Economic Management

经管学术文库

随机森林组合预测理论及其在金融中的应用

方匡南/著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

经管学术文库

随机森林组合预测理论及其在金融中的应用

方匡南/著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

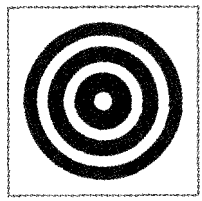
随机森林组合预测理论及其在金融中的应用/方匡南著.
—厦门:厦门大学出版社,2012.5
ISBN 978-7-5615-4210-1

Ⅰ.①随… Ⅱ.①方… Ⅲ.①非参数统计—应用—金融学
Ⅳ.①F830

中国版本图书馆 CIP 数据核字(2012)第 075469 号

厦门大学出版社出版发行
(地址:厦门市软件园二期望海路 39 号 邮编:361008)
http://www.xmupress.com
xmup @ xmupress.com
沙县方圆印刷有限公司印刷
2012 年 5 月第 1 版 2012 年 5 月第 1 次印刷
开本:889×1240 1/32 印张:7.5 插页:2
字数:200 千字
定价:20.00 元

作者的话



统计预测方法是利用历史数据建立有关统计模型，并通过统计模型对未来的经济活动作出估计、描述、分析和判断，揭示有关变量之间的规律性联系，用于预测和推测未来发展变化情况。统计预测在经济领域得到了广泛的应用，各国相继成立各种专门的预测研究机构或咨询机构，为各国政府以及企业提供宏观经济预测、行业经济预测和其他经济预测服务，并同时对统计预测的理论和方法进行了深入的研究，促进了统计预测方法的发展。统计预测方法发展迅速、日新月异，经历了结构计量模型阶段、时间序列模型阶段、非线性非参数计量模型阶段、数据挖掘智能预测与组合预测等四个阶段。

近年来，各学科间不断地融合，研究方法相互渗透已成为现代科学发展的一大趋势。金融理论、数理统计、计量经济学、计算机技术、数据挖掘、机器学习等学科的融合为经济金融的研究提供了新的研究方法和思路。我们注意到，起源于数据挖掘领域的非参数随机森林方法，以非参数决策树方法为基础，借助于

机器学习的组合预测思想，结合计算机技术，不仅可以很好地处理非线性、非高斯问题，而且具有较高的预测精度。此外，在非参数随机森林的基础上，不断发展出了分位数回归森林、随机生存回归森林等，并在医学、市场营销、物理、考古等领域都有众多应用。

本书的主要研究成果是由作者的博士学位论文以及后续相关研究组成。本书主要深入研究了最新的非参数随机森林以及由此衍生出来的相关理论和算法，并重点探讨这些方法在经济金融中的应用，尤其是在我国信用卡信用违约预测、基金股票市场的趋势预测、房屋抵押贷款违约预测、保险客户利润率预测，以及金融市场风险预测等的应用。

本书的内容安排如下：

第 1 章　从统计预测方法的发展历程出发，指出了目前统计预测方法的发展趋势，并将统计预测方法的发展划分成四个阶段，探讨了经济理论导向模型和数据导向模型的优劣，揭示了非参数随机森林方法在金融市场预测中的重要意义和重大应用前景，由此确立了本书研究的目的、内容、框架和思路。

第 2 章　详细介绍与讨论决策分类回归树(decision tree)的原理、分割变量的选择、算法、误差估计、过拟合问题，以及模型性能的评估方法。

第 3 章　详细介绍、讨论了随机森林的原理、收敛性、泛化误差、强度和相关系数、随机特征选取的两种方法，并利用大量数据集进行实证分析。

第 4 章　详细介绍、讨论了分位数回归和分位数回归森林

的原理和其参数回归模型、非参数回归模型、半参数回归模型的三类参数估计方法。

第 5 章　探讨了随机森林分类方法在基金超额收益方向预测中的应用，并在方向预测的基础上构建了不同的投资策略。

第 6 章　探讨了基于非参数随机森林方法的我国商业银行贷款违约预测模型，通过对我国某银行 17 469 个房屋抵押贷款样本的研究，发现随机森林的预测准确率较高。此外，对利率调整政策进行了模拟，发现利率对违约率的影响存在不对称性和非线性特征。

第 7 章　提出了基于随机森林方法的信用卡信用风险识别模型，利用随机森林变量重要性度量方法筛选合适的评价指标体系，建立可靠的分析模型，对信用卡用户的行为进行风险识别和预测。

第 8 章　针对保险业的特殊性，提出了保险客户利润贡献度(ICP)模型，综合考虑了客户历史购买行为和未来现金流、责任准备金。以某保险公司的数据为例，建立了基于非参数随机森林回归的预测模型。并与 SVM、CART、GLM、回归、NN 这五种算法进行比较，其中准确度最高的是随机森林模型。

第 9 章　构造了基于滞后收益率模型、基于星期效应模型以及基于滞后均值方差模型的分位数回归森林法求 VaR。本章选取了我国上证综指、深证成指、SP500 指数、FTSE100 指数、香港恒生指数(HSI)、日经 225(N225)等国际股票市场主要指数的收益率进行了实证研究，得出了一些很有意义的结论。

本书适合于经济、金融、统计、管理等专业的高校教师、科研

人员以及学生，也适合于经济分析师、企事业、政府经济管理有关预测分析人员阅读。本书理论与应用分别阐述，各自成章，脉络清晰，不同的读者可以根据自己的需要进行有选择地阅读。

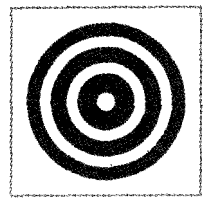
本书的出版得到了厦门大学出版社的大力支持，在此表示衷心地感谢！本书的出版还得到了厦门大学统计系朱建平教授和台湾辅仁大学谢邦昌教授的鼓励。本书的编写得到了中央高校基本科研业务费专项资金(2010221040)、福建省社会科学基金青年项目(2011C042)、国家统计局全国统计科研计划项目(2011LD002)以及国家社科基金(11BTJ001)的资助。

由于作者水平有限，成书匆匆，难免有疏漏或错误之处，恳请读者多提宝贵意见！

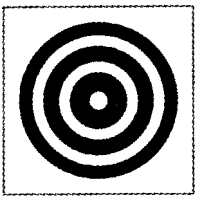
作者

2012 年 5 月于厦门

目 录



第1章 绪 论



本章从统计预测方法的发展过程出发，对统计预测方法的发展阶段进行了合理地划分，指出了目前统计预测方法的发展趋势，揭示了非参数统计方法、随机森林方法在金融市场预测中的重要意义和重大应用前景，由此确立了本书研究的目的、内容、框架和思路。本章还对非参数随机森林的研究框架和相关研究现状进行了分析和回顾。

1.1 统计预测方法发展历程

何为预测？《韦伯斯特辞典》中预测的定义为“以现有的相应资料的理论研究和分析成果来计算或预报未来的某些事件或情况”。预测的历史源远流长，早在春秋战国时期，越国范蠡就提出了根据商品的供求数量来预测价格的思想。预测存在于自然现象和社会现象的各个领域，比如社会预测、经济预测、气象预测、军事预测等。传统的预测方法以定性预测为主，主要是凭经验进行主观推断，往往精确度不高。为了适应

社会经济发展的需要，综合运用数学、统计、逻辑和计算机技术的统计预测方法（也称定量预测方法）得到快速的发展。统计预测方法是利用历史数据建立有关统计模型，通过统计模型对未来的经济活动作出估计、描述、分析和判断，揭示有关变量之间的规律性联系，用于预测和推测未来发展变化情况。统计预测在经济领域得到了广泛的应用，各国相继成立各种专门的预测研究机构或咨询机构，为各国政府以及企业提供宏观经济预测、行业经济预测和其他经济预测服务，并同时对统计预测的理论和方法进行了深入的研究，促进了统计预测方法的发展。

陈诗一（2008）把统计预测（quantitative forecast）方法的发展根据出现时间的先后大体分为三个阶段：结构计量模型阶段、时间序列分析阶段和人工智能预测阶段。虽然这种阶段划分一定程度上描述了统计预测方法的发展历程，但本书认为这种划分还不够完善，没有很好地把很多统计学家最近几十年研究的非线性非参数计量模型的成果包括在内，或者说没有突出非参数非线性计量模型在统计预测中的重要性。因此，本书在借鉴陈诗一（2008）对统计预测方法划分的基础上，把统计预测方法划分为四个阶段：结构计量模型阶段、时间序列模型阶段、非线性非参数计量模型阶段、数据挖掘智能预测与组合预测阶段。这四个阶段并非界限明显、完全区隔，后一阶段的预测方法也并不一定优于前一阶段的预测方法，也不是说前面的方法就不能使用了。对于一个复杂系统，可能需要借助多种预测方法。下面逐一详细介绍统计预测方法的四个阶段演进过程。

1.1.1 结构计量模型阶段

所谓结构计量模型(structural econometric model)就是先找到关于某个问题的经济理论,并先假定这个理论是正确的,然后根据该理论来设定具体的统计模型以用于估计和预测(陈诗一,2008)。它本质上是一种理论导向型方法。这类模型最常见的是古典线性回归模型(classical linear regression model),包括单方程模型和多方程模型(如联立方程)。

在20世纪80年代之前,由于结构计量模型能把经济理论和数学方法很好地融合在一起,利用经济学中比较成熟的经济和金融理论帮助人们进行模型设定,并且模型估计出来的系数都具有很好的经济含义,因此结构计量模型深受经济学家喜爱。美国计量经济学家Klein是利用结构计量经济模型进行经济预测的代表人物,他把凯恩斯主义宏观经济理论和计量经济方法结合起来,建立了美国宏观经济模型,用来定期从事经济预测,为美国政府的经济决策提供有力的支持。他所建立的最庞大的经济模型达150多个回归方程,上千个经济变量。Klein因在经济领域的卓越贡献而获得了1980年美国诺贝尔经济学奖。

经济理论可以指导计量模型的设定,反过来计量模型的实证结果也可以用来检验经济理论的正确性。经济理论在20世纪30年代至70年代发展的经典计量经济学模型的设定中起着导向作用。计量经济学根据已有的经济理论进行总体模型的设定,将模型估计和模型检验看作自己的主要任务。经济理论可以被认为是嵌入计量经济学模型的,相对经验数据而言具有先验性。

经典计量经济学模型理论方法体系是基于截面数据建构的(李子奈,2007)。截面数据的关键特征是,数据来源于随机抽样,数据顺序与计量分析无关,随机抽样隐含了待界定的特定总体。在经典的 Gauss-Markov 假设和随机扰动项服从正态分布假设下,基于来自总体的一个随机抽样,按照最大可能性(极大似然估计)或最小偏差(最小二乘法)的统计法则,对总体模型参数进行统计推断,得到估计的总体模型,称为样本回归模型。只要 Gauss-Markov 假设隐含的总体模型足够现实,只要样本容量足够大,大数定律就保证了估计量的一致性,即渐近无偏性,而中心极限定理则为大样本下随机扰动项渐近服从正态分布提供了支持,并保证了估计量的渐近有效性。于是,估计得到的总体模型方程与自在的原型方程的偏差是可以忽略的。因此,按照计量分析规则建立的知识是可靠的(李子奈,2007)。

结构计量模型对 20 世纪 70 年代的经济衰退和滞胀的预测与政策分析失效,引来了著名的"卢卡斯批判"。Lucas(1976)认为使用计量经济模型预测未来经济政策的变化所产生的效用是不可信的,提出了结构模型参数是否随时间变化的问题。Sargent(1976)以货币政策为例,重新解析了卢卡斯批判,认为结构模型对于评价政策似乎是无能为力的。Sims(1980)指出,为使结构方程可以识别而施加了许多约束,而这些约束是不可信的,建议采用向量自回归(VAR)模型而避免结构约束问题。

1.1.2 时间序列模型阶段

1927 年英国的 Yule 将自回归模型用于太阳黑子数据分析和俄国的 Slutzky 创建滑动平均模型标志着时间序列作为一种分析方法诞生。但由于缺乏经济含义,时间序列分析方法自提