浙江金融职业学院
"985 工程"项目成果

基于约束的关联规则挖掘

Constraint-based Association Rule Mining

陶再平 著

浙江工商大学出版社
ZHEJIANG GONGSHANG UNIVERSITY PRESS

浙江金融职业学院
“985工程”项目成果

基于约束的关联规则挖掘

Constraint-based Association Rule Mining

陶再平 著

浙江工商大学出版社

图书在版编目(CIP)数据

基于约束的关联规则挖掘 / 陶再平著.
— 杭州 : 浙江工商大学出版社, 2012.3
ISBN 978-7-81140-476-0

Ⅰ. ①基… Ⅱ. ①陶… Ⅲ. ①数据采集—
研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2012)第 030831 号

基于约束的关联规则挖掘

陶再平 著

责任编辑 刘 韵 赵 丹
责任校对 周敏燕
封面设计 流 云
责任印制 汪 俊
出版发行 浙江工商大学出版社
(杭州市教工路 198 号 邮政编码 310012)
(E-mail:zjgsupress@163.com)
(网址:http://www.zjgsupress.com)
电话:0571-88904980,88831806(传真)
排 版 杭州朝曦图文设计有限公司
印 刷 杭州杭新印务有限公司
开 本 850mm×1168mm 1/32
印 张 4.75
字 数 123 千
版 印 次 2012 年 3 月第 1 版 2012 年 3 月第 1 次印刷
书 号 ISBN 978-7-81140-476-0
定 价 15.00 元

前　言

关联规则挖掘是数据挖掘领域中一个非常重要的研究课题。在R. Agrawal等人提出该问题之后，关联规则的挖掘就越来越受到人们的关注。约束条件的引入使得关联规则的挖掘更具有针对性，用户能够更好地参与和控制整个数据挖掘进程。本书重点讨论约束条件下的关联规则的挖掘问题，对不同类型约束条件下的关联规则的挖掘问题进行了深入的分析和研究，提出并验证了若干效率较高的挖掘算法。

本书共分为七章，其中第一章主要介绍了数据挖掘的概念、数据挖掘系统的结构、数据挖掘的分类情况以及算法的组成、常见的数据挖掘工具、典型的数据挖掘系统和数据挖掘的应用情况。

第二章主要介绍了关联规则的定义及其发现步骤、关联规则的分类以及挖掘算法的分类情况，以经典的关联规则挖掘算法——Apriori算法为研究对象分析了关联规则的挖掘过程，并分析了Apriori算法存在的不足。另外，还介绍了典型的改进算法和关联规则的生成算法，讨论了基于约束条件的关联规则的挖掘问题以及关联规则在各个领域的应用情况。

第三章重点讨论了周期性关联规则的挖掘问题。对固定周期和可变周期两种情况下的周期性关联规则的挖掘进行了分析，并针对

这两种情况分别提出了相应的周期性关联规则的挖掘算法。算法结合了支持度剪枝和周期性剪枝技术，较大程度地提高了规则的挖掘效率。

第四章分析了基于项约束条件下的关联规则挖掘问题，提出了解决基于项约束的关联规则挖掘问题的 ARMIC 算法。该算法结合了 Reorder 算法和 Direct 算法的优点，并对挖掘的目标数据集进行了有效的过滤，从而减小了数据扫描空间，提高了算法的效率。

第五章对基于属性约束的关联规则挖掘问题进行了研究，重点分析和归纳了不同类型的属性约束条件的性质，给出了解决不同类型属性约束条件下的关联规则挖掘算法，提出了一个基于约束的关联规则挖掘系统的模型。

第六章主要对基于时间窗口约束的关联规则增量式更新问题进行了研究，给出了基于时间窗口的关联规则增量式更新的 TW_EIUP 算法。该算法充分利用了以前的挖掘结果，并对候选项集进行有效的剪枝，减少了对各数据集的扫描次数，测试结果表明 TW_EIUP 算法具有较高的挖掘效率。

第七章分析了数据挖掘的发展趋势以及面临的挑战。

目　录

第一章　数据挖掘综述

近年来，随着信息技术的日益普及，各个领域的数据量都在急剧增加。成千上万的数据库被广泛应用于商业管理、行政管理、科学与工程数据管理以及其他各个不同的应用领域，并且这些数据库应用的数量正在像雨后春笋般不断涌现。这主要归因于数据存储技术的快速发展以及强大且经济的数据库系统的开发。现代计算机技术与数据库技术使我们能够将“数据洪流”转换为“整齐有序”但却是“堆积如山”的数据集合。

面对“堆积如山”的数据集合，无论在时间意义上还是空间意义上，传统的数据分析手段都难以对付。人们无法理解并有效地使用这些数据，这就迫使决策者不得不采用两种可能导致“数据灾难”的对策：其一是“穷于应付”；其二是“置之不理”。事实上，无论哪一种对策都是出于无奈，并由此导致越来越严重的“数据灾难”。无怪乎奈斯伯特（John Naisbett）惊呼：“We are drowning in the information, but starving for knowledge.”（人类被数据所淹没，但知识依然匮乏。）在需要对大量的数据进行分析之后才能作出正确决策的领域中，这是一个普遍存在的问题。

传统的数据分析方法，如电子报表、统计分析、特定查询等，都只能获得数据的表层信息，而不能获得数据各属性之间的内在联系和隐含信息，即不能获得重要的知识。于是快速的数据产生和收集技术与拙劣的数据分析方法之间就形成了尖锐的矛盾。这就需要新的技术来“智能地”和“自动地”分析这些原始数据，使消耗了大量财力

与物力所收集与整理的宝贵数据资源能得以充分利用。这种需求引起了数据库、人工智能、统计学、数据仓库、在线分析处理、专家系统、高性能计算和数据可视化等研究领域学者的广泛重视。于是就使得数据挖掘和数据库知识发现技术应运而生。数据挖掘和数据库知识发现实际上是由上述各个学科相互交叉、融合所形成的一个新兴的研究领域，目前已成为国际上数据库、人工智能界研究的热点。

一、数据挖掘与知识发现

(一)基本概念

什么是数据挖掘？简单地说，数据挖掘是从大量数据中提取或"挖掘"知识。更确切地说，数据挖掘(Data Mining，简称 DM)是指从大量数据中提取潜在的有用的信息，并从不同角度观察或浏览它们，从中发现的知识可以用于辅助决策、信息管理以及查询处理。它主要是基于人工智能、机器学习、统计学等技术高度自动化地分析数据，并作出归纳性的推理，从中挖掘出潜在的模式，从而帮助决策者作出正确的决策。与数据挖掘类似或相近的术语还有如数据库中知识挖掘、知识提取、数据/模式分析、数据考古和数据捕捞。

长期以来，在知识发现领域，数据挖掘(DM)和数据库知识发现(Knowledge Discovery in Databases，简称 KDD)这两个术语的范畴和使用界限一直不是很清晰。知识发现术语在 1989 年的第一届 KDD 专题讨论上首次被采用，用于表示在数据库中发现知识的广泛进程。这个术语强调了知识是数据发现的最终产品，并很快在人工智能和机器学习领域得到了广泛应用。

在 1996 年 KDD 国际会议上，Fayyad、Piatetsky Shapiro 和 Smyth 对 KDD 下了最新的定义：数据库知识发现(KDD)是指识别出存在于数据库中有效的、新颖的、具有潜在效用的、最终可理解的模式的非平凡过程。

KDD过程是一个多步骤的处理过程(如图1-1)。它主要包括以下一些处理步骤。

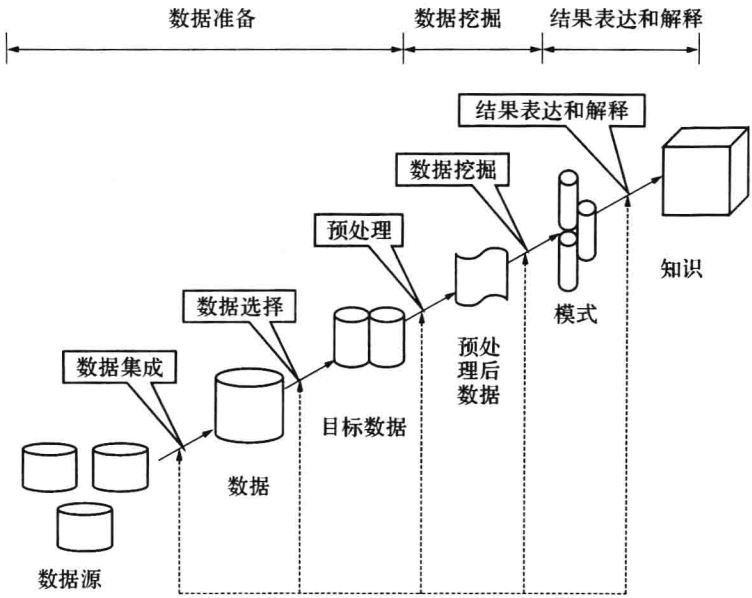

图1-1　数据库知识发现过程

1. 数据准备

了解KDD相关领域的有关情况,熟悉有关的背景知识,并弄清楚用户的需求。

2. 数据选择

根据用户的需求从数据库中提取与KDD相关的数据。KDD将主要从这些数据中进行知识提取。在此过程中,经常会利用一些数据库操作对数据进行处理。

3. 数据预处理

主要是对前一阶段产生的数据进行加工,检查数据完整性和数据的一致性,对其中的噪音数据进行处理,对丢失的数据可以利用统

计的方法进行填补。

4. 数据缩减

对经过预处理的数据，根据知识发现的任务对数据进行再处理，主要通过投影、降维或数据转换等操作来减少数据量，降低数据的复杂性。

5. 确定 KDD 目标

根据用户的需求，确定 KDD 是发现何种类型的知识。因为 KDD 要求不同，在具体的知识发现过程中将会采用不同的知识发现算法。

6. 确定知识发现算法

根据步骤 5 所确定的任务，选择合适的知识发现算法，这包括选取合适的模型和合理的参数，并使得知识发现算法与整个数据库知识发现(KDD)的评判标准保持一致。

7. 数据挖掘(DM)

运用选定的知识发现算法，从数据中提取出用户所关心的知识。这些知识可以用一种特定的方式表示或使用一些常用的表示方式，如产生式规则等。

8. 模式解释

对发现的模式进行解释。在此过程中，为了取得更为有效的知识，可能会返回到前面处理过程中的某一步骤(环节)进行反复提取，从而提取出更加有效的知识。

9. 知识呈现和评价

将发现的知识以用户能理解的方式呈现给用户。这期间也包括对知识一致性的检查，以确保本次发现的知识不与以前发现的知识相抵触。

整个发现过程并不是简单的线性流程，各步骤之间包含了循环和反复。这样就可以对所发现的知识不断地求精和深入，并使其易于理解。

Fayyad、Piatetsky Shapiro 和 Smyth 等人认为：KDD 是指从数据库中发现知识的全部过程，是应用特定的数据挖掘算法和评价解释模式的一个循环反复的过程，并要对所发现知识进行不断地求精和深化，使其易于理解；而 DM 只是这一过程中的一个特定步骤。即利用挖掘算法从数据中抽取模式，它并不包括数据的预处理、领域知识的结合以及发现结果的评价等步骤。

但也有很多学者如 Han 和 R. Agrawal 等人则认为：数据挖掘(DM)和数据库知识发现(KDD)是两个等价的概念。从数据挖掘定义可以看出，作为一个学术领域，数据挖掘和数据库知识发现具有很大的重合度。数据挖掘从理论和技术实现上继承了知识发现领域的成果，同时又有着独特的内涵。数据挖掘更着眼于设计高效的算法以达到从海量数据中快速发现知识的目的。数据挖掘充分利用了机器学习、人工智能、模糊逻辑、人工神经网络、分形几何等方面的理论和方法。因此，在很多情况下，数据库中的知识发现和数据挖掘两个术语往往不加以明确区分。

与数据挖掘关系密切的研究领域包括归纳学习(Inductive Learning)、机器学习(Machine Learning)和统计分析(Statistics)等。特别是机器学习被认为和数据挖掘的关系最为密切。两者的主要区别在于：数据挖掘的任务是发现可以理解的知识，挖掘的对象一般是大规模的数据库；而机器学习关心的是提高系统的性能，一般来说其处理的数据集要小得多。因此，效率问题对数据挖掘是至关重要的。

(二)数据挖掘与 OLAP

数据挖掘(DM)和联机分析处理(OLAP)都属于分析型工具，但两者之间有着明显的区别。

数据挖掘是一种有效地从大量数据中发现潜在的数据模式，并作出预测性分析的分析型工具。它是现有的一些人工智能、统计学等成熟技术在特定的数据库领域中的应用。它与其他分析工具最大

不同之处在于:它的分析过程是自动的,它能自动地发现隐藏在数据中的模式。一个成熟的数据挖掘系统除了具有良好的核心技术外,还应该具有开放性的结构和友好的用户接口。数据挖掘不需要用户提出确切的问题,而由数据挖掘去挖掘隐藏的模式并预测未来的趋势,这样更有利于发现未知的事实。

联机分析处理(OLAP)是一种自上而下、不断深入的分析工具。首先用户提出问题或假设,然后由 OLAP 负责自上而下地提取出关于该问题的详细信息,并以可视化的形式呈现给用户。与数据挖掘相比,OLAP 更需要依靠用户输入的问题或假设。但用户先入为主的局限性往往会限制问题和假设的范围,从而影响最终的结论。因此,作为验证型分析工具,OLAP 更需要对用户的需求有全面而深入的了解。显然,从对数据分析深度的角度来看,OLAP 位于较浅的层次,而数据挖掘所处的位置则较深。两者最关键的区别在于信息发现过程是否是自动的。

需要说明的是,数据挖掘并非是一些人想象中的魔法。数据挖掘工具也不是驻扎在数据库中的幽灵,监视着数据库里发生的一举一动,当发现任何有趣的趋势或变化时,便会自动发出一封电子邮件来提醒用户注意。数据挖掘仍然需要用户的主动参与和控制。只有计算机和用户各自承担其最擅长的工作,才能使数据挖掘工具真正发挥出最大的作用。

二、数据挖掘系统的结构

典型的数据挖掘系统结构如图 1-2 所示,它由以下几个主要成分组成。

1. 数据源

数据源可以是一个或多个数据库、数据仓库、电子表格或其他类型的数据库。它们应该是经过集成、过滤和抽取以进一步处理的数据集合。

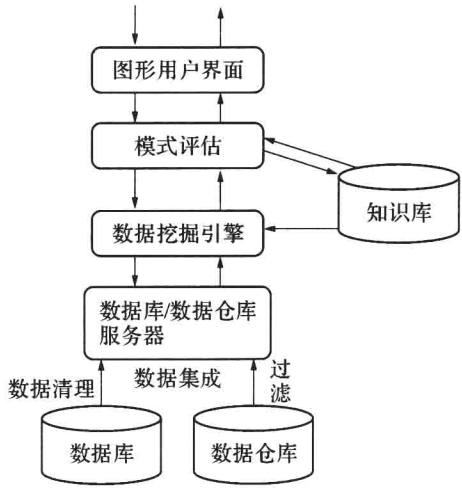

图 1-2　数据挖掘系统结构

2. 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。

3. 知识库或模式库

知识库或模式库是存储挖掘出来的中间或最终结果模式和知识，也存储用于指导搜索或评估结果模式的领域知识。这样的知识可能包含不同的抽象层次，适用于不同粒度的知识、多种模式、多种知识的表达形式等。

4. 数据挖掘引擎

这是数据挖掘系统最基本的部分，由一组功能模块组成。它具有知识推理机的功能，能反复利用已获得的知识和用户互动，达到最终形成知识模式的目的。用于特征化、关联规则、分类规则、聚类分析以及演变和偏差分析等。

5. 模式评估模块

使用兴趣度度量(包括支持度、置信度等)，并与数据挖掘模块交

互，以便将搜索聚焦在有趣的(Interesting)模式上。为提高挖掘效率，模式评估工作应尽可能深入到挖掘的不同层次中，这样可以保证搜索限制在感兴趣的模式中。

6. 可视化用户界面

数据挖掘系统必须允许用户能够指定数据挖掘查询或任务。对有经验的用户，理想的情况应允许他们使用约束条件等形式指导不同阶段的挖掘工作。通过用户和知识库的互动，帮助系统聚焦搜索。在需要的时候，也可以根据数据挖掘的中间结果进行探索式挖掘。此外，用户还可以通过直观而且形式多样的显示方式展现挖掘结果。也可以允许用户浏览数据库和数据仓库模式或数据结构，评估挖掘的模式，帮助用户了解系统的状况。

一般来说，数据挖掘可以在任何类型的信息存储上进行，这包括关系数据库、数据仓库、事务数据库、高级数据库系统、面向特殊应用的数据库和 Web 网页等。高级数据库系统包括面向对象数据库和对象-关系数据库，面向特殊应用的数据库包括空间数据库、时间序列数据库、文本数据库和多媒体数据库等。

一个实用的数据挖掘系统必须具有以下特征：

(1)快速的响应能力，即便是系统在处理长时间的操作时，也能及时处理用户的指导，并反馈给用户；

(2)处理大数据集的能力，即实现的方法具有较好的时间复杂度；

(3)具有良好的人机交互界面，采用多种形式输出挖掘结果；

(4)具有良好的扩充性，方便加入新的数据挖掘算法；

(5)自适应的选择和建议较好参数及模型的能力，因为用户往往不是专家，无法知道哪种模型适合目前的数据。

三、数据挖掘的分类

近年来，大量的数据挖掘技术和挖掘系统得以开发。从不同的

角度出发，可以对数据挖掘进行不同的分类。分类方法主要有根据挖掘的数据库类型分类、根据所采用的技术分类、根据所发现的知识类型分类等几种。

1. 根据挖掘的数据库类型分类

数据挖掘系统可以根据所处理的数据库类型进行分类。例如数据库为关系型的，它就是关系型的数据挖掘系统；数据库为面向对象型的，它就是面向对象型的数据挖掘系统。一般来讲，一个数据挖掘系统可以根据不同的数据库类型进行分类：如关系(Relational)型数据库挖掘、事务(Transactional)数据库挖掘、面向对象(Object-Oriented)数据库挖掘、演绎(Deductive)型数据库挖掘、空间(Spatial)数据库挖掘、多媒体(Multimedia)数据库挖掘、异质(Heterogeneous)数据库挖掘、主动(Active)数据库挖掘、时间(Temporal)序列数据库的数据挖掘、基于因特网的数据挖掘、数据仓库(Data Warehouse)挖掘等。

2. 根据挖掘的知识类型分类

根据挖掘的知识类型分类可以分为：关联(Association)规则、分类(Classification)规则、特征(Characterization)规则、区分(Discrimination)规则、聚类(Clustering)规则、汇总(Summarization)规则、预测(Prediction)分析、趋势(Trend)分析、演化和差异(Deviation)分析等。

(1)关联(Association)规则

关联规则是发现数据对象间的相互依赖关系。一条关联规则的形式为：$A_1 \wedge A_2 \wedge \cdots \wedge A_i \Rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_j$，如果 $B_1, B_2, \cdots, B_j$ 出现，则 $A_1, A_2, \cdots, A_i$ 就一定出现。这表明数据 $A_1, A_2, \cdots, A_i$ 和数据 $B_1, B_2, \cdots, B_j$ 有着某种联系。例如，在疾病症状的研究过程中，人们也许会发现某些症状的出现一定会伴随着其他一些症状的出现，通过对这种现象的深入研究，将会有助于疾病的诊断。

购物篮分析问题就是一个非常典型的关联规则挖掘的应用实