高维稀疏聚类知识发现

武森　高学东
M.巴斯蒂安　著

冶金工业出版社

高维稀疏聚类知识发现

武　森　高学东　M.巴斯蒂安　著

北　京
冶 金 工 业 出 版 社
2003

图书在版编目(CIP)数据

高维稀疏聚类知识发现/武森等著. —北京:冶金工业出版社,2003.1
ISBN 7-5024-3134-9

Ⅰ.高… Ⅱ.武… Ⅲ.计算机应用—经济管理 Ⅳ.F224-39

中国版本图书馆 CIP 数据核字(2002)第 085735 号

出版人 曹胜利(北京沙滩嵩祝院北巷 39 号,邮编 100009)
责任编辑 张 卫 美术编辑 王耀忠 责任校对 栾雅谦 责任印制 牛晓波
北京鑫正大印刷有限公司印刷;冶金工业出版社发行;各地新华书店经销
2003 年 1 月第 1 版,2003 年 1 月第 1 次印刷
850mm×1168mm 1/32; 4.125 印张; 107 千字; 119 页; 1-3000 册
14.50 元
冶金工业出版社发行部 电话:(010)64044283 传真:(010)64027893
冶金书店 地址:北京东四西大街 46 号(100711) 电话:(010)65289081
(本社图书如有印装质量问题,本社发行部负责退换)

前　言

许多企业已经成功地开发并有效地应用了计算机管理系统，这不仅提高了企业的管理效率，而且对企业决策活动也起到了一定的支持作用。但是，在实际的应用过程中，从传统的数据库中提取用于辅助决策的高度综合信息，不仅需要投入昂贵的人力和机器资源，而且花费的时间亦较长。随着数据库的应用和信息量的飞速增长，我们难以从巨量的数据中找到真正有用的决策信息，有人称之为“数据爆炸，而信息贫乏”。因此，企业迫切需要新的技术和方法从巨量数据中抽取有价值的信息或知识。

自 20 世纪 70 年代以来，人们一直在寻求决策支持系统解决方案，但实际应用并不很成功。数据库知识发现是对企业数据应用理念的一种转变，数据库知识发现的过程与传统的决策支持系统及专家系统截然不同。传统的方法是由专家或信息技术人员总结并描述知识或规则，从外部输入系统，形成知识库、模型库，进而用于辅助决策活动。由于知识大多具有复杂性和模糊性，难以准确描述，模型又难以精确地表述具体的应用，导致许多决策支持系统走向失败。数据库知识发现技术（Knowledge Discovery in Database，KDD）是从一个系统内部自动获取知识，从大量数据中寻找蕴涵其中但尚未发现的知识，这种数据应用技术的出现，必然会更有力地支持企业的战略决策。数据库知识发现还是一个很新的研究领域，但由于其具有为企业创造巨大经济效益的潜力，因而国内外

研究人员都给予了高度的关注。

本书针对数据库知识发现的聚类任务，介绍了高属性维稀疏聚类及聚类趋势发现的理论和方法。全书共分 8 章，内容包括数据库知识发现、聚类知识发现、高属性维稀疏数据聚类的 SFC 方法、高属性维稀疏数据聚类的 CABOSFV 算法、高属性维稀疏类的特征表示、聚类趋势发现、聚类知识发现的数据模型与数据准备、应用实现技术等。

本书适用于数据库知识发现领域的研究和应用人员，也可作为相关领域博士生、硕士生和本科生的参考书。

喻斌、崔焕荣和金海燕参加了本书第 8 章的写作，在此表示感谢！在本书的写作过程中，参阅了大量文献，在此向文献的作者表示感谢！

编 者

2002 年 10 月

目　录

1 数据库知识发现

1.1 数据库知识发现的产生与发展

数据库的广泛应用和数据量的飞速增长，使人们迫切地感到需要新的技术和工具以支持从大量的数据中智能地、自动地抽取出有价值的知识或信息，为此数据库知识发现(Knowledge Discovery in Database，KDD)技术应运而生。

数据库知识发现是从大量原始数据中挖掘出隐含的、有用的、尚未发现的信息和知识，它不仅被许多研究人员看作是数据库系统和机器学习等方面一个重要的研究课题，而且被许多工商界人士看作是一个能带来巨大回报的重要领域。“数据库知识发现”一词第一次出现是1989年8月在美国底特律召开的第11届国际人工智能联合会议的专题讨论会上。1991年、1993年和1994年又分别举行过数据库知识发现专题讨论会。由于参加会议的人数逐渐增多，所以从1995年开始，每年都要举办一次数据库知识发现国际会议。随着研究的不断深入，人们对数据库知识发现的理解越来越全面，对数据库知识发现的定义也不断修改，下面是对数据库知识发现比较公认的一个定义：数据库知识发现是从数据集中识别出可信的、有效的、新颖的、潜在有用的以及最终可理解模式的高级处理过程。

由于数据库知识发现的研究还处于初期阶段，而且是一门交叉性学科，受到来自各种不同研究领域学者的关注，所以有很多不同的名称。其中，最常用的术语是“数据库知识发现”和“数据挖掘”(Data Mining，DM)。相对来讲，后者主要流行于统计界，数据分析、数据库和管理信息系统界；而前者则主要流行于人工智能和机器学习界。随着数据库知识发现的迅速发展和逐渐为各界所了

解，较为普遍的观点认为："数据挖掘"是数据库知识发现中专门负责发现知识的核心环节；而"数据库知识发现"是一个交互式、循环反复的整体过程，除了包括数据挖掘外，还包括数据准备和发现的结果解释、评估等诸多环节。

1.2　数据库知识发现的处理过程

数据库知识发现过程如图 1.1 所示，整个过程可以理解为三个阶段：数据准备(Data Preparation)、数据挖掘、挖掘结果的解释与评估(Interpretation and Evaluation)。

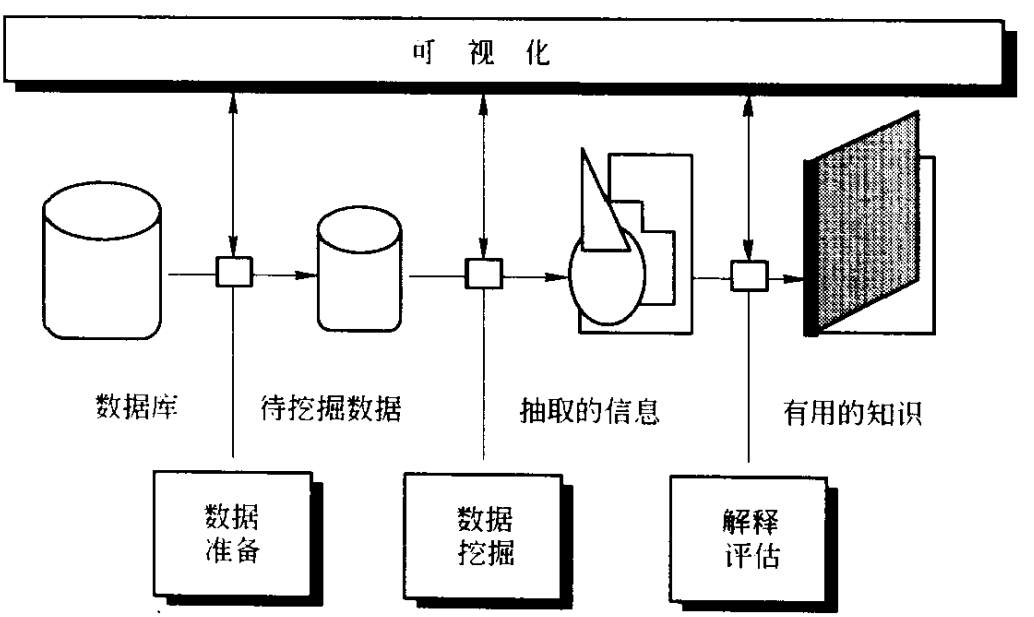

图 1.1　数据库知识发现的过程

数据准备阶段的工作包括四个方面的内容：数据的净化、数据的集成、数据的应用变换和数据的精简。

数据净化是清除数据源中不正确、不完整或其他方面不能达到数据挖掘质量要求的数据，例如推导计算缺值数据、消除重复记录等。数据净化可以提高数据的质量，从而得到更正确的数据挖掘结果。

数据的集成是在数据挖掘所应用的数据来自多个数据源的情况下，将数据进行统一的存储，并需要消除其中的不一致性。

数据的应用变换就是为了使数据适用于计算的需要而进行的数据转换。这种变换可能是现有数据不满足分析需求而进行的，也可能是所应用的具体数据挖掘算法对数据提出的要求。

数据精简是采用一定的方法对数据的数量进行缩减，或从初始特征中找出真正有用的特征来消减数据的维数，从而提高数据挖掘算法的效率与质量。

数据挖掘阶段首先要确定挖掘的任务或目的，如数据总结、分类、聚类、关联规则发现或序列模式发现等。确定了挖掘任务后，就要决定使用什么样的挖掘算法。同样的任务可以用不同的算法来实现，一般要考虑多方面的因素来确定具体的挖掘算法，例如：不同的数据有不同的特点，因此需要用与之相关的算法来挖掘；用户对数据挖掘有着不同的要求，有的用户可能希望获取描述型的、容易理解的知识，而有的用户（或系统）的目的是获取预测准确度尽可能高的预测型知识。

需要指出的是，尽管数据挖掘算法是数据库知识发现的核心，也是目前研究人员主要努力的方向，但要获得好的挖掘效果，必须对各种挖掘算法的要求或前提假设有充分的理解。

数据挖掘阶段发现得到的模式，经过用户或机器的评估，可能存在冗余或无关的模式，这时需要将其剔除；也有可能模式不满足用户要求，这时则需要整个发现过程退回到挖掘阶段之前，如重新选取数据、采用新的数据变换方法、设定新的数据挖掘参数值，甚至换一种挖掘算法。另外，数据库知识发现由于最终是面向用户的，因此可能要对发现的模式进行可视化，或者把结果转换为用户易懂的方式表达。

1.3 数据库知识发现的任务

数据库知识发现的核心部分——数据挖掘按挖掘任务分为：分类知识发现、数据总结、数据聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常发现和趋势预测等。

分类知识发现是数据挖掘中最常见的任务，其目的在于根据

样本数据寻求相应的分类规则，然后根据获得的规则来确定某一非样本个体或对象是否属于某一特定的组或类。在这种分类知识发现中，样本个体或对象的类标记是已知的。数据挖掘的任务在于从样本数据的属性中发现个体或对象分类的一般规则，从而根据该规则对非样本数据对象进行分类应用。

数据聚类是用于发现在数据库中未知的数据类。这种数据类划分的依据是“物以类聚”，即按个体或数据对象间的相似性，将研究对象划分为若干类。由于在数据挖掘之前，数据类划分的数量与类型均是未知的，因此在数据挖掘后需要对数据挖掘结果进行合理的分析与解释。

关联规则发现是在数据库中寻找数据对象间的关联模式，例如“在购买个人电脑的顾客中，有 90% 的顾客购买了打印机”，就是一种关联模式。关联模式发现在研究和应用的早期主要用于零售业交易数据分析，以便进行物品更合理的摆放，最终提高销售量。因此，该方法有时也直接称为“货篮分析”。

数据总结是将数据库中的大量相关数据从较低概念层次抽象到较高概念层次的过程。计数、求和、求平均值、求最大值和最小值等计算都是数据总结的具体化。由于数据库中的数据所包含的信息往往是最原始、最基本的信息，而有时人们需要从较高的层次上浏览数据，这就要求从不同的层次上对数据进行总结以满足分析需要。

序列模式发现是在数据库中寻找基于一段时间区间的关联模式，例如“在某一时间购买个人电脑的所有顾客中，有 60% 的顾客会在三个月内购买应用软件”，这就是一序列模式。序列模式同关联模式非常相似，区别在于序列模式表述的是基于时间的关系，而不是关于数据对象间的关系。因此，在有些文献中也称其为基于时间的关联规则发现。

依赖关系或依赖模型发现是通过对数据库中数据的分析，获取数据间的某种因果联系。这种因果联系既可能是内在的某种概率分布关系的描述，也可能是数据对象间存在的确定的函数关系。

异常发现用于在数据库中发现数据中存在的偏差或异常，例如下列几种偏差或异常就应引起人们的关注：不符合任何一标准类的异常，有时可能意味着严重的错误或欺诈；相邻时间段内信息的异常变动，如二月份与一月份相比，二月份销售收入的骤然升高。

趋势预测是根据数据库中的历史信息对未来信息作出估计。实际上，预测这一数据挖掘任务并不一定是独立的。一般来讲，上述几种数据挖掘任务的结果，皆可以在分析后用于趋势预测。

1.4 数据仓库与数据挖掘

数据挖掘的研究涉及机器学习、数据库、模式识别、统计学、人工智能、管理信息系统、知识获取、数据可视化等许多领域。其中，数据挖掘与数据库新技术——数据仓库有着密切的关系。

数据仓库技术是近十几年来信息技术领域迅速发展起来的一种数据组织和管理技术。按照数据仓库技术的创始人依曼(W. H. Inmon)的定义，“数据仓库是面向主题的、综合的、不同时间的、稳定的数据集合，它主要用于支持经营管理中的决策制定过程。”数据仓库中存储的是面向分析型应用的集成的数据，一般包含 5～10 年的历史数据。

数据仓库技术源于数据库技术，它的主要设计思想是将分析决策所需的大量数据从传统的操作环境中分离出来，把分散的、难以访问的操作数据转换成集中统一的、随时可用的信息而建立的一种数据库存储环境。可以说，数据仓库是一种专门的数据存储，用于支持分析型的数据处理。如何将数据仓库与数据挖掘结合在一起，更好地支持分析决策，已引起了研究领域的普遍关注。数据仓库为数据挖掘提供了数据基础，在数据仓库中进行数据挖掘亦对数据挖掘提出了更高的要求：

(1) 数据仓库中集成和存储着来自若干异构的数据源的信息。这些数据源本身就可能是一个规模庞大的数据库，有着比一般数据库系统更大的数据规模。这就要求在数据仓库中进行数据挖掘的算法必须更有效、更快速。

(2) 在一般的数据库中,为了提高系统的效率,一般会尽可能少地保留历史信息。而数据仓库具有一个重要的特征,就是一般具有长时间的历史数据存储。存储长时间历史数据的目的就是进行数据长期趋势的分析。数据仓库为决策者的长期决策行为提供了有力的数据支持。然而,数据仓库中的数据在时间轴上的特征,在一定程度上增加了数据挖掘的难度。

另一方面,数据仓库也为数据挖掘创造了更方便的数据条件,体现在:

(1) 从一个企业的角度看,数据仓库集成了企业内各部门的全面的、综合的数据。数据挖掘要面对的是企业全局模式的知识发现。从这一点上讲,基于数据仓库的数据挖掘能更好地满足高层战略决策的要求。而且,数据仓库大大地降低了数据挖掘的障碍。数据挖掘一般要求大量的数据准备工作,而数据仓库中的数据已经被充分收集起来,进行了整理、合并,并且有些还进行了初步的分析处理。这样,可以集中精力于数据挖掘核心处理阶段。另外,数据仓库中对数据不同粒度的集成和综合,更有效地支持了多层次、多种知识的挖掘。

(2) 数据仓库是面向决策支持的,因此它的体系结构努力保证了查询和分析的实时性;而一般的联机事务处理系统则主要要求更新的实时性。一般的数据仓库设计成只读方式,最终用户不能更新数据。数据仓库中的数据更新由专门的一套机制来实现。数据仓库对查询的强大支持使数据挖掘效率更高,挖掘过程可以做到实时交互,使决策者的思维保持连续,有可能发现更深入、更有价值的知识。

综上所述,可以说数据仓库在纵向和横向都为数据挖掘提供了更广阔的空间。数据仓库完成了数据的收集、集成、存储、管理等工作,数据挖掘面对的是经过初步加工的数据,使得数据挖掘能更专注于知识的发现;另一方面,由于数据仓库所具有的新的特点,又对数据挖掘技术提出了更高的要求。所以说,数据挖掘技术与数据仓库技术结合起来,能够更充分地发挥潜力。

2 聚类知识发现

2.1 聚类知识发现概述

聚类(Clustering)是数据挖掘领域最为常见的技术之一,用于发现在数据库中未知的对象类。通过聚类过程形成的每一个组称为一个类(Cluster)。在数据挖掘之前,对象类划分的数量与类型均是未知的,因此在数据挖掘后一般需要对数据挖掘结果进行合理的分析与解释。聚类是现实世界中普遍存在的现象,其应用也非常广泛。据文献所载,在破产预测、手写体字符的计算机识别、交通管理与塞车状况预测等方面都有过成功的应用。

在20世纪70年代,对聚类分析已经有了比较深入的研究。聚类的方法主要有统计学方法和机器学习的方法。在统计学中,聚类一般称为聚类分析,主要研究基于几何距离的聚类。在使用上,首先要定义多维空间,在多维空间中计算对象间的距离,然后以距离作为对象间的相似性判别标准。在机器学习中,聚类称为无监督学习(Unsupervised Study),主要体现在聚类学习的数据对象没有类别标记,需要由聚类学习算法自动计算。

近十年左右,随着数据库知识发现技术的兴起,对聚类的研究掀起了新的热潮。除了统计学和人工智能领域的研究人员以外,数据库领域的人员也加入到了这一研究队伍中,并取得了可喜的成果。从数据库知识发现的角度来讲,对聚类问题的研究是要从大量的数据集中智能地、自动地抽取出有价值的聚类知识,在本书中将称其为聚类知识发现。

在下面的几节中,将首先介绍对象间相似性的计算方法,即差

异度的计算，然后讨论几种主要的聚类方法，最后从数据库知识发现的角度总结了聚类问题的研究难点。

2.2 差异度的计算方法

聚类的主要依据是对象间的相似性。确定对象之间是否相似，是通过计算对象之间的差异度来完成的。在描述对象的属性取值类型不同时，差异度的计算方法也不相同。下面给出对象属性取值分别为区间变量、二态变量时差异度的计算方法。

2.2.1 区间变量及其差异度

区间变量是一种连续变量，一般取值为线性度量值，例如：高度、长度、宽度、重量等，都是区间变量。

假设有 n 个对象，描述第 i 个对象的 m 个属性值分别对应于区间变量值 $x_{i1}, x_{i2}, \cdots, x_{im}$，描述第 j 个对象的 m 个属性值分别对应于区间变量值 $x_{j1}, x_{j2}, \cdots, x_{jm}$，$i, j \in \{1,2,\cdots,n\}$，那么对象 i 与 j 之间的差异度一般以它们之间的距离 $d(i,j)$ 来表示。距离越近，表明对象 i 与 j 之间越相似，差异越小；距离越远，表明对象 i 与 j 之间越不相似，差异越大。

距离 $d(i,j)$ 的计算主要有如下三种方法。

A 欧几里德(Euclidean)距离

欧几里德距离是比较常用的距离计算方法。实际上，在人的肉眼可以辨识的三维空间中物体之间距离的计算，采用的就是欧几里德距离，其具体计算方法为：

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{im}-x_{jm}|^2} \tag{2.1}$$

$$i\in\{1,2,\cdots,n\}; \qquad j\in\{1,2,\cdots,n\}$$

B 绝对值(Manhattan)距离

绝对值距离也是一种比较常见的距离计算方法，其具体的计算方法为：