基于关联规则和粗糙集的数据挖掘方法研究

任志波/著

科学出版社

基于关联规则和粗糙集的数据挖掘方法研究

任志波　著

本书受河北大学双一流学科管理科学与工程项目资助

科学出版社

北京

内 容 简 介

本书是一本深入研究数据挖掘领域中关联规则挖掘和可变精度模糊粗糙集理论的著作，其中关联规则挖掘以频繁集挖掘为主要内容，研究包括如何充分利用模糊约束进行频繁集的挖掘，高效的最大频繁集挖掘算法，以及可变精度模糊粗糙集的性质和算法。本书强调理论性和技术性的统一，在理论研究的同时提供了技术的实现。

本书可供数据挖掘领域内的研究人员、技术人员参考，也可以用做高年级本科生或者研究生的参考书目。

图书在版编目（CIP）数据

基于关联规则和粗糙集的数据挖掘方法研究 / 任志波著. —北京：科学出版社，2017.10

ISBN 978-7-03-053008-0

Ⅰ. ①基… Ⅱ. ①任… Ⅲ. ①数据采集–研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 119024 号

责任编辑：马 跃 方小丽 / 责任校对：杜子昂

责任印制：吴兆东 / 封面设计：无极书装

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

北京京华虎彩印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 10 月第 一 版 开本：720×1000 B5

2017 年 10 月第一次印刷 印张：8 1/2

字数：171 000

定价：58.00 元

（如有印装质量问题，我社负责调换）

前　言

当今社会已经进入了大数据时代，数据挖掘（data mining）是大数据应用中的核心技术。本书主要研究数据挖掘技术中关联规则（association rule）和可变精度模糊粗糙集（rough set）的一些问题，主要内容如下。

（1）总结数据挖掘的主要内容，论述数据挖掘的概念、步骤和任务。

（2）分析关联规则挖掘和模糊粗糙集的研究现状以及存在的问题，重点讨论关联规则挖掘中的模糊约束问题、最大频繁集问题和模糊粗糙集研究中存在的问题。

（3）提出模糊约束的概念和性质。模糊约束更符合人类的思维习惯。引入模糊约束的最大挑战如下：很多模糊约束具有分段性质，如何在挖掘过程中同时使用单调和反单调的约束是一个研究重点，同时也是一个难点。根据这种情况，我们给出了两种算法：第一种算法在 Apriori 算法的基础上，使用单调约束裁减事务空间，使用反单调约束裁减项集空间，同时为了得到最优的模糊集，我们使用了遗传算法寻优；第二种算法采用双向搜索策略，在自上而下搜索和自下而上搜索过程中，同时应用两种约束裁减项集空间，这充分使用了模糊约束的裁减能力，大大缩减了搜索空间，进而提高了算法效率。

（4）针对当前最大频繁集挖掘算法存在的问题，提出了一种新的算法。当前算法大都对某一类数据有效，如稠密数据或稀疏数据。我们提出的算法对稠密数据和稀疏数据都基本有效。该算法主要的创新之处在于：采用有向

图表示项集，将频繁集发现转化为有向图中的搜索。另外，该算法采用了位图表示数据以及压缩、多种裁减技术。实验表明其达到了预期目的。

（5）模糊粗糙集模型比经典粗糙集模型更适合处理不确定的数据，为了处理包含噪声的数据，Rolka 等定义了可变精度模糊粗糙集，但没有给出该定义对应的代数性质和约减算法。我们证明该种定义满足的一些代数性质，并推广模糊粗糙集的约减算法，形成可变精度模糊粗糙集的约减算法。

任志波

2017 年 6 月于河北大学

目　录

第1章 绪 论

1.1 研究意义

关联规则和粗糙集技术是数据挖掘领域重要的两个工具，这一节主要对这两个工具的研究意义进行简单的阐述。

1.1.1 研究关联规则的意义

关联规则的研究是从购物篮数据分析开始的。随着条形码技术的应用，大型超市聚集了海量数据，如美国的零售商沃尔玛 1998 年搜集了一个 11 太比特的客户交易数据库[1]。企业希望从这些数据中找到能够帮助其进行决策的规则或模式，如反映用户购买倾向的模式，“啤酒→尿布”规则就是一个例子，即购买啤酒的一部分人有一种购买尿布的倾向，超市管理人员据此可以调整货价摆放、联合促销，增加销售量。当前，关联规则挖掘在很多领域都有应用，如网络链接分析、广告邮寄分析、生物信息挖掘等。

关联规则挖掘是数据挖掘的一个核心问题，由于其面对的是海量的数据，传统的统计方法或人工智能方法都难以处理这么大规模的数据，为此，关联规则挖掘和其他数据挖掘技术面临相同的问题，即面对海量数据时挖掘算法的效率问题。本书对关联规则的探讨都涉及效率问题，我们主要讨论了

带有模糊约束的频繁集挖掘问题和最大频繁集挖掘的算法问题。模糊约束允许用户以符合人们思维习惯的方式提出约束，以此表达用户关注的焦点，模糊约束包含用户的背景知识，是挖掘算法充分利用所研究问题的领域知识的一个有效途径。通过模糊约束，算法可以充分裁减搜索空间，提高算法的效率，这样不仅节省了挖掘需要的时间，而且最终得到的也是用户关心的结果。最大频繁集挖掘是为了提高算法效率而提出的。由于最大频繁集包含了所有频繁集，且最大频繁集数量又远远小于频繁集的数量，这意味着挖掘最大频繁集可以有更高的效率。总之，无论是带有模糊约束的频繁集挖掘还是最大频繁集挖掘，都会有更高的挖掘效率，因此这是我们研究关联规则问题时重点关注的两个问题。

1.1.2 研究可变精度模糊粗糙集的意义

粗糙集理论是波兰数学家 Pawlak 在 1982 年提出的一种处理含糊和不精确问题的数学工具[2]，该理论可以通过对属性数据的约减，最终得到决策规则。经典的粗糙集理论仅能处理离散的数据，对连续的数据，需要离散化处理，但离散会导致信息的丢失，影响最终决策规则的准确性。对连续数据的模糊化处理也是一种常见的处理方法，模糊化处理比离散化丢失的信息要少，另外，在很多情况下，描述对象最终形成的数据也是取值为模糊的数据，为此，Dubois 和 Prade 提出了结合模糊集理论和粗糙集理论的模糊粗糙集理论[3]，用于处理这种情况时数据的属性约减问题。由于现实数据中经常存在噪声数据，而 Dubois 和 Prade 定义的模糊粗糙集对噪声数据很敏感，这无疑会影响决策的最终结果，为此，需要将可变精度思想引入模糊粗糙集理论中来。可变精度模糊粗糙集理论是一个描述现实数据的更理想的模型，该理论尚有很多问题尚待解决，因此我们重点研究该模型，包括该模型的代数性质和约减算法。

1.2 研究现状与分析

1.2.1 关联规则研究现状及分析

自从1993年Agrawal等提出关联规则问题[4],经过二十多年的相关研究,现已取得丰硕成果。总体来看,有以下研究方向:以提高算法效率为目标的研究以及由算法效率衍生的一些内容、关联规则的扩展性研究、约束问题、关联规则的兴趣度问题、关联规则的应用等。下文对其研究脉络进行梳理和分析,对其发展趋势进行预测。

1. 关联规则算法的效率问题

数据挖掘面对的是海量数据,因此效率问题是数据挖掘首要考虑的问题。当前提出了很多关联规则的算法,其中最基础的算法是Apriori算法[4]。Apriori算法分为两步:①发现频繁集(frequent intemset);②根据频繁集生成关联规则。第一步是整个算法的关键。Apriori算法的基础是频繁集的下闭合(downward closure)性质,即任何一个频繁集的子集都是频繁的,而不频繁集的超集都是不频繁的。Apriori算法采用宽度优先,需要反复生成候选项集,再扫描数据以得到候选项集的支持度,其代价主要是生成候选项集需要的CPU时间和扫描数据库需要的I/O时间。

后续的很多关联规则挖掘算法都是在Apriori算法的基础上改进的。文献[5,6]引进散列的技术,在生成频繁1项集的同时根据事务生成2项集,并放入hash表中,这样可以大大压缩要考察的2项集,在引进散列技术的同时,压缩数据库。文献[7]针对文献[6]中存在的问题,提出使用更好的编码技术,以减少散列表中的冲突。划分技术将数据划分成多块,在每块中寻找频繁集,

这些局部频繁集形成全局候选项集，再扫描一次数据库可以得到全局频繁集[8]。文献[9]使用取样技术。取样技术是一种以精度换取效率的算法，数据挖掘可以在样本空间上进行，搜索的空间因此大大减小，并且样本可以放在内存中，提高扫描速度。文献[10]使用取样技术进行关联规则的更新。文献[11，12]研究了采用取样技术得到的近似规则的精度以及取样的大小问题。Brin 等提出了对候选集进行动态计数算法（dynamic itemset counting，DIC）[13]，即在同一遍数据库扫描过程中分段增加候选项集。一般来讲，长度为 k 的项集的支持数要在第 k 遍数据库扫描时才进行统计。动态计数算法在确定一个项集的所有子集都是频繁后，就开始其支持率的统计，而不是等到下一轮数据库扫描。该算法需要更少的数据库扫描，因此效率更高。

虽然取样算法、分区算法、动态计数算法都希望减少对数据集合的搜索次数，压缩搜索空间，但都存在一些问题。取样算法从原数据集合中随机抽样，利用样本来减少算法的搜索空间，但是由于数据集合中数据分布经常不均匀，有时数据也会很倾斜，所以随机抽样无法保证能够抽取到有代表性的样本，另外可能会丢失一些频繁集；分区算法虽然通过对数据集合分块挖掘，最后进行汇总的方法来减轻 I/O 的负担，但它增加了 CPU 的负担；动态计数算法采用动态计数的策略来减少搜索次数，但该算法基本思想与 Apriori 算法相同，也是一个宽度优先层层搜索的算法。这些算法生成候选项集时，产生许多不必要的候选项集，计算量大，对海量数据集合来说，以上算法只有在一定限制下才有效，否则频繁集的大量组合可能超过机器的存储和计算能力。因为这些算法都必须生成项集，并扫描其支持度，所以影响算法效率的是生成项集及其支持度的计算问题。每一次的计算不仅耗用大量 CPU 时间，而且需要大量 I/O 的请求，因而只有从本质上减少生成不必要的项集，减少要得到项集的支持度所需的计算时间，才能缩短数据挖掘时间，这是当前很多算法的出发点。

Han 等提出的频繁模式增长（FP-growth）算法[14]是针对 Apriori 算法中需要产生大量候选项集的问题，提出将数据库压缩成一棵频繁模式树（FP-tree），但仍保留项的关联信息，避免生成大量的候选项集，然后，在频

繁模式树上挖掘频繁集。FP-growth 算法比 Apriori 算法快大约 1 个数量级。对于大量的数据，建立频繁模式树的时间和空间代价也是可观的。

文献[15]提出了树投影（Tree Projection）算法，它将频繁集挖掘转化为逐步构造一种模式字典树（the lexicographic tree）的过程，同时将一个大型数据库投影为一系列子库。Tree Projection 算法对数据库进行投影，策略有宽度优先、深度优先、混合投影三种。采用深度优先投影可以很好地控制需要同时检查的组合或需要同时保存的投影形成的数据子库，但是扫描数据库的次数较多；采用宽度优先扫描，可以减少数据库扫描次数，但同时保存的投影子库或检查的组合较多，仍有可能面临组合爆炸问题；混合投影策略试图在两者之间找到平衡。

另外一些和效率有关的算法包括并行分布式挖掘算法[8][16~19]等。因为并行挖掘是指在多 CPU 共享内存和外存的条件下的挖掘，因此 CPU 间的通信和对内存的协调是主要考虑的问题。分布式挖掘是在网络环境下的挖掘，分布式挖掘中，节点间的通信量是主要考虑的一个问题。

在研究算法效率问题的过程中，人们开始考虑另外两种频繁集，即最大频繁集和闭项集，之所以研究这两类频繁集，其原因是频繁集的总数十分庞大，发现所有的频繁集以及分析这些项集都很困难。而通常最大频繁集（maximum frequent itemset，MFI）的数量会比闭项集（frequent closed itemset，FCI）少几个数量级，而闭项集比所有频繁集少几个数量级，同时，只要发现了所有最大频繁集和闭项集，则就可以找到所有的频繁集，因此发现这两种频繁集代价较小，于是提出了多种最大频繁集和闭项集挖掘算法。

典型的闭项集挖掘算法包括 A-Close 算法[20]、Closet 算法[21]和 Charm 算法[22]等。A-Close 算法类似于 Apriori 算法，算法策略有宽度优先、裁减候选集、扫描数据库确定支持度。该算法的核心是生成子的概念，即一个项集 p 是闭项集 c 的生成子，如果 p 的闭包等于 c 而 p 的任意子集的闭包都不等于 c。确定了生成子，就可以确定闭项集。Closet 算法使用 FP-tree 结构，采用分而治之的策略和投影技术挖掘闭项集。该算法需要构造 FP-tree，然后对大

型数据库分区，求解每个分区的闭项集，最后需要扫描整个数据库求解全局闭项集。Charm 算法通过构造 IT-tree 对项集和事务双向搜索以发现闭项集，另外该算法还采用了动态排序和 diffset 技术。

典型的挖掘最大频繁集算法有 MaxMiner 算法[23]、Pincer-Search 算法[24]、Mafia 算法[25]、GenMax 算法[26]和基于 FP-tree 的算法[14]等。关于最大频繁集算法的详细总结请参照第 4 章。

2. 关联规则的扩展性研究

关联规则挖掘有许多扩充。序列模式挖掘[27,28]是在一系列的序列数据中发现序列模式，如“50%购买啤酒的人会在一周内购买尿布”。这些挖掘算法的基础也是 Apriori 算法，这些算法都是针对离散值的数据的挖掘，连续数据的挖掘需要先离散化。

挖掘负关联规则（或反向关联规则）[29~31]。负关联规则考虑的是类似“如果购买了啤酒就不会购买尿布，或者说不购买啤酒也就意味着不会购买尿布”这样的规则。显然，由于负项的引入，原来问题的规模增大，负关联规则挖掘主要考虑效率和复杂性问题，可以采用领域知识限制问题规模、定义兴趣度等方法来降低复杂性。

多层关联规则挖掘在文献[32，33]中研究，这种挖掘以概化关联规则的形式研究，如普通关联规则“啤酒⇒尿布”可以概化为“饮料⇒婴儿用品”。因为概化如“啤酒”为“饮料”，显然“饮料”的支持度要明显高于“啤酒”的支持度，因此如何处理这种不同层次的支持度是多层关联规则挖掘重点考虑的问题。

挖掘量化关联规则。Apriori 算法仅用于挖掘布尔型属性，而实际属性很多是数值型的，即大都为连续的数据，如年龄、工资等。量化关联规则挖掘首先需要解决的问题是把数值型数据转化为离散型数据，再通过一般关联规则算法挖掘。文献[34]提出了根据规则聚类挖掘量化关联规则的系统。基于 x-单调和直线区间挖掘量化规则的技术由文献[35，36]提出。文献[37]使用非

栅格的技术使用完全性度量挖掘量化关联规则。使用量化属性的静态离散化和数据立方体，挖掘多维关联规则的技术在文献[38]中研究。

3. 约束问题

通过引入适当的约束可以充分利用用户的背景知识，缩小挖掘的范围，提高算法的效率，向用户提交他们所关心的结果。

从广义上讲，任何由用户指定的用于指导挖掘的条件都可以看做约束，根据约束的性质不同，Han 和 Kamber 将约束分为以下五类[39]：①知识类型约束，用来指定待挖掘的模式的类型，如关联、聚类、分类等。这类约束不同于其他类型的约束，通常在挖掘开始之前便已经由用户指定。②数据约束，用于指定待挖掘的数据对象，因为用户通常对数据集中的一个子集的行为更感兴趣，这类约束也称项目约束[40, 41]。③维数及层次约束，用于指定待挖掘的数据库或数据仓库（data warehouse）中的数据的维数或概念层次。④规则有趣度约束，用于指定所发现的规则的有趣程度的统计度量，最常见的有趣度约束即最小支持度和最小信任度约束。⑤规则约束，该约束用于指定待挖掘的规则的具体形式，这种约束可以用元规则（规则模板）表示，如可以出现在规则前件或后件中谓词的最大或最小个数，或属性、属性值和聚集之间的联系[42, 43]。

文献[44~46]研究了带有时态约束下关联规则的挖掘问题，挖掘得到的规则带有时间性，这样的规则对决策有针对性的指导意义。

文献[47~51]研究了约束的性质，指出哪些约束可以“推进”（push）到挖掘过程，并且仍然保证解空间的完全性。对于频繁集挖掘，规则约束可以分为五类，即反单调的、单调的、简洁的、可转变的、不可转变的。

其他有关约束的研究包括：文献[52]提出一种新的算法 Dense-Miner，该算法挖掘过程中充分使用了最小支持度约束、最小置信度约束，另外还提出了一种最小改进度约束（minimum improvement），充分使用这些约束既提高了挖掘的性能，也使挖掘得到的规则更易于被用户接受和理解。文献[53]提

出了面向用户为中心的算法，用户可以使用约束表达其关注的焦点，在挖掘过程中用户可以动态改变约束，改变最小支持度。该算法主要使用的技术为分段支持映射（segment support map），并使用 delta 隶属生成函数（delta member generating function）生成满足新约束的项集。文献[54]在 Apriori 算法的基础上提出一种支持度阈值可变的挖掘算法，该算法可以在运行过程中动态确定一个项集对应的最好的支持度阈值。文献[55]提出了 DualMiner 算法，该算法的新颖之处在于：其可以在算法中同时使用反单调性约束和单调性约束。文献[56]研究了基于利润约束的关联规则发现。

4. 关联规则的兴趣度问题

关联规则的兴趣度问题一直是人们关注的问题，在文献[57~59]中对其进行了讨论，文献[59]提出了使用相关性来代替支持度从而度量关联规则的有趣性。

5. 关联规则挖掘的其他问题

当前算法大多关注于挖掘算法的效率，但从用户的角度看，再有效率的算法得到的规则如果难以理解或解释，这个算法也是无效的。当前算法对海量数据挖掘，通常返回的规则数量非常庞大，而且还可能包含错误，大量的规则用户难以使用，因而需要对规则进行后处理（post-process）。一些研究采用聚类技术，对发现的关联规则进行聚类分组以发现用户感兴趣的规则[60,61]。文献[62]从闭项集挖掘的角度研究了挖掘更精简的规则。

关联规则在应用方面得到了拓展，如在 Web 挖掘方面[63~65]、空间数据库方面[66, 67]、评价系统方面[68, 69]等。随着大数据技术的成熟和应用，人们也关注到数据挖掘对隐私的侵犯，于是在关联规则挖掘的同时进行隐私保护也成了研究的一个方向[70]。

近年来，人们更关注频繁模式的挖掘，2014 年 Aggarwal 和 Han 在 Springer

上的论文集总结了频繁模式挖掘的研究成果，论述了模式增长、长模式、基于约束的模式、感兴趣的模式、数据流中的模式挖掘等多个方面[71]。

1.2.2 可变精度模糊粗糙集研究现状及分析

粗糙集模型提出以后，模糊集和粗糙集的关系曾引起人们的争论，最终人们趋向于将两者合在一起，产生一种新的理论。多位学者曾经提出了几种模糊粗糙集的定义，但是从目前的文献来看，人们主要认可的是 Dubois 和 Prade 给出的定义[3]。文献[72]在 a-截集基础上构造了模糊粗糙集，但其本质上还是 Dubios 模型。进一步研究，得出了泛化模糊粗糙集理论，这是模糊粗糙集理论的推广。文献[73]定义了一种模糊粗糙集，其主要基于在模糊自反关系上的模糊近似。文献[74]探讨了从上下近似运算角度的模糊粗糙集的公理性质。很重要的研究是文献[75]，其主要讨论了泛化的模糊粗糙集，该泛化的模糊粗糙集主要基于三种模糊蕴涵运算，即 S 蕴涵、R 蕴涵和 QL 蕴涵。在该文献中，讨论了这三种模糊粗糙集满足的一些来自于经典粗糙集的公理性质。从目前文献来看，该研究把模糊粗糙集的讨论推到了一个新的高度。

总体来看，对模糊粗糙集的讨论主要集中在其代数性质方面，经典粗糙集的核心是约减和提取决策规则，目前对模糊粗糙集的约减和规则提取的研究很不充分，可以看到的研究只有两篇文献[76，77]，其主要还是推广了经典粗糙集的约减思想，给出了模糊粗糙集的约减算法，但是关于如何提取决策规则，至今尚未见到研究报告。

模糊粗糙集模型无法处理数据中存在的噪声，为此需要引入可变精度思想。在目前的文献中，文献[78]提到了一种可变精度模糊粗糙集模型，其主要基于聚集运算，但该文献中并没有给出具体的形式定义。文献[79]在 Dubios 模糊粗糙集模型基础上定义了可变精度模糊粗糙集，其做法是使用截集去掉了数据中被认为是噪声的数据。

1.3 主要研究内容和创新点

本书的主要研究内容分为两大部分，即对关联规则部分的研究和对可变精度模糊粗糙集的研究。关联规则部分的研究分为模糊约束问题和最大频繁集问题。可变精度模糊粗糙集的研究内容主要包括可变精度模糊粗糙集的性质及约减算法。

1.3.1 关联规则挖掘中的模糊约束问题

引入适当的约束可以充分利用用户的背景知识，提高算法的效率，同时由于得到的规则数量减少，便于用户分析理解发现的规则，当然其最直接的目的是向用户提交他们最关心的结果。从广义上讲，任何由用户指定的限定条件都可以看做约束。Han 和 Kamber 详细讨论了约束的分类及性质[39]，提出将约束推进到发现关联规则的挖掘过程中，裁减了需要搜索的空间，从而提高了算法的效率。

但是目前的研究都是针对清晰的约束，实际情况下用户难以提出清晰精确的约束，而可能提出模糊的约束，如价格高的商品等，而如何将模糊的约束嵌入挖掘过程中，是一个尚待解决的问题。有些时候，数量型属性在离散化的时候为了软化划分的边界以及符合人们的思维习惯，也会引入模糊概念，这也会导致模糊约束的问题。

引入模糊约束以后会带来一些清晰约束中没有的问题。首先，很多模糊约束同时具有分段的单调性和反单调性，即在一个区间上具有单调性，在另一个区间上具有反单调性，其性质以及如何利用该性质尚需要研究，而模糊约束带来的最大挑战是如何将两种性质的约束同时推进到挖掘过程中。在关联规则挖掘算法中，以 Apriori 算法[4]为代表的大多数算法使用了支持度约束