知识管理丛书

基于数学规划的数据挖掘分类算法研究及应用

魏利伟　主编

中国质检出版社
中国标准出版社

知识管理丛书

基于数学规划的数据挖掘分类算法研究及应用

魏利伟 主编

中国质检出版社
中国标准出版社
北京

图书在版编目(CIP)数据

基于数学规划的数据挖掘分类算法研究及应用/魏利伟主编. —北京:中国标准出版社,2012.12

ISBN 978-7-5066-7039-5

Ⅰ. ①基… Ⅱ. ①魏… Ⅲ. ①数学规划—应用—数据采集—分类—研究 Ⅳ. ①TP274②0221

中国版本图书馆 CIP 数据核字(2012)第 239875 号

中国质检出版社
中国标准出版社 出版发行

北京市朝阳区和平里西街甲2号(100013)
北京市西城区三里河北街 16 号(100045)

网址:www.spc.net.cn
总编室:(010)64275323 发行中心:(010)51780235
读者服务部:(010)68523946
中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销

*

开本 787×1092 1/16 印张 8 字数 179 千字
2012 年 12 月第一版 2012 年 12 月第一次印刷

*

定价 **30.00** 元

编委会名单

主　　任　李爱仙

副 主 任　汪　滨　卢丽丽

主　　编　魏利伟

编写人员　陈震宇　楚　琳　陈　颖　陈　兵
甘克勤　高　燕　洪　凌　李　景
李　菁　李　波　刘恬渊　刘莫尘
刘春卉　刘亚中　旻　苏　潘　薇
孙东霞　万夏青　闻　伟　谢启伟
周　洁　赵　萍　赵　奇　张梅荣

序

如果说知识经济带来了经济形态变革的话，那么知识管理(Knowledge Management，KM)是信息管理适应知识经济发展的必然结果。作为一种跨学科的新型管理模式与管理技术，知识管理正在引发一系列传统学科从研究范围到逻辑体系乃至核心思想的变迁，如管理学、经济学、企业管理、信息管理等。一向以知识的组织形式和知识的交流内容为核心研究对象的图书馆学、情报学，受知识管理的影响尤其深刻。知识管理已成为国内外情报学研究机构、教育机构的重要研究领域，备受广大专家学者们的关注。这种关注的坚实基础在于知识管理与情报学研究对象在本质上的一致性，即对知识内容的挖掘、整理、传递、利用。其中，数据挖掘技术是以上四个环节的重要技术手段之一，它可以从数据库积累的大量数据中挖掘出有价值的知识，增强组织的决策能力和智能。数据挖掘通过数据总结、数据分类、数据聚类和关联规则来发现组织中的显性知识和隐性知识。

数据挖掘技术已经成为知识管理的一种有力工具。不仅在资源密集型的图书文献机构是一种不可或缺的资源管理和对外服务的重要手段，而且在各种组织中也被充分利用。政府机构、企业、事业单位及社会团体，均有意或无意识地采用数据挖掘技术进行知识管理。通过数据挖掘技术，政府机构及事业单位提高了工作效率，提升了服务能力和形象；企业提升了运用集体智慧的能力和创新能力，创造了巨大的效益，扩大了品牌影响力。

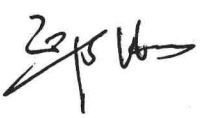
鉴于知识管理特别是数据挖掘技术的显著作用，相关的研究也蓬勃开展，这些研究涉及到多个方面，取得了很多有价值的成果，不同角度实现知识管理的数据挖掘技术类书籍陆续出版，知识管理的专业教育和培训体系已经逐步形成。但目前利用数据挖掘分类技术实现图书馆知识管理的研究方面仍处于一种自发、零散、无序状态，缺乏系统性的理论升华和实践指导，

深入探讨如何从科研实践的角度把数据挖掘分类技术应用到图书馆知识管理的工作实践中去，将具有理论的前瞻性和实践的导向性。

魏利伟博士及团队长期从事标准文献知识管理和数据挖掘方法的研究，不仅承担了多项科研项目，而且近年来在该领域的研究成果颇丰。《知识管理丛书——基于数学规划的数据挖掘分类算法研究及应用》一书是在多年研究和工作实践的基础上完成的，该书吸收了国内外数据挖掘技术的前沿内容，对知识管理中的数据挖掘分类方法的概念、理论、方法和技术等进行了详细的分析和介绍。全书共有八章，内容全面、观点前沿、案例丰富、实用性强，涉及知识挖掘基础知识、机器学习分类算法改进、多目标规划分类算法改进、自适应模型选择、信用风险评价、文本挖掘中文本分类任务的实现等问题。

该书有两个鲜明的特点：一是该书归纳总结了不同资源类别的特征，提出了一系列模型，而且针对理论模型给出了具体操作流程建议；二是理论和实际紧密结合，全书既有详尽深刻的理论阐述，又辅以生动新颖的实际应用案例进行剖析，使原本枯燥无味的理论知识生动形象、易读易懂，增加了读者学习的兴趣。

《知识管理丛书——基于数学规划的数据挖掘分类算法研究及应用》是一部对企业、科研人员、广大管理工作者、专业标准文献知识管理和图书专业情报工作者有价值的工具书和参考书。相信该书的问世，必将为知识管理学科的发展带来新的活力，给知识管理学界、知识管理工作者以及广大的标准化科研工作者，尤其是标准文献情报工作者带来一些帮助和启迪。



2012 年 10 月

前　言

随着信息时代、网络时代、知识经济，特别是全球经济一体化的深入发展，建设创新型国家日益受到高度重视，标准已成为世界各国发展贸易、保护民族企业、规范市场秩序、推动技术和实现高新技术产业化的重要手段，在经济和社会发展中发挥着越来越重要的作用。在此情况下，标准文献知识管理和知识服务研究工作的需求也日益迫切，标准文献知识管理也逐渐形成潮流。同时，近年来知识管理在图书馆、企业等各种组织中的实践也越来越广泛，信息化的推进让各种组织积累了大量的数据，建立充分利用这些数据的意识，从凌乱的数据中挖掘有用知识，这意味着组织开始向知识管理迈进。数据挖掘技术可以从组织数据中挖掘出有价值的知识，增强组织决策能力和智能。

数据挖掘通过数据总结、数据分类、数据聚类和关联规则来发现组织中的显性知识和隐性知识。分类问题是实际应用中普遍存在的问题，也是数学规划的重要应用领域之一，快速发展的信息技术对其在理论研究和实际应用中提出了许多新的难题和挑战。基于数学规划的分类方法借助其强有力的理论基础来实现分类任务，并表现出很多优越的性能。支持向量机是基于统计学习理论，借助最优化方法来解决机器学习问题的有力工具，目前，已成为研究的热点。但是这些分类方法依然存在诸多问题需要解决。准确率、速度、鲁棒性、可伸缩性、可解释性是评估分类方法的五条标准，其中准确率又是重中之重。本书基于多目标规划的视角在这五个方面深入研究、剖析，并改进最小二乘支持向量机、多目标规划分类模型。

本书以深入探讨分类问题为研究目标，立足于对基于多目标规划的分类模型和其在实际中的应用进行完善、推广和创新。本书的主要研究内容

如下：利用多核函数技术以及不同范数的特点，并以进化策略模型为优化工具，对现有多目标规划分类方法进行改进，提出五个改进模型。实际应用案例证明，这些模型实现了样本和特征的双重稀疏性，不仅提高了准确性，还大大提高了计算速度和解释能力，同时增强了鲁棒性。

在研究过程中，本书沿着 $L_2 \rightarrow L_1 \rightarrow L_p (p \in [1,2])$ 的路线逐步研究数据挖掘模式分类方法的改进，并根据问题导向层层递进，不断地发现问题，尝试着解决问题，尝试着从更加合理的角度提出更加贴切实际问题的最终模型：自适应惩罚函数选择的分类模型。该模型更加适合现实数据库的特点，能够获得很好的分类结果。希望读者通过对这些内容的学习，可以将基于数学规划的数据挖掘分类技术更好地应用到知识管理工作中。

本书获得质检公益性行业科研专项课题（编号：201210011）的资助，在此表示感谢。

鉴于基于数学规划的数据挖掘分类技术这一领域的技术研究时间比较短，因此作者在该领域的知识运用和实践经验有限，书中难免存在不足之处，敬请读者批评、指正。

编　者

2012 年 7 月于北京

符号说明

$\mathbf{R}$	实数集合
R^n	n 维欧氏空间
$\mathbf{R}^{m\times n}$	实数 $m\times n$ 维矩阵
$\boldsymbol{e}$	元素全为 1 的列向量
0	元素全为 0 的列向量
$\boldsymbol{I}$	单位矩阵
$\|\boldsymbol{Q}\|_p$	矩阵 $\boldsymbol{Q}$ 的 p 阶范数
L_p	p 阶范数
α	拉格朗日乘子
sgn a	a 的符号函数
ω	分类超平面的法向量
b	分类超平面的阈值
ξ	松弛变量
$k(,)$	核函数
$\boldsymbol{K}$	核矩阵
ϕ	非线性映射
γ	惩罚参数
H	分类超平面

目　录

1 导　论

近年来，知识管理飞速发展，学者对知识管理的研究越来越深入，知识管理在图书馆、企业等各种组织中的实践也越来越广泛。而且随着计算机技术、通信技术、网络技术的飞速发展，以电子格式存储的数据和信息出现了急剧的增长。组织可以广泛搜集到其所掌握的技术诀窍、业务资料和长期实践经验等数据资料。但如何对这些数据资料进行科学地分析、处理，从而发掘出对管理和决策有价值的信息和知识，却是这些组织面临的主要挑战。而组织要在激烈的市场竞争中获胜，必须对组织中的知识进行整理或收集，形成组织的核心竞争能力的知识资本，从而提高组织的市场竞争力。数据挖掘技术可以有效地解决这一问题，并且被广泛应用于各种组织的知识管理中。

数据挖掘技术可以从组织数据中挖掘出有价值的知识，增强组织决策能力和智能。信息化的推进让各种组织积累了大量的数据，建立充分利用这些数据的意识，从凌乱的数据中挖掘有用知识，这意味着组织开始向知识管理迈进。数据挖掘通过数据总结、数据分类、数据聚类和关联规则来发现组织中的显式知识和隐式知识。

数据挖掘和知识发现利用方法和技术从内容丰富、蕴藏大量信息的数据库中提取有用的模式和知识。分类模型是数据挖掘的重要方法之一，已经广泛地应用到商业和科学领域。在各种各样的分类模型中，基于数学规划的方法已经被证实在分类准确率、鲁棒性和有效性方面是非常好的[1-6]。支持向量机方法就是非常著名的基于数学规划的分类方法，起源于统计学习理论[15]。Vapnik 等人提出的统计学习理论(statistical learning theory，SLT)是一种针对小样本情况研究统计学习规律的理论，该理论的核心思想是通过引入结构风险最小化准则(SRM)来控制学习机器的容量，从而刻画了过度拟合与泛化能力之间的关系。在这一理论基础上产生的支持向量机(support vector machines，SVM)学习方法近年来受到广泛重视。在支持向量机中，两类之间间隔的最大化和误差损失之和最小化是建立分类模型的两个目标[16]。由支持向量机决定的两个超平面来分隔数据集的两类。分离超平面在输入空间和非线性的高维特征空间中均是线性的[7]。

但是，这些分类方法依然存在诸多问题需要解决。学者们最近比较关注基于多目标规划的模式分类研究领域的如下几个问题。首先，如何改善模型的可解释性，对于用户来说只知道分类结果是很难从挖掘系统得到的结果中发现隐藏在其中的有价值的知识；第二，如何提高模型的抗噪声能力；第三，如何使得模型根据数据的特点进行分类，从而更好地适应现实世界中各种各样的数据结构；第四，应该慎重选择模型自由参数优化工具从而帮助模型各方面的性能得到充分发挥；第五，如何拓宽这些方法的实践应用。满足组织的需求，真正能够实现知识管理，增强组织的核心竞争力。本书就是围绕以上五个问题进行研究的。

1.1 本书的研究背景与意义

随着计算机和数据获取技术的不断发展以及互联网和各种局域网的普及，人们获得的

数据正以前所未有的速度急剧增加。近几十年产生了很多超大型数据库，遍及超级市场销售、银行存款、天文学、粒子物理、化学、医学以及政府统计等领域。例如，美国著名零售商沃尔玛(Wal-Mart)每天要做 2 千万次交易；美国电报电话公司(AT&T)每天有 1 亿多用户在远程网络上呼叫 2 亿多次；美孚石油公司计划存储的有关石油开采数据将达 10^{14} 字节；美国国家宇航局(NASA)的地球观测系统每小时产生 5 万兆字节的数据；人类基因组计划也已收集了几千兆的相关数据。

如此超大规模的数据库将使人们不得不面对一个越来越突出的问题，即如何从海量数据中快速获取有效信息。数据库作为一种资源，本身并没什么直接利用价值，有价值的是从中抽取到的知识和信息。但是，与这种巨大的“海量”数据相比，人们分析处理它们的能力以及从中获取知识的能力都存在着相当大的差距，形成所谓“数据过剩”而又“信息匮乏”的被动局面。对于什么是信息，什么是知识，恐怕迄今还没有一个精确的定义来描述它们。但可以用这样一个例子来说明，例如，“我吃了一个苹果”应该只能算是信息，而“苹果是可以吃的”就是一条知识。信息和知识的关系，正如 Churchman 早在 1971 年就明确地指出，知识并不是简单地存在于信息集合中。因此，如何从这些大型数据库中，确切地说从大量的信息中挖掘出有用模式和知识，如何开发有效的挖掘方法，已成为众多科技工作者共同关注的焦点，变成了一个具有重要意义的研究领域[8-12]。

建立在数据库系统之上的计算机决策支持系统的出现，为进行高层次数据决策分析提供了一种思路和方法。但由于决策支持系统在数据的采集、分析方法上的灵活性等方面存在局限性，使得人们不得不寻求更有效的途径去开拓数据决策分析的思路。基于数学规划的数据挖掘方法的出现提供了一套新的思路，它能够模拟人类的学习方式，通过对数据对象之间关系的分析，提取出隐含在数据中的模式，即知识。

由于实际工作的需求及相关技术的发展，利用数据库技术来存储管理数据，利用机器学习方法来分析数据，从而挖掘出大量的隐藏在数据背后的知识，这两种思想的结合形成了现在深受人们关注的研究领域：数据库知识发现(knowledge discovery in databases，KDD)。其中，数据挖掘是 KDD 中一个最为关键的环节[13]。

数据挖掘(data mining，DM)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取出潜在的、可信的、新颖的、有价值的知识(模型或规则)的过程，是一类深层次的数据分析方法[10,13-14]。它综合利用了统计学、模式识别、人工智能、机器学习、数据库技术以及高性能并行计算等领域的知识技术，已在经济、商业、金融、天文等行业得到了成功的应用。

数据挖掘的任务包括分类(classification)、聚类(clustering)、描述和可视化(description and visualization)、相关性分组或关联规则(affinity grouping or association rules)等。其中，分类是数据挖掘的一项非常重要的任务，它通过对已有海量数据学习得到分类器(也称作分类函数、分类模型、分类规则、假设)。

分类技术在很多领域都有应用，例如可以通过客户分类构造一个分类模型来对银行贷款进行风险评估。当前的市场营销中很重要的一个特点是强调客户细分，客户类别分析的功能也在于此。采用数据挖掘中的分类技术，可以将客户分成不同的类别。比如，呼叫中心设计时可以分为：呼叫频繁的客户，偶然大量呼叫的客户，稳定呼叫的客户，其他。帮助呼叫中心寻找出这些不同种类客户之间的特征。这样的分类模型可以让用户了解不同行为类别

客户的分布特征。其他分类应用，如文献检索和搜索引擎中的自动文本分类技术；安全领域有基于分类技术的入侵检测等。机器学习、专家系统、统计学和神经网络等领域的研究人员已经提出了许多具体的分类预测方法。

分类的目的是提出一个分类函数或分类模型（即分类器），通过分类器将数据对象映射到一个给定的类别中。数据分类可以分为两步进行。第一步建立模型，用于描述给定的数据集合。通过分析由属性描述的数据集合来建立反映数据集合特性的模型。这一步也称作有监督的学习，导出模型是基于训练数据集的，训练数据集是已知类标记的数据对象。第二步使用模型对数据对象进行分类。首先应该评估模型的分类准确度，如果模型准确度可以接受，就可以用它来对未知类标记的对象进行分类。分类模型过程的示意图如图 1.1 所示。

根据分类的定义可知分类方法（分类器）是分类的核心，其性能决定了分类结果在多大程度上成为对人们有用的知识。

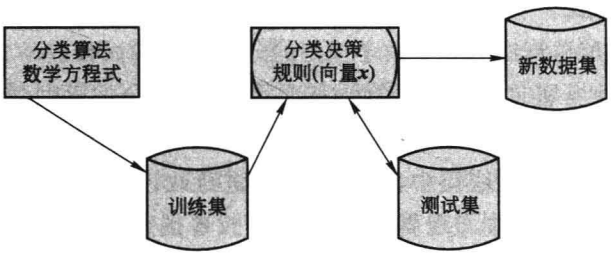


图 1.1 分类模型过程示意图

分类方法有很多种，它们各有所长，也各有所短。通过对它们进行研究和比较评估，一来可以扬长避短，在特定情况下使用最优的分类方法；二来可以对其加以改进，使其性能得到优化。对于分类方法公认的比较和评价的标准如下。

（1）准确率

指模型正确地预测新的或未见过的数据的类标号的能力，这也是模型的首要能力。如果一个模型的分类准确率小于 50%，那么可以认为其结果是无价值的。在其他条件等同的情况下，当然首选准确率高的分类方法。

（2）速度

指产生和使用模型的时间复杂度。产生模型的试验数据集通常是巨量的，因为一般情况下其数量和分类准确率成正比。如果产生和使用模型的时间过长，将严重影响用户的使用。

（3）鲁棒性

指给定噪声数据或具有空缺值的数据，模型正确预测的能力。现实中的数据库通常有噪声，有时还很大。如果一个分类器不善于消除噪声的影响，将严重影响分类准确率。

（4）可伸缩性

指给定大量数据，有效的构造模型的能力。有些分类器在数据量很小的情况下可以有效的构造模型，随着数据量的增大，其构造模型的能力显著下降，这最终也会影响分类准确率。

(5) 可解释性

指学习模型所提供的理解和洞察的层次。

综上可知,模型的预测精度是非常重要的,举一个例子,分类的准确率是分类器的最重要的指标,在其他情况等同的情况下,如果分类器 A 的分类准确率是 80%,分类器 B 的分类准确率是 90%,那么分类器 B 当然就比 A 有价值得多。假设分类器 A 的分类速度比 B 快,对 A 进行改进使其分类准确率达到 90%,那么分类器 A 就比 B 有价值了。因此,数据挖掘模型的微小进步,对决策者而言就可能意味着丰厚的利润。

1.2 本书主要研究内容

数据挖掘是数据库和信息决策领域的一个理论前沿,是知识发现的核心部分。数据挖掘技术可以快速有效地分析和处理来自组织内外部的大量的数据和信息,从而为组织的预测和决策提供科学依据。

本书旨在对数据挖掘算法中基于数学规划的分类模型所面临的一些问题进行研究。因此,重点分析基于数学规划的分类模型的研究现状,并按照分类模型的 5 个衡量标准,以多目标规划为主线来研究和改进一些基于多目标规划的分类模型,如最小二乘支持向量机和多目标规划分类方法。在实际应用中,把这些模型进行改进使其更加适应当今世界和数据库技术的发展,从大量的数据中发现更多的知识,不仅要提高分类精度,增强模型的鲁棒性,减少计算的复杂度,还要使模型本身更具有解释性。而且对于分类模型而言,要对大多数的数据库都能达到这些目标才能算是具有不错的性能。

本书主要有以下研究内容:

第 1 部分对研究意义和背景做了分析,论证了本书研究主题具有积极的现实和理论意义。介绍了本书的研究内容以及重要结论,并对研究方法和研究思路、技术路径进行了描述。对于本书创新点进行了阐述。

第 2 部分对于相关文献进行了综述,分别对于数据挖掘分类算法的重要性、数据挖掘分类算法的发展历史、以及基于数学规划的分类算法的研究现状进行了回顾,接着对基于数学规划的三个分类算法(SVM、LS-SVM 和多目标规划方法)研究的相关文献进行了整理和评述,总结出目前基于数学规划的分类模型所面临的问题,从而得到本书要对 LS-SVM 和多目标规划方法着手进行改进研究的研究要点。

第 3 部分从数据挖掘的数据准备入手,首先谈到数据挖掘数据准备的重要性,以及必要的处理手段,防止垃圾数据进垃圾数据出的情况出现;接着从模型挖掘结果评价,以及优化工具进行介绍,从而为本书所介绍的模型提供最基本的参数优化工具的原理和方法;最后介绍几款常用的数据挖掘开源软件供大家在实战演习时使用。

第 4 部分重点对机器学习分类方法进行改进,首先从改进最小二乘支持向量机模型的可解释性入手,从应用多核函数代替单个核函数为突破口,提出多核最小二乘支持向量机可解释性模型。通过引入多核函数,将特征选择问题转化为普通的参数学习问题,从而大大提高了模型的可解释能力,同时减少了计算的复杂度。医学数据库的试验结果很好地证明了这一点。接着重点改进机器学习分类模型的鲁棒性,兼顾模型的可解释性提出稀疏最小二乘支持向量机模型,这个模型中引入凸多核函数,将问题表示成为等价于过完备词典中的独