数据的特定类型分类,可以分为空间的、时间序列的、文本的、多媒体的、Web 的数据挖掘系统。

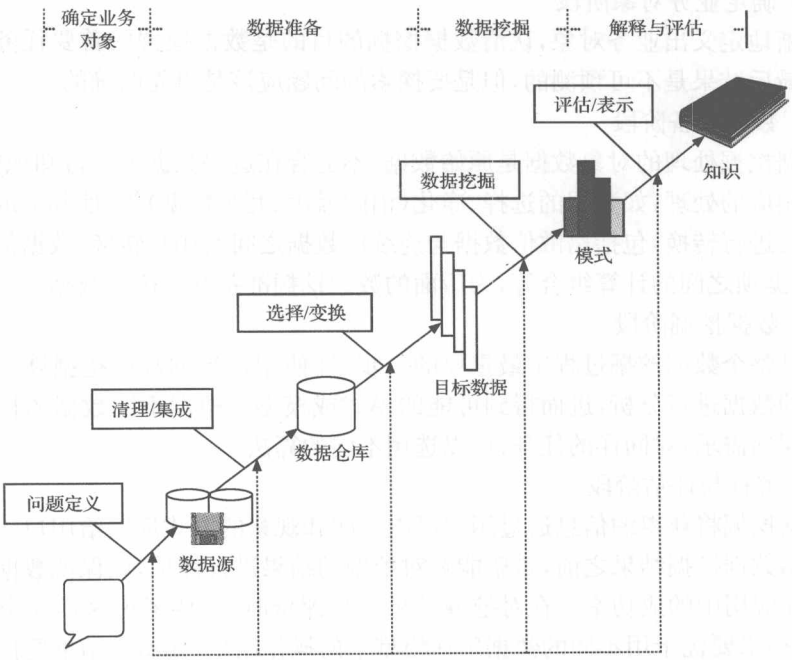

图 1.1 数据挖掘的过程

(2) 根据挖掘的知识类型分类

此类数据挖掘系统依据以下规则分类，这些规则有关联规则、分类规则、特征规则、序列模式、相似模式、混沌模式、聚类分析、孤立点分析、时间序列模式分析等。

(3) 根据挖掘方法分类

根据挖掘方法可以分为神经网络方法、统计方法、模糊集方法、支持向量机方法、粗糙集方法、公式发现方法、基于证据理论和元模式的方法等。

6. 数据挖掘的功能

数据挖掘是一门交叉学科，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。数据挖掘的主要功能和研究内容有[34]：

(1) 概念描述

概念描述本质上就是对某类对象的内涵特征进行概括。概念描述分为特征化描述和区别性描述。前者是描述目标类数据的一般特征和特性的汇总，后者是将目标类对象的一般特性与对比类对象的特性进行比较。获得概念描述的方法主要有以下两种[1]：利用更为广义的属性，对所分析数据进行概要总结[35,36]；对比目标

数据集和数据集的数据特点，并对对比结果给出概要性总结[37,38]。

（2）分类

分类是根据数据集的特点构造一个分类器，然后利用分类器给未知类别的样本赋予类别。构造分类器的过程一般分为训练和测试两个步骤。在训练阶段，建立模型来描述一个预先确定的数据类或概念的集合，即进行有监督的学习。在测试阶段，首先是采用测试数据集检验模型的准确度，如果达到预定要求，则可将模型用于预测未来数据对象的类别[39]。

（3）聚类

将物理或抽象对象的集合分组为由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇是一组数据对象的集合，同一个簇中的对象彼此相似，不同簇中的对象相异。聚类是在数据对象没有预定类别的前提下，根据类内相似性最大化、类间相似性最小化的原则，自动对数据分类是无监督的学习[40]。

（4）关联分析

关联分析用于发现事物间的关联规则，它是数据对象属性项之间的相互依赖关系。一个关联规则的形式为：$A_1 \wedge A_2 \wedge \cdots \wedge A_m \rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_n$。

如果 $B_1 \wedge B_2 \wedge \cdots \wedge B_n$ 出现，那么 $A_1 \wedge A_2 \wedge \cdots \wedge A_m$ 一定出现，这表明 $A_1 \wedge A_2 \wedge \cdots \wedge A_m$ 和 $B_1 \wedge B_2 \wedge \cdots \wedge B_n$ 有着某种联系。关联规则是 1993 年由 Agrawal 等提出的，然后扩展到从关系数据库、空间数据库和多媒体数据库中挖掘关联关系，并且要求挖掘通用的、多层次的、用户有兴趣的关联规则。随着应用和技术的发展，近年来对挖掘关联规则技术提出了更新、更高的要求，如在线挖掘、提高挖掘大型数据库的计算效率、减小 I/O 开销、挖掘定量型关联规则等。

（5）偏差检测

偏差检测就是对数据库中的偏差数据进行检测和分析。数据库中的数据常有一些异常记录，在某些特征上与数据库中的大部分数据有显著的不同。这些数据记录就是偏差，也叫孤立点。偏差检测方法主要有基于统计的方法、基于距离的方法和基于偏移的方法。孤立点数据的发现可以在信用卡使用、金融欺诈、医学数据分析等领域应用。

（6）预测

预测是从历史数据中找出变化规律、建立模型，并由此模型对新的样本数据类别及特征进行预测。一般意义上的预测是指利用回归方法预测连续值或有序值。预测关心的是精度和不确定性，通常用预测方差来度量。

（7）时序模式

时序模式是指通过时间序列搜索出重复发生概率较高的模式。与回归一样，它也是用已知的数据预测未来的值，但这些数据的区别是数据变量所处的时间点不同。