信息科学技术学术著作丛书

关联数据聚类
——模型、算法及应用

Bo Long　Zhongfei Zhang　Philip S. Yu　著

龙　波　刘喜昂　译

科学出版社

信息科学技术学术著作丛书

关联数据聚类
——模型、算法及应用

Bo Long　Zhongfei Zhang　Philip S. Yu　著
龙　波　刘喜昂　译

科学出版社
北京

图字:01-2011-3734 号

内 容 简 介

本书展示了数据挖掘研究领域中一个新的热点方向——关联数据聚类,是作者研究工作的总结和提炼。主要内容有集体聚类、异质关联数据聚类、同质关联数据聚类、一般关联数据聚类、多视图关联数据聚类、演化数据聚类的算法模型,算法的具体实现和应用。

本书可以作为计算机、通信、信息等相关专业高年级本科生和研究生学习数据挖掘或机器学习专题的参考书,也可以作为相关专业的工程技术人员、研究人员、教师的参考用书。

图书在版编目(CIP)数据

关联数据聚类:模型、算法及应用/龙波(Long, B.)等著;龙波,刘喜昂译. —北京:科学出版社,2015
(信息科学技术学术著作丛书)
书名原文:Relational Data Clustering: Models, Algorithms, and Applications
ISBN 978-7-03-045093-7

Ⅰ. 关… Ⅱ. ①龙… ②刘… Ⅲ. 关系数据库-研究 Ⅳ. TP311.132.3

中国版本图书馆 CIP 数据核字(2015)第 132982 号

责任编辑:魏英杰 / 责任校对:郭瑞芝
责任印制:张 倩 / 封面设计:陈 敬

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

三河市骏杰印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2015 年 6 月第 一 版 开本:720×1000 1/16
2015 年 6 月第一次印刷 印张:10 1/2
字数:208 000

定价:90.00 元

(如有印装质量问题,我社负责调换)

《信息科学技术学术著作丛书》序

21世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的新动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术、超级计算机、软件和信息存储技术，数据知识化和基于知识处理的未来信息服务业，低成本信息化和用信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前沿交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性；体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国工程院院士
原中国科学院计算技术研究所所长
李国杰

作者中文版序

《关联数据聚类——模型、算法及应用》一书是我们在关联数据聚类，这一当今数据挖掘领域热点方向上多年研究的积累，也是国际学术界在这一方向上的首部专著。作为原版书的作者，欣喜地得知我们的英文图书将被科学出版社正式翻译出版，并作为科学出版社跟踪最新国际学术成果的学术著作丛书之一。在此我们要感谢本书的主要翻译者刘喜昂老师和负责该书翻译出版的科学出版社魏英杰老师。他们付出了辛勤的劳动和汗水终于使得本书的翻译工作得以圆满的结束。我们希望本书的翻译出版能起到抛砖引玉的作用，从而进一步推动中国学术界在数据挖掘领域研究的发展，尤其是在关联数据聚类方向上的发展。

在21世纪的今天，我们正处在互联网和大数据时代。无论在我们的工作，还是生活中，都在被无处不在的数据所包围。这些数据每天不仅在量上以指数爆炸的方式增长，同时其形式也越来越趋于复杂和多变。更重要的是，所有这些数据都存在着彼此之间显性或隐性的关联性。有效挖掘复杂数据间的关联关系及由此展开的知识发现是目前成功解决许多实际问题的关键所在。这也正是本书所探讨的主要内容。因此，本书不仅为相关领域的研究奠定了理论基础，也是一本解决关联数据挖掘的实际问题的参考工具书。

本书是三位作者在美国工作时完成的。当科学出版社决定把原作翻译成中文，以及在整个翻译过程中，恰逢原作者之一张仲非教授在美国纽约州立大学留职停薪全职在中国浙江大学工作。因此，在本书的翻译版中，张教授同时使用了其美国和中国的两个工作单位。我们再次感谢所有关心和支持本书翻译出版工作的朋友。同时，我们也在此衷心祝愿本书的翻译出版能为中国在数据挖掘领域研究和应用的迅速发展作出贡献。

作　者

2015年6月

前　言

今天，我们生活的世界充满了有关联的数据——互联网、社会网络、电信、顾客的购物模式，以及在生物信息学研究的微阵列数据。这导致产生了一个活跃的研究领域，称为数据挖掘研究领域中的关联数据挖掘。在许多现实世界应用中，我们或者不能奢侈地得到任何训练的数据，或者得到训练数据会非常的昂贵。在相关学术界，关联数据聚类引起研究者极大的重视，并因此作为关联数据挖掘领域新热点问题出现。本书是关于关联数据聚类的第一本专著，其中的理论自成体系，涉及关联数据聚类的基本原理和应用，包括理论模型、算法，以及应用这些模型和算法解决实际问题的典型应用。

多年来，本书作者一直致力于关联数据聚类的研究，本书是这些研究成果的总结。对这个研究课题有兴趣的研究者，可以将本书作为相关研究札记的汇总，对实际工作者或工程师可以作为一本参考书，也可以作为研究生关于关联数据聚类主题研讨会的教材。本书还可以用作研究生或高年级本科生的入门课程。本书收集的参考文献可以作为读者进一步阅读的参考资料。

虽然本书涉及关联数据挖掘中的许多问题，但并不刻意收集所有的相关资料。对那些从事关联数据挖掘研究的读者，或希望了解关联数据挖掘领域的读者，本书的目的是对这一领域的最新进展作一个系统的阐述。对新读者而言，本书的目的是对该领域作一个系统的介绍。

没有一个大团队和组织的支持，我们完成这本书是不可能的。特别要感谢 Taylor & Francis/CRC 出版社给我们机会完成这本书；作为 Chapman & Hall/CRC 数据挖掘和知识发现系列之一，明尼苏达大学的 Kumar 教授作为本系列的主编给我们的指导。我们要感谢本书编辑 Cohen 的热情和耐心的支持；项目编辑 Donley 和匿名校对员；国际排版 Kumar。我们要感谢在伊利诺伊州大学厄巴纳香槟校区的 Han 教授和亚利桑那州立大学的 Ye 教授，以及其他匿名审稿人的艰苦努力审查本书和他们有价值的评论。这本书来源是本书作者的贡献，部分资料也是他们的同事谷歌研究室的 Wu 和在宾汉姆顿的 Xu 贡献的。本书项目得到美国国家科学基金会的资助(授权号 IIS-0812114)，项目经理 Zemankova 博士的支持。本材料中表达的任何意见、研究成果和结论，或者建议都是作者的，并不代表美国国家科学基金会的观点。

最后，感谢我们的家庭，因为对本书的完成离不开他们的爱和支持。

目　录

第三部分 算 法

第四部分　应　　用

第五部分 总 结

第一部分
引　言

第1章 引 言

1.1 研究领域

聚类问题是数据挖掘和机器学习研究领域的基本问题。聚类分析是这样一个过程，它把数据对象集合划分成聚类，这样来自同一聚类的对象相似，来自不同聚类的对象不相似[105]。

在文献中，很多聚类方法集中在“单一”的数据，这些数据对象用一个固定长度的属性矢量表示[105]。然而，很多实际数据集合在结构上更加丰富，涉及多种类型对象，它们之间互相联系。例如，在一个文献库中，文献与关键词；在一个网页搜索系统中，网页、查询和网页用户；在营销体系中，商店、顾客、供应商、股东和广告媒体。我们称这些数据为关联数据，即数据对象之间互相有关联。

因为关联数据在不同重要应用中的惊人作用，关联数据得到越来越多的关注，如文本分析推荐系统、Web 挖掘、网络广告、生物信息学、引用分析和流行病学。在不同的数据挖掘领域，不同的关联学习问题有不同的称呼。关联学习的一个重要任务是在关联数据中发现隐含团体（聚类），即关联数据聚类。下面是关联数据聚类的例子。

① 文本分析。从双类型关联数据，文献-关键词数据中，学习文献聚类和关键词聚类。

② 推荐系统。基于用户聚类（团体）和电影聚类的电影推荐系统，从涉及用户、电影、演员的关联数据中学习。

③ 网上广告。基于关联数据库，其中广告人、投标人条件和关键字是内在互相联系的，广告聚类和投标人条件聚类可以从投标条款建议中学习。

④ 生物信息学。从基因、环境、标注的关键词的关联数据中，自动识别基因团（聚类）。

⑤ 研究团体挖掘和主题识别。从作者、论文、关键词组成的关联数据中，识别研究团体（作者聚类）和研究主题（论文聚类）。

一般而言，关联数据库包含三种类型的数据信息，即个体对象的属性、同类型对象的同质关联、不同类型对象的异质关联。例如，在一个论文和作者的科学出版物关联数据集合中，个人信息，如作者的联系方式是属性，论文间的引用关联是同质关联，论文与作者间的著作权关联是异质关联。在机器学习和统计学中，这些数据违反了经典的独立同分布假设，对传统的聚类方法提出巨大的挑战。

对关联数据聚类的挑战问题有两种结构，一种是单独聚类结构，另一种是集体聚类结构。

在单独聚类结构中，我们把关联数据转换成平面数据，然后单独聚类每个类型的对象。这个结构的一个直观方法是把所有的关联都转换成特征，然后直接应用传统聚类算法。另一方面，在集体聚类结构中，我们同时聚类不同类型的数据对象。与集体聚类结构比较，单独聚类结构有如下欠缺的地方。

首先，转换导致关联和结构信息丢失[48]。第二，在聚类关联数据时，传统聚类算法不能解决传播影响，即对象不同类型的隐含模式可能互相直接影响和间接影响(通过关联链)。第三，在一些数据挖掘应用中，用户不仅关心对象不同类型的隐含模式，而且关心涉及对象多种类型的互动模式。例如，在文献聚类中，除文献聚类和关键词聚类之外，文献聚类与关键词聚类之间的关联也是有用的信息。通过单独聚类对象的每个类型，发现这些互动模式是困难的。

另一方面，基于所有三种类型信息，学习局部和全局聚类结构。集体聚类结构有明显的好处，在集体聚类结构下，与关联数据的不同类型有不同的重点。这里有不同关联数据聚类的子领域：基于双类型异质关联数据库的集体聚类、多类型异质关联数据的异质关联数据聚类、同质关联数据的同质关联数据聚类、一般关联数据的一般关联数据聚类。

另一个有兴趣的注意点是重要聚类的数目问题，它在相关文献中有重要的意义，可以看作关联聚类的特殊情况。例如，图聚类分割[28,75,113]可以看作单个类型关联数据库的聚类，它由同质类型关联(代表如图的亲密矩阵)的集体聚类组成[11,44]。它出现在重要的应用中，如文件聚类和微阵列数据聚类。在联合聚类结构下，这可以规划为只有异质关联组成的双类型关联数据聚类。半监督聚类[14,124]是一个特别的聚类类型，既使用已标记的数据，又使用未标记的数据。在 11.3 节中，我们将说明半监督聚类可以规划为聚类在单个类型关联数据上，数据由属性和同质类型关联组成。

虽然本书内容主要集中在集体聚类结构，但是也包括我们关于单独聚类结构的研究，特别是多视图关联数据聚类。因为在一些应用中，当在一个关联数据集中，大量的对象类型以复杂的方式互相联系时，我们希望关注数据对象的特定类型来减少模型的复杂度。

图 1.1 说明了关联数据聚类不同领域之间的关联。总之，作为蓬勃发展的研究领域，关联数据聚类在广泛的应用范围兴起，在文献中也涉及一些重要的聚类问题，所以非常需要有一个关联数据聚类的实用算法推导和理论结构构建，这也是本书的主要目的。

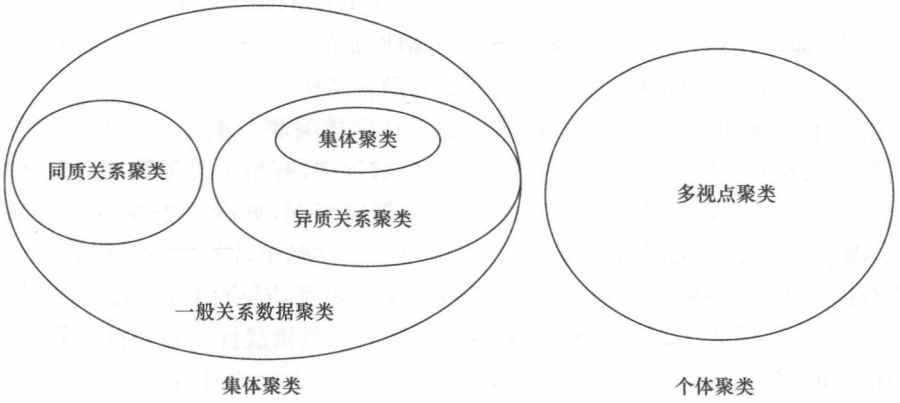

图 1.1 关联数据聚类不同领域之间的关联

1.2 本书的内容和组织

本书的目标是为一个新的数据挖掘领域，关联数据聚类和一族不同关联聚类问题的新算法介绍一个新颖的理论结构。

本书的组织如下。全书包含引言、模型、算法和应用。引言部分定义关联数据聚类的领域并概述全书的内容；模型部分介绍不同类型关联数据聚类的模型公式；算法部分说明对应模型的不同算法；应用部分用广泛的实验结果展现模型和算法的应用。本书集中在 6 个关联数据聚类专题。

第一个专题是双类型异质关联数据库，即数据对象的双类型之间是异质关联。例如，一个文献集可以形式化为文献和关键词的双类型关联数据集，其中文献和关键词是异质关联。在文献中，双类型关联数据聚类也被称为集体聚类。对双类型异质关联数据，我们提出一个新的集体聚类结构，即块值分解（block value decomposition，BVD）。对双类型关联数据，块值分解把关联数据矩阵分解为三个部分（行系数矩阵、块值矩阵、列系数矩阵）。在这个结构下，我们关注一个特殊已经非常流行的情况——非负关联数据，并提出具体新颖的集体聚类算法。算法基于更新规则，迭代计算三个分解矩阵。

第二个专题是关于双类型异质关联数据的更一般情况。多类型异质关联数据库可以形式化为不同结构的 k 部图。事实上，许多真实世界数据的例子涉及数据对象的多种类型，它们彼此互相联系，自然地形成数据对象异质类型的 k 部图。例如，分类挖掘中的文献、关键词、类别；网页搜索系统中的网页、查询要求、网页用户；科学出版档案室中的论文、关键词、作者和出版地点。我们提出一个通用模型，即网络关联摘要，从 k 部异质关联图发现隐含结构（局部聚类结构和全局团体结

构)。模型提出非监督学习不同结构 k 部异质关联图的主要结构。在一个广泛失真标准下,通过构建一个网络关联摘要来近似原始的 k 部异质关联图,我们可以得到一个新颖的算法来确认 k 部异质关联图的隐含结构。

第三个专题部分是同质关联数据聚类。在异质关联数据中,我们有数据对象不同类型的异质关联。另一方面,在同质关联数据中,有单个类型数据对象之间的同质关联。同质关联数据同样在重要的应用领域兴起,如网页挖掘、社会关联分析、生物信息学、超大规模电路设计、任务调度。在文献中,图分割被看做一个同质关联数据聚类的特例。基本上图分割寻找稠密的聚类,对应内部连接非常强的子图。另一方面,同质关联数据聚类的目标更普遍,更具挑战性。它要确认稠密的聚类和稀疏的聚类。在这一章中,我们提出一个基于图近似的通用模型,从图中学习基于聚类结构的关联模式。模型推广了传统图分割方法,并适用于学习各种聚类结构。在这个模型下,我们得到一族算法,可以灵活地学习各种聚类结构,并易于吸收聚类结构较早的知识。

第四个专题是在关联数据最常见情况下的聚类,包含三种类型的信息,即个体对象的属性、同类型对象之间的同质关联、不同类型对象之间的异质关联。对聚类多类型相关的对象,如何同时利用好三种类型的信息是一个巨大的挑战,因为这三种类型信息有不同的形式和非常不同的统计特性。我们对关联聚类提出一个概率模型,也提供了一个主要结构以统一各种重要的聚类问题,包括传统以属性为基础的属性聚类、半监督聚类、集体聚类和图聚类。提出的模型旨在为数据对象的每个类型确定聚类结构和不同类型对象之间的互动模式。在这个模型下,我们提出在一个数目大的指数族分布下参数硬或软的关联聚类算法。

第五个专题是关于单独聚类结构。对于这个专题,我们对多视图非监督学习提出一个通用模型。提出的模型采用映射函数的概念,使不同模式空间的不同模式可以比较,因此最佳的模式可从多重表征的多种模式中学到。在这个模式下,我们为两种重要的情况,即非监督学习、聚类和谱降维,规划出两种具体的模型,为多视图聚类推导出一个迭代算法和一个简单算法,它能提供一个全局最优的多谱降维。随着边界信息,我们还扩展了进化聚类和非监督学习已提出的模型和算法。

第六个专题是关于我们在进化聚类方面的研究,可以把时间效应也纳入到关联聚类中。在数据挖掘领域,进化聚类是一个相对较新的研究方向。进化聚类是指场景,在场景中数据集在时间上演化;每个时间,数据集有一定数目的聚类;当数据集从一个时间演化到另一个时间,新的数据项可能加入到数据集中,数据集中存在的数据项可能消失。类似的,新的聚类可能出现,同时存在的聚类可能消失,所以数据项和聚类集合都可能随时间变化。与传统聚类算法比较,这对进化聚类算法问题提出一个巨大的挑战。在本书中,我们提出基于 Dirichlet 过程的进化聚类模型和算法。

1.3 本书的读者

本书的预期读者包括该领域的研究人员和工程师，包括但不限于数据挖掘、机器学习、计算机视觉、多媒体数据挖掘、模式识别、统计学，也包括其他使用关联数据聚类技术的应用领域，如 Web 挖掘、信息检索、营销和生物信息学的人。由于本书资料的介绍是自包含的，同样可以作为对关联数据聚类这一新领域有兴趣人的理想参考书。另外，也包括任何对它有兴趣，或者工作领域需要这本参考书的人。最后，这本书还可以作为数据挖掘或机器学习课程的参考书。

1.4 进一步的阅读

作为数据挖掘和机器学习领域新出现的研究热点，关联数据聚类可以说还在起步阶段。目前还没有专门的、重要的地点来出版这个领域的研究工作，所以作为本书进一步阅读的补充信息，这个领域的相关工作还可以在两个上级领域的文献中发现。

在数据挖掘领域，相关工作可能在重要的会议中发现，例如 ACM 关于知识发现和数据挖掘国际会议、IEEE 关于数据挖掘国际会议和 SIAM 关于数据挖掘国际会议。特别地，相关工作可能发现在专注于关联学习领域的专题讨论会中，如统计关联学习专题讨论会。在期刊方面，数据挖掘领域的主要期刊可能包含关联数据挖掘相关工作，如 IEEE 关于知识和数据工程师的学报、ACM 关于数据挖掘的学报。

在机器学习领域，有关工作可以在重要的会议中发现，如机器学习国际会议、神经信息处理系统、欧洲机器学习会议、欧洲在数据库中知识发现原则和实践会议、国际人工智能联合会议和学习理论会议。在期刊方面，机器学习方面的主要期刊也包含关联数据聚类的相关工作，包括机器学习研究和机器学习杂志。