北京市教委特色专业建设资助项目
首都经济贸易大学统计学前沿文库

多目标线性规划分类方法业绩分析与改进研究

DUOMUBIAO XIANXING GUIHUA FENLEI FANGFA
YEJI FENXI YU GAIJIN YANJIU

朱梅红/著

首都经济贸易大学出版社
Capital University of Economics and Business Press

北京市教委特色专业建设资助项目
首都经济贸易大学统计学前沿文库

多目标线性规划分类方法业绩分析与改进研究

DUOMUBIAO XIANXING GUIHUA FENLEI FANGFA YEJI FENXI YU GAIJIN YANJIU

朱梅红/著

首都经济贸易大学出版社
Capital University of Economics and Business Press
·北 京·

图书在版编目(CIP)数据

多目标线性规划分类方法业绩分析与改进研究/朱梅红著. —北京:首都经济贸易大学出版社,2011.6

(首都经济贸易大学统计学前沿文库)

ISBN 978-7-5638-1912-6

Ⅰ.①多… Ⅱ.①朱… Ⅲ.①企业管理—研究 Ⅳ.①F270

中国版本图书馆 CIP 数据核字(2011)第 082447 号

多目标线性规划分类方法业绩分析与改进研究

朱梅红 著

出版发行	首都经济贸易大学出版社
地　　址	北京市朝阳区红庙(邮编 100026)
电　　话	(010)65976483　65065761　65071505(传真)
网　　址	http://www.sjmcb.com
E-mail	publish@cueb.edu.cn
经　　销	全国新华书店
照　　排	首都经济贸易大学出版社激光照排服务部
印　　刷	北京地泰德印刷有限责任公司
开　　本	787 毫米×960 毫米　1/16
字　　数	149 千字
印　　张	8.75
版　　次	2011 年 6 月第 1 版第 1 次印刷
书　　号	ISBN 978-7-5638-1912-6/F·1086
定　　价	18.00 元

出版总序

社会发展离不开数据，而数据必须使用统计方法来加以分析。自威廉·配第《政治算术》始，历史上几乎每一次对社会经济发展的深刻理解都是建立在统计分析方法变革的基础上的。正是这种变革所提供的各种数据分析工具加深了人们对社会经济本质的理解，使得人们的认识能够还原真实世界并与之无限接近。统计学数百年的发展历经两次方法上的“革命”：从最初不完整的全面调查方法到大样本统计推断，是统计方法的第一次革命；以大样本统计推断方法为基础，进一步发展出小样本统计推断方法，是统计方法的第二次革命。这两次革命都是施于用样本数据推断总体特征这一思想，而抽样误差的干扰导致统计方法日益复杂，使其应用受到限制。目前，以数据挖掘方法为代表的统计学的第三次革命即将到来。数据挖掘是在继承已有统计理论的基础上，与计算机技术紧密结合，充分发挥计算机运算速度快、存储量大的特点，将统计方法从抽样推断向海量数据分析推进，是统计学、计算机技术、仿真计算、机器学习、人工智能甚至哲学思想相融合的新学科，体现了科学发展“螺旋式上升”的哲学内涵。

统计学的发展过程不只是方法上的创新过程，更是一个统计学应用领域不断拓展的过程。从宏观经济计量、宏观经济统计分析，到涵盖消费、收入分配、投资、对外贸易等领域的宏观经济统计专题分析、博弈论等传统经济统计学以及国民经济核算等，再到如今在金融统计、财政统计、精算与风险管理、管理统计或商务统计（含企业微观统计、微观金融、微观核算、微观经济计量等）、市场调查、数据挖掘、质量控制与试验设计等社会经济生活方方面面的广泛应用，统计学的思想和方法无处不绽放出闪耀的光芒，为引领人类社会朝着客观、公平、公正的方向发展提供了必不可少的工具，同时也

为政策制定和决策提供了广泛而坚实的数据基础。

首都经济贸易大学统计学院是全国首家将统计学和数学融合在一起的统计学院，注重从方法的深入研究和应用领域的积极开拓两个方面进行发展，是实现统计学全面而迅速地与国际实质性接轨的先行者。近年来，该学院不断吸收国外统计前沿思想的精髓，引进高端人才，加强国际交流与合作，产生了一大批具有重要价值的前沿成果。本套丛书就是从中挑选出来的优秀成果之一，充分展示了首都经济贸易大学统计学院当前的研究方向和科研实力。

本套丛书的执笔人均为首都经济贸易大学统计学院处于教学科研一线岗位的具有博士学位的青年教师，他们思维活跃，在充分掌握当代数理统计理论与方法的基础上，对实际问题展开了深入而具体的研究，在翔实的数据基础上，对社会经济生活发展的多方面提出了谏言。本套丛书的研究范围具体涵盖了收入分配、宏观经济分析、金融计量、金融数学、数理统计、市场调查、数据挖掘等方向的前沿内容。从这些著作当中，可以清晰地看到他们深厚的统计理论功底与解决实际应用问题的能力，从而为推动统计学真正与国际接轨和良好发展打下了坚实的学术基础。

纪　宏

2009 年 11 月

前 言

20世纪80年代末开始，随着数据库、互联网技术的发展，人们获得的数据正以前所未有的速度急剧增加，产生了很多大型或超大型数据库。这些数据库不仅记录条数多，而且通常维数很高。这就迫切需要人们提出新的计算理论和工具，以便从这些“海量”数据中提取有用的信息或称知识，数据挖掘(Data Mining，DM)这个新的研究领域便应运而生。作为一门交叉学科，它会聚了数据库、人工智能、机器学习、模式识别、统计学、运筹学、可视化、高性能计算等不同学科和领域。在该领域，研究者们已经为实业界提供了许多有效的数据处理和分析工具。目前，它仍然是上述多个领域中最前沿的研究方向之一。

对数据进行分类，是数据挖掘领域的一项重要任务，也是一个重点研究问题。来自不同领域的众多研究者提出了不同的分类方法。其中，石勇教授提出的多目标线性规划(Multiple-Criteria Linear Programming，MCLP)分类方法，是最优化理论应用于数据挖掘领域的重要成果。

石勇教授一直从事最优化理论和方法的研究工作，在其前期研究的基础上，于20世纪90年代末提出了MCLP分类方法。它基于深厚的理论基础，是一种年轻而又具有广阔发展前景的分类方法。作为一种线性分类方法，MCLP的原理易于理解，模型求解也比较容易。与其他方法相比，它不需要对数据的假设，是一种非参数的方法。

自MCLP被提出以来，其理论和方法体系不断发展完善，目前已经形成了一个算法家族，并在银行、电子商务、门户网站、能源、国土安全等领域得到了广泛应用和高度评价。MCLP在国际数据挖掘研究和应用领域被广泛认同，被视为一种有效的分类方法。它已被嵌入到美国国际商业机器公司(International Business Machine，IBM)的智能挖掘软件(Intelligent Miner)和统

计分析系统(Statistics Analysis System,SAS)的企业挖掘软件(Enterprise Miner)的分类程序中。

笔者在中国科学院攻读博士学位期间,有幸师从石勇教授对MCLP进行相关研究。期间,笔者从最初对MCLP的一无所知,到逐渐地理解和掌握了它的思想和方法,并完成了博士学位论文,并且博士毕业两年来一直从事这方面的研究。本书是笔者在博士论文的基础上修改扩充而成的,是笔者这些年对MCLP研究的主要结果。

本书主要运用了统计学和运筹学的知识以及分类领域的通用技术,展开两个方面的研究。第一方面,对MCLP的几个重要特性,包括分类精度、稳定性、泛化能力、对数据集不平衡程度的敏感性,进行了大量的实证分析,总结出了这些特性所呈现出的统计规律,并进行了相应的理论分析。第二方面,根据MCLP的特性,提出了针对MCLP的业绩改进方法。对MCLP业绩的改进是从两个角度进行的。一是直接对模型进行改进。二是从数据入手,通过对数据的处理,使MCLP在新的数据上训练出的模型在原数据上有更好的业绩表现。具体地,根据对MCLP分类精度和稳定性的研究结果,实证分析了两种基本组合技术——Bagging和Boosting对MCLP的有效性和存在的问题,并进行了理论分析;在此基础上提出了改进的MCLP模型和几种适用于MCLP的改进的组合技术。根据MCLP在不平衡数据集上的表现,从数据角度,实证分析了常用的数据平衡技术——抽样或称再抽样技术(sampling or re-sampling)对MCLP的有效性及存在的问题,并提出了适用于MCLP的改进的数据平衡技术;从模型角度,提出了改进的MCLP模型,并进行了理论和实证分析。虽然本书的研究对象是MCLP,但本书的研究框架和思路也适用于其他分类方法的研究。如果本书的研究能为数据挖掘领域作一点贡献,此书的目的也就达到了。

由于MCLP提出较晚,因而国内数据挖掘界的同行对其了解并不多。笔者通过参加学术会议或发表学术论文来介绍MCLP分类方法,以便国内同行了解和使用这一年轻而又优秀的分类方法。

当然,笔者能对MCLP进行比较深入细致的研究,直接得益于恩师石勇教授的悉心指导,也受益于学习期间与师弟师妹们的相互切磋。

在此，对所有为本书的出版作出贡献和提供帮助的专家学者表示衷心的感谢！

首都经济贸易大学统计学院对本书的出版资助，使得我能有更多机会与同行们分享它；在本书出版过程中，首都经济贸易大学出版社提供了大力支持。在此，一并表示感谢！

由于本人能力有限，书中难免有疏漏和谬误之处，责任全部由我个人承担，恳请专家同行批评指正。

朱梅红

2011 年 5 月 25 日

目 录

1 绪 论

1.1 研究背景

随着数据库、互联网技术的发展，自20世纪80年代末开始，人们获得的数据正以前所未有的速度急剧增加，产生了很多大型或超大型数据库，如金融、航空、交通和天文学、地理学、化学、医学、生物学以及政府统计等领域都积累了“海量”数据。这些数据不仅记录条数多，而且通常维数也很高。这就迫切需要人们提出新的计算理论和工具，以便从数据中提取有用的信息或知识，知识发现（Knowledge Discovery in Databases，KDD）和数据挖掘（Data Mining，DM）这两个概念应运而生。其中，知识发现这个概念产生于1989年的第一次KDD学术会议，被人工智能和机器学习界广泛采用；而数据挖掘则主要流行于统计学界、计算分析人员中和管理信息系统等领域。菲亚德（Fayyad）等人对它们的定义是：知识发现是从数据库中发现有用知识的全部过程，包括数据准备、数据选择、数据清理、数据转换、数据挖掘、对挖掘结果的解释和评价、形成对用户有用的知识等环节，目的是把低水平的数据映射成更复杂、更抽象、更有用的其他形式的知识；而数据挖掘则是其中的特定步骤，它用专门算法从数据中提取模式（patterns）。本书正是基于这个观点展开后续的分析。知识发现自提出以来，因其在应用领域取得的巨大成功，一直受到各研究领域的广泛关注。作为一门交叉学科，它会聚了数据库、人工智能、机器学习、模式识别、统计学、运筹学、可视化、高性能计算等不同学科和领域的知识。尤其是人工智能、机器学习、模式识别、统计学、运筹学等领域，为数据挖掘阶段提供了很多有用的方法。目前，知识发现和数据挖掘仍然是上述多个领域中最前沿的学术研究方向之一。

建模方法可以分为基于理论的方法和基于数据的方法。传统的基

于理论的建模方法在海量数据面前大部分已经失效。数据挖掘是典型的数据驱动方法，它不需要足够的先验信息和理论，而是着重于从数据中得到结论。

目前，数据规模不断扩大，数据格式越来越多样化，数据结构越来越复杂，用户的要求越来越高也越来越个性化，这对所有的知识发现方法提出了挑战，迫使知识发现的理论和方法体系不断完善和发展。本书主要从统计学和运筹学的角度，研究数据挖掘中基于运筹学的分类方法的相关问题。

1.2 问题的提出

数据分类是数据挖掘的一项重要任务，也是数据挖掘中一个重要的研究和应用领域。简单地说，分类是指通过建立一个分类模式(pattern)，将数据分派到不同的类中。具体而言，假定有 n 个样本数据，称之为训练集(training set)。其中每个样本数据称为一条记录，每条记录由 $r+1$ 个字段值组成。我们把这些字段称作属性(attribute)，把 r 个用于预测的属性称为预测属性，把用于分类的属性称作标签(label)，即训练集的类别标记。第 i 个样本可以表示为 $(A_{i1}, A_{i2}, \cdots, A_{ij}, \cdots, A_{ir}; c_i)$，其中，$A_{ij}$ 表示第 i 个样本的第 j 个预测属性的值，c_i 表示第 i 个样本所属的类别，$i=1,2,\cdots,n; j=1,2,\cdots,r$。所谓分类，就是根据已有的 n 个样本数据，学习一个分类函数或分类规则，统称为分类模式，有时也称为分类模型或分类器(classifier)。该分类器能够把训练集内的每条记录映射到给定的某一类别中，从而能对现有的 n 个数据进行分类。另外，它还可以对不包含标签属性的其他记录进行分类预测。在分类问题中，预测属性可以是定性的，也可以是定量的，而标签属性必须是定性的或离散的定量属性，且标签属性取值的数目越少越好。标签值的数目越少，构造出来的分类器的错误率越低。

目前，已经有来自不同领域的多种分类方法，如统计学领域的 Fisher 线性判别分析、Bayes 分类、Logistic 回归、最近邻分类，机器学习领域的决策树和支持向量机，人工智能领域的神经网络，运筹学领域的多目标线性规划分类方法，等等。这里借用菲亚德等人的图示来直观地说明这几种方法的分

类思想。

假定有一个信用卡用户数据集，有两个预测变量：收入和受教育年限；一个目标变量即客户的类别，包括两类客户，即好客户(○)和坏客户(+)，两类客户的分布情况如图 1.1 所示。

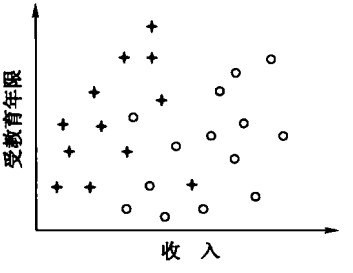

图 1.1 两类客户的分布情况

一些线性分类方法，如统计学中的 Fisher 线性判别分析、机器学习中的线性支持向量机、运筹学中的多目标线性规划分类方法等，都是寻找一个线性分类边界，将两类数据点分割开来，如图 1.2 所示。一些非线性分类方法，如神经网络，则是寻找一个非线性分类边界，将两类数据点进行分割，如图 1.3 所示。而决策树等，则是基于规则的分类方法，如图 1.4 所示，其分类规则可以表示为：如果收入$\geq t$，该客户就是“好客户”。

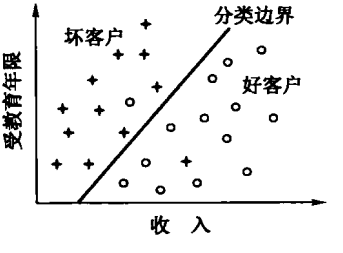

图 1.2 线性分类方法对两类客户的分类结果

由于现实问题的日渐复杂化，各种方法都仍在不断完善和发展中。本书以多目标线性规划分类方法为研究对象，提出相应的业绩改进技术。

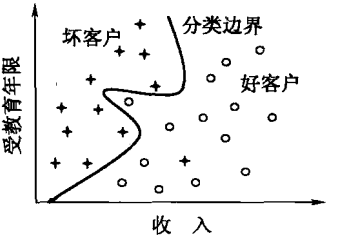

图 1.3 非线性分类方法对两类客户的分类结果

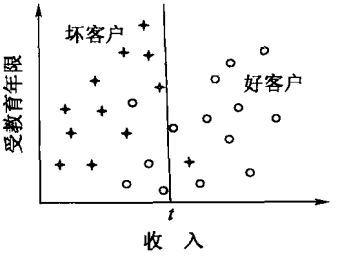

图 1.4 基于规则的分类方法对两类客户的分类结果

1.3 基本概念

为展开分析,这里先介绍用到的一些基本概念。

1.3.1 精度

分类精度(accuracy)是指运用特定的分类方法建立的分类模型在数据集上的分类正确率。精度高的分类方法称为强学习算法,精度低或仅比随机猜测好一点的分类方法称为弱学习算法。在分类问题中,人们一般要同时关注训练集上和测试集上的精度,而更关注的是测试集上的精度;同时考察总的精度和各个类别的精度;还需要研究单个分类器的精度和组合分类器的精度。

1.3.2 稳定性

这里用由训练数据集的差异引起的测试集上分类结果的差异来衡量分

类方法的稳定性(stability)。如果训练集数据较小的变化会带来测试集上分类结果的较大变化,则此分类方法不够稳定,否则就认为该分类方法比较稳定。这里所谓的训练集的差异,不仅指在同一训练样本量下不同训练样本之间的差异,也包括训练样本量之间的差异。分类方法的稳定性依赖于很多因素,比如,分类方法本身的复杂性、分类方法对训练集数据构成变化的敏感性、训练数据的分布情况等诸多因素。

1.3.3 泛化能力

泛化能力(generalization ability)是指运用特定分类方法、根据已有数据建立的分类模型在独立的检验数据集上的预测能力。如果根据已有数据建立的模型能很好地适应独立的检验数据集上的新数据,在新数据上表现出与已有的训练数据上相同或更好的分类业绩,称分类方法对数据的泛化能力较强,否则称分类方法的泛化能力不强。在训练过程中将数据分为训练集和测试集,则模型在测试集上的表现就是泛化能力的具体体现。泛化能力既反映了分类方法的分类精度,也反映了分类方法的稳定性,是两者的综合表现。分类方法的泛化能力可以用模型在测试集上的泛化误差来衡量。一般来说,随着模型越来越复杂,它能够适应更复杂的数据,但可能会出现对训练集的过度拟合(over fitting)现象,导致分类方法在测试集上的泛化能力减弱;不稳定性分类方法的泛化能力也较弱;如果训练集和测试集的结构相差较大,分类方法的泛化能力也会较弱。

1.3.4 类分布

一个数据集上,各类数据的数量占总数据量的比重,或各类数据之间数据量的比例,称为类分布(class distribution)。以二分类问题为例,其中一类数目占总数的30%,或表示为两类的比例为3:7,都反映类分布。原始数据集的类分布也称为自然分布。

1.3.5 不平衡数据

如果一个数据集内,各类数据量不相等,就称为不平衡(unbalanced or

imbalance)。以二分类问题为例,其中,数目多的一类称为多数类或大类、负类;数目少的一类称为少数类或小类、正类。根据两类之间数据量的差异程度,分为严重不平衡(或严重偏斜)、一般不平衡和基本平衡。

1.3.6 分类方法对数据不平衡程度的敏感性

如果一个分类方法在不平衡数据集上运用自然分布的训练集进行训练,分类结果明显偏向大类,或者改变类分布则分类结果明显变化,就称该方法对数据不平衡程度比较敏感(sensitive)。

1.4 本书的研究内容与方法

在众多分类方法中,多目标线性规划(Multiple-Criteria Linear Programming,MCLP)分类方法,是最优化技术应用于数据挖掘领域的最重要的成果之一。它基于深厚的理论基础,是一种年轻而又具有广阔发展前景的分类方法。自从2000年石勇提出该分类方法以来,其理论和方法体系不断完善和发展,目前已经形成了一个算法家族,并在金融、电子商务、门户网站、能源、国土安全等领域得到了广泛应用,获得了高度评价。

尽管MCLP有较完美的理论支持和良好的业绩表现,但在现实应用中,也存在一些尚待完善的地方。这主要表现在以下几个方面:第一,有待对MCLP的特性进行深入分析。第二,有待进一步提高算法的运行速度。第三,有待进一步提高分类的精度。对于此目标,有两种主要思路。一是改进原有分类模型。但分类模型的改进难以跟上数据的变化,而且任何一个分类模型都不可能适用于所有类型的数据,所以这种方法的作用是有限的。二是采用辅助手段。比如,从数据入手,对原有的数据进行加工,使分类方法在加工过的数据上训练出的模型在原数据上有更满意的结果。第四,有待进一步提高分类方法对不同数据的适应性。比如,对于不平衡数据,虽然实践中经常使用1:1的训练集,但还可以考虑从分类模型改进和数据方面设计适用于不平衡数据的分类模型和合适的数据处理技术。第五,有待向更深更广的方向拓展。比如,从二分类问题扩展到多分类问题,从线性模型扩展到非线性模型,从提出模型到寻找背后的理论支持。

本书的研究重点有两个方面:一是分析 MCLP 的特性;二是针对 MCLP 的特性,提出相应的业绩改进方法。所有的研究都从定性和定量两个方面进行。

对于 MCLP 特性的分析,本书主要集中在对它的分类精度、稳定性、泛化能力、对数据不平衡程度的敏感性的分析上。首先,对 MCLP 的分类精度、稳定性和泛化能力进行了理论和实证分析。基于多明戈斯(Domingos)的期望预测误差(Expected Prediction Error)分解框架,在三个数据集上,对 MCLP、线性判别分析(Linear Discrimination Analysis,LDA)和决策树 C5.0 的偏差和方差特性进行了比较分析,从而在同一框架下同时分析了三种分类方法的分类精度、稳定性和泛化能力,也便于在更广阔的背景下理解 MCLP 的特性。其次,研究了 MCLP 在不平衡数据集上的业绩表现。对于不平衡数据分类问题,给出了两个业绩评价标准;考察了同一样本量不同类分布下 MCLP 的业绩变化情况,以及不同样本量不同类分布下 MCLP 的业绩变化特点;指出了 MCLP 对数据不平衡程度的敏感性;给出了两个业绩评价标准下的最优类分布;对所选择的最优类分布的可靠性进行了分析。

对于 MCLP 的业绩改进,一是对分类模型进行了改进;二是从数据入手,通过对数据的处理,使 MCLP 在新的数据上训练出的模型在原数据上有更好的业绩表现。第一,根据对 MCLP 分类精度和稳定性的研究结果,提出了相应的组合(ensemble)分类器方法,以期对 MCLP 的分类业绩进行改进。明确了组合分类器业绩评价标准;针对 Adaboost 的特点,对标准的 MCLP 模型进行了修改,提出了对错分点加权的 MCLP 模型;运用两种基本的组合技术——Bagging 和 Adaboost 进行试验,分析了 Bagging 的作用,也发现了 MCLP Adaboost 组合方法在一些特殊数据集上的不良表现,并从不同角度进行了原因分析;针对 MCLP Adaboost 在噪声较严重的数据集上的失败,提出了相应的平滑噪声的 MCLP Smooth Boosting 方法,并在有关数据集上验证了其有效性;针对 MCLP Adaboost 在类间交叠严重且不平衡的数据集上的不良表现,提出了一种序贯的 MCLP Bagging 技术,并在相应的数据集上进行了实证分析;通过不同组合方法对 MCLP 业绩的影响分析,对 MCLP 的特性有了更进一步的认识,也指出了不同组合方法对 MCLP 业绩的影响机制及影响程