

数据驱动的金融时间 序列预测模型研究

A Data-driven Research on the Model of Financial
Time Series Forecasting

张贵生/著



科学出版社

数据驱动金融时间序列 预测模型研究

张贵生 著

科学出版社

北京

内 容 简 介

以非线性动力学的观点看来,现代金融理论中金融系统的不确定性恰源于其自身就是一个受多种因素综合影响的具有开放性质的复杂巨系统,相应地,作为系统观测值的金融时序数据则从形式上表现了该系统的复杂运动规律。基于此,本书借鉴复杂系统视角建模的思想,结合智能计算、计算实验金融、数据挖掘及控制论等相关领域的最新研究成果,“自底向上”地展开金融时序数据经验知识融合下的机器学习预测建模创新研究,以探索金融系统的复杂演化规律。

本书可供财务、金融、管理科学与工程、系统工程与智能计算等领域或有相关知识背景的本科生、研究生及行业从业者参考。

图书在版编目(CIP)数据

数据驱动的金融时间序列预测模型研究 / 张贵生著. —北京:科学出版社, 2018. 1

ISBN 978-7-03-054250-2

I. ①数… II. ①张… III. ①金融-时间序列分析 IV. ①F830

中国版本图书馆CIP数据核字(2017)第210679号

责任编辑:李 敏 杨逢渤 / 责任校对:彭 涛

责任印制:张 伟 / 封面设计:无极书装

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

<http://www.sciencep.com>

北京京华虎彩印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018年1月第 一 版 开本:720×1000 B5

2018年1月第一次印刷 印张:8 3/4

字数:200 000

定价:68.00元

(如有印装质量问题,我社负责调换)

前 言

1997 年诺贝尔经济学奖获得者美国经济学家罗伯特·C. 莫顿 (Robert Carhart Merton) 提出, 现代金融理论的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。而按照非线性动力学的观点来看, 现代金融理论中金融系统的不确定性恰恰源于其自身就是一个受多种因素综合影响的具有开放性质的复杂巨系统, 相应地, 作为系统观测值的金融时序数据则从形式上表现了该系统的复杂运动规律。相关金融时序可预测性的文献研究表明, 无论是线性范式下的传统统计方法, 还是非线性的计算智能方法, 以及多种不同类型方法的组合模型都在一定范围内提升或改善了人们对于金融时序数据预测的精确性和稳定性, 但大多缺乏对不同类型金融时序数据内部时间相关性知识、价格变化趋势信息及不同市场间互信息等经验知识的有效融合, 制约了其预测性能的进一步提高。

基于此, 本书借鉴复杂系统视角建模的思想, 针对各种不同类型的金融时序数据, 结合智能计算、计算实验金融、数据挖掘及控制论等相关领域的最新研究成果, “自底向上”地展开金融时序数据经验知识融合下的机器学习预测建模创新研究, 以探索金融系统的复杂演化规律。主要研究成果和创新概括如下。

1) 针对单变量金融时序数据变化趋势信息和市场隔夜跳空开盘信息的重要性, 在借助跟踪微分器提取数据近似微分的基础上, 分别构造了基于微分信息和基于梯度信息的 ARMA-GARCH 单预测模型, 旨在增强模型在高噪声扰动环境下对于时序数据变化趋势的判别能力; 从分阶段混合模型构造的角度, 提出了一种新的基于 ARIMA 和泰勒展开的预测模型, 改进传统方法对于时序数据内部经验知识学习不够充分的问题。

2) 针对多变量金融时序数据所具有的高维复杂性, 考虑到金融市场间日益显著的联动及传染效应, 提出了一种基于近邻互信息的 SVM-GARCH 模型, 旨在融合市场联动行为信息的基础上, 提高传统单模型对于时序数据非线性成分的处理能力; 从构造混合模型提高预测稳定性的角度, 借助时间测地线距离的概念构造新的混合预测模型, 改善模型对于时序数据内部时间相关性知识的学习和泛化

能力，提高预测结果的精确性和有效性。

数据驱动的“自底向上”的建模方法就是要通过对复杂系统的观测数据进行分析，来展开对金融系统的反向研究，有利于克服规范分析假设过于严格、实证分析难以进行灵敏度测试的弱点，是金融理论与实践的高度统一。本书中无论是相关单变量金融时序数据的分析，还是针对高维金融时序面板数据的研究，数据驱动的模型创新都体现了复杂系统视角建模思想在金融领域的应用，旨在从不同的角度充分挖掘并发现复杂金融现象背后的客观规律，降低金融系统的不确定性，提高金融市场效率，并为相关金融理论创新研究提供新的动力和方向。

目 录

第 1 章 绪论	1
1.1 研究的背景和意义	1
1.2 文献回顾与评述	3
1.3 主要研究成果及创新	13
1.4 研究方法及技术路线	17
第 2 章 相关理论基础	21
2.1 ARIMA 模型	21
2.2 GARCH 模型族	23
2.3 支持向量机	27
第 3 章 基于微分信息的 ARMAD-GARCH 股票价格预测模型	33
3.1 ARMAD-GARCH 模型	35
3.2 实证研究	38
3.3 本章小结	46
第 4 章 基于梯度因子的 G-ARMA-GARCH 股票价格预测模型	48
4.1 G-ARMA-GARCH 模型	50
4.2 实证研究	52
4.3 本章小结	60
第 5 章 基于近邻互信息的 SVM-GARCH 股票价格预测模型	61
5.1 SVM-GARCH 模型构造	63
5.2 实证研究	66
5.3 本章小结	78
第 6 章 基于 ARIMA 和时间测地线距离 SVM 的股票价格时序数据混合预测模型	80
6.1 时间相关性经验知识	80
6.2 基础模型介绍	82

6.3	混合预测模型构造	83
6.4	实证研究	84
6.5	本章小结	90
第7章	基于 ARIMA 和泰勒展开的金融时序混合预测模型	92
7.1	模型构建背景知识	94
7.2	基于跟踪微分器的泰勒展开预测模型	97
7.3	混合预测模型 ARIMA-TEF 构建	103
7.4	本章小结	118
第8章	结论与展望	120
8.1	结论	120
8.2	不足与展望	122
参考文献		124

第1章 绪 论

1.1 研究的背景和意义

金融业是现代经济的核心,关系着整个经济社会发展的全局。通过方法论的不断创新,探索金融市场内部的客观规律以提高对其监管的科学性和有效性、改善统一市场体系整体效率是所有金融市场参与者和各级管理部门不断追求的目标之一^[1]。经济全球化发展的趋势下,国际金融市场一体化的加深在推动世界经济高速发展的同时,不可避免地带来了其自身稳定性下降的潜在风险。不同国家或地区金融市场的局部波动有时会引起强烈的连锁反应,尤其是在经济萧条的背景下,金融市场间的联动或传染效应往往演变成全球范围内的金融危机,给世界各国的经济造成严重破坏^[2]。1982年拉美国家债务危机最终导致了大规模的经济萧条;1992年欧洲货币体系风波迭起,汇率机制出现剧烈动荡直至爆发大规模危机;1994年墨西哥发生了比索汇率狂跌、股票价格暴泻的金融危机;1997年亚洲金融风暴席卷多个国家,打破了亚洲经济急速发展的景象;1998年俄罗斯经历了由三次金融大风波构成的金融危机,最终导致两届政府垮台,甚至产生全球效应;2007年美国次贷危机、欧洲债务危机引发的金融海啸席卷整个世界金融市场进而影响了实体经济。这些危机发生期间,全球股市均应声下落,市场投资者损失惨重。众所周知,信息是金融市场中所有参与者进行交易的主要依据,信息效率的高低直接影响资产的价格并最终决定市场资源配置效率的高低。对于金融危机的相关研究发现,这些特殊的极端金融现象背后都有其特有的规律,而这些规律往往都隐藏在之前很长一段时间内的复杂市场表象之后。假如我们能够通过数据建立相关的分析和预测模型,尽早发现其背后潜在的客观规律,就可以通过模型判断,及时发出预警信号,减少或避免大规模金融灾难的发生。金融市场是一个多因素综合影响下的具有高度不确定性的复杂巨系统,尤其是当今国际金融创新日新月异、系统不确定性日益增强的经济金融环境,使得对于金融市场

背后潜在规律的分析与研究更加富有挑战性。

在有效市场环境中,价格实时反映各种信息,是市场资源配置的风向标。金融时间序列的预测性研究就是要深入挖掘金融市场背后的客观规律信息,改善市场信息的传递速度和市场交易透明度,进而促进价格机制的良性运转,实现金融市场对于资源的有效配置。虽然对于市场有效性的研究争论不断,但是致力于辅助获取市场超额收益的金融时间序列可预测性的研究却从未间断^[3]。近年来,无论是“市场有效假说”研究范式下的传统线性分析研究模型,还是强调证券市场信息接受程度和投资时间尺度对投资者行为影响的“分形市场假说”研究范式下的非线性动力学分析研究模型等,不同角度的大量研究已经证明,尽管内部产生原因众说纷纭,但金融市场收益的可预测性却是不争的事实。正如经济学家 Timmermann 和 2003 年诺贝尔经济学奖获奖者克莱夫·格兰杰 (Clive W. J. Granger) 教授^[4]所指出的,不能简单对立地看待旨在获取超额收益的金融市场预测性研究与市场有效性理论。实质上,在各种超额收益获取模式被市场参与者学习与应用的过程中,尤其是在投资者通过各种方法挖掘金融市场背后客观规律以改善对于市场运动变化趋势的预测性研究的推进过程中,短期的利润获取会带来更多市场参与者的效仿和跟进,促使股票价格不断向真实价值恢复,进而提升了市场的价值发现功能,而这种伴随不断适应技术革新现象而得到的市场效率改善恰恰也就是金融时序数据预测模型不断推陈出新的原动力所在。格罗斯曼-斯蒂格利茨悖论 (Grossman-Stiglitz paradox)^[5]也指出,如果价格完全反映市场信息,则市场参与者就不会有去获取并分析信息的积极性,但若大家都不去理会市场信息,那么市场价格就不可能是信息有效的。所以,市场中的价格形成机制乃至整个市场效率的高低依赖于市场有效信息的获取,换言之,依赖于市场参与者企图获取超额收益时对各种不同投资模式的大胆尝试和运用。

金融时间序列是一种特殊的时间序列数据,内容既可以是股票价格、期货价格或各种债券价格及汇率利率数据,也可以是其他微观或宏观经济类的相同时间间隔采集的一系列数据。形式上可以是传统的一维(单变量)时序数据,也可以是由多个属性所构成的高维(多变量)面板数据。与其他时间序列数据相比,金融时序数据由于受多种因素的综合影响,体现出与人类社会活动密切相关的属性,具有信噪比低且不易分离、非平稳且非线性显著的特点^[2],而且表现出很强的多周期性,如“周末效应”“一月效应”“假日效应”等^[6,7]。“本质决定现象,现象反映本质”,这些金融时序所表现出来的高度复杂性背后必定蕴藏着金

融市场内部的潜在规律信息,亟待广大研究人员展开相应的分析与深层次研究,揭开其神秘的面纱^[8]。Shafer 和 Sylla 指出,预测学研究的目的就在于减少人类社会生活中由于各种不确定性的存在所带来的决策风险^[9,10]。金融时序数据的预测性研究,就是要因地制宜地在深入分析数据高维复杂性的基础上,创新相关预测体系研究,获取其背后客观规律信息,通过降低金融市场中的各种不确定性指导投资决策,改善市场效率^[11]。

1.2 文献回顾与评述

在过去的 40 年中,金融时间序列的分析主要在 Fama 有效市场假说(efficient markets hypothesis, EMH)的指导下受到一个线性研究范式的主宰。不可否认,有效市场假说的研究框架推动了金融市场的发展。然而,随着“过度交易”“羊群效应”“周末效应”等一系列金融异象的出现,很多学者为了克服市场有效论的不足,展开了行为金融学(behaviour finance, BF)和非线性动力学基础上的分形市场理论(fractal market hypothesis, FMH)研究。不仅取消了市场有效理论中的理性投资者假设,而且不再对数据分布做正态分布和线性关系的前提性假设,采用大量非线性的模型对金融时序数据展开研究,另外,经济学家开始求助新兴的计算实验金融(computational experiment finance, CEF)方法,采用自下而上的建模方式,在一定的交互机制作用下,观察整体的宏观涌现状态,以期通过对具体的技术规则进行统计检验来寻求时间序列收益可预测性(time series returns predictability, TSRP)问题的答案。这些新理论和技术出现,不仅很好地解释了金融市场中的各种“异象”,而且也使得对于金融时序数据的分析应用模型更加贴近了真实的市场环境^[12]。

对于金融时间序列可预测性的研究 20 世纪下半叶就已经开始,迄今为止已有 70 年左右的历史了。按照其不同发展阶段的理论基础和结构特点来分,金融时序数据预测模型的研究大致可以分为三类:一是以自回归为代表的基于传统统计学方法的模型;二是以人工智能算法为代表的计算智能方法;三是对于组合预测模型的研究。

1.2.1 传统的统计学方法研究现状

对于金融时序数据预测模型的研究中,传统的统计方法属于建立在大数定理之

上的渐进理论,是最常使用的也是最早应用于时序预测模型的一类方法。其中包括随机游走模型 (random walk, RW)、自回归移动平均模型 (autoregressive integrated moving average, ARIMA)、最小二乘线性回归模型 (ordinary least square, OLS)、自回归条件异方差模型 (autoregressive conditional heteroskedasticity, ARCH) 及广义自回归条件异方差模型 (generalized ARCH, GARCH) 等。

ARCH 模型于 1982 年由 Engle^[13] 首次提出,采用自回归的形式刻画时序数据可用信息和条件方差的时变性,很快在计量经济学领域得到了广泛的应用并取得了良好的效果。1986 年, Bollerslev^[14] 在 Engle 工作的基础上创造性地提出了 GARCH 模型,用于分析时间序列的异方差特性。GARCH 模型不再认为单变量金融时序数据通过线性回归后得到的残差是白噪声,而是具有 ARCH 效应的,应该在 GARCH 模型处理异方差效应之后,才能利用线性回归方程对时序数据进行分析 and 预测。GARCH 模型在很多方面尤其是在经济金融领域有了广泛的应用,而且取得了良好的效果。此外, GARCH 族模型与其他时变的波动模型相比具有使用方便、形式简洁的优点。French 等^[15] 运用 GARCH 方法在 *Journal of Financial Economics* 发表了关于美国股票市场收益与市场波动相互关系的文章。实证研究表明预期市场风险溢价与可预测收益波动呈正相关关系,而非预期股票收益与市场收益波动的非预期变化呈负相关关系。Robert 等^[16] 构建了 GARCH-M 模型对 3 组利率时序数据展开研究,最后的实证结果说明该模型很好地解释了利率期限结构假说在经济计量方面的失败。另外,还有大量文献,如 Nelson^[17]、Glosten 等^[18],以及 Engle 和 Ng^[19] 的相关研究均证实了条件异方差模型在解释收益率波动过程中的适应性。为了克服“数据窥探”问题,White^[20] 开展了大量工作,对 GARCH (1, 1) 模型和其他备择模型进行了比较研究,研究结果证实了 GARCH (1, 1) 模型的广泛适应能力,以及在对 IBM 股票的收益波动进行预测时 APARCH^① 模型更优的预测性能。Hansen 和 Lunde^[21] 比较了 330 种不同的波动率模型对于德国马克与美元汇率及 IBM 股票收益率数据的预测效果,结果证实了 GARCH (1, 1) 模型是所有异方差模型中的最优模型。Mohammadi 和 Su^[22] 强调了 MA (1)-GARCH(1, 1) 模型在对条件均值和波动进行建模及预测时的适应性优点。Liu 和 Shi^[23] 利用 ARMA-GARCH 及其改进的 ARMA-GARCH-M 模型对电价序列进行了建模及 1h 单步预测,实证结果证明了该组模型的有效性。

① APARCH, 即 the asymmetric power ARCH。

在学习国外学者对 GARCH 类模型研究成果的基础上,国内研究机构和广大学者也开展了大量的关于 GARCH 类模型在经济金融领域的理论与实证研究。魏巍贤和周晓明^[24]利用 GARCH 模型对中国股市进行了预测。唐齐鸣和陈健^[25]对我国股市的 ARCH 效应进行了相关分析,并对股指波动进行了拟合。另外,陈千里^[26]、陈千里和周少甫^[27]和岳朝龙^[28]等针对国内上海证券综合指数(简称上证综指)收益率的波动特征做了一系列的研究,证实了我国股市收益具有类似发达国家股市中“杠杆效应”“波动聚集”“不对称性”等相似的特点。张世英和柯珂^[29]对国内外 ARCH 模型研究的进展进行了综述研究,讨论了其长记忆性能和参数估计问题,而且针对复杂的 ARCH 模型还提出运用遗传算法的思想进一步解决其参数估计和检验问题。楼迎军^[30]与李胜利^[31]分别利用 EGARCH 模型和 VSGARCH 模型研究国内股市的不对称性,对不同样本的实证研究均表明国内股市存在显著的“杠杆效应”,投资者对负面影响敏感度高,明显存在“羊群效应”。张永东和毕秋香^[32]对国内上证综指股市的周波动性进行了样本外预测的比较研究,排除评价因子不同所带来的影响,实证结果表明了指数平滑模型预测效果的优越性。黄海南和钟伟^[33]对 GARCH 类模型预测波动率的精度进行了评价,对已实现波动率对不同的 GARCH 模型进行了比较性研究,并利用损失函数评价模型的预测精度,相关实证表明了 GARCH 模型的预测稳定性和不同分布下 GJR-GARCH (1, 1) 模型预测性能的优越性。邓超和曾光辉^[34],王佳妮和李文浩^[35]分析了上证综指波动率的特点,并用 EGARCH、GARCH-M、GARCH 及 ARCH 等几个常用模型进行比较研究,还讨论了不同的精确度衡量指标对于模型预测效果的评价结果。龚锐等^[36]为了避免参数 VaR 方法在分布假设及模型选择中的人为因素,提出用基于正态分布、 t 分布及 GED^① 等不同分布的多种衍生 GARCH 模型对风险进行测量。并针对国内股市上证综指和深圳证券交易所成分股价指数(简称深证成指)的实证结果进行了比较和分析,也得出了一些相关的结论。郑振龙和黄慧舟^[37]对香港恒生指数期权市场所含信息的研究发现,在预测期限较短(1周)时,GARCH (1, 1) 模型所含信息较多,预测能力要强于基于期权价格的隐含波动率模型。赵树然等^[38]利用非参数 GARCH 模型开展美元与日元对人民币汇率日收益率数据的波动性的预测研究,实证结果证实了非参数 GARCH 模型预测结果的精确性和稳健性。吴恒煜等^[39]建立风险中性 Levy-ARMA-

① GED, 即 genelized error distribution, 广义误差分布。

GARCH 模型,对香港恒生指数的期权定价进行了实证研究。方立兵等^[40]运用蒙特卡洛模拟方法考察了偏斜参数设定对 GARCH 族模型估计结果的影响效应,结果发现,若标准化扰动项偏斜而且厚尾,则基于对称厚尾分布的极大似然估计量是渐近有偏的。侯利强等^[41]针对金融时序数据的非线性和不确定性,提出一种模糊 FEGARCH 模型,对 2006~2011 年上证综指收益率序列的高频波动性进行预测研究,实证结果表明,模糊 FEGARCH 模型对于具有高偏度、尖峰厚尾和杠杆效应的非线性特征波动数据的预测效果更佳。于志军等^[42]按照以误差校正来提高股票收益率预测精度的思路,首先,利用灰色神经网络模型对股票收益率进行初步预测;其次,引入 EGRACH 模型来分析预测误差序列的内部信息;最后,根据误差预测结果对股票收益率的初始预测值进行校正。相关研究还有很多,但随着金融市场的发展及对金融时序数据研究的不断深入,人们发现其内在的非线性属性表现得越来越突出,然而基于传统统计学的各种常用的方法却由于模型自身线性结构的约束,影响了对于时序数据预测精度的进一步提高,在实际应用中往往不能达到很满意的效果。

1.2.2 计算智能方法研究现状

典型的计算智能技术也有很多,其中常用到的包括:人工神经网络 (artificial neural network, ANN)、模糊逻辑 (fuzzy logic, FL)、遗传算法 (genetic algorithm, GA)、进化算法 (evolutionary algorithm, EA) 和支持向量机 (support vector machine, SVM) 等。从 20 世纪末开始,人工神经网络逐步得到推广和应用,成为对于时间序列预测最常用的非线性方法之一,且被广泛应用于金融时序数据的分析和处理。例如, Gingburg 和 Horn^[43]提出了混合几种前馈神经网络的方法用来提高对于金融时序数据预测的精度。Luxhøj 等^[44]通过组合计量经济学模型和 ANN 模型构造出新的混合模型对销售进行了预测。Li 和 Cios^[45]构造了一个径向基函数 (radical basis functions, RBF) 和 Box-Jenkins 模型的混合方法用于时间序列数据的预测。Kyrtou 和 Terraza^[46]采用主成分回归 (principal components regression, PCR) 和径向基函数对法国 CAC40 指数收益率序列展开预测性研究,与传统的 ARCH 模型的比较实证结果表明这些方法的预测精度并无显著差异。Jasic 和 Wood^[47]用基于单变量的神经网络对未经转换的标准普尔 500 指数、法兰克福 DAX、东证股价指数及英国富时指数真实时序数据展开其收益的

短期预测, 这些指数从 1965 ~ 1999 年的实证研究结果的多个评价指标一致表明, 神经网络模型优于作为基准模型的线性自回归模型, 而且由神经网络所预测的买卖信号显著不同于单日无条件平均收益的结果, 对于指导投资者获取超额收益具有更科学的指导意义。Cao 等^[48]将神经网络和 Fama-French 三因子模型对中国股票市场的预测效果进行对比, 发现神经网络在金融领域的预测效果优于其他模型。Lee 和 Ko^[49]构建了基于非线性时间进化粒子群算法的神经网络预测模型, 相比其他普通的粒子群径向基函数优化模型, 该模型在预测准确性和计算效率方面表现更优。Khashei 等^[50]将 ARIMA 方法分别和人工神经网络及模糊逻辑系统进行组合, 以克服 ARIMA 线性模型对于时序数据量的限制。另外, Zhang^[51]提出混合 ARIMA 和 ANN 的模型的思路, 用于解决金融时序数据中的线性成分和非线性成分, 实证结果验证了混合算法性能显著提高。Shen 等^[52]在鱼群算法 (artificial fish swarm algorithm, AFSA) 优化径向基函数的基础上构建了新的神经网络模型对上证综指展开预测, 并与基因算法、粒子群优化的 RBF 算法、ARIMA 算法、BP 神经网络算法及支持向量机算法进行比较研究, 验证了所构建模型具有更高的预测精确度。Babu 和 Reddy^[53]通过研究时序的波动, 构建了一种新的 ARIMA 和 ANN 的混合多步预测模型, 并通过多种不同的真实数据对模型进行了实证研究, 结果表明新模型的预测精度有显著提高。研究人员 Akbilgic 等^[54]构建新的混合神经网络模型研究伊斯坦布尔 ISE 国家 100 指数的日收益波动规律, 实证结果表明将回归树、岭回归与 RBF 神经网络结合起来的混合模型对于高度独立且非线性的数据具有较好的预测性。Oliveira 和 Ludermir^[55]通过组合指数平滑、ARIMA 模型、ANN 模型和支持向量回归机, 提出了一种新的混合算法, 用以金融时序数据的预测, 而且实证结果表明了混合算法预测的精确性和有效性。

何芳和陈收^[56]给出一种新颖的基于扩展 Kalman 滤波的神经网络学习算法, 与传统的 BP 算法相比, 在用于股价预测时, 该方法具有更好的收敛率和学习能力。崔建福和李兴绪^[57]针对股价数据非平稳及非线性特点, 对 GARCH 模型和 BP 模型的预测效果展开比较分析, 实证结果表明, 这两种模型均优于简单的随机游走模型, 且 BP 模型无论在训练阶段还是针对样本外的预测结果, 其预测误差均显著小于 GARCH (1, 1) 模型, 但有时也会陷入局部极小值的困境。肖冬荣和杨子天^[58]也使用了神经网络模型展开时序预测研究, 而且还在模型的训练阶段运用了粒子群优化算法, 用优化后的模型对股票时序数据进行预测。实证结

果表明,基于粒子群训练的神经网络学习算法不仅易于实现,而且准确率较高。王文波等^[59]构建了一种针对中国股票市场建模及预测的 EMD^①神经网络模型,提高了模型对多种目标函数的学习能力,有效提高了预测精度。祝华凤^[60]以上证综指的波浪曲线底部和顶位时间窗口期为对象展开研究,对 1991 年以来的上证综指的特性,对其底部和顶位的时间窗口期利用 RBF 神经网络模型进行分析预测。刘海玥和白艳萍^[61]研究发现相比自回归模型, RBF 模型和 GRNN^②模型不仅预测精度高而且结果稳定。但大量文献尤其是基于统计学习理论的研究结果表明,在实际应用中 ANN 由于受到诸如学习算法收敛速度慢、存在局部极小值、网络结构难以很好地确定等缺陷的影响,制约了其算法性能进一步提高。

而支持向量机基于结构风险最小化原则克服了神经网络的缺陷,针对股票价格的时变性和非线性在实际应用中展现了很好的泛化能力,而且 SVM 可以不用明确解析函数的方式,直接从数据分析的角度出发来构造高维非线性时间序列分析模型,已在经济、金融研究领域得到了广泛的应用^[62-64]。Kim^[65]采用支持向量机回归模型来预测股票价格,实证结果表明与回馈 BP 神经网络预测模型相比,文中所构建的支持向量机回归模型具有更高的预测精确度。Huang 等^[66]采用日本经济平均指数(简称日经 225 指数)证实, SVM 比 BP 神经网络、线性判别分析(linear discriminant analysis, LDA)及二次判别分析(quadratic discriminant analysis, QDA)具有更高的准确率。Henham^[67]在利用遗传算法和显著性分析(significance analysis, SA)方法进行特征选择的基础上,构建 SVM 时序数据的回归预测模型,最终的预测精度和训练速度都有所提高。Ince 和 Trafalis^[68]利用 SVM 回归预测股票价格,实证结果表明相比 MLP^③和 ARIMA 模型, SVM 的预测更加准确,效果更好。Huang 和 Tsai^[69]、Hsu 等^[70]均是通过自组织特征映射(self-organization feature mapping, SOFM)技术对训练时序数据进行同属性数据的聚类来更好地刻画金融时序数据的非线性特征,然后再利用支持向量回归机展开预测。Huang 还在文章中对输入向量事先进行了特征选择工作,以提高模型训练和预测的效率。这种分阶段的预测模型在针对不同的实证数据研究过程中体现出了更高的预测精度,而且也大大缩短了模型训练的时间。Wen 等^[71]通过独立频谱分析将时序数据分解为趋势成分、波动成分和噪声干扰等部分,再用 SVM

① EMD, 即 empirical mode decomposition, 经验模态分解。

② GRNN, 即 generalized regression neural network, 广义回归神经网络。

③ MLP, 即 multi-layer perceptron, 多层神经网络。

对时序进行分析,结果提高了模型预测的准确性。

杨建辉和李龙^[72]把非参数方法 SVR^①与改进的期权定价方法相结合,提出了一种期权价格预测模型,并利用 SVR 来解决应用中的非线性问题,进而减小模型预测的误差。对长虹 CWB1 权证及随机 10 支认购权证日价格数据的实证研究表明,在预测精度方面,非参数方法要优于传统的参数方法,而改进后的期权定价方法比传统的方法更符合实际应用中的具体情况。姚潇和余乐安^[73]提出了一种融合模糊隶属度思想的模糊近似 SVM,降低了奇异点和数据噪声对模型的干扰,能够显著提高信贷风险分类的预测准确性,具有一定的实际应用价值。赛英等^[74]基于国内股指期货相关重要指标,在用遗传算法和粒子群算法对模型不同核函数进行优化的基础上,用支持向量机开展股指期货回归预测研究。最后的实证表明,基于粒子群算法优化的多项式 (polynomial) 核函数支持向量机具有更高的预测精度。陈懿冰等^[75]提出了改进的支持向量机回归模型,将近期的预测错误施加更严厉的惩罚因子 W ,与传统的支持向量机回归模型相比,该模型预测效果更好。耿立艳^[76]构建了一种融合相空间重构技术、最小二乘支持向量机 (least squares support vector machine, LSSVM) 及变加速系数粒子群优化算法的新的波动率预测模型,对上证综指股指收益的波动率预测实证结果表明,该方法获得了较高的波动率预测精度。李坤和谭梦羽^[77]结合小波理论与支持向量机方法,提出了一种小波支持向量机回归的股票预测模型,3 种大盘指数和 13 类不同行业的股票进行测试的结果一致证明,经小波基函数改造的支持向量机核函数预测效果良好。黄润鹏等^[78]以行为金融理论为依据,通过格兰杰 (Granger) 因果关系检验上证综指时间序列数据与情绪倾向时间序列数据之间的相互关系,并在此基础上建立支持向量机模型来预测股票价格的变化,实证结果再次表明融合了微博情绪信息的预测模型在实际预测时能够获得更高的准确性和有效性。

美国数学家 L. Zadeh 于 1965 年创立了模糊 (fuzzy) 数学,旨在精确表述和处理具有模糊性质的对象。该方法已经在金融时序数据的预测研究领域有了广泛而成功的应用,而且大量研究文献表明,基于模糊逻辑的预测模型在很多情形下的预测精确度显著高于传统的方法^[79-81]。模糊时间序列属于模糊数学与时间序列研究相结合的产物,模糊逻辑应用于金融时序数据的预测就会产生模糊金融时序数据,虽然直接结果不是精确的值,但可以利用模糊逻辑表达不确定知识或经

① SVR, 即 support vector regression, 支持向量回归。

验知识的长处,模拟人类的思维模式得出基本的判断,然后再通过相应变换得到精确的预测值指导金融实践。FL 方法的局限性之一是如何生成合适的模糊规则。进化算法与传统的优化算法相比,是一种具有鲁棒性的全局优化方法,善于解决不确定性系统下的高维复杂问题。根据不同的进化方式可以分为遗传算法、进化规划 (evolutionary programming, EP)、遗传规划 (genetic programming, GP) 和进化策略 (evolutionary strategy, ES)。鉴于其出色的优化功能,EA 簇算法已广泛地被应用于金融时序数据的分析预测模型中,而且大多都是在和其他算法相结合以提高智能计算方法的预测精确性和稳定性^[82-84]。

1.2.3 组合预测方法研究现状

实证研究表明,单一模型很难全面反映和捕捉预测对象,尤其是具有复杂运动规律的预测对象的整体变化规律。面对具有高度不确定性的金融系统时序数据,很难找到一个单预测模型能对这一复杂巨系统的高维非线性运动轨迹拟合得非常紧密,并对变动背后的客观规律信息做出确定性的解释。组合预测方法就是通过几种定性或定量方法的组合,综合各单项预测模型擅长处理的信息片段侧重点,并通过集成学习的思想得到新的有价值信息,从提高模型预测结果确定性和稳定性的角度,显著改善预测效果。

Bates 和 Granger^[85]在 *A Quarterly Journal of Operations Research* 上首次系统性地研究了组合预测方法,之后便引起了国际上相关研究人员的重视,1970 年后涌现出了大量的关于组合模型的研究成果。不少学者在 *International Journal of Forecasting* 杂志相继发表了一系列研究专集^[86-88],Armstrong、Granger 和 Clemen 不仅回顾了之前的组合预测模型、方法及研究现状,而且还对进一步研究的方向和重点进行了专门的评价和展望,为学者的后续研究奠定了坚实的基础。20 世纪 90 年代之后,关于组合预测模型的研究已经得到了广泛的应用并取得了丰硕的研究成果。与此同时,国内也涌现出了一大批优秀的研究成果。唐小我和曹长修^[89],唐小我等^[90],谢开贵和周家启^[91]分别对组合预测误差问题、最优组合问题及收敛性等几个基础性问题进行了较为深入的研究,得到了对组合模型结构特征及变权重组合模型系数估计和收敛性研究的一些基本准则和理论证明。基于不同数据的实证研究均证明了文章对于几个问题的认识和设想,为国内组合预测模型研究奠定了一定的基础。王应明从 1997 年开始分别对广义加权多重平均组合