量化投资与对冲基金

Research on Quantitative Investment by Machine Learning

学习富可敌国的华尔街对冲基金的赚钱秘诀
深度解读金融大鳄的核心投资策略

机器学习在量化投资中的应用研究

汤凌冰 著

量化投资与对冲基金丛书

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

机器学习
在量化投资中的应用研究

汤凌冰
著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内容简介

本书是国内少有的研究机器学习在量化投资中应用的专著。主要运用多层感知器神经网络、广义自回归神经网络、模糊神经网络与支持向量机对证券时间序列进行回归分析。特别是在支持向量机框架下构造了小波、流形小波与样条小波三种核函数，并在此基础上建立了股指收益与波动预测两类新的量化投资模型。与经典高斯核相比，具备多分辨分析特性的新模型能较好地捕捉曲线性状，各预测指标在模拟数据与真实数据上均占优，表明其具有良好的适用性与有效性。

图书在版编目（CIP）数据

机器学习在量化投资中的应用研究 / 汤凌冰著. —北京：电子工业出版社，2014.11
（量化投资与对冲基金丛书）
ISBN 978-7-121-24494-0

Ⅰ. ①机… Ⅱ. ①汤… Ⅲ. ①机器学习－应用－投资－经济策略－研究
Ⅳ. ①F830.59-39

中国版本图书馆 CIP 数据核字（2014）第 233235 号

策划编辑：李　冰
责任编辑：葛　娜
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编 100036
开　　本：720×1000　1/16　　印张：11　　字数：185 千字
版　　次：2014 年 11 月第 1 版
印　　次：2014 年 11 月第 1 次印刷
定　　价：59.00 元

量化投资与对冲基金丛书简介

最近十年来，量化投资已成为欧美资本市场发展的热点与焦点，一举成为国际投资界兴起的一个新方法，发展势头迅猛，和基本面分析、技术面分析并称为三大主流方法。由于量化投资交易策略的业绩稳定，其市场规模和份额不断扩大，得到国际上越来越多投资者的追捧。

过去 20 年收益率最高的基金，是文艺复兴科技公司的大奖章，其客户平均年收益率高达 35%；而过去 4 年高盛旗下的量化基金规模翻了一倍，超过 1000 亿美金。由此可见，量化投资已经成为机构投资者的重要利器。

量化投资对于基金公司/资产管理公司而言，有着非常明显的价值：

首先是容易冲规模。一个有效的量化模型是可以在多个产品上进行快速复制，从而迅速做大规模。这一点在巴克莱的指数增强系列产品上得到最明显的体现。截至 2011 年底，巴克莱量化基金管理规模超过 1.6 万亿美金，超过富达基金，成为全球最大的资产管理公司。

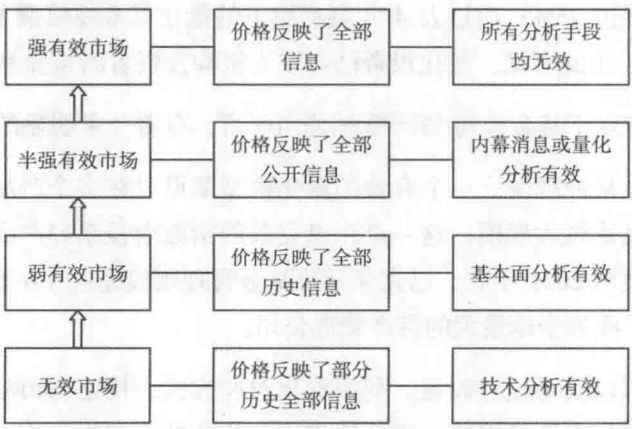
其次是可以获得绝对收益。利用量化对冲方式，构建与市场涨跌无关的产品，赚取市场中性的策略，适合追求稳健收益的大机构客户，例如保险资金、银行理财等。这个产品的代表性公司就是目前全球最大的对冲基金 BridgeWater，旗下的旗舰产品 Pure Alpha 过去 5 年共赚取超过 350 亿美金。

再次是杜绝了内幕消息和老鼠仓。量化投资只利用公开数据，通过数学

模型的运算，挖掘出隐藏在公开数据后面的信息，从而战胜市场，从方法论上就杜绝了内幕消息的可能。在交易过程中利用复杂的 IT 系统进行程序化交易，使得老鼠仓也无法成为可能。在国内金融市场监管日趋规范的情况下，量化投资必然会成为投资研究的主要方法。

量化投资的理论基础

说到量化投资的理论基础，就要从市场有效性假说说起，下图是市场有效性假说的一个层次图。从最右边的来看，技术分析、基本面分析和量化分析代表了有效市场的三个不同层次。在无效市场，技术分析是充分有效的，这在中国资本市场最初的 10 年得到了很好的体现；当市场进入弱有效市场后，可以依靠基本面分析获得超额收益，2000 年—2010 年这十年基本上属于这个时代；我们可以观察到，当市场进入半强有效市场后，也就是从 2010 年开始，大部分基本面分析的产品已经无法获得超额收益，此时国内市场已经进入半强有效市场。当然当市场进入强有效市场后，则无论哪种方法均无法战胜市场，那时候只能被动指数化投资。



传统的有效市场假说认为，在半强有效市场，只能依靠非公开信息（内幕消息或者私人消息）来获得超额收益。但是我们可以知道的是，非公开信息并不是只有内幕消息和私人消息，还有另外一个获得非公开信息的方法，就是利用数据挖掘的方法，从公开的数据中挖掘出非公开信息，也就是量化

投资的方法。这也就是在美国等成熟市场（基本上进入半强式有效市场状态），量化投资可以得到蓬勃发展的原因。

随着中国市场有效性的提高，中国开始进入半强式有效市场阶段，再加上监管层对内幕消息的监管越来越严厉，使得通过这种方法获得非公开信息的方式越来越难，因此量化投资就成为一个最好的获得非公开信息的科学理论与技术。

很多人会问："量化投资仅仅是一个昙花一现的概念，还是一个可以长期有效的科学理论？"通过上述对有效市场假说的分析，已经得到了明确的答案：**量化投资是在半强式有效市场中的最佳分析理论，也几乎是唯一可行的分析理论。**

量化投资与对冲基金丛书

2011 年年底，电子工业出版社出版了丁鹏博士的专业著作《量化投资——策略与技术》一书，其前身只是为了培训新来的金融工程研究员而撰写的讲义，但是出版后一个月内第一版就脱销了。后来又多次印刷，依然供不应求，整个市场对于量化投资方面的书籍和读物的渴求非常大。在这种情况下，电子工业出版社向中国量化投资学会发出邀请，由丁鹏博士作为主编，联合策划了"量化投资与对冲基金丛书"，立刻得到了业内专业人士的热烈响应，并且积极参与其中，丛书涉及策略、技术、理论、外版经典、人物传记等方向。

强大的技术后援团——中国量化投资学会

放眼国内，2009 年开启了"中国量化投资元年"，2010 年 4 月股指期货出台，2011 年年底，上海交通大学金融工程研究中心举办了第一届中国量化投资高峰论坛，2012 年 1 月，作为国内第一本全面介绍量化投资策略方面的教材《量化投资——策略技术》正式上架，2012 年 6 月，国内第一本《量化投资与对冲基金》杂志刊发。从 2012 年开始，量化投资与对冲基金的各类论

坛、会议、组织如雨后春笋，层出不穷。而随着国债期货、转融通、期权等一系列产品的推出已经提上监管层的日程安排，未来中国量化投资与对冲基金时代，已如滚滚长江，浩荡而来。

2012 年 3 月中国量化投资学会（http://chinaqi.org）成立，仅仅半年，就从十几个人的兴趣小组发展成为中国最大的量化投资的研究团体，会员遍及全国各地，甚至北美、欧洲、我国港台地区均有相应分会成立。

中国量化投资学会的宗旨是：构建研究合作、资源分享的平台，让热爱量化投资的研究员、投资者、产品设计人员分享经验，搭建人脉，共同推动量化投资在中国的发展。

美好前景

未来 10 年，量化投资与对冲基金这个领域是少有的几个可以诞生个人英雄的行业，无论出生贵贱，无论学历高低，无论有无经验，只要你勤奋、努力。脚踏实地地研究模型，研究市场，开发出适合市场稳健赢利的交易系统，实现财务自由，并非遥不可及的梦想。

在中国目前的很多领域，赚钱已经变成一件非常困难的事情，在量化投资与对冲基金领域，完全依靠自己的勤奋与努力。一个持续稳定赚取的模型，不是靠关系和背景就可以实现的，而要靠着自己的聪明才智和脚踏实地的工作。

让我们一起拥抱中国量化投资与对冲基金黄金时代的到来！

前　言

诺贝尔经济学奖得主罗伯特·默顿（Robert Merton）认为现代金融理论由资金的时间价值、资产定价与风险管理三大支柱构成，其核心问题就是如何在不确定的环境下对资源进行跨期的最优配置。基于这一理解，斯坦利·R·普利斯卡从整个数理金融领域归纳出了随机过程与随机控制两类基本模型。显然，前者是后者的前提与基础。因而，作为离散随机过程的金融时间序列必然是金融模型研究的基石与关键。同时，鉴于股指收益序列与波动率序列在投资组合和风险规避中的重要作用，本书拟围绕其展开研究。

与传统统计学相比，统计学习理论（Statistical Learning Theory，SLT）是一种专门研究小样本情况下机器学习规律的新型理论。该理论针对小样本统计问题建立了一套全新的理论体系，其统计推理规则不仅考虑了对渐近性能的要求，而且追求在现有有限信息的条件下得到最优结果。因此，本书拟基于统计学习理论，以机器学习为工具，进行股指收益序列与波动率序列的建模研究。

本书讨论了模糊神经网络在股价预测中的应用。模糊神经网络克服模糊规则产生对专家的依赖性及模糊集的非自适应性，隶属函数的自适应和模糊规则的自组织通过神经网络的自学习和竞争获得。通过一个股价预测实例验证了该方法的有效性。接着，本书将支持向量回归机这一新型神经网络应用

于收益序列预测的回归分析，力求在克服数据过拟合现象的基础上寻找问题的全局最优解。通过交叉验证选择学习参数。实验表明基于二次规划与核函数理论的高斯核函数支持向量回归机能准确捕捉动态股票收益序列的波形特征，其预测性能与多层感知器以及广义回归神经网络进行比较，具有较为明显的优势。

基于小波理论，本书提出了小波核的一种新型构造方法。用高维母小波函数直接生成小波框架，通过缩放与平移产生平方可积空间中的一个完备基，从而构造出满足 Mercer 条件的小波核函数。该核在理论上具有任意逼近平方可积空间中目标函数的优点。实验表明与高斯等核函数相比具有多分辨率特性的小波核确实能较好地逼近目标函数。

基于流形理论，本书提出了一种新的流形小波核。该核借鉴了 Amari 提出的依据数据流形几何特征修改核函数进而增进分类性能的思想方法，通过缩减超平面附近的黎曼距离处理回归问题。该核具有融入支持向量数据依赖知识的优点。实验表明流形小波核能比高斯等核函数更好地捕捉曲线性状。

基于样条理论，本书提出了一种新的样条小波核。用一维样条母小波通过平移与缩放产生一维样条小波核函数，接着依据乘法原理，生成高维样条小波核函数。该核具有函数形式简单与支集小等优点。实验表明样条小波核解析波动特征的能力比高斯等核函数要强。

针对金融时间序列自身的高噪声、动态与混沌等特性，本书提出了新型小波支持向量机-股价动力学模型。该模型具有所需样本小、泛化性能好、全局最优与高容噪性等优点。与高斯等核函数相比，其多分辨特性使得该模型各主要预测性能指标在模拟数据与真实股指数据实验中占优，因而能较好地分析股指收益。

针对波动率序列高峰、厚尾与长效依赖等特性，本书提出了新型小波支持向量机-广义自回归条件异方差模型。该模型同样具有所需样本小、泛化性能好、全局最优与高容噪性等优点。借助采用多尺度分析核的小波支持向量

机能有效捕捉波动率的聚集特性，从而对股指波动进行较为准确的预测。通过模拟实验与真实股指数据分析，该模型在波动率分析中的适用性与有效性获得了证实。

本书可供计算机、信息管理与金融类专业高年级本科生与研究生使用，也可供从事机器学习技术与应用研究的科研人员、金融市场数据分析人员以及机器学习软件开发人员参考。

目　录

第1章

绪　论

1.1　背景与意义

（1）基于研究对象的视角

随着各国金融市场的不断开放，国际资本在世界范围内自由流动。金融全球化以超乎寻常的方式影响着整个人类的经济生活，有力地推动了世界经济的一体化进程。然而，它在极大地增进经济活力的同时也不可避免地孕育出巨大的潜在风险。各主要金融市场间的依赖性随着全球资源的动态配置而日益增强，与此相反，世界金融体系的稳定性却在显著下降。这使得任何局部地区的金融波动都会迅速扩散到其他地方，从而引发金融危机，给世界经济造成严重伤害。

20世纪70年代布雷顿森林体系崩溃导致了国际货币体系的瓦解，接着美联储以货币总量管理代替利率管理的体制调整又引起了世界经济的剧烈动荡。进入80年代以后，国际金融市场经历了前所未有的迅猛发展，全球化浪潮更是不可阻挡，世界范围内的金融波动便层出不穷——1982年的拉美国家债务危机、1992年的欧洲货币体系危机、1994年的墨西哥金融危机、1997年

的东南亚金融危机、1998 年的俄罗斯金融危机和 2002 年由巴西和乌拉圭金融动荡引起的拉美金融危机等无不伴随着汇率动荡、货币贬值、股市暴跌、公司破产、银行倒闭等衰退现象。特别是最近爆发延续至今的世界金融危机也深刻地印证了这一点。

因此，科学地预测金融市场的波动特征，掌握金融市场的波动规律及其结构对风险的规避防范与管理监控有着重要意义。而作为世界经济晴雨表的各主要股票指数，其收益分析与波动率预测是诸多金融模型研究的前提与基础。这是本书选择股指收益及波动率作为研究对象的金融意义与初衷。

（2）基于研究方法的视角

传统统计学研究的是样本数目趋于无穷大时的渐近理论，现有的学习方法也多是基于此一假设。而现实中样本数往往有限，因此一些理论上很优秀的学习方法实际表现却不尽人意。同时，预先知道样本分布形式的要求，代价很大也很苛刻。这些都制约了计量金融的实际应用范围并阻碍了它的进一步发展。与传统统计学相比，统计学习理论（Statistical Learning Theory，SLT）是一种专门研究小样本情况下机器学习规律的新型理论。该理论针对小样本统计问题建立了一套全新的理论体系，该体系下的统计推理规则不仅考虑了对渐近性能的要求，而且追求在现有有限信息的条件下得到最优结果。因此，基于统计学习理论的计算金融，研究面向数据驱动的人工智能与机器学习算法在金融分析中的应用，以其良好的适用性与有效性而逐渐成为国际上金融领域的热点和重点。

支持向量机是建立在统计学习 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折中，以期获得最好的推广能力（Generalization Ability）的新型神经网络。V. Vapnik 等人从 20 世纪 60、70 年代开始致力于此方面的研究，直到 90 年代才使抽象的理论转化为通用的学习算法。随着ε不敏感损失函数的引入，支持向量机从原来只处理分类问题逐步扩展到也能胜任回归任务。尤为值得一提的是，通过构造核函数能在无须知道映射具体形式的情况下将非线性问题映射到高维线性空间，并对 SVM 的预测性能起到决定性作用。

小波分析是当前数学中一个迅速发展的新领域，它同时具有理论深刻与应用广泛的双重意义。与窗口傅里叶变换相比，小波变换是一种灵活的时频局域化变换。通过伸缩和平移等运算功能对函数或信号进行多尺度细化分析，可以逼近任意函数，解决了傅里叶变换不能解决的许多难题，被誉为“数学显微镜”，是调和分析发展史上的重要里程碑。样条小波具有构造简单、短支集、半正交等优良特性，而且作为平滑函数具有很高的正则性，其显示表达便于对问题进行深入分析和估计，也易于构造高维小波并为计算机编程实现。流形学习是一种新的非监督学习方法，近来引起越来越多机器学习和认知科学研究人员的重视。20 世纪 80 年代末，在 Pattern Analysis and Machine Intelligence 杂志上就有了流形模式识别研究。2000 年 Science 上的三篇论文从认知论上讨论了流形学习，并引入了 manifold learning 术语，强调了认知过程的整体性。而基于微分几何、信息论以及统计学的信息几何理论也着重整体结构研究。两者均假定观测空间的数据集由某些内在变量控制生成，在仅存观测空间数据集的前提下恢复其内在整体结构和相应映射关系，因而更能体现事物本质，解决机器学习中一些完成不好或无法解决的问题，是研究非结构化、非线性空间更合理的方法与手段。

本书希望根据金融时间序列的有关特性，研究基于机器学习的金融时间序列建模。结合支持向量机、小波分析与流形理论的各自优点，构造出高维小波核函数，证明其满足 Mercer 条件，进而提出优于 Gaussian 等核函数的小波支持向量回归机，从而为金融时间序列特别是股指收益与波动率序列提供一种可选方案。下一节将对全文提出的三个小波核所涉及的有关模型进行综述，以期为各章构造相应的小波核提供背景支持。

1.2　国内外研究现状

1.2.1　金融时间序列方法

金融时间序列指描述不同金融产品诸如股票、汇率与基金等的时间序列。它与金融市场中人类的各种经济活动密切相关，呈现出复杂多变的状态。因而与其他的时间序列相比，金融时间序列具有以下特性[1][2]。