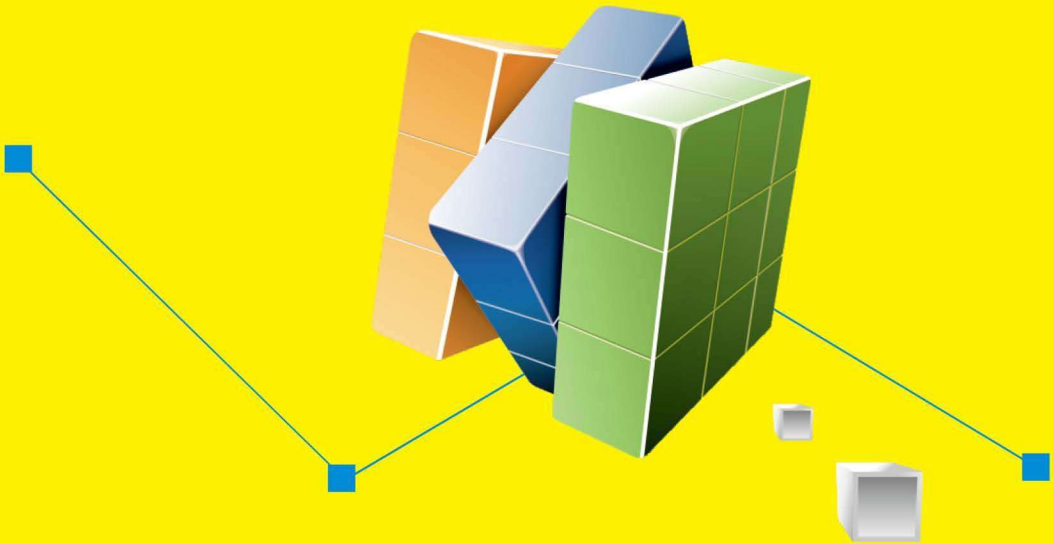
基于多目标决策的

数据挖掘方法评估与应用

MCDM BASED
DATA MINING METHODS
EVALUATION AND APPLICATIONS

邬文帅 / 著



四川大学出版社

责任编辑：唐　飞
责任校对：蒋　玮
封面设计：墨创文化
责任印制：王　炜

图书在版编目(CIP)数据

基于多目标决策的数据挖掘方法评估与应用 / 邬文帅著. 成都：四川大学出版社，2016.3
ISBN 978-7-5614-9359-5

Ⅰ.①基… Ⅱ.①邬… Ⅲ.①数据采集－研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2016）第 056561 号

书名　**基于多目标决策的数据挖掘方法评估与应用**
Jiyu Duomubiao Juece de Shuju Wajue Fangfa Pinggu yu Yingyong

著　　者　邬文帅
出　　版　四川大学出版社
地　　址　成都市一环路南一段 24 号 (610065)
发　　行　四川大学出版社
书　　号　ISBN 978 7 5614 9359 5
印　　刷　四川永先数码印刷有限公司
成品尺寸　148 mm×210 mm
印　　张　7.125
字　　数　192 千字
版　　次　2016 年 3 月第 1 版
印　　次　2016 年 3 月第 1 次印刷
定　　价　30.00 元

前　言

方法或模型评估问题普遍存在于数据挖掘、机器学习和人工智能等领域。在管理学领域，方法或模型评估问题同样不可忽视，是一个具有挑战性的研究热点和难点问题。没有免费的午餐定理指出不存在普适性的最优方法或模型，而决策者往往又十分关注最优方法或模型以实现最优决策，这一矛盾该如何调和呢？如何针对给定的目标问题或数据集，科学地选择合适、高效的评估方法或模型以找寻最优决策呢？另外，在数据挖掘领域，许多研究者大多将精力聚焦在设计新方法或开发新模型上，很少能够对挖掘出的结果进行深入的处理、分析和展示。用户难以理解挖掘出的结果，可操作性的概率更低，从而易造成知识及数据资源的无形浪费。

鉴于上述问题，本书基于群决策理论、多目标决策理论和数据挖掘理论，通过加强领域知识、专家经验与多目标决策方法相结合，针对数据挖掘分类和聚类方法的评估问题进行探讨和深入研究。主要研究内容如下：

(1) 提出基于多目标决策的数据挖掘方法评估理论框架。通过领域知识、专家经验将群决策和多目标决策理论引入到数据挖掘中，提出基于多目标决策的数据挖掘方法评估理论框架。该方法评估理论框架由三大评估阶段和六大模块的组件构成。

(2) 建立基于多目标决策的数据挖掘方法评估理论框架的

实证应用。基于方法评估理论框架，展开分类方法评估和聚类方法评估，并分别提出对应的层次分析模型（AHM）和共识支持模型（CDMECA），开展基于多目标决策的数据挖掘的二次挖掘和知识发现，以增强挖掘结果的易理解性与实用性。

（3）通过把领域知识、专家经验与多目标决策方法 AHP 相结合，提出 IAHP-GDM 和 EWAHP-GDM 方法。在书中，首次把集结个体判断矩阵（AIJ）和集结个体排序（AIP）统一在 AHP 群决策模型中，扩展和丰富了 AHP 群决策方法的集结技术。实证阶段，通过与传统的 AHP 群决策方法对比分析，验证了所提方法的有效性。同时，提出的 EWAHP-GDM 方法还被进一步扩展为一种确定准则权重的方法，为后续章节的分类方法评估和聚类方法评估奠定扎实的研究基础。

（4）针对不同决策方法或模型的评估绩效往往不一致甚至存在矛盾这一重难点问题，本书基于二八定律展开二次挖掘，提出一个共识融合模型来选择最佳聚类方法。在该模型中，决策参与者的整体满意度被充分考虑，并进行量化研究，且该模型能调和评估结果绩效不一致的差异。

邬文帅　著

华东交通大学

2016 年 1 月

目　录

第1章 绪 论

1.1 研究背景与意义

随着物联网、移动互联网、互联网金融技术的突飞猛进，社会产生的数据正以前所未有的增长速度激增[1-2]。商业、科研和政府机构相继建立起许多大型的数据库，积累了海量的异构数据。伴随计算机技术的迅猛发展，我们已经步入了大数据时代，怎么从数量巨大且复杂异构的数据中更好地提取出有用的信息，成为一个愈发重要且亟待解决的难点问题[3-5]。数据挖掘近年来作为信息处理的一门新兴的核心骨干技术[6]，其主要原理是从海量数据中挖掘、提取和识别出有价值的模式、知识和规律，并将其进一步高效地指导商业决策和进行科学研究[7-10]。目前数据挖掘已经在金融领域、医疗领域、通信领域、制造领域、司法领域、软件工程、生物工程等领域得到了广泛应用[11-15]。

方法或模型评估问题在许多学科领域都是一个活跃的且具有挑战性的研究热点问题，并且该问题将一直存在。没有免费的午餐定理（No Free Lunch）说明和指出：性能完全最优的方法或模型是永远不存在的[16]，也就是说，不存在具有普适性的最优方法。而决策者往往又十分关注和重视最优决策，如何针对给定的目标问题或数据集，来选定合适的评估方法或模型以找寻最优

决策，建立一套高效实用的方法评估机制，是一个极具挑战性的难题。近十几年来，许多研究者侧重于为各种数据挖掘任务（如关联规则挖掘、分类、聚类等）和数据类型（如文本、图形、多媒体等）建立新方法或新模型[17]。同时通过对在1944年到2005年期间发表的数据挖掘期刊、会议及学位论文进行文献调研分析，1600多篇论文中关于方法或模型的研究高达70%[18]。由于这些研究的核心在于设计和开发鲁棒的、高效率的新方法或新模型，所以学者们把其称为“方法驱动的数据挖掘”[19]。方法驱动的数据挖掘是数据挖掘的技术基础，推动了数据挖掘学科的进步。然而，由于许多研究者大多都将精力聚焦在设计新方法和开发新模型上，很少能够对挖掘出的结果进行深入的处置与分析，造成用户难以理解挖掘出的结果，能操作性的概率就更低，使得用户不能够轻松、有效地掌握和使用它们，造成知识及数据资源的无形浪费。

2007年，“知识驱动的数据挖掘”最早由Graco等在国际数据挖掘的会议上提出[19-20]。在相同的年份，“富含知识的数据挖掘”由Domingos在数据挖掘的权威期刊上提出[17-18]。知识在这里是指领域知识、专家经验等。知识驱动的数据挖掘和富含知识的数据挖掘的提出，表明知识越来越被受到重视。从数据挖掘项目决策者的角度来看，其关注的核心问题仍然是知识发现的问题，强调的是能够为企业创造利润、创造价值、提升竞争优势的可行动知识[11]。由于行业背景不同，对数据挖掘结果的展现、理解方式、运行时间、经济成本和质量指标要求等均有差异，如何缩小挖掘的结果与用户心理预期之间的差距，提高挖掘结果的准确性和实用性，是当前数据挖掘，同时也是基于多目标决策的数据挖掘研究的热点和难点问题。

Rokach[21]认为方法或模型的评估和选择需要考虑多个度量指标，如方法或模型的预测精度、方法或模型的稳定性、方法或

模型的泛化能力等，因此可以被看作多目标决策问题[21]。而多目标决策方法不仅能够基于多个相互矛盾乃至冲突的度量指标进行方案评估，而且还可以很好地反映决策者对评价指标的主观偏好，因此多目标决策方法在方法或模型评估领域具有很大的潜在优势。现有的基于领域知识和专家经验的研究成果同样适用于基于多目标决策的数据挖掘。而多目标决策在数据选取、方法构建、参数设置、结果表达这些步骤中所具有的特性，对知识驱动的数据挖掘又提出了新的要求。

本书通过把领域知识、专家经验和多目标决策与数据挖掘相结合，突出交叉学科的融合优势，整合优势资源，对基于多目标决策的数据挖掘的方法评估问题展开深入研究和探讨，建立基于多目标决策的数据挖掘的方法评估理论框架，并开发 EWAHP-GDM 方法来确定准则权重。基于建立的方法评估理论框架，针对分类方法评估和聚类方法评估问题建立实证应用，并开展二次挖掘与知识发现，提高数据挖掘的效率和结果的可理解性。

1.2　研究问题

方法或模型评估问题普遍存在于数据挖掘、机器学习和人工智能等领域，是一个具有挑战性的研究热点问题。在管理学领域同样不可忽视，如信用风险评分方法评估、决策方法评估等。没有免费午餐的定理明确指出不存在普适性的最优方法或模型，而决策者往往又十分关注最优决策，这一矛盾该如何调和呢？针对给定的目标问题或数据集，如何科学地来选择合适的、高效的评估方法或模型以找寻最优决策，这是本书重点研究和亟待解决的核心问题。

在管理学领域，如信用风险管理、投资策略管理、应急管理、软件缺陷管理等领域，均涉及方法或模型的评估和选择问

题，科学工作者及实业家也十分关注和重视如何针对给定的目标问题或数据集，来选定合适的、高效的评估方法或模型以找寻最优决策。比如在信用风险管理领域，随着全球经济多样化及金融市场的高速发展，涌现出了一些新兴的研究热点、难点问题，如互联网金融、金融大数据、量化投资等，从而体现了信用在社会经济中的地位也越来越重要。信用风险是商业银行等金融机构面临的主要风险之一，是当前构建“信用中国”“和谐中国”所面临的重要风险，也是当前社会管理中的关键问题。而信用风险评估是信用风险管理的首要工作和关键环节，事关银行等金融机构的生存和社会的稳定[22]。结合领域知识和专家经验，可知信用风险评估最为关键的步骤是信用评分方法的确定，而最为有效的信用评分方法则是数据挖掘领域的分类方法，如贝叶斯分类方法、回归分析方法、决策树分类方法、k 最近邻分类方法、神经网络分类方法、支持向量机分类方法等[23-27]。

然而，按照“经纪人”的模式，人们在对各种备选方案进行评价和选择时，通常采用“最优化原则”，即人们总是希望通过对各种备选方案进行比较，从中选择一个最好的方案作为可行的方案[28]。但是，最优化原则在复杂的现实生活中往往难以实现。决策理论学派提出用“满意原则”来代替“最优化原则”[29]。所谓满意原则，就是寻找能使决策者感到满意的决策方案的原则[30]。而多目标决策理论是依据决策背景，综合考虑多个相互间可能存在分歧甚至矛盾的评价指标，对多个备选方案进行选优和排序，利用统计学原理、运筹学方法、管理学理念以及最优化理论等作出决策或者获得妥协解的方法理论体系[31]。通过多目标决策可以寻找到最满意解而进行最优决策，正好契合这一研究范畴。

针对管理学领域的重难点问题，如信用风险管理、投资策略管理、应急管理和软件缺陷管理等领域，在方法或模型评估的问

题上，没有免费午餐的定理明确指出不存在普适性的最优方法或模型。而决策者往往又十分关注最优决策，希望找出最优方法或模型，这一矛盾该如何调和呢？本人结合领域知识和专家经验，在多目标决策、数据挖掘和群决策理论与技术的基础上，进行了一些探索性的研究工作，并取得了一些初步的研究成果[32-36]。因此，本书在已取得的研究成果的基础上，结合领域知识和专家经验，面向管理学领域的重难点问题，如信用风险管理、投资策略管理、应急管理、软件缺陷管理等，探究对于给定的目标问题或数据集，如何科学地选择合适且高效的评估方法或模型以找寻最优决策。

1.3　文献综述

1.3.1　AHP 群决策综述

层次分析法（AHP），是由美国学者 Saaty 提出的一种用于解决多目标、多方案的优选和排序问题的一种决策分析方法[37-38]。在实际的管理决策过程中，该方法被广泛地应用在各个领域，如生产制造、组织管理、商业分析、业务流程评估、物流优化、供应链管理、金融、保险、风险评估、科技成果评价等[36,39-40]。AHP 的基本原理是根据目标问题建立决策层级结构，通过专家咨询和专家评分获得两两对比判断矩阵，据此计算其最大特征值和其特征向量，通过一致性测试，诱导而获得各备选方案的最终排序。

随着科技的迅猛发展和社会的长足进步，决策问题变得越来越复杂，影响决策的因素也越来越多，并越发频繁。期待由某一个决策者作出客观、准确、科学的决策是十分困难的，理由归纳如下[41]：①决策往往具有时间压力或时效性，需要在某一段时

间内甚至当即就要作出决策；②许多决策属性、准则很难定量化研究；③单一的专家或决策者会受到自身专业知识、经验背景、个人偏好的影响，尤其在处理现实的复杂问题时。

由于专家的个人偏好、专业知识和经验背景不同，由专家评分而获得的两两对比矩阵也不尽相同，存在一定的主观性。在面对现实中复杂的决策问题时，可能会造成对决策理解的偏差，甚至产生矛盾的结论。考虑多个专家的综合意见、发挥群体智慧以消除个人决策者的主观偏好，是十分必要的。AHP 的本质是将目标问题分解为一个决策层级结构，通过专家咨询和专家打分，经过科学计算，最终合成一个优先权向量，并进行方案优选和排序。AHP 的决策过程是先分解再合成[42-43]。因此，学者们基于 AHP 原理与群决策理论，把专家意见融合到 AHP 的分解与合成的过程中，充分发挥交叉学科的融合优势，展开 AHP 群决策的研究。由于知识主要是指领域知识和专家经验，因此，本书把对 AHP 群决策的研究也定义为知识驱动的 AHP 研究。

群决策的本质是研究如何有效集结个体意见以达成群体共识。Ishizaka and Labib[44] 在其发表的对 AHP 综述的研究论文中指出 AHP 群决策集结个体偏好最常用、最有效的方法有集结个体判断矩阵（AIJ）和集结个体排序（AIP）两种，其具体内容如表 1-1 所示。集结个体判断矩阵是指集结个体的每一组判断矩阵形成群体判断矩阵，集结个体排序是指集结每个个体的排序以形成群排序。然而，这两种方法被学者认为是彼此独立的两个集结方法[44-46]，割裂了它们彼此的联系，并没有考虑相互依赖的关系。

表 1-1　AHP 群决策中集结个人决策意见的方式[44]

集结方法	集结方式（是否通过数学方法集结）	
	是	否
集结个体判断矩阵	几何均值法	投票集结

续表

集结个体排序	加权算数均值法	投票集结

1.3.2　知识驱动的数据挖掘研究综述

知识驱动的数据挖掘最早由 Graco 等在 2007 年的国际数据挖掘顶级会议上提出[19-20]。同年，富含知识的数据挖掘也在数据挖掘的权威期刊上由 Domingos 提出[17-18]。知识在这里是指领域知识、专家经验等。知识驱动的数据挖掘和富含知识的数据挖掘，尽管名称有所不同，但其核心都是将领域知识、专家经验融入数据挖掘的理论与技术中，用来提高挖掘结果的质量和效率。知识驱动的数据挖掘与方法驱动的数据挖掘有很大不同，方法驱动的数据挖掘的核心在于设计、开发鲁棒的、高效率的新方法和新模型。由于许多研究者大多将精力聚焦在设计新模型和开发新方法上，而忽视了对挖掘出的结果进行深入的处置与分析，用户难以理解挖掘出的结果，能操作性的概率就更低，使得用户不能够轻松有效地掌握和使用它们，造成知识及数据资源的无形浪费。

把领域知识、专家经验融入数据挖掘中的难点问题是如何更好地协调人机之间知识的交流和互动。对我们人类来讲，表达越丰富，沟通越畅通的语言如人类语言，越有助于促进知识的交流和互动；而相对计算机而言，便于计算机自动识别和处置的语言如机器语言，则越能促进知识的交流和互动。但当前的技术实力和技术水平还难以直接让计算机自动处理人类语言。针对这一难题，科学研究者作出了不懈的努力，也取得了一些研究成果：考虑到一阶逻辑，也叫一阶谓词演算，能够简洁地表达清楚大多数人类语言的含义，并且提供了比较完备和准确的推理功能[47]。于是，科学研究者通过将概率图形的表达与一阶逻辑相结合建立马儿可夫逻辑网，代表性的成果是 Richardson 和 Domingos 提出

的马可夫逻辑网[48]，能较好地实现计算机以自动处理的方式理解、表达和交流领域知识。Bogorny 等[49]提出的基于领域知识约束的最大频繁模式挖掘方法，通过与经典的 Apriori 方法对比分析，验证了该方法可以减少 80%的频繁模式，从而大大地增强了挖掘结果的效率。同时美国华盛顿大学还设计和开发了一系列基于马儿可夫逻辑网的模型、方法以及软件工具来实现知识驱动的数据挖掘[50]。

国内在知识驱动的数据挖掘方面也作了不少的探索性研究工作，并取得了一些成果。蓝荣钦和杨晓梅[51]根据空间数据挖掘的特性和内在需求，把领域专家知识分成三类，进而剖析和概述了领域专家知识和领域专家在空间数据挖掘中的突出作用。鲍洪庆、石冰和王石[52]指出成功的数据清洗往往都需要考虑领域知识，于是开发和提出了一个基于领域知识的数据清洗框架。李雄炎等[53]则通过结合油层水淹领域的相关领域知识，基于数据挖掘的技术和方法，从领域驱动的角度建立储集层水淹程度的预测模型。

本书的核心是基于领域知识、专家经验和多目标决策对数据挖掘方法评估问题展开深入研究，由于起步较晚，还未形成较体系化的研究成果。而知识驱动的数据挖掘的最新研究成果正好可以很好地融入我们的研究当中。

1.3.3 基于多目标决策的数据挖掘研究综述

1951 年，Kuhn 和 Tukcer 利用数学规划模型研究目标函数极大化问题，并给出了“有效解”存在的最优条件，该“有效解”被称为 Kuhn-Tukcer 有效解[54]，为多目标最优化理论和实证研究奠定了重要的基础。到 20 世纪 70 年代后，对多目标最优化问题的研究热潮才在国内外的学者中逐渐兴起[55-56]。目前多目标最优化在数据挖掘中的研究主要集中在分类问题上，例如，健康

保险欺诈分析中理赔申请的分类、信用卡用户行为的分类、电讯用户管理中用户的分类[57]。典型的分类方法有逻辑回归、贝叶斯网络分类、SVM、KNN、遗传方法、决策树等[58]。各种分类方法从不同的角度对训练数据集（已标注了类别的数据）进行分析，找出训练数据集中存在的普遍规律，经过验证后，将其用来对具有类似数据结构的未知数据的类别进行预测。

2011 年以来，有学者把多目标决策和数据挖掘技术结合起来进行一些探索性的研究工作。Peng 等[59]基于数据集成、多目标决策方法和数据挖掘技术提出一个能够有效应对突发事件的信息管理框架。该框架由三个主要模块组成：第一个模块是高级别数据集成模块，为了保障大量异构的源数据以统一的方式集成和输出；第二个模块是数据挖掘模块，使用数据挖掘方法来识别有用的模式，并为突发事件事前和事后的信息管理提供差异化的服务；第三个模块是多目标决策模块，其利用多目标决策方法来评估突发事件当前态势，找出满意的解决方案，并及时作出恰当的应对。Kou 等[34]通过集成多目标决策方法和数据挖掘技术来评估软件可靠性问题。文章首先应用数据挖掘分类方法对软件缺陷数据进行分类预测，再生成方法评估绩效的性能指标，然后通过多目标决策方法选出最佳的分类器。Kou 和 Wu[32]基于多目标决策和数据挖掘理论与方法，针对信用风险数据，提出一个层次分析模型对分类方法进行评估和优选。该优化模型可以快速准确地识别出最鲁棒的信用评分方法，进而能够进一步有效地指导决策者规避信用风险，并且该模型很好地解决了没有免费午餐的定理指出的经典问题。

由于多目标决策在数据挖掘中的研究起步较晚，目前还尚未形成较体系化的研究成果。而且对基于多目标决策的数据挖掘交叉集成研究，在建模、评估、决策和结果表现上，要求决策者既要具备多目标决策方面的理论知识和技术，又要了解数据挖掘方

面的理论和技术，从而导致了在建模过程中用户参与度低、结果的可理解性和可操作性低等问题。同时，没有免费午餐的定理指出无法找到一个普遍性的数据挖掘方法。每种数据挖掘方法有其适用的条件及各自的特点，为给定的目标问题选定恰当的方法是具有挑战性的工作，它直接关系着挖掘结果的质量和知识发现的效率。而结合领域知识、专家经验的数据挖掘方法为解决这些问题提供了可行的研究方向。

1.3.4 方法评估研究综述

方法或模型评估问题普遍存在于数据挖掘、机器学习、商业分析和人工智能等领域，是一个具有挑战性的研究热点问题，并将一直存在[32]。方法或模型评估通常需要综合考虑多方面的因素，例如方法或模型的预测效果能否达到要求，方法或模型的运行效率是否在可接受的范围，方法或模型的稳定性是否能够满足条件以及方法或模型的输出结果是否容易被决策者理解等。

当一组方法或模型的评估结果被获得后，应该在公平、公正的环境下评估方法或模型对数据的学习能力及预测能力，并且还应该评价方法或模型的泛化能力，进而验证和识别出最优的方法或模型。数据挖掘方法或模型的学习能力是指方法或模型学到隐含在目标数据中信息或规律的能力；泛化能力则是方法或模型在新鲜样本上的适应能力，也就是方法或模型对新输入的数据进行科学合理的响应能力。方法或模型的学习能力和泛化能力越高，则该方法或模型的理论价值和应用价值也越大。但要提高方法或模型的泛化能力毋庸置疑是非常困难的，因为未来的数据结构完全无法知晓，可能是和当前的数据一致，也可能存在很大的差异，甚至可能完全不同。因此，目标数据的结构和分布特征对方法或模型的评估和选择极为重要。为了消除数据本身的影响，文章是假定我们基于给定的数据集而展开方法评估的研究。