

学校代号_____10731_____

学 号_____172085212013_____

分 类 号_____TP181_____

密 级_____公 开_____



兰州理工大学
LANZHOU UNIVERSITY OF TECHNOLOGY

全日制工程硕士学位论文

基于 XGBoost 的混合模型在 股票预测中的应用研究

学位申请人姓名_____郭元凯_____

培 养 单 位_____计算机与通信学院_____

导师姓名及职称_____王燕 教授_____

学 科 专 业_____软件工程_____

研 究 方 向_____模式识别与人工智能_____

论文提交日期_____2020 年 4 月 13 日_____

学校代号：10731

学 号：172085212013

密 级：公 开

兰州理工大学工程硕士学位论文

基于 XGBoost 的混合模型在股票预测中的应用研究

学位申请人姓名：郭元凯

导师姓名及职称：王燕 教授

培 养 单 位：计算机与通信学院

专 业 名 称：软件工程

论文提交日期：2020 年 4 月 13 日

论文答辩日期：2020 年 5 月 27 日

答辩委员会主席：沈玉琳 研究员

Based on the hybrid model of XGBoost
Applied Research in Stock Forecasting

by

GUO Yuankai

A thesis submitted in partial satisfaction of the

Requirements for the degree of

Master of Engineering

in

Software Engineering

in the

School of Computer and Communication

Lanzhou University of Technology

Supervisor

Professor Wang Yan

May, 2020

兰州理工大学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的
研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或
集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均
已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：郭元凯

日期：2020年 5 月 29 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保
留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。
本人授权兰州理工大学可以将本学位论文的全部或部分内容编入有关数据库进行
检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密□，在___年解密后适用本授权书。

2、不保密☒。

(请在以上相应方框内打“√”)

作者签名：郭元凯

日期：2020 年 5 月 29 日

导师签名：王亚

日期：2020 年 6 月 3 日

目录

摘 要.....	I
Abstract.....	II
第 1 章 绪 论.....	1
1.1 课题背景与意义.....	1
1.2 国内外研究现状.....	1
1.3 主要研究内容.....	3
第 2 章 相关概念与理论基础.....	5
2.1 股票预测分析基本理论.....	5
2.1.1 股票预测的理论基础.....	5
2.1.2 股票预测面临的问题.....	6
2.2 集成学习算法.....	7
2.2.1 决策树原理.....	7
2.2.2 CART 原理.....	8
2.2.3 GBDT 原理.....	8
2.2.4 XGBoost 原理.....	10
2.3 时间序列模型.....	11
2.3.1 自回归模型（AR）.....	11
2.3.2 移动平均模型（MA）.....	11
2.3.3 自回归移动平均模型（ARMA）.....	12
2.3.4 差分整合移动平均自回归模型（ARIMA）.....	12
2.4 离散小波变换.....	12
2.5 网格搜索算法.....	13
2.6 本章小结.....	14
第 3 章 基于 GS-XGBoost 模型的股票预测研究.....	15
3.1 引言.....	15
3.2 GS-XGBoost 模型的构建.....	16
3.2.1 GS-XGBoost 模型.....	16
3.2.2 GS-XGBoost 模型流程图.....	16
3.3 实验结果及分析.....	17
3.3.1 实验评价指标.....	17
3.3.2 实验数据.....	18
3.3.3 XGBoost 模型实验结果分析.....	18

3.3.4 GS-XGBoost 模型实验结果分析.....	19
3.3.5 实验结果对比分析.....	20
3.4 本章小结.....	22
第 4 章 基于 DWT-GS-XGBoost 模型的股票预测研究.....	23
4.1 引言.....	23
4.2 DWT-GS-XGBoost 模型的构建.....	23
4.2.1 DWT-GS-XGBoost 模型.....	23
4.2.2 DWT-GS-XGBoost 模型流程图.....	25
4.3 实验结果与分析.....	25
4.3.1 实验数据.....	25
4.3.2 离散小波变换实验分析.....	26
4.3.3 DWT-GS-XGBoost 模型实验分析.....	28
4.3.4 模型实验对比分析.....	29
4.4 本章小结.....	32
第 5 章 基于 DWT-ARIMA-GSXGB 模型的股票预测研究.....	33
5.1 引言.....	33
5.2 DWT-ARIMA-GSXGB 模型的构建.....	33
5.2.1 DWT-ARIMA-GSXGB 模型.....	33
5.2.2 DWT-ARIMA-GSXGB 模型流程图.....	34
5.3 实验研究与分析.....	35
5.3.1 实验数据.....	35
5.3.2 实验评价指标.....	36
5.3.3 DWT-ARIMA-GSXGB 模型实验结果对比与分析.....	36
5.4 本章小结.....	45
总结与展望.....	46
参考文献.....	48
致 谢.....	54
附录 A 攻读硕士学位期间发表的学术论文.....	55

摘 要

股票市场是一个国家经济市场的重要组成部分之一。股票实际上是公司筹集运转资金的最重要方式。随着中国特色社会主义市场经济的发展,不仅投资者在时刻关注股票,普通民众也将其视为投资理财的一种渠道。随着时代的不断进步,人民生活水平日益提高,在解决温饱问题之余,有了可供投资的余财,越来越多的人将目光转向股市投资,为股市发展提供了资金条件。然而在纷繁复杂的股票市场,如何寻找最优股成为亟待解决的问题。这不仅是投资者单方面的困惑,也是股票价格预测领域中学者们所关心的重点。因此,对股票市场预测系统的设计和实施不仅具有深刻的理论意义,而且具有非常重要的使用价值。

近年来由于人工智能的快速发展,推动了机器学习理论的发展,并被广泛的应用于各种实际应用当中,进而刮起了一股机器学习的热潮。本文将以机器学习理论为基础,探究了XGBoost与ARIMA的混合模型在股票预测中的应用。然后在XGBoost与ARIMA模型的基础上,结合机器学习相关理论,提出相应的模型结构改进和优化,并做出相应的模型对比。

本文的主要研究和创新点包含以下几点:

(1) 提出网格搜索算法优化的XGBoost金融预测模型(GS-XGBoost)。首先,根据网格搜索算法的思想,先设定将要选择的参数组合区间,基于Xgboost算法,在参数寻优的过程中,结合网格搜索算法的思想,不断地训练模型,通过评价函数对每个参数组合得到的分类结果进行评价,最终得到最优参数组合,最后将最优参数组合代入Xgboost算法,从而使预测性能得到提升。

(2) 提出离散小波变换与优化的XGBoost算法结合的股价预测模型(DWT-GS-XGBoost)。综合考虑了DWT与XGBoost模型的优点,采用离散小波变换进行数据降噪处理与分解,然后使用网格搜索优化的XGBoost模型对处理后的股票数据集进行训练和预测,并与GS-XGBoost模型预测结果进行对比分析。通过实验预测结果表明DWT-GS-XGBoost模型预测效果优于GS-XGBoost模型。

(3) 提出了一种离散小波变换、ARIMA和优化的XGBoost的混合模型(DWT-ARIMA-GSXGB)来解决股票价格预测问题。实验结果表明,DWT-ARIMA-GSXGB股价预测模型具有较好的拟合能力和泛化能力,极大地改善了单个ARIMA模型或单个XGBoost模型在预测股票价格方面的预测性能。

关键词: 股票预测; XGBoost; ARIMA; 离散小波变换; 网格搜索

Abstract

The stock market is one of the important components of a country's economic market. Stocks are actually the most important way for companies to raise working capital. With the development of the socialist market economy with Chinese characteristics, not only investors are always concerned about stocks, but ordinary people also regard it as a channel for investment and financial management. With the continuous advancement of the times, people's living standards have been increasing. In addition to solving the problem of food and clothing, there is surplus money available for investment. More and more people are turning their attention to stock market investment, which provides financial conditions for the development of the stock market. However, in the complicated stock market, how to find the optimal stock has become an urgent problem to be solved. This is not only a unilateral confusion for investors, but also a focus of scholars in the field of stock price forecasting. Therefore, the design and implementation of the stock market forecasting system not only has profound theoretical significance, but also has very important use value.

In recent years, due to the rapid development of artificial intelligence, the development of machine learning theory has been promoted, and it has been widely used in various practical applications, which has ignited a wave of machine learning. Based on machine learning theory, this paper explores the application of XGBoost and ARIMA hybrid models in stock forecasting. Then based on the XGBoost and ARIMA models, combined with the theory of machine learning, the corresponding model structure improvement and optimization are proposed, and the corresponding model comparison is made. The main research and innovations of this paper include the following:

(1) The XGBoost financial prediction model (GS-XGBoost) optimized by grid search algorithm is proposed. First, according to the idea of grid search algorithm, first set the parameter combination interval to be selected. Based on the Xgboost algorithm, in the process of parameter optimization, combined with the idea of grid search algorithm, the model is continuously trained, and each function is evaluated by the evaluation function. The classification results of each parameter combination are evaluated, and finally the optimal parameter combination is obtained. Finally, the optimal parameter combination is substituted into the Xgboost algorithm, thereby improving the prediction performance.

(2) Propose a stock price prediction model (DWT-GS-XGBoost) combining discrete wavelet transform and optimized XGBoost algorithm. Considering the advantages of DWT

and XGBoost models comprehensively, discrete wavelet transform is used for data denoising and decomposition, and then the XGBoost model optimized by grid search is used to train and predict the processed stock data set, and predict with GS-XGBoost model. The results were compared and analyzed. The experimental prediction results show that the prediction effect of DWT-GS-XGBoost model is better than that of GS-XGBoost model.

(3) A hybrid model of discrete wavelet transform, ARIMA and optimized XGBoost (DWT-ARIMA-GSXGB) is proposed to solve the stock price prediction problem. The experimental results show that the DWT-ARIMA-GSXGB stock price prediction model has good fitting ability and generalization ability, which greatly improves the prediction performance of a single ARIMA model or a single XGBoost model in predicting stock prices.

Keywords: Stock forecasting; XGBoost; ARIMA; Discrete wavelet transform; Grid search

第 1 章 绪 论

1.1 课题背景与意义

股票市场是一个国家经济市场的重要组成部分之一。实际上股票是公司筹集运转资金的最重要方式。随着中国特色社会主义市场经济的发展，不仅投资者在时刻关注股票，普通民众也将其视为投资理财的一种渠道。随着时代的不断进步，人民生活水平日益提高。在解决温饱问题之余，有了可供投资的余财。越来越多的人将目光转向股市投资，为股市发展提供了资金条件。然而在纷繁复杂的股票市场，如何寻找最优股成为亟待解决的问题。这不仅是投资者单方面的困惑，也是股票价格预测领域中学者们所关心的重点。因此，对股票市场预测系统的设计和实施不仅具有深刻的理论意义，而且具有非常重要的使用价值。

1.2 国内外研究现状

股票价格预测是金融市场中的一个重要问题，同时也是一个经典而有趣的话题。由于合理准确的预测有可能产生高额的经济利益，因此吸引许多研究人员参与到股价涨跌预测的研究中。股票价格是一种非常不稳定的时间序列，受多种因素的影响。影响股市的外部因素很多，主要有经济因素，政治因素和公司自身因素三个方面的情况^[1]。自股票市场出现以来，研究人员采用各种方法研究股票价格的波动。从经济角度来看，投资者普遍使用传统的基本面分析，技术分析和演化分析来预测^[2-4]。而这些传统的分析方法过于理论化，不能充分反映数据之间的相关性。随着数理统计的深入和机器学习的广泛应用，越来越多的人将现代预测方法应用于股票预测中。

在 1992 年，Yiu Kuen Tse 比较了基于历史样本方差的朴素方法、指数加权移动平均(EWMA)方法和广义自回归条件异方差(GARCH)模型的预测性能。根据月收益率方差的对比，表明 EWMA 方法在波动性股价预测中是有效的^[5]。2000 年，郑丕谔提出一种基于 RBF 神经网络的股市预测建模方法，采用递阶遗传算法训练 RBF 网络的参数、权重和结构，由预测结果可知，该模型具有很强的学习与泛化能力，且具有很高的应用价值^[6]。2001 年，吴微利用 BP 网络对沪市综合指数涨跌的预测进行初步探讨。结果表明，人工神经网络应用于中国股票市场的预测是可行和有效的^[7]。2003 年，K Kim 将 SVM 用于股票价格指数的预测，并与反向传播神经网络和基于案例的推理方法进行了比较。结果表明，SVM 是一种很有前途的股票市场预测方法^[8]。2004 年，张燕平提出一种改进的覆盖学习算法(简称 SLA),并将该算法应用于金融股市的预测。实验结果表明了 SLA 算法的可行性

和应用前景^[9]。2006 年, Eric F. Oteng-Abayie 利用随机游走(RW)、GARCH(1,1)、EGARCH(1,1)和 TGARCH(1,1)模型对加纳证券交易所的波动(条件方差)进行建模和预测。结果表明,在假设数据集服从正态分布的情况下, GARCH(1,1)模型优于其他模型^[10]。2007 年, K. Miao 提出了一种新的径向基函数神经网络(RBFNN)算法,用于预测股价,并在预测中引入技术指标模型。与传统的 RBFNN 算法相比,新算法可以获得更高的训练效率和更好的预测结果^[11]。同年, Xueshen Sui 采用软阈值去噪模型和 SVM 模型相结合对股票市场趋势进行预测,从而提高股票预测性能^[12]。2009 年, Zhe Liao 提出了一种改进的神经网络随机时效神经网络模型,并将该模型应用于股票指数的波动分析中,表现出了不错的效果^[13]。2013 年, Ticknor 将贝叶斯正则化与人工神经网络相结合进行股价涨跌预测,降低了过度拟合和过度训练的可能性,提高了网络的预测质量和泛化能力,确定了该模型的有效性^[14]。2015 年, LA Laboissiere 利用人工神经网络(ANNs)对巴西配电公司股票收盘价进行实际预测,通过平均绝对误差(MAE)、平均绝对百分比误差(MAPE)和均方根误差(RMSE)计算对神经网络的性能进行评价。表明人工神经网络在时间序列预测中具有比较好的预测性能^[15]。同年, Ariyo 通过构建 ARIMA (2, 1, 0) 股价预测模型,并将其用在股价的短期预测中,发现该模型具有较强潜力^[16]。Mu-Yen Chen 采用一种新的模糊时间序列模型对股票市场价格进行预测。通过与三种不同的支持向量机模型方法相比,该预测模型具有较高的预测精度^[17]。2017 年, Avraam Tsantekidis 采用卷积神经网络(CNNs)的深度学习方法,对大规模高频时间序列进行价格走势预测^[18]。比较表明, CNNs 更适合这类任务。2018 年, Shuheng Wang 对传统的支持向量机模糊预测算法进行了改进,提出模糊支持向量机模型并对股票进行预测,发现该模型具有很好的预测性能^[19]。

根据以上的研究不难发现对于时间序列预测研究中,以往大多采用的是对单一模型的改进与应用。由于股票数据中存在线性部分,同时也存在非线性部分^[20-21],因此采用组合模型进行预测也成为众多学者的热门研究。

自 Bates 和 Granger 的早期工作以来,已经探索了几种组合预测的结构^[22]。Clemen 在该领域进行了全面的书目评论^[23]。Menezes 等为组合预测提供了良好的指导^[24]。他们得出结论,组合预测的问题是实施多标准过程并判断错误规范的属性。2001 年, Lam 等人提出了一种目标规划模型,以获得组合预测模型的最优权重^[25]。2003 年, G.Peter Zhang 利用 ARIMA 和 ANN 模型在线性和非线性建模中的独特优势,构建 ARIMA 和 ANN 相结合的股价预测模型^[26]。结果表明,组合模型可以有效地提高单独使用的两种模型的预测精度。2013 年, Zhang, Yan 提出了一种新的股票指数预测方法,基于小波分析结合自回归综合移动平均(ARIMA)和人工神经网络^[27]。将非平稳股价指数序列分解重构为一个低频信号和多个高频信号;利用 ARIMA 预测模型对近似平稳低频信号进行预测,利用 Elman 神经网络

模型对高频信号进行预测;利用径向基函数(RBF)神经网络对各层的预测结果进行混合,得到最终的预测结果。实例表明,该组合预测模型的预测精度较高。同年,Shuzhen Shi 提出了一种 ARMA、BPNN 和马尔可夫模型相结合的股票价格预测模型^[28]。结果表明,ARMA-BPNN 模型优于单一 ARMA 模型和 BPNN 模型。2017 年,Ye 等提出了一种基于小波分析和 ARIMA-SVR 的股票预测模型^[29]。通过小波分解和小波重构,将股票价格分解为重构部分和误差部分。然后分别用 ARIMA 模型和 SVR 模型对重构部分和误差部分进行预测,并将 ARIMA 模型和 SVR 模型的预测结果结合起来,得到最终预测结果。实验结果表明,与单一预测模型相比,该模型是一种有效的股票价格预测方法,大大提高了预测精度。2018 年,Wang, Chengzhang 采用了新的误差准则和权值更新规则对股价进行预测,结果表明,在股票价格预测方面,Boosting-ANN 模型比其他模型具有更好的预测效果^[30]。2019 年, Kim Taewook 提出特征融合长短期记忆卷积神经网络(LSTM-CNN)模型,研究表明,使用来自相同数据的时间和图像特征的组合,而不是单独使用这些特征,可以有效地降低预测误差^[31]。除此之外,预测方法通常采用经典决策算法进行分类或者回归预测分析^[32-34],该算法能有效提高股票预测结果,但由于检测速率相对较慢,为寻求快速且精确度较高的预测方法一些学者将目标锁定在集成学习算法上^[35-36]。

由于人工智能的快速发展,推动了机器学习理论的发展,并被广泛的应用于各种实际应用当中,进而刮起了一股机器学习的热潮。在机器学习技术日趋成熟以及人工智能日益发展的时代下,本文选择基于机器学习理论为基础,探究了 XGBoost 模型与 ARIMA 模型在股票预测中的应用^[37-39]。在 XGBoost 与 ARIMA 模型的基础上,结合机器学习相关理论,提出相应的模型结构改进和优化,并做出相应的模型对比。

1.3 主要研究内容

通过研究机器学习的理论知识,分析了机器学习中 XGBoost 和 ARIMA 两种模型结构,阐述了 XGBoost 模型的优越性。以 XGBoost 模型为基础,设计和构建了基于 XGBoost 的股票预测模型;然后在 XGBoost 模型的基础上提出了 DWT-GS-XGBoost 股票预测模型;最后再结合差分整合移动平均自回归模型(ARIMA)理论知识,设计出 DWT-ARIMA-GSXGB 股票预测模型。

本文的主要研究和创新点包含以下几点:

(1) 提出网格搜索算法优化的 XGBoost 金融预测模型(GS-XGBoost)。首先,根据网格搜索算法的思想,先设定将要选择的参数组合区间,基于 XGboost 算法,在参数寻优的过程中,结合网格搜索算法的思想,不断地训练模型,通过评价函数对每个参数组合得到的分类结果进行评价,最终得到最优参数组合,最后将最

优参数组合代入 Xgboost 算法，从而使预测性能得到提升。

(2) 提出离散小波变换与优化的 XGBoost 算法结合的股价预测模型 (DWT-GS-XGBoost)。综合考虑了 DWT 与 XGBoost 模型的优点，其中为了降低股票数据集中的噪声，采用在去噪方面表现良好的离散小波变换进行数据降噪处理与分解，然后使用网格搜索优化的 XGBoost 模型对降噪处理后的股票数据集进行训练和预测，并与 GS-XGBoost 模型预测结果进行对比分析。通过实验预测结果表明 DWT-GS-XGBoost 模型预测效果优于 GS-XGBoost 模型。

(3) 提出了一种离散小波变换、ARIMA 和优化的 XGBoost 的混合模型 (DWT-ARIMA-GSXGB)来解决股票价格预测问题。其中混合模型采用离散小波变换将数据集拆分为近似部分和误差部分，ARIMA 模型处理近似部分数据，网格搜索改进的 XGBoost 模型处理误差部分数据。实验结果表明，DWT-ARIMA-GSXGB 股价预测模型具有较好的拟合能力和泛化能力，极大地改善了单个 ARIMA 模型或单个 XGBoost 模型在预测股票价格方面的预测性能。

本文共分为 5 章，结构安排如下：

第 1 章，绪论。主要叙述了股票预测的研究背景与意义；然后对现有的股票预测方法进行简单介绍，阐述了股票预测的研究现状和方法理论；最后列举本文的主要研究内容和章节组织安排。

第 2 章，股票预测、机器学习基础理论介绍。对股票预测、机器学习进行了简单的介绍；然后介绍了 XGBoost 的发展历程和 ARIMA 模型的基础理论；接着介绍了机器学习中常用的一些优化算法；最后介绍了机器学习中常用的模型 XGBoost 和 ARIMA 的基础理论。

第 3 章，基于 GS-XGBoost 模型的股票预测研究，首先介绍了股票数据的预处理工作，然后详细阐述了 GS-XGBoost 模型的构建，最后将 GS-XGBoost 应用到股票收盘价的预测中。

第 4 章，基于 DWT-GS-XGBoost 模型的股票预测研究，提出离散小波变换算法和 GS-XGBoost 模型并做出相应的介绍，将离散小波变换算法与 GS-XGBoost 模型相结合，构建 DWT-GS-XGBoost 模型，并对股票进行预测研究。

第 5 章，基于 DWT-ARIMA-GSXGB 模型的股票预测研究，首先介绍了 ARIMA 模型的理论基础，然后提出 ARIMA 与 DWT-GS-XGBoost 进行结合，得到 DWT-ARIMA-GSXGB 模型，最后将 DWT-ARIMA-GSXGB 模型应用到股票预测研究当中。

总结与展望。对本文的主要研究内容进行了总结归纳，并对下一步的研究工作进行了展望。

第2章 相关概念与理论基础

2.1 股票预测分析基本理论

股票市场是一个国家经济市场的重要组成部分之一。股票实际上是公司筹集运转资金的最重要方式。随着中国特色社会主义市场经济的发展,不仅投资者在时刻关注股票,普通民众也将其视为投资理财的一种渠道。随着时代的不断进步,人民生活水平日益提高。在解决温饱问题之余,有了可供投资的余财。越来越多的人将目光转向股市投资,为股市发展提供了资金条件。然而在纷繁复杂的股票市场,如何寻找最优股成为亟待解决的问题。这不仅是投资者单方面的困惑,也是股票预测领域中学者们所关心的重点。投资者为了避免投资风险的同时得到较大的收益,开始对股价走势的精确分析产生了浓厚的兴趣。前文中我们提到,股票价格是一种非常不稳定的时间序列,受多种因素的影响。主要有经济因素,政治因素和公司自身因素三个方面的情况^[40-42]。自股票市场出现以来,研究人员采用各种方法研究股票价格的波动。从经济角度来看,投资者普遍使用传统的基本面分析,技术分析和演化分析来预测。而这些传统的分析方法过于理论化,不能充分反映数据之间的相关性。随着数理统计的深入和机器学习的广泛应用,越来越多的人将现代预测方法应用于股票预测中,如神经网络预测,决策树预测,支持向量机预测,逻辑回顾预测,深度学习预测等^[43-47]。

2.1.1 股票预测的理论基础

股票价格预测是金融市场中的一个重要问题,同时也是一个经典而有趣的话题。由于合理准确的预测有可能产生高额的经济利益,因此吸引许多研究人员参与到股价涨跌预测的研究中。自股票市场出现以来,研究人员采用各种方法研究股票价格的波动,主要包括传统的基本面分析法、技术面分析法,基于人工智能算法的预测方法等。从经济角度来看,投资者普遍使用传统的基本面分析,技术分析和演化分析来预测。而这些传统的分析方法过于理论化,不能充分反映数据之间的相关性。随着数理统计的深入和机器学习的广泛应用,越来越多的人将人工智能算法应用于股票预测中以提高股票收益。下面对基本面分析法、技术面分析法,基于人工智能算法的预测方法进行简单讲解。

(1) 基本面分析法

基本面分析又称基本分析,主要是对证券价格变动的一般规律进行分析与研究,从而为投资者提供可供参考的科学投资方案。基本面分析可以分为宏观社会经济类、行业类和公司类三个影响因素,即宏观经济分析、行业分析和微观企业

(公司) 分析。根据这三类因素的判断, 投资者能够比较简单且全面的把握证券价格的基本走势, 但是对于短线投资的指导作用比较薄弱, 预测精度相对较低。

(2) 技术面分析法

技术面分析又称技术分析, 主要是借助于各种技术图线根据股票以往的价格和交易量数据的变动, 捕获主要和次要的趋势来进行未来价格走向的预测。与评估证券内在价值的基本面分析不同, 技术分析专注于价格变动图表和各种分析工具来评估证券或商品的优势或劣势。

(3) 人工智能算法的预测方法

人工智能预测方法主要是通过机器学习算法构建选股模型或者股票预测模型来提高股票预测精度, 从而使投资者得到比较好的收益回报。随着数理统计的深入和机器学习的广泛应用, 越来越多的人将现代预测方法应用于股票预测中, 如神经网络预测, 决策树预测, 支持向量机预测, 逻辑回顾预测, 深度学习预测等。

2.1.2 股票预测面临的问题

在股票预测过程中, 由于存在多种复杂因素的影响, 导致预测过程存在一定的困难。因此需要通过行之有效的分析与模型的建立来降低股票投资的风险。目前在股票预测过程中基本面分析法、技术面分析法以及基于人工智能算法的预测方法被人们综合考虑用于股票预测中, 但是在预测过程中还存在许多问题。其中主要包含以下几个方面:

股票数据中存在着较高的噪声^[48]。在股票市场中, 由于股票价格波动会受到一些财团的操纵和一些重大事件的影响, 从而导致其出现很多不在原有轨迹上的数据, 影响股票数据的平滑性。这些通过人为操纵和重大事件影响的因素被人们定义为股票数据中的噪声。由于股票市场受到噪声的影响, 致使股票价格起伏波动, 具有不稳定、非线性的特点。这样的高噪声股票数据中很有可能形成数据冗余, 从而导致股价预测模型出现预测失误、精度降低等特点。因此为了使股票投资达到比较好的收益, 对股票数据中的噪声进行处理, 并通过适合股票的非线性关系的模型对股票进行预测是投资者比较关心的话题, 也是研究人员需要克服的问题。

投资行为的不确定性^[49]。每个投资者都有自己的主观能动性, 在股票投资的过程中需要经过自己的主观判断来确定股票的买进卖出, 然而由于人的主观意识容易受到专业能力、群体效应等因素的影响, 从而导致自己在股票投资时出现不确定性。由于我国证券市场起步比较晚, 形成于 20 世纪 90 年代初期, 经历的时间不长, 投资者的投资心态也不是很成熟, 市场经济体制还有待改善, 从而导致我国股市很容易受到政策影响与财团操控, 很难用一个比较准确的模型进行股票

市场的精确预测。

2.2 集成学习算法

集成学习算法本身不算一种单独的机器学习算法，而是通过构建并结合多个机器学习器来完成学习任务。作为一种比较热门的机器学习算法，拥有较高的准确率。由于基学习器的数量的增加，从而导致了模型的训练过程比较复杂、效率下降等问题。目前常见的集成学习算法主要有 2 种：基于 Bagging 的算法和基于 Boosting 的算法^[50-51]，基于 Bagging 的代表算法有随机森林，而基于 Boosting 的代表算法则有 Adaboost、GBDT、XGBoost 等^[52-54]。

提升算法(Boosting)作为常用的统计学习算法，通过迭代思想，使用一个弱学习器弥补前一个弱学习器的“不足”，来串行地构造一个较强的学习器，从而使目标函数值变得足够的小。基本思想：先赋予每个训练样本相同的概率；然后进行 T 次迭代，每次迭代后，对分类错误的样本进行重采样，使得在下一次的迭代中更加关注这些样本。具体流程如图 2.1 示。

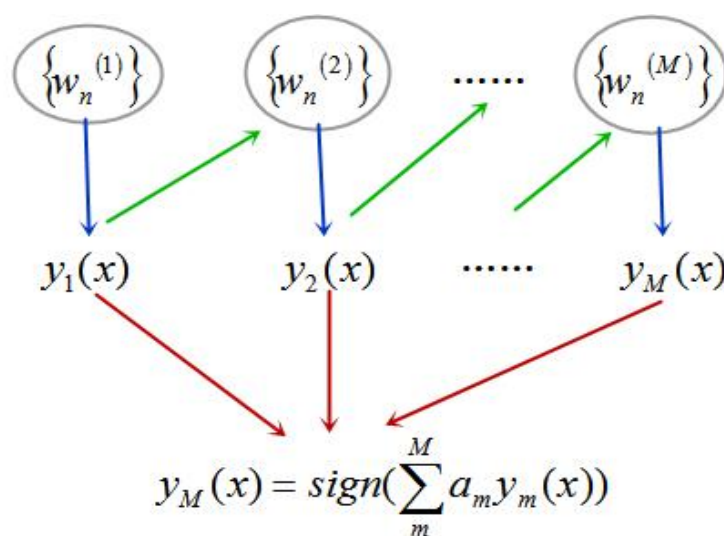


图 2.1 提升算法(Boosting)流程图

2.2.1 决策树原理

决策树是一种分而治之(Divide and Conquer)的决策过程。作为一种树形结构，根据树的分支节点特性，被划分为两个或两个以上较为简单的子集，从结构上划分为不同的子问题。通过递归选择最优特征，根据所选取的特征对训练数据进行分割，从而使每一个子数据集达到一个最优的分类过程，这一过程反应了决策树模型的构建。决策树在构建的过程中，随着树的深度不断增加，分支节点的子集越来越小，当分支节点的深度或者问题的简单程度满足一定的停止规则时，该分

支节点会停止劈分，此为自上而下的停止阈值法，有些决策树也使用自下而上的剪枝法。在机器学习中，决策树作为一种预测模型，被学者广泛应用。

2.2.2 CART 原理

分类回归树(Classification And Regression Tree, CART)模型，由特征选择、树的生成以及剪枝三部分组成，可以用于分类也可以用于回归，作为一种应用广泛的决策树学习方法，在 1984 年由 Breiman 等人提出^[55]。CART 算法通过使用一种二分递归分割的技术，将样本分成两个子样本集，使得生成的非叶子节点都有两个分支。CART 既可以作为分类树，同时也可当作回归树。当 CART 是分类树的时候，采用 GINI 值作为分裂节点的依据，当 CART 作为回归树的时候，使用样本的最小方差作为分裂节点的依据。CART 算法由决策树的生成和决策树的剪枝两部分组成。

决策树的生成就是递归地构建二叉决策树的过程，对回归树用平方误差最小化准则，对分类树用基尼指数最小化准则，进行特征选择，生成二叉树。分类与回归树，是二叉树，可以用于分类，也可以用于回归问题。分类树的输出是样本的类别，回归树的输出是一个实数。CART 回归树是假设树为二叉树，通过不断将特征进行分裂。

CART 回归树产生的目标函数为：

$$\sum_{x_i \in R_m} (y_i - f(x_i))^2 \quad (2.1)$$

为了求解最优的切分特征 j 和最优的切分点 s ，选择变量 x_j 为切分变量，它的取值 s 为切分点，那么就会得到两个区域 R_1 和 R_2 ，如公式 (2.2)：

$$R_1(j, s) = \{x | x^j \leq s\}, R_2(j, s) = \{x | x^j > s\} \quad (2.2)$$

当 j 和 s 固定时，找到两个区域的代表值 c_1, c_2 使各自区间上的平方差最小如公式 (2.3)，目标函数就转化为公式 (2.4)：

$$\min_{j, s} [\min_{c_1} \sum_{x_i \in R_1(j, s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j, s)} (y_i - c_2)^2] \quad (2.3)$$

$$\hat{c}_1 = \text{ave}(y_i | x_i \in R_1(j, s)), \hat{c}_2 = \text{ave}(y_i | x_i \in R_2(j, s)) \quad (2.4)$$

2.2.3 GBDT 原理

GBDT (Gradient Boosting Decision Tree) 梯度提升迭代决策树，是 Boosting 算法的一种。其算法模型采用加法模型，利用前一轮的弱学习器的误差来更新样本权重值，然后一轮一轮的迭代。作为前向分步算法，其基函数必须为 CART 树。

在 GBDT 模型训练的时候，要求模型预测的样本损失尽可能的小。GBDT 模型具体训练流程如图 2.2 所示。

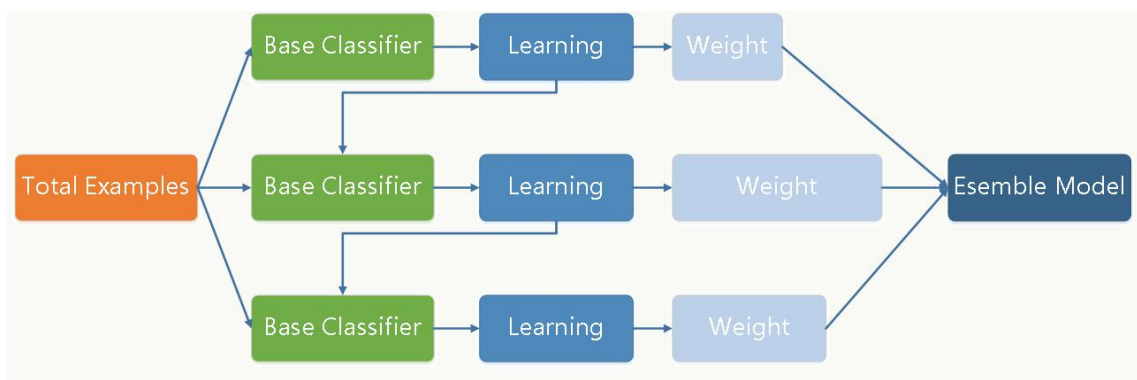


图 2.2 GBDT 模型的训练流程图

GBDT 加法模型表示如公式 (2.5) 所示，

$$f_M(x) = \sum_{m=1}^M T(x; \Theta_m) \quad (2.5)$$

其中， $T(x; \Theta_m)$ 为决策树， Θ_m 为决策树的参数， M 为树的个数。采用前向分步算法，让初始提升树 $f_0(x)=0$ ，则第 m 步如公式 (2.6) 所示，

$$f_m(x) = f_{m-1}(x) + T(x; \Theta_m) \quad (2.6)$$

当残差尽可能小时，最优划分点参数如公式 (2.7) 所示，其中， L 为正则函数。

$$\hat{\Theta}_m = \arg \min_{\Theta_m} \sum_{i=1}^N L(y_i, f_{m-1}(x_i) + T(x; \Theta_m)) \quad (2.7)$$

GBDT 无论是用于分类和回归，采用的都是回归树。图 2.3 展示了 GBDT 模型最终将拟合值转换为概率来进行分类的计算过程。

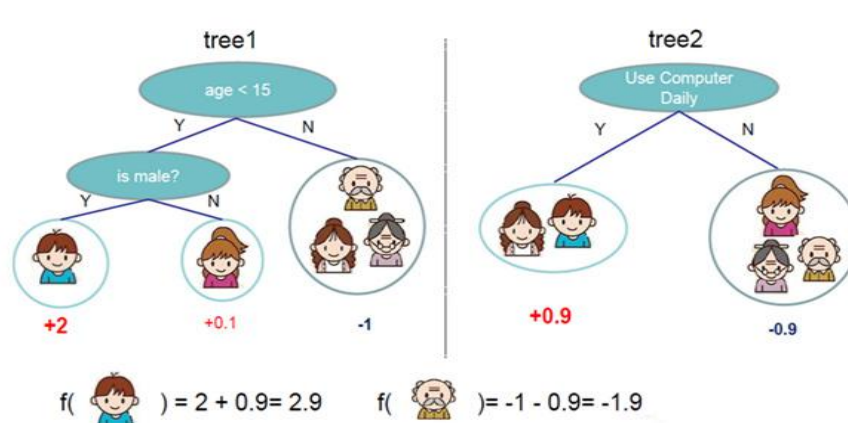


图 2.3 GBDT 模型的分类型概率计算

2.2.4 XGBoost 原理

XGBoost 是一种基于梯度提升决策树的改进算法,可以有效地构建增强树并且并行运行,优化目标函数的价值是 XGBoost 的核心。目标函数,即损失函数,通过加上表示模型复杂度的正则项和最小化损失函数来构建最优模型,且 XGBoost 对应的模型包含了多个 CART 树,因此, XGBoost 模型的目标函数可以表示为公式 (2.8),

$$\hat{y} = \sum_{t=1}^T f_t(x_i), f_t \in F \quad (2.8)$$

其中, T 为树的棵数, F 表示所有可能的 CART 树, f_t 表示一棵具体的 CART 树。

XGBoost 模型的目标函数可以表示为公式 (2.9),

$$\text{Obj}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \quad (2.9)$$

其中, $\hat{y}^{(t-1)}$ 表示保留前面 $t-1$ 轮的模型预测, f_t 为一个新的函数, C 为常数项,将目标函数进行泰勒二阶展开,针对原来的目标函数为了方便进行计算,定义两个变量。如公式 (2.10) 所示,

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)}) \quad (2.10)$$

可以看到这时候的目标函数可以改成公式 (2.11) 的形式。

$$\text{Obj}^{(t)} \approx \sum_{i=1}^n [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + C \quad (2.11)$$

模型训练时,目标函数可以用公式 (2.12) 表示

$$\text{Obj}^{(t)} = \sum_{j=1}^J [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \quad (2.12)$$

定义公式 (2.13),

$$G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i \quad (2.13)$$

将公式 (2.13) 带入公式 (2.12) 中,得到公式 (2.14)

$$\begin{aligned} \text{Obj}^{(t)} &= \sum_{j=1}^J [G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2] + \gamma T \\ &= - \frac{1}{2} \sum_{j=1}^J \frac{G_j^2}{H_j + \lambda} + \gamma T \end{aligned} \quad (2.14)$$

公式 (2.14) 也称为打分函数(scoring function),它是衡量树结构好坏的标准,

值越小，代表这样的结构越好。我们用打分函数选择最佳切分点，从而构建 CART 树。由于打分函数是衡量树结构好坏的标准，因此，可用打分函数来选择最佳切分点。首先确定样本特征的所有切分点，对每一个确定的切分点进行切分，切分好坏的标准如公式（2.15）所示。

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (2.15)$$

$Gain$ 表示单节点 $obj(t)$ 与切分后的两个节点的树 $obj(t)$ 之差，遍历所有特征的切分点，找到最大 $Gain$ 的切分点即是最佳分裂点，根据这种方法继续切分节点，得到 CART 树。

2.3 时间序列模型

时间序列是指将同一统计指标按时间变化且相互关联的数据进行时间顺序排列而形成的数列^[56]。时间序列通常通过多种变化形式的叠加或耦合得到，可以分为主要的四部分，包含：长期趋势变动、季节变动、循环变动和不规则变动。时间序列分析作为数据分析的一个重要领域，根据所研究的依据不同，可有不同的分类。其中，根据研究的对象的多少划分可分为一元时间序列和多元时间序列；根据时间的连续性可分为离散时间序列和连续时间序列；根据序列的统计特性可分为有平稳时间序列和非平稳时间序列；根据时间序列的分布规律可分为高斯型时间序列和非高斯型时间序列。

2.3.1 自回归模型（AR）

AR(Auto regressive)自回归模型，一种线性预测模型^[57]。作为一种处理时间序列的方法，通过时间序列过去时点的线性组合加上白噪声即可预测当前时点，是随机游走的一个简单扩展。根据第 N 点的数据与第 $N-1$ 点的数据的自相关性建立回归方程，并假设它们是线性关系。

AR 模型如公式（2.16）所示，

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + u_t \quad (2.16)$$

其中， y_t 为预测值， u_t 为白噪声， $\varphi_p (p=1,2,3,\dots,p)$ 为自回归系数， p 为自回归阶数。当只有一个时间记录点时，称为一阶自回归过程 AR(1)，如公式（2.17）所示。

$$y_t = \varphi_1 y_{t-1} + u_t \quad (2.17)$$

2.3.2 移动平均模型（MA）

MA 模型(moving average model)称为滑动平均模型，MA 模型和 AR 模型大同小异，与 AR 模型不同的是，MA 模型是通过将一段时间序列中白噪声序列进行

加权并求和得到^[58]。

MA 模型如公式 (2.18) 所示,

$$y_t = \varepsilon_1 \theta_{t-1} + \varepsilon_2 \theta_{t-2} + \cdots + \varepsilon_q \theta_{t-q} + \theta_t \quad (2.18)$$

其中, θ_t 表示不同时间点的白噪声。

2.3.3 自回归移动平均模型 (ARMA)

ARMA(auto regressive moving average)自回归滑动平均模型是由自回归(AR)和滑动平均(MA)模型两部分组成^[59]。

ARMA(p, q)模型如公式 (2.19) 所示,

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} - \varepsilon_1 \theta_{t-1} - \varepsilon_2 \theta_{t-2} - \cdots - \varepsilon_q \theta_{t-q} \quad (2.19)$$

式中前半部分为自回归部分, p 为自回归阶数, φ_p 为自回归系数, 后半部分为滑动平均部分, q 为滑动平均阶数, ε_i 为滑动平均系数; y_t 为消耗股票数据相关序列, θ_q 为随机误差。该模型大致可以分为三个部分: 模型识别、模型定阶和模型检验。

2.3.4 差分整合移动平均自回归模型 (ARIMA)

由 Box 和 Jenkins 开创的 ARIMA 模型在时间序列预测中是最受欢迎的预测方法之一^[60]。其中 ARIMA(p,d,q)称为差分自回归滑动平均模型, AR 是自回归, MA 为滑动平均, p 、 q 分别为对应的阶数, d 为时间序列成为平稳时所做的差分次数。该模型是一个线性回归模型, 适用于跟踪平稳时间序列数据中的线性趋势, 其中时间序列的未来值是根据过去观察的线性函数生成的。ARIMA(p,d,q)模型实质是先对非平稳的股票历史数据 y_t 进行 d 次差分处理得到新的平稳的股票历史序列 X_t , 将 X_t 拟合 ARMA(p,q)模型, 然后再将原 d 次差分还原, 便可以得到 y_t 的预测数据。

根据公式 (2.19), 在经过一次差分后, 形成 ARIMA(1,1,1)模型, 其公式可以表示为:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} - \theta_t - \theta_1 \varepsilon_{t-1} \quad (2.20)$$

2.4 离散小波变换

基于傅里叶变换的频谱分析是用于频域分析的最常用工具^[61]。根据傅里叶理论, 信号可以表示为一系列正弦和余弦的总和。然而, 傅立叶变换的严重限制是它不能提供关于时间的频谱变化的任何信息。小波变换(wavelet transform, WT)类似于傅里叶变换^[62], 是一种新的变换分析方法, 它继承和发展了短时傅立叶变换局部化的思想, 同时又克服了窗口大小不随频率变化等缺点, 能够提供一个随频率改变的"时间-频率"窗口, 通过伸缩平移运算对信号(函数)逐步进行多尺度细

化，最终达到高频处时间细分，低频处频率细分，是进行信号时频分析和处理的理想工具。离散小波变换（DWT）分析金融中的时间序列数据时，可以将原始序列数据分为细节部分数据和近似部分数据^[63-64]，同时保持正交性，进行多分辨率分析，能自动适应时频信号分析的要求，从而可聚焦到信号的任意细节。下面对小波变换的基本理论进行阐述。

给定一个基本函数 $\varphi(t)$ ，如公式（2.21）令

$$\varphi_{a,b}(t) = \frac{1}{\sqrt{a}} \varphi\left(\frac{t-b}{a}\right) \quad (2.21)$$

式中 a 表示移动因子， b 表示位移因子， a, b 均为常数，且 $a>0$ 。 $\varphi_{a,b}(t)$ 是基本函数 $\varphi(t)$ 先做移位再做伸缩得到的。若 a, b 不断地变化，可以得到一簇函数 $\varphi_{a,b}(t)$ ，给定平方可积的信号 $f(t)$ ，即 $f(t) \in L_2(R)$ ，则 $f(t)$ 的小波变换如公式（2.22）所示：

$$\begin{aligned} W_f(a,b) &= \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) \varphi^*\left(\frac{t-b}{a}\right) dt \\ &= \int_{-\infty}^{+\infty} f(t) \varphi_{a,b}^*(t) dt = \langle f, \varphi_{a,b}(t) \rangle \end{aligned} \quad (2.22)$$

式中 $\langle *, * \rangle$ 表示内积， $*$ 表示复数共轭， a, b 和 t 均为连续变量，因此该式也被称为连续小波变换（CWT）^[65]。

在实际应用中，需要对尺度因子 a 和位移因子 b 进行离散化处理，可以取： $a=a_0^m, b=nb_0^m a_0^m$ ，其中 m, n 为整数； a_0, b_0 为大于0的常数； a, b 的选取与小波 $\varphi(t)$ 的具体形式有关。离散小波函数如公式（2.23）表示为：

$$\varphi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} \varphi\left(\frac{t-nb_0a_0^m}{a_0^m}\right) = \frac{1}{\sqrt{a_0^m}} \varphi(a_0^m t - nb_0) \quad (2.23)$$

相应的离散小波变换如公式（2.24）表示为：

$$W_f(m,n) = \langle f, \varphi_{m,n}(t) \rangle = \int_{-\infty}^{+\infty} f(t) \varphi_{m,n}^*(t) dt \quad (2.24)$$

特别的，当 $a_0=2, b_0=1$ 时，离散小波变换称为二进离散小波变换。这种二进离散小波变换简单方便，在实际时间序列处理中被广泛应用。

2.5 网格搜索算法

网格搜索是指定参数值的一种穷举搜索方法，通过将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。将各个参数的可能取值进行排列组合，列出所有的组合结果生成“网格”。然后将各组合参数用于模型训练中，并使用交叉验证对表现进行评估。在拟合函数尝试了所有的参数组合后，返回一个合适的分类器，自动调整至最佳参数组合^[66]。

2.6 本章小结

本章首先对股票预测的相关理论及基础知识进行介绍，详细介绍了集成学习算法和时间序列模型，并对理论算法中存在的问题，给出了相应解决问题的方法或参考文献。最后介绍了离散小波变换与网格搜索算法的相关概念。本章理论知识的介绍为接下来三章模型的提出及求解奠定了基础。

第 3 章 基于 GS-XGBoost 模型的股票预测研究

3.1 引言

随着时代的不断进步，人民生活水平日益提高。在解决温饱问题之余，有了可供投资的余财。越来越多的人将目光转向股市投资，为股市发展提供了资金条件。然而在纷繁复杂的股票市场，如何寻找最优股成为亟待解决的问题。这不仅是投资者单方面的困惑，也是股票价格预测领域中学者们所关心的重点。

股票价格是一种非常不稳定的时间序列，受多种因素的影响。影响股市的外部因素很多，主要有经济因素，政治因素和公司自身因素三个方面的情况。自股票市场出现以来，研究人员采用各种方法研究股票价格的波动。从经济角度来看，投资者普遍使用传统的基本面分析，技术分析和演化分析来预测。而这些传统的分析方法过于理论化，不能充分反映数据之间的相关性。随着数理统计的深入和机器学习的广泛应用，越来越多的人将现代预测方法应用于股票预测中，如神经网络预测，决策树预测，支持向量机预测，逻辑回归预测，深度学习预测等。文献^[67]采用一种基于卷积神经网络（CNN）的深度学习集成算法，通过与多层神经网络和支持向量机进行比较，说明合适的神经网络模型对股票价格的预测精度有一定的提高效果。文献^[68]采用支持向量机（SVM）对股票买卖点进行预测，说明了 SVM 在股票预测中具有较好的表现。文献^[69]采用 Adaboost 集成算法对股票收益进行预测，体现了机器学习算法在股票预测中具有很好的预测性能。文献^[70]采用贡献度与相关分析相结合的方法，利用梯度增强决策树(GBDT)对股票走势进行预测。结果表明，GBDT 组合模型在预测精度上优于线性回归组合模型和随机森林组合模型。

XGBoost 是由 Tianqi Chen 在 2016 年提出来，并在文献^[37]中证明了其模型的计算复杂度低，运行速度快，准确度高等特点。XGBoost 是 GBDT 的高效实现。在分析时间序列数据时，GBDT 虽然能有效提高股票预测结果，但由于检测速率相对较慢，为寻求快速且精确度较高的预测方法，采用 XGBoost 模型进行股票预测，在提高预测精度同时也提高预测速率。本章将 XGBoost 模型与网格搜索相结合，发挥各自的优势，提高股票价格预测的准确性。实验中采用封装及其简便的 sklearn 框架下搭建 XGBoost 网络模型，然后利用网格搜索对其进行优化，构建网格搜索优化的 XGBoost 模型（本章称为 GS-XGBoost），利用改进的 XGBoost 网络模型对中国平安、中国建筑、中国中车、科大讯飞和三一重工股票历史数据的收盘价进行分析预测，将真实值和预测值进行对比，最后通过中国平安股票的

预测结果来评判 GS-XGBoost 模型对股价预测的效果。

3.2 GS-XGBoost 模型的构建

3.2.1 GS-XGBoost 模型

网格搜索是指定参数值的一种穷举搜索方法，通过将估计函数 XGBoost 的参数用交叉验证的方法进行筛选来得到最优的学习算法。网格搜索在进行 XGBoost 参数优化的过程中将各个参数的可能取值进行排列组合，并列出所有参数的组合结果生成“网格”。然后将各个组合参数用于 XGBoost 模型训练，并使用交叉验证对模型的表现进行评估。在拟合函数尝试了所有的参数组合后，自动调整至最佳参数组合，最终返回一个适合该数据集的分类器。为解决参数值随机选取的不确定性，本章构建了 GS-XGBoost 金融预测模型。首先，根据网格搜索算法的思想，先设定将要选择的参数组合区间，基于 XGboost 算法，在参数寻优的过程中，结合网格搜索算法的思想，不断地训练模型，通过评价函数对每个参数组合得到的分类结果进行评价，最终得到最优参数组合，最后将最优参数组合代入 XGboost 算法，从而使预测性能得到提升。在构建好 GS-XGBoost 模型后，进行多步预测，将该模型应用于股票连续 30 天的收盘价的预测中，然后将预测结果分别与原始 XGBoost 模型、GBDT 模型和 SVM 模型进行比较，最后根据模型评价指标进行验证^[71]。

3.2.2 GS-XGBoost 模型流程图

GS-XGBoost 模型具体步骤如下：

（1）获取股票的历史数据，进行缺失值处理，并将数据集分为训练集和测试集；

（2）构建 XGBoost 预测模型；

（3）使用网格搜索算法对 XGBoost 预测模型进行优化。首先将 XGBoost 预测模型各个参数的可能取值进行排列组合，并列出所有参数的组合结果，然后将各个组合参数用于 XGBoost 模型训练，并使用交叉验证对模型的表现进行评估。构建 GS-XGBoost 预测模型，并使用训练集对 GS-XGBoost 模型训练进行训练；

（4）然后使用测试集对 GS-XGBoost 预测模型进行测试；

（5）对比真实结果与预测结果之间的差异。

具体实验流程图如图 3.1 所示。

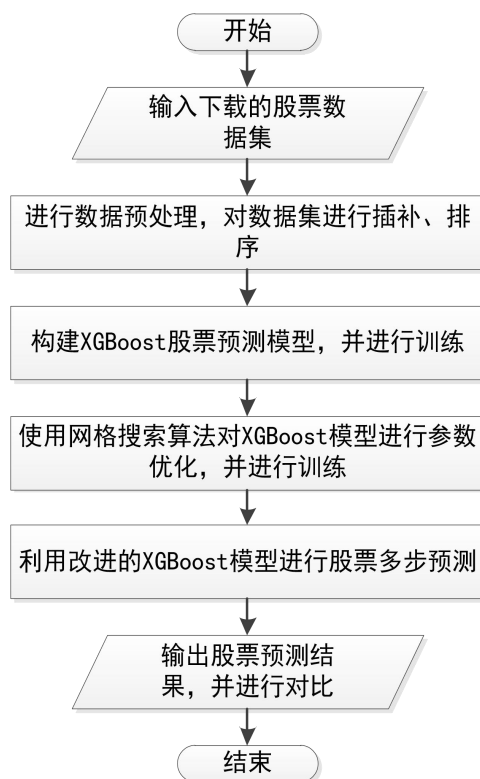


图 3.1 实验流程图

3.3 实验结果及分析

3.3.1 实验评价指标

股票预测模型的预测性能评价指标采用均方误差（MSE）、均方根误差（RMSE）、和平均绝对误差（MAE）三个评价指标对实验结果进行对比。

（1）均方误差是线性回归模型拟合过程中，最小化误差平方和（SSE）与代价函数的平均值。预测效果越好，值越接近于 0，反之，值越远离 0，其计算公式如公式（3.1）所示。

$$MSE = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 \quad (3.1)$$

式中， y 为预测的真实值， \hat{y} 为预测值。

（2）均方根误差计算公式如公式（3.2）所示。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2} \quad (3.2)$$

（3）平均绝对误差计算公式如公式（3.3）所示。

$$MAE = \frac{1}{n} \sum_{i=1}^n |y^{(i)} - \hat{y}^{(i)}| \quad (3.3)$$

3.3.2 实验数据

本章实验在英特尔 I7 3.1GHz 双核四线程 CPU，4G RAM，Windows8 操作系统的计算机进行，仿真平台为 pycharm，使用 python 语言进行编程，分别用到了 python 中的 sklearn、pandas、numpy、Tushare 等包。选取中国平安作为预测对象，利用 python 自带的 Tushare 包，下载中国平安 2005 年 1 月 4 日至 2018 年 12 月 28 日中的开盘价、收盘价、最高价、最低价、交易量等时序数据，共 3253 条。中国平安的收盘价涨跌图如图 3.2 所示。

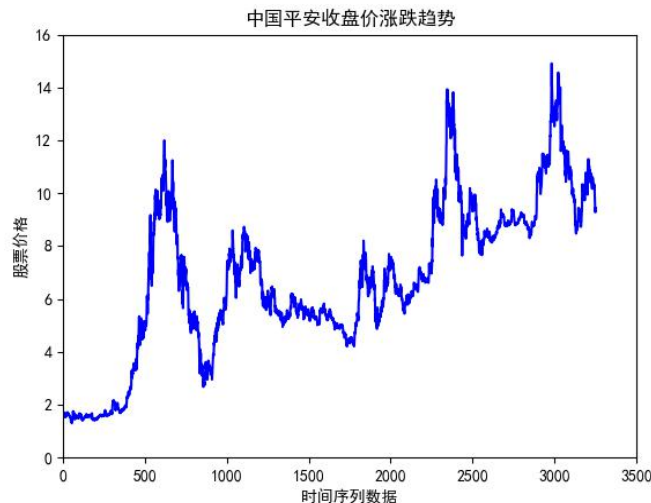


图 3.2 中国平安收盘价涨跌图

3.3.3 XGBoost 模型实验结果分析

XGBoost 模型在构建之前，为了防止原始数据可能存在乱序及缺值的情况，首先对数据集进行插值、排序等操作^[17]，从而得到规整的股票时序数据，进一步构建完整有效的数据集。在搭建完成模型后，对于 3253 条中国平安的股票数据进行数据集拆分，将前 3223 条数据作为训练集，最后 30 条数据作为测试集。测试 XGBoost 模型的预测性能，使用中国平安数据集的训练集训练 XGBoost 模型，将训练过后的 XGBoost 模型测试中国平安数据集的测试集，计算中国平安收盘价预测的均方误差（MSE）、均方根误差（RMSE）和平均绝对误差（MAE）。

在实验中经过多次调试与测试，在权衡计算量与模型的综合得分后，为了与本章提出的 GS-XGBoost 模型进行对比，以展示出 GS-XGBoost 模型自动寻优高于随机选取参数的精度与效率，故将 XGBoost 模型中比较重要的几个参数进行随机设定，其中学习率 `learning_rate` 设置为 0.1，树的深度 `max_depth` 设置为 2，树的棵树 `n_estimators` 设置为 45，最小叶子权重 `min_child_weight` 设置为 4，其余参数都设置为默认参数。在实验一中主要探索 XGBoost 模型对短期股价的预测性能。其实验结果如表 3.1 和图 3.3 所示。

表 3.1 XGBoost 模型的预测结果

模型	MSE	RMSE	MAE
XGBoost	0.0170	0.1305	0.1157

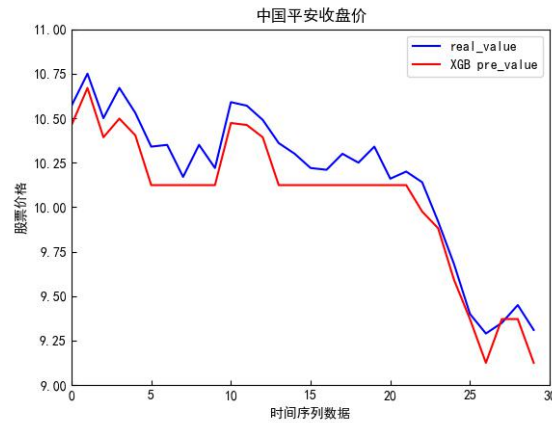


图 3.3 XGBoost 收盘价预测结果图

由表 3.1 和图 3.3 可以看出 XGBoost 模型在预测中的均方误差为 0.017，均方根误差为 0.1305，平均绝对误差为 0.1157。发现该模型的预测效果不太理想，图 3.3 中可以清楚的看到预测效果虽然在趋势上有所接近实际股价趋势，但是总体上股价预测值普遍低于实际值，因此在股价短期预测中该模型还需要改进。

3.3.4 GS-XGBoost 模型实验结果分析

结合网格搜索算法与 XGBoost 模型，通过采用网格搜索算法寻得 XGBoost 模型的最优解，构建 GS-XGBoost 模型。其中 XGBoost 模型优化后的参数为 $\text{learning_rate}=0.1$, $\text{n_estimators}=74$, $\text{max_depth}=15$, $\text{min_child_weight}=4$, $\text{gamma}=0$, $\text{subsample}=0.8$, $\text{reg_alpha}=0.1$, $\text{colsample_bytree}=0.75$ 。本实验采用与 XGBoost 模型实验一致的测试集和训练集，在模型进行训练和测试后，与 XGBoost 模型实验结果进行对比分析，本实验预测结果如图 3.4 和表 3.2 所示。

表 3.2 GS-XGBoost 模型的预测结果

模型	MSE	RMSE	MAE
GS-XGB	0.0007	0.0268	0.0212

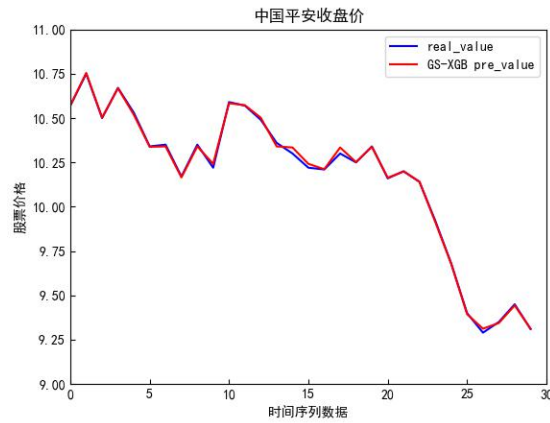


图 3.4 GS-XGBoost 收盘价预测结果图

3.3.5 实验结果对比分析

从 GS-XGBoost 模型的实验结果可知, GS-XGBoost 模型的预测性能要明显高于 XGBoost 模型的预测性能, 图 3.4 显示出中国平安 30 天收盘价的预测值与实际值的拟合度较高。对比表 3.1 和表 3.2 可知 MSE、RMSE 与 MAE 的值分别减少了 0.0163、0.1037、0.0945。对比发现本章提出的 GS-XGBoost 模型在短时股价预测中比 XGBoost 模型具有更高的拟合度。模型预测效果更好。

为了更进一步验证 GS-XGBoost 模型在短时股价预测中的有效性, 分别与同类型的 GBDT 模型比较, 同时与在股价预测中表现良好的支持向量机模型进行对比。实验数据采用与 XGBoost 模型一致的测试集和训练集, 同时为了验证模型的泛化能力和预测性能, 分别将该模型应用于中国建筑、中国中车、科大讯飞和三一重工连续 30 天的收盘价预测中, 其中 4 支个股的数据集为 2005 年 1 月 4 日至 2018 年 12 月 28 日中的开盘价、收盘价、最高价、最低价、交易量等时序数据, 在模型进行训练和测试后, 对实验结果进行对比分析。其实验对比结果如图 3.5~3.9 和表 3.3~3.7 所示。

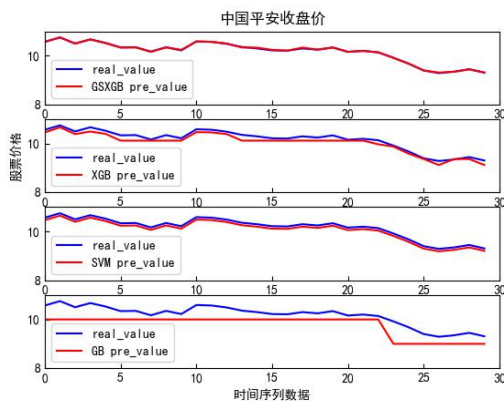


图 3.5 GS-XGBoost 与 GB、SVM 的对比

表 3.3 中国平安预测结果对比

模型	MSE	RMSE	MAE
GS-XGBoost	0.0007	0.0268	0.0212
XGBoost	0.0170	0.1305	0.1157
SVM	0.0099	0.9984	0.9984
GB	0.3978	0.6307	0.5396

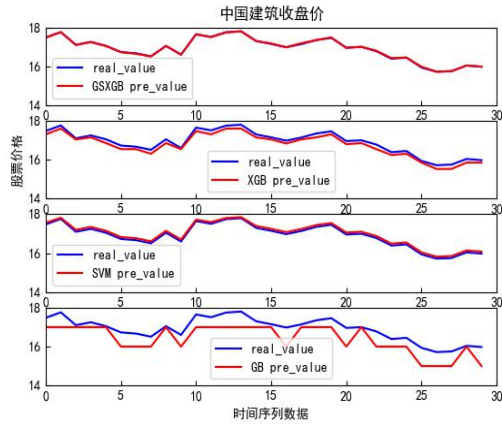


图 3.6 GS-XGBoost 与 GB、SVM 的对比

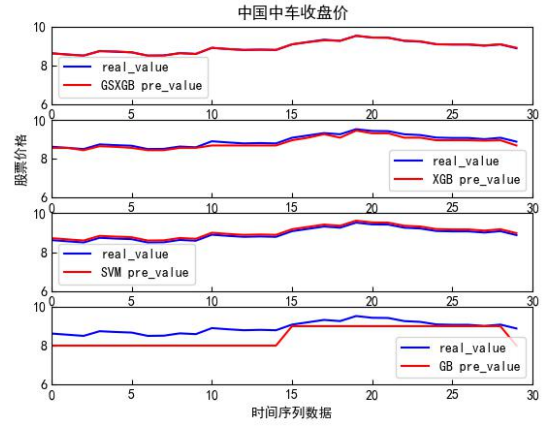


图 3.7 GS-XGB 模型与 GB、SVM 的对比

表 3.4 中国建筑预测结果对比

模型	MSE	RMSE	MAE
GS-XGBoost	0.0013	0.0368	0.0145
XGBoost	0.0259	0.1612	0.1546
SVM	0.0088	0.0938	0.0928
GB	0.3537	0.5947	0.5101

表 3.5 中国中车预测结果对比

模型	MSE	RMSE	MAE
GS-XGBoost	0.0001	0.0102	0.0069
XGBoost	0.0135	0.1164	0.1063
SVM	0.0099	0.0997	0.0997
GB	0.2958	0.5438	0.4687

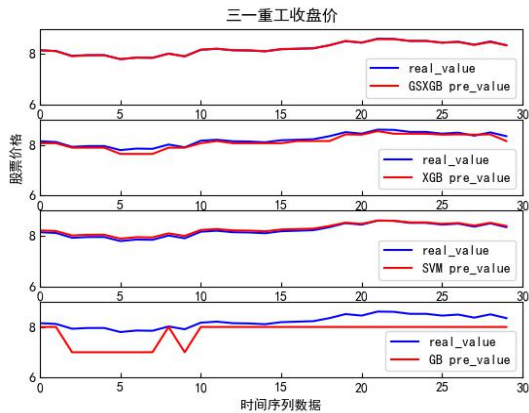


图 3.8 GS-XGB 与 GB、SVM 的对比

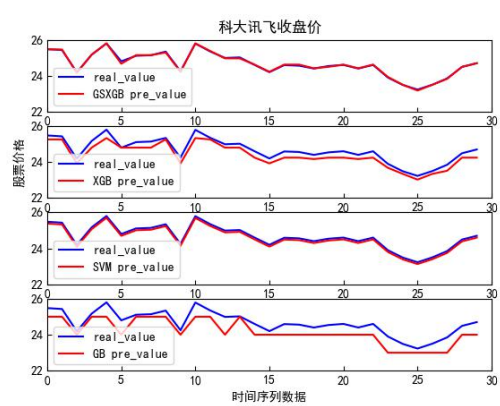


图 3.9 GS-XGB 模型与 GB、SVM 的对比

表 3.6 科大讯飞预测结果对比

模型	MSE	RMSE	MAE
GS-XGBoost	0.0013	0.0361	0.0238
XGBoost	0.0829	0.2879	0.2664
SVM	0.0099	0.0999	0.0999
GB	0.2970	0.5450	0.4841

表 3.7 三一重工预测结果对比

模型	MSE	RMSE	MAE
GS-XGBoost	0.0001	0.0124	0.0087
XGBoost	0.0107	0.1036	0.0887
SVM	0.0043	0.0662	0.0594
GB	0.2915	0.5399	0.4564

从图 3.5~3.9 中可知，GS-XGBoost 模型在预测值与实际值的拟合度上表现较好。其预测性能明显高于 GBDT 模型的预测性能，与 SVM 模型进行比较也具有相对优势。表 3.3~3.7 显示 GS-XGBoost 模型在 MSE、RMSE、MAE 都具有出色的表现。

经过以上对比，说明 GS-XGBoost 模型在短期股价预测中比同类型的 GBDT 模型和股价预测中表现良好的支持向量机模型的预测性能都要高。从而验证本章提出的 GS-XGBoost 在短期股价预测中的可行性，以及其出色的预测性能。

3.4 本章小结

本章提出了一种基于网格搜索算法改进的 XGBoost 的金融时间序列模型，即 GS-XGBoost 模型。利用网格搜索算法对 XGBoost 模型进行参数优化来提高 XGBoost 模型在股票短期预测中的拟合度，采用中国平安、中国建筑、中国中车、科大讯飞和三一重工的收盘价对该模型进行验证。通过评价指标均方误差 MSE、均方根误差 RMSE 和平均绝对误差 MAE 的对比，发现本章提出的模型在短期股价多步预测上具有较高的拟合度。给股民如何掌握股票的总体趋势带来更有价值的参考。

第 4 章基于 DWT-GS-XGBoost 模型的股票预测研究

4.1 引言

虽然第三章中提出的 GS-XGBoost 模型在一些股票预测上具有较高的拟合度，但是并没有考虑到股票数据中存在的噪声影响，为了更进一步提高模型的泛化能力，除了改进算法外，同时还需要考虑股票数据中存在的噪声。为了去除可能对系统性能产生负面影响的不相关数据样本，同时仍然保留数据的主要结构，需要对股票数据集进行降噪处理，提高数据的平滑性，从而使模型的预测性能得到提高。本章在 XGBoost 模型对股价涨跌预测研究的基础上，为提高股价预测的拟合度，在原有模型基础之上结合离散小波分解和网格搜索算法对 XGBoost 模型进行优化改进，提出离散小波变换与优化的 XGBoost 算法结合的短期股价预测模型（DWT-GS-XGBoost），并对中国石油、中国建筑、中国中车、上汽集团和三一重工开盘价进行股价预测。将真实值和预测值进行对比来评判 DWT-GS-XGBoost 模型的股价预测效果，从而验证该模型在短期股价预测中具有较好的预测性能。

4.2 DWT-GS-XGBoost 模型的构建

4.2.1 DWT-GS-XGBoost 模型

数据去噪的主要目的是尽可能地降低噪声并恢复原始数据，尽可能少地丢失重要信息^[72]。由于本章选取的股票指数数据存在一定的噪声影响，并且在对股票历史数据集进行训练时数据噪声会影响模型预测性能，为了减少噪声影响，提高模型预测性能，本章采用 DWT 对其进行降噪处理。使用公式（2.24）将各股票开盘价指数的数据集分为低频部分（趋势）和高频部分（细节），同时保持正交性，进行多分辨率分析，自动适应时频信号分析的要求，从而可聚焦到信号的任意细节^[73]。采用 DWT 将原始数据集分为低频部分和高频部分，具体分解过程如图 4.1 所示，其中 cD1 和 cA1 是小波单尺度分解后的低频部分和高频部分，cD2 和 cA2 是从 cA1 小波分解出来的低频和高频部分。

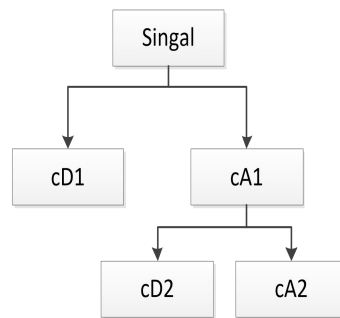


图 4.1 小波变换示意图

由于小波变换中基函数和分解层数对数据集的降噪与分解有很大的影响，故对小波变换基函数与分解层数要进行确定。DWT 最佳小波基的选择取决于要分析的原始信号的特征和期望的分析目标，目前对 DWT 最佳小波基的选择与最佳分解层数的确定还没有具体的寻找方法，为了避免最佳小波基函数的遗漏，本章采用穷举的方法进行小波基函数的选取^[74]。一般来说，对于小波函数的选择和分解层数的确定通常都在 3 至 5 个左右。因此，本章将在分解层数为 3 至 5 层的情况下，测试 N 分别为 1 至 8 的 dbN 离散小波函数构建模型的效果，经过对比后选择较优的小波函数和分解层数。

在确定好小波基函数与分解层数后，采用 XGBoost 模型对离散小波变换分解后的低频部分进行训练和预测。由于 XGBoost 的参数影响着 XGBoost 算法的性能，为解决 XGBoost 算法最优参数难寻这一问题，本章采用网格搜索算法（GS）进行参数优化。GS 是指定参数值的一种穷举搜索方法，通过将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。将各个参数的可能取值进行排列组合，然后将各组合参数用于 XGBoost 训练，并使用交叉验证对表现进行评估。在拟合函数尝试了所有的参数组合后，返回一个合适的分类器，自动调整至最佳参数组合。为解决参数值随机选取的不确定性，首先根据网格搜索算法的思想，先设定将要选择的参数组合区间，基于 XGBoost 算法，在参数寻优的过程中，结合网格搜索算法的思想，不断地训练模型，通过评价函数对每个参数组合得到的分类结果进行评价，最终得到最优参数组合，最后将最优参数组合代入 XGBoost 算法，提高模型的预测性能。

综合考虑了 DWT 与 XGBoost 模型的优点，其中为了降低股票数据集中的噪声，采用在去噪方面表现良好的离散小波变换进行数据降噪处理与分解，然后使用网格搜索优化的 XGBoost 模型对降噪处理后的股票数据集进行训练和预测。构建离散小波变换与优化的 XGBoost 算法结合的股价预测模型（DWT-GS-XGBoost），将该模型应用于中国石油、中国建筑、中国中车、上汽集团和三一重工五只股票的开盘价预测当中，最后根据评价指标来评判该模型的性能。

4.2.2 DWT-GS-XGBoost 模型流程图

DWT-GS-XGBoost 模型构建过程如下：

- (1) 获取股票指数的历史数据，并进行缺失值处理；
- (2) 采用离散小波变换对股票历史数据进行降噪与分解，通过穷举的方法分别得到最优小波基函数与分解层数；
- (3) 采用 sklearn 包中的 XGB 方法，实现 XGBoost 算法，对股票历史数据集进行训练，构建 XGBoost 股价预测模型；
- (4) 采用 XGBoost 股价预测模型对小波分解后的数据集进行预测，其中 XGBoost 模型采用网格搜索算法进行参数优化，然后构建 DWT 与网格搜索优化的 XGBoost 算法结合的股价预测模型，进行股价预测；
- (5) 将预测结果进行小波重构，重构后的结果为最终预测结果。
- (6) 对比真实值与预测值之间的差异，判断 DWT-GS-XGBoost 模型的预测性能。

具体模型构建过程图如图 4.2 所示：

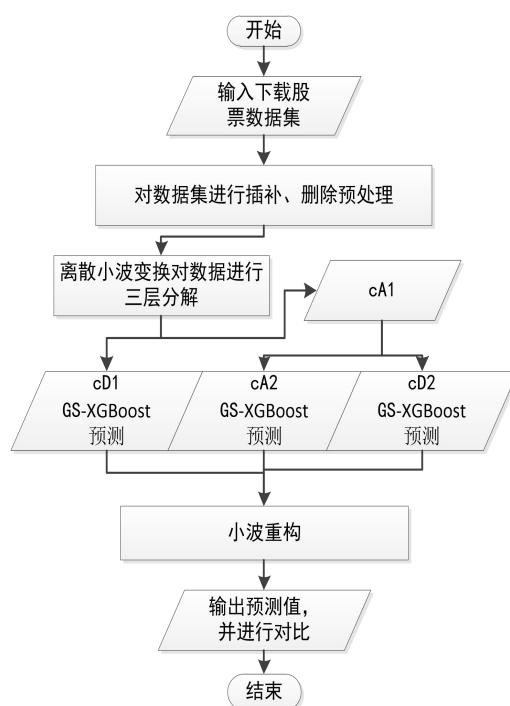


图 4.2 模型构建过程图

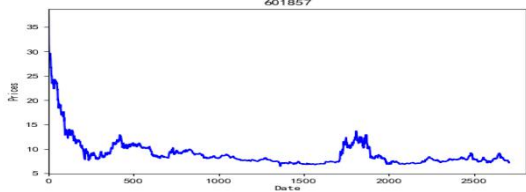
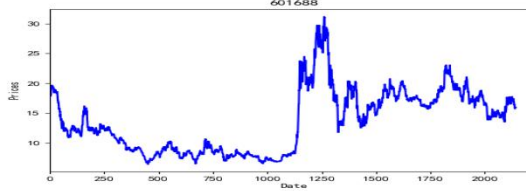
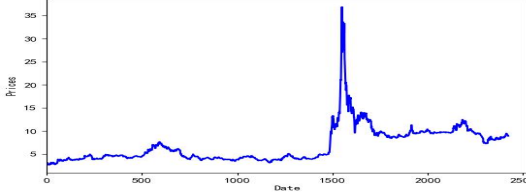
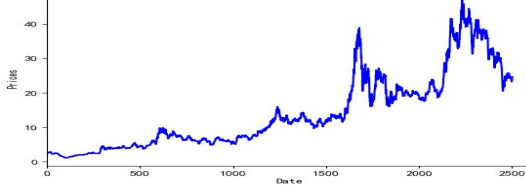
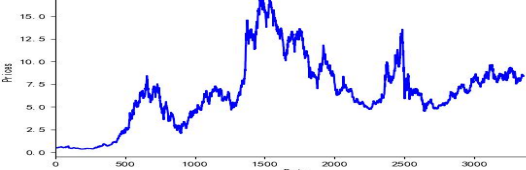
4.3 实验结果与分析

4.3.1 实验数据

本章实验在英特尔 I7 3.1GHz 双核四线程 CPU，4G RAM，Windows8 操作系统的计算机进行，仿真平台为 pycharm，使用 python 语言进行编程，分别用到

了 python 中的 sklearn、pandas、numpy、Tushare、pywt 等包。选取中国石油、中国建筑、中国中车、上汽集团和三一重工的开盘价作为预测对象，利用 python 自带的 Tushare 包，下载五只股票 2005 年 1 月至 2018 年 12 月的开盘价、收盘价、最高价、最低价、交易量等时序数据。其中各个股票开盘价的涨跌图如表 4.1 所示。

表 4.1 五只股票开盘价涨跌图

股票	开盘价
中国石油（601857）	
中国建筑（601688）	
中国中车（601766）	
上汽集团（600104）	
三一重工（600031）	

4.3.2 离散小波变换实验分析

为测试离散小波变换的降噪性能，寻找离散小波变换的最优分解层数与小波函数。使用中国石油的开盘价进行离散小波分解，计算中国石油开盘价分解后的 RMSE。在实验中经过多次调试与测试，首先对数据集进行插值、排序等处理，然后对数据集进行离散小波变换，分别将离散小波变换的分解层数设定在在 3 至

5 个左右，小波基函数 dbN 中的 N 设置为 1 至 8，采用网格搜索的方法对小波变换的基函数与分解层数进行寻优。不同小波基函数和分解层数下模型的拟合误差（RMSE）和预测误差（RMSE）如图 4.3 和图 4.4 所示。

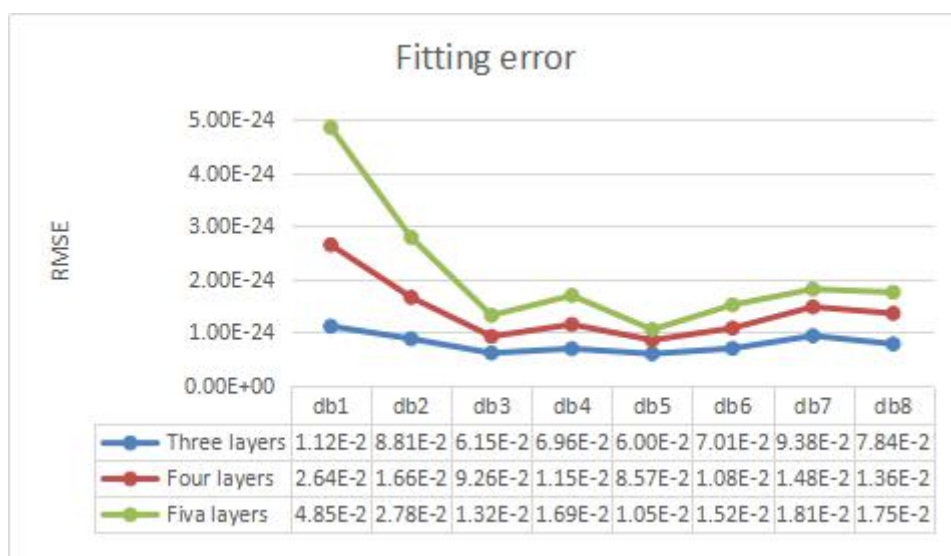


图 4-3 不同小波函数和分解层数下模型的拟合误差对比

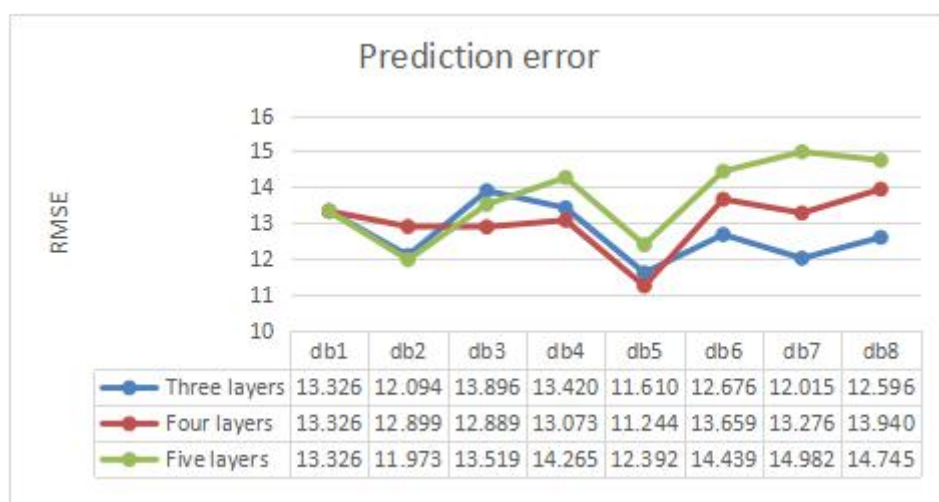


图 4.4 不同小波函数和分解层数下模型的预测误差对比

由图 4.3 可以看出，当分解层数为 3 层时，模型的拟合误差相对最小，拟合效果优于 4 层分解和 5 层分解；同时，当小波函数为 db5 时，预测误差最小，为 11.24。由图 4.4 可知，当小波函数为 4 层分解时，模型的预测误差最低，但考虑到此时拟合误差较高，因此最终选用的组合预测模型中小波函数为 db5 的三层分解模型。以中国石油开盘价为例，采用小波函数为 db5 的三层分解的离散小波变换进行分解，分解结果如图 4.5 所示。其中 A1 序列表示经过 db5 小波函数三层

分解后的逼近信号，即股指的整体趋势，而 D1 和 D2 表示投资者的交易行为对中国石油开盘价的影响。

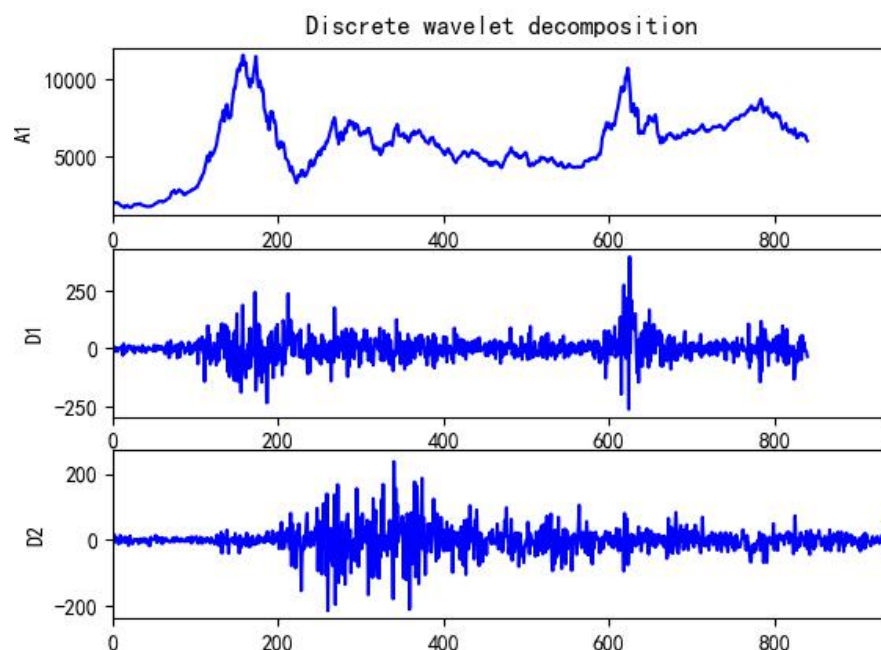


图 4.5 db5 小波函数的三层分解

4.3.3 DWT-GS-XGBoost 模型实验分析

测试 XGBoost 模型的预测性能，分别使用五只股票开盘价的训练集训练 XGBoost 模型，将训练过后的 XGBoost 模型测试各个数据集的测试集，计算五只股票开盘价预测的均方根误差（RMSE）、平均绝对误差（MAE）和决定系数 R-Square(R^2)。

XGBoost 模型在构建之前，为了防止原始数据可能存在乱序及缺值的情况，首先对数据集进行插值、排序等操作，从而得到规整的股票时序数据，进一步构建完整有效的数据集。对于五只股票数据进行数据集拆分，将每只股票的最后 50 条数据作为测试集，其余作为训练集。在实验中经过多次调试与测试，在权衡计算量与模型的综合得分后将 XGBoost 模型参数都设置为默认参数，最后经过评价指标评判模型的预测性能。其实验结果如表 4.2 所示。

表 4.2 XGBoost 模型预测结果

Stock	中国石油	中国建筑	中国中车	上汽集团	三一重工
RMSE	0.0069	0.0494	0.0100	0.1742	0.0093
MAE	0.0553	0.1749	0.0803	0.3404	0.0741
R-Square	0.9621	0.9621	0.9155	0.9566	0.8949

结合离散小波变换与 XGBoost 模型，根据离散小波变换的拟合和预测结果，

当小波函数为 db5，三层分解的小波时，模型的拟合效果较好，因此构建小波基函数为 db5 的三层分解的离散小波变换与 XGBoost 算法结合的短期股价预测模型。其中，XGBoost 模型的参数采用网格搜索算法进行优化，将 XGBoost 模型和离散小波变换与优化的 XGBoost 算法的组合模型进行性能对比。五只股票开盘价的预测结果如表 4.3 所示。

表 4.3 DWT-GS-XGBoost 模型预测结果

Stock	中国石油	中国建筑	中国中车	上汽集团	三一重工
RMSE	0.0003	0.0012	0.0005	0.0279	0.0002
MAE	0.0151	0.0151	0.0166	0.1208	0.0134
R ²	0.9981	0.9981	0.9955	0.9930	0.9971

对比预测结果，DWT-GS-XGBoost模型的预测性能要明显高于XGBoost模型的预测性能。根据表4.2和表4.3可知，DWT-GS-XGBoost模型在RMSE、MAE和决定系数R²均表现出较好的性能。经过对比发现本章提出的DWT-GS-XGBoost模型在股票预测中比XGBoost模型具有更高的拟合度。模型预测效果更好。

4.3.4 模型实验对比分析

为了进一步验证DWT-GS-XGBoost模型在股票预测中的有效性，将该模型分别与XGBoost原模型、网格搜索优化的XGBoost模型（GS-XGBoost）、SVR模型和GBDT模型进行对比。实验数据采用与实验二一致的测试集和训练集，在模型进行训练和测试后，对实验结果进行对比分析。其实验的对比结果如表4.4~4.8所示。由表4.4~4.8可知，DWT-GS-XGBoost模型在预测值与实际值的拟合度上表现较好。其预测性能明显高于XGBoost模型和GS-XGBoost模型的预测性能，与SVR模型和GBDT模型的预测性能进行比较也具有相对优势。且DWT-GS-XGBoost模型在RMSE、MAE和R²评价指标上都具有较好的表现。

图4.6~4.10展示了五只股票开盘价在XGBoost模型、GS-XGBoost模型、SVR模型、GBDT模型和DWT-GS-XGBoost模型的预测值与真实值的对比结果，发现DWT-GS-XGBoost模型在预测价格上更接近与实际价格，进一步说明DWT-GS-XGBoost模型在股票短时预测中具有较高的拟合度。

经过以上对比，说明 DWT-GS-XGBoost 模型在短时股票预测中比 XGBoost 模型、GS-XGBoost 模型、SVR 模型和 GBDT 模型的预测性能都要高。从而验证本章提出的 DWT-GS-XGBoost 在股票预测中的可行性，以及其出色的预测性能。

表 4.4 中国石油预测结果对比

模型	RMSE	MAE	R2
DWT-GS-XGBoost	0.0003	0.0151	0.9981
GS-XGB	0.0028	0.0381	0.9841
XGBoost	0.0069	0.0553	0.9621
SVR	0.0041	0.0551	0.9771
GBDT	0.0057	0.0504	0.9686

表 4.5 中国建筑预测结果对比

模型	RMSE	MAE	R2
DWT-GS-XGBoost	0.0012	0.0305	0.9957
GS-XGB	0.0631	0.1838	0.7881
XGBoost	0.0494	0.1749	0.8341
SVR	0.0326	0.1399	0.8905
GBDT	0.0461	0.1756	0.8453

表 4.6 中国中车预测结果对比

模型	RMSE	MAE	R2
DWT-GS-XGBoost	0.0005	0.0166	0.9955
GS-XGB	0.0087	0.0774	0.9264
XGBoost	0.0100	0.0803	0.9155
SVR	0.0051	0.0581	0.9562
GBDT	0.0105	0.0802	0.9113

表 4.7 上汽集团预测结果对比

模型	RMSE	MAE	R2
DWT-GS-XGBoost	0.0279	0.1208	0.9930
GS-XGB	0.1680	0.3126	0.9581
XGBoost	0.1742	0.3404	0.9566
SVR	1.0127	0.5102	0.7478
GBDT	0.1930	0.3561	0.9519

表 4.8 三一重工预测结果对比

模型	RMSE	MAE	R2
DWT-GS-XGBoost	0.0002	0.0134	0.9971
GS-XGB	0.0063	0.0608	0.9284
XGBoost	0.0093	0.0741	0.8949
SVR	0.0051	0.0572	0.9422
GBDT	0.0085	0.0734	0.9037

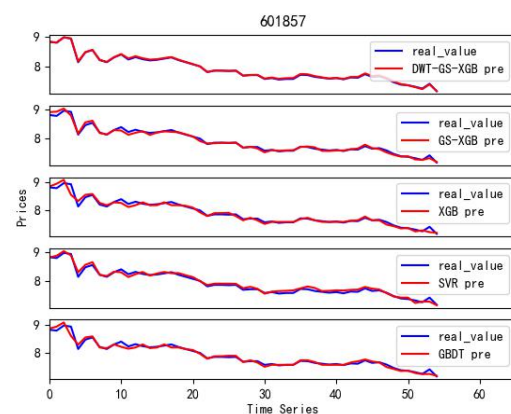


图 4.6 中国石油预测结果

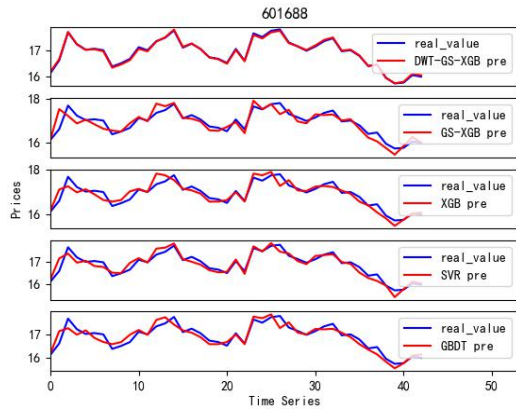


图 4.7 中国建筑预测结果

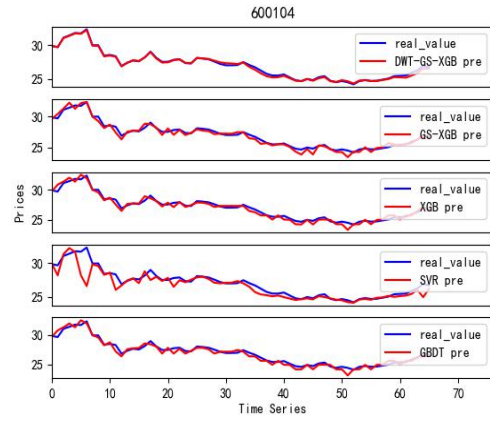


图 4.8 上汽集团预测结果

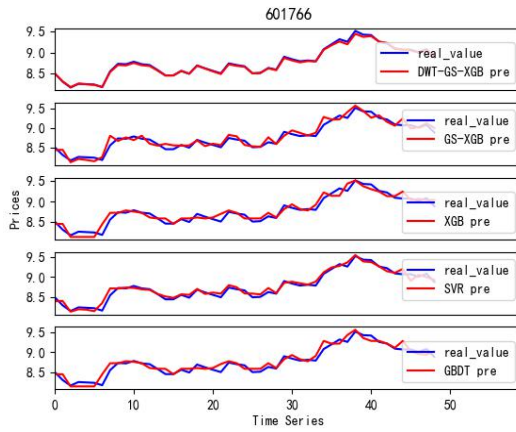


图 4.9 中国中车预测结果

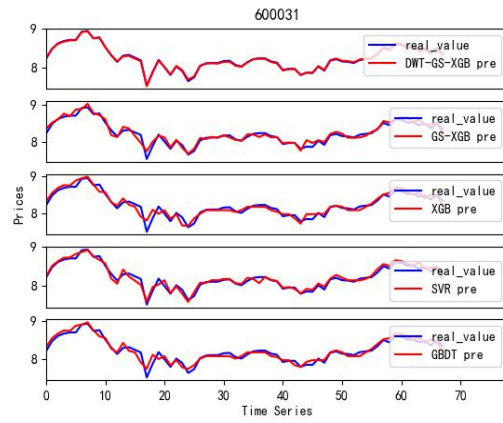


图 4.10 三一重工预测结果

4.4 本章小结

本章采用对数据集具有降噪功能的离散小波变换（DWT）和使用网格搜索算法优化的 XGBoost 模型，即 DWT-GS-XGBoost 模型对短期股价进行预测。利用基函数为 db5 离散小波变换对数据集进行小波三层分解，然后使用 XGBoost 模型对分解后的部分进行训练、预测，最后将预测结果进行小波重构，重构后的结果为最终的预测值，进而提高股价预测的拟合度。采用中国石油、中国建筑、中国中车、上汽集团和三一重工的开盘价对该模型进行验证。通过评价指标均方根误差 RMSE、平均绝对误差 MAE 和决定系数 R^2 的对比，发现本章提出的模型在股价短期预测上具有较好的拟合度。进一步为如何掌握股价的总体涨跌带来更有价值的参考。

第 5 章 基于 DWT-ARIMA-GSXGB 模型的股票预测研究

5.1 引言

早在 1991 年，McQueen, Grant 采用马尔科夫链验证了股价预测的可行性，从而为后续的研究奠定坚实的基础^[75]。在之后的几十年中学者们使用了一些先进的机器学习技术对股价进行预测。目前，已经开发了不同的股票价格预测技术，其中在股价预测模型的构建中，时间序列预测，如自回归整合移动平均线（ARIMA）、自回归条件异方差（ARCH）和广义自回归条件异方差（GARCH）^[76-77]，以及基于机器学习的技术，如神经网络、深度学习、支持向量机（SVM）和决策树都被广泛使用。

由于第四章提出的 DWT-GS-XGBoost 模型在股票预测过程中对噪声部分进行了剔除，而影响到数据的完整性，故本章将噪声部分也加以考虑来提高模型的预测精度与泛化能力，从而提出了一种离散小波变换、ARIMA 和优化的 XGBoost 的混合模型（DWT-ARIMA-GSXGB）来解决股票价格预测问题。提出的混合模型（DWT-ARIMA-GSXGB）采用离散小波变换将数据集拆分为近似部分和误差部分，其中 ARIMA 模型处理近似部分数据，网格搜索改进的 XGBoost 模型（GSXGB）处理误差部分数据。根据 10 只股票数据集的实验对比，发现 DWT-ARIMA-GSXGB 模型的预测误差均小于 ARIMA、XGBoost、GSXGB 和 DWT-ARIMA-XGBoost 四种预测模型。仿真结果表明，本章提出的 DWT-ARIMA-GSXGB 股价预测模型具有较好的逼近能力和泛化能力，能够很好地对股指开盘价进行分析预测。

5.2 DWT-ARIMA-GSXGB 模型的构建

5.2.1 DWT-ARIMA-GSXGB 模型

股票价格的行为不容易被捕获，混合模型可以模拟股票价格行为模式，提高整体预测性能。因此，具有股票价格建模能力的混合策略是预测股票价格的良好选择。其中，ARIMA 和 XGBoost 模型都具有捕获股票价格中数据特征的不同能力，因此本研究中提出的混合模型由 ARIMA 组件和 XGBoost 组件组成。首先采用基函数为 db4 的离散小波变换将股票数据集（ Y_t ）分为近似（ L_t ）和误差（ N_t ）两个部分，如公式（5.1）所示，其中 t 为时间。

$$Y_t = L_t + N_t \quad (5.1)$$

然后对已经分离好的数据集 L_t ，采用 ARIMA(p,d,q)模型对其训练预测，得到近似部分预测数据集 (\hat{L}_t)，模型中的 p,d,q 的确定，分别根据 ACF 图和 PACF 图进行确定。数据集 N_t ，采用 XGBoost 模型对其训练预测，得到误差部分预测数据集 (\hat{N}_t)，由于 XGBoost 算法中的参数影响着 XGBoost 算法的性能，为解决 XGBoost 算法最优参数难寻这一问题，本章采用网格搜索算法 (GS) 进行超参数优化，其中 GS 中的 CV 根据人为经验设置为 5。最后将 \hat{L}_t 与 \hat{N}_t 进行小波重构，得到最终预测结果 \hat{Y}_t ，如公式(5.2)所示。

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t \quad (5.2)$$

5.2.2 DWT-ARIMA-GSXGB 模型流程图

DWT-ARIMA-GSXGB 模型的具体构建过程如表 5.1 所示，流程图如图 5.1 所示。

表 5.1 DWT-ARIMA-GSXGB 模型的具体构建过程

<p>1、获取股票历史数据，并进行缺失值删除处理；</p> <pre>data = tushare.get_k_data(Stock code, start=start, end=end) data= data.replace(to_replace='?', value=np.nan).dropna()</pre> <p>2、采用基函数为 db4 的离散小波变换根据公式 (2.23) 对股票历史数据 (Y_t) 进行降噪与分解，将数据集的开盘价分为近似部分 (L_t) 与误差部分 (N_t)；</p> <pre>L_t, N_t = pywt.wavedec(Y_t, 'db4', mode='sym', level=1)</pre> <p>3、对分解后的近似部分采用 ARIMA(p,d,q)模型进行训练预测，根据 ACF 图和 PACF 图进行模型 p,d,q 的确定；</p> <pre>params=ARIMA(L_t,order=(1,1,0)).fit() \hat{L}_t = ARIMA(L_t,order=(1,1,0)).predict(params=params,start=1,end=len(L_t))</pre> <p>4、采用 sklearn 包中的 XGB 方法，实现 XGBoost 算法，对分解后的误差数据集进行训练，其中 XGBoost 模型采用网格搜索算法进行参数优化，构建 GSXGB 股价预测模型；</p> <pre>param_dist = {'max_depth': [x for x in range(1, 20, 1)], 'n_estimators': [x for x in range(1, 200, 1)], 'min_child_weight': [x for x in range(1, 50, 1)],} \hat{N}_t = GridSearchCV(XGBRegressor(max_depth , n_estimators , min_child_weight), param_grid,cv=5).fit().predict(N_t)</pre> <p>5、将 ARIMA(p,d,q)模型预测的近似部分 (\hat{L}_t) 与 GSXGB 模型预测的误差部分 (\hat{N}_t) 根据公式 (5.2) 进行小波重构，重构后的结果为 (\hat{Y}_t) 最终的预测结果；</p> <pre>\hat{Y}_t = pywt.waverec([\hat{L}_t , \hat{N}_t], 'db4')</pre> <p>6、对比真实值与预测值之间的差异，判断 DWT-ARIMA-GSXGB 模型的预测性能。</p>

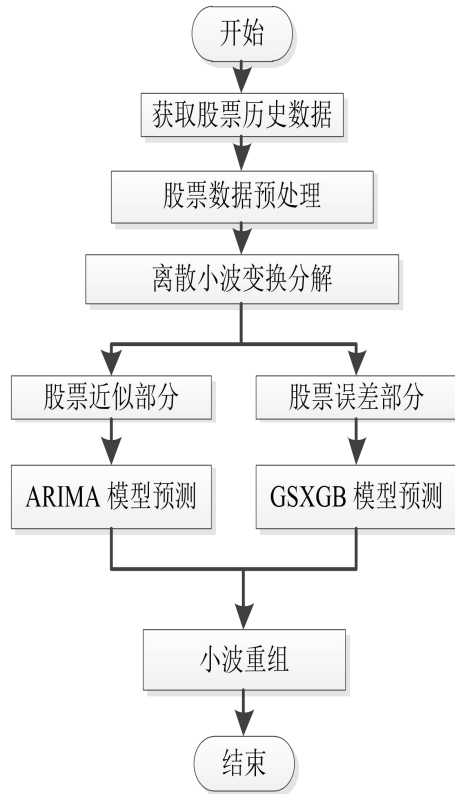


图 5.1 DWT-ARIMA-GSXGB 模型流程图.

5.3 实验研究与分析

5.3.1 实验数据

在这项研究中使用了 10 只股票来检验所提出模型的性能。收集每只股票 2015-2018 年的每日开盘价作为库存数据集。其中每只股票开盘价的后 100 个数据被用作测试数据集，其余的数据集被用作训练数据集。在这项研究中，为了防止估计样本预测误差与前期累积误差对预测的影响，只考虑提前一步预测。10 只股票数据集的具体信息如表 5.2 所示。

表 5.2 股票数据集

股票	中国 石油	平安 保险	贵州 茅台	上汽 集团	海天 集团	中国 建筑	顺丰 控股	福耀 集团	福晶 科技	中信 证券
代码	601857	601318	600519	600104	603288	601688	002352	600600	002222	600030
日期	1/5/201	1/5/201	1/5/201	1/5/201	1/5/201	1/5/201	1/5/201	1/5/201	1/5/201	1/5/201
	5-12/28	5-12/28	5-12/28	5-12/28	5-12/28	5-12/28	5-12/28	5-12/28	5-12/28	5-12/28
	/2018	/2018	/2018	/2018	/2018	/2018	/2018	/2018	/2018	/2018
数据	975	975	975	955	975	975	930	975	975	971

5.3.2 实验评价指标

五个指数， $RMSE$ （平均绝对百分误差）、 MAE （平均绝对误差）、 R^2 （决定系数）、 AUC 和 $Accuracy$ （准确率）被用作预测准确度的度量。指数如下所示

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2} \quad (5.3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y^{(i)} - \hat{y}^{(i)}| \quad (5.4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5.5)$$

$$AUC = \frac{\sum_{ins_i \in positiveclass} rank_i - \frac{M \times (M+1)}{2}}{M \times N} \quad (5.6)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.7)$$

其中，公式（5.3~5.7）中， y 为真实值， \hat{y} 为预测值。公式（5.6）中， $rank_i$ 代表第 i 条股票数据的序号， M 和 N 分别为正样本（股价上涨）的个数和负样本（股价下跌）的个数。公式（5.7）中， TP 和 FP 为预测股价上涨正确与错误的个数， TN 和 FN 为预测股价下跌正确与错误的个数。

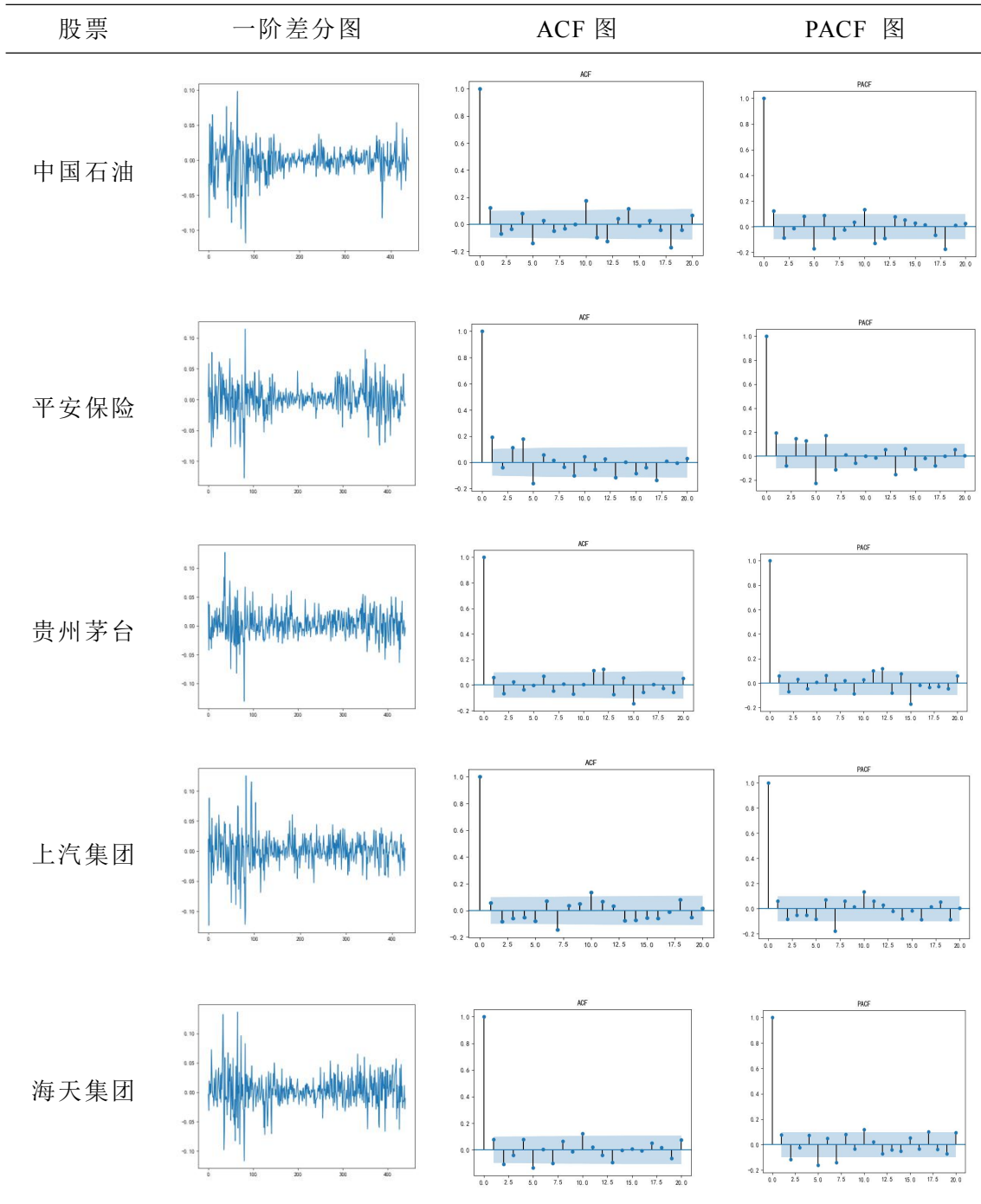
5.3.3 DWT-ARIMA-GSXGB 模型实验结果对比与分析

在本研究中，首先采用基函数为 $db4$ 的离散小波变换将股票历史数据集分别拆分为近似和误差两个部分。然后分别采用 $ARIMA$ 模型，改进的 $XGBoost$ 模型对拆分后的近似部分和误差部分进行预测。最后再将预测的结果进行小波重构，重构后的预测结果为最终的预测值。 $ARIMA$ 模型有三个阶段：模型识别，参数估计和诊断检查。在拟合 $ARIMA(p, d, q)$ 模型之前，必须指定模型的 p, d, q 。 ACF 图和 $PACF$ 图有助于 $ARIMA$ 模型中 p, q 的决策过程。

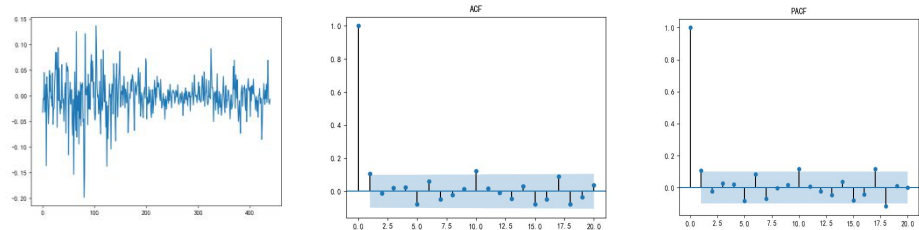
表 5.3 中的 Data Plot 部分表示 10 组股票开盘价近似部分进行一阶差分后的数据图，从图中的数据部分可以看出，10 只股票的近似部分已基本趋于平稳，趋势包括增加/减少趋势，偶尔的大倾角和稳定的相关系数，以及混合的振荡周期，大多数数据集显示出似乎接近白噪声的振荡趋势。 ACF Plot 和 $PACF$ Plot 两部分分别表示前 20 个数据的自相关与偏相关图，根据 10 组股票数据开盘价的 $ACF / PACF$ 图，可以确定 (p, q) 分别为 $(0,1)$ ， $(1,0)$ ， $(2,0)$ ， $(2,1)$ ， $(3,0)$ ， $(3,1)$ ， $(3,2)$ ， $(4,0)$ $(4,1)$ $(4,2)$ $(4,3)$ 时均适用于预测本章中的所有数据。为了更好的对混合模型的性能进行检测，这里不考虑 $ARIMA$ 最优模型的选择，只是根据经验选择 p, q 的值。本章采用 $ARIMA(0,1,1)$ 、 $ARIMA(1,1,0)$ 、

ARIMA (2,1,1)、ARIMA (3,1,0) 随机游走模型分别对股票历史数据集开盘价的近似部分进行预测。在 XGBoost 模型对误差部分进行预测时，由于参数的选取影响着模型的预测结果，为实现参数的自动寻优，故采用网格搜索算法分别对参数 \max_depth , $n_estimators$ 和 \min_child_weight 进行优化。

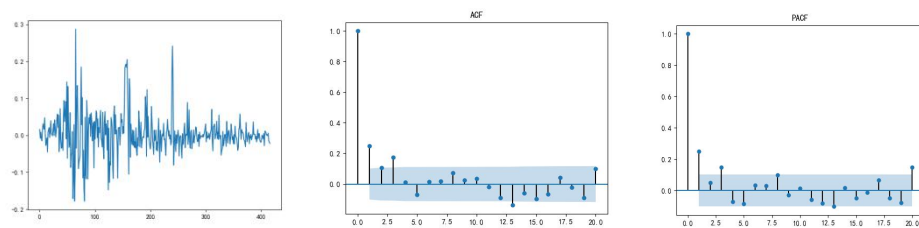
表 5.3 一阶差分 ACF 与 PACF 图



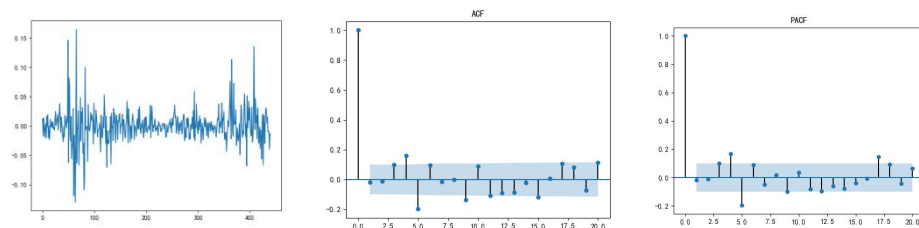
中国建筑



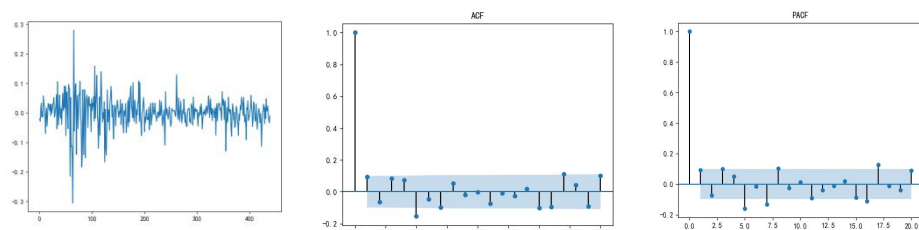
顺丰控股



福耀集团



福晶科技



中信证券

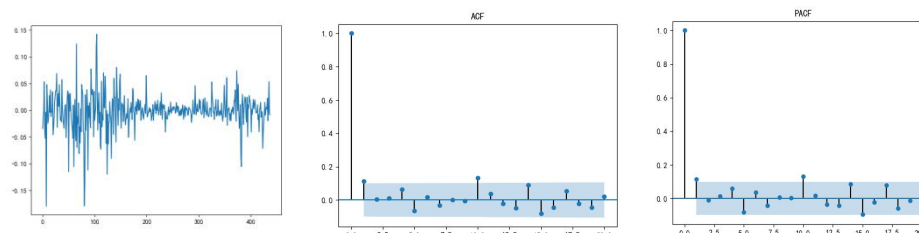


表 5.4~5.7 中 XGBoost 模型与 DWT-ARIMA-XGBoost 模型中的 XGBoost 参数一致，均采用默认参数。在表 5.4~5.7 中比较了不同模型的预测结果。这些结果表明，混合模型 DWT-ARIMA-GSXGB 在五个指数方面分别优于 ARIMA 模型、XGBoost 模型和网格搜索优化的 XGBoost 模型（GSXGB）以及 DWT-ARIMA-XGBoost 模型。为了更清楚的看到 DWT-ARIMA-XGBoost 的预测效果，本章只展示 DWT-ARIMA（1,1,0）-GSXGB 模型实际值和预测值之间的对比进行图示。如图 5.2~5.11 所示，更加清楚的看到本章提出的模型在股票预测中的拟合度高于其他几个对比模型。进一步表明本章所提出的 DWT-ARIMA-GSXGB 混合模型，与单一的 ARIMA 模型、XGBoost 模型、GSXGB 模型以及

DWT-ARIMA-XGBoost 相比，在股价预测方面具有更高的拟合度和预测性能。

表 5.4 DWT-ARIMA(0,1,1)-GSXGB 模型预测指标的比较

	RMSE	MAE	P ²	AUC	Accuracy
ARIMA(0,1,1) model					
1. CNPC	0.67501	0.78710	-2.0283	0.54582	54.55%
2. PING AN	471.001	21.2465	-54.547	0.48506	48.48%
3. KMCL	134707.	365.869	-45.662	0.46481	46.46%
4. SAIC Motor	108.213	10.3056	-29.141	0.53518	53.54%
5. HT-Food	1027.71	31.2866	-71.497	0.43428	43.43%
6. CSCEC	6.39467	2.12402	-3.6265	0.56591	56.57%
7. S.F	190.090	13.5408	-21.072	0.46481	46.46%
8. Fuyao Group	128.236	10.5442	-18.771	0.54551	54.55%
9. CASTECHINC	5.39963	2.01520	-5.7946	0.52512	52.53%
10. CITIC	1.63360	0.92211	-1.9490	0.57591	57.58%
XGBoost model					
1. CNPC	0.00539	0.05300	0.97581	0.73261	73.74%
2. PING AN	0.54991	0.57517	0.93514	0.72258	71.72%
3. KMCL	81.3232	7.08542	0.97182	0.64814	63.64%
4. SAIC Motor	0.13253	0.29011	0.96308	0.75490	75.76%
5. HT-Food	2.27602	1.14520	0.83944	0.56408	56.57%
6. CSCEC	0.05220	0.18098	0.96222	0.75714	75.76%
7. S.F	0.32150	0.41549	0.96266	0.56362	55.56%
8. Fuyao Group	0.22360	0.37932	0.96552	0.65714	65.66%
9. CASTECHINC	0.02889	0.13483	0.96363	0.71384	71.72%
10. CITIC	0.03054	0.13427	0.94486	0.72775	72.73%
GSXGB model					
1. CNPC	0.00388	0.04475	0.98255	0.75695	75.76%
2. PING AN	0.58152	0.59318	0.93141	0.72606	72.73%
3. KMCL	78.2017	6.62409	0.97256	0.70000	69.70%
4. SAIC Motor	0.13021	0.28482	0.96373	0.73772	73.74%
5. HT-Food	2.39683	1.19438	0.83092	0.54448	54.55%
6. CSCEC	0.04667	0.16319	0.96622	0.77816	77.78%
7. S.F	0.28832	0.40951	0.96652	0.61067	60.61%
8. Fuyao Group	0.19286	0.34883	0.97026	0.67653	67.68%
9. CASTECHINC	0.02893	0.13524	0.96358	0.72365	72.73%
10. CITIC	0.02854	0.12601	0.94847	0.73714	73.74%
DWT+ARIMA(0,1,1)+XGBoost model					
1. CNPC	0.00237	0.03740	0.98932	0.85720	85.86%
2. PING AN	0.35650	0.47561	0.95795	0.82835	82.83%
3. KMCL	36.3691	4.89102	0.98740	0.77777	77.78%
4. SAIC Motor	0.05886	0.19827	0.98360	0.80053	79.80%
5. HT-Food	0.48033	0.56755	0.96611	0.80795	80.81%
6. CSCEC	0.02662	0.13392	0.98073	0.77734	77.78%
7. S.F	0.14146	0.26035	0.98357	0.83694	83.84%
8. Fuyao Group	0.23138	0.40342	0.96432	0.73734	73.74%
9. CASTECHINC	0.02660	0.13404	0.96652	0.82843	82.83%
10. CITIC	0.02483	0.12650	0.95516	0.80775	80.81%
DWT + ARIMA(0,1,1) + GSXGB model					
1. CNPC	0.00174	0.03115	0.99218	0.87745	87.88%
2. PING AN	0.22161	0.37932	0.97386	0.87745	87.88%
3. KMCL	20.2947	3.66994	0.99296	0.79999	79.80%
4. SAIC Motor	0.03635	0.15905	0.98987	0.81014	80.81%
5. HT-Food	0.23330	0.41443	0.98354	0.83836	83.84%
6. CSCEC	0.01616	0.10369	0.98830	0.84836	84.85%
7. S.F	0.09425	0.21617	0.98905	0.89668	89.90%
8. Fuyao Group	0.11304	0.28878	0.98257	0.79775	79.80%
9. CASTECHINC	0.01710	0.10620	0.97848	0.87928	87.88%
10. CITIC	0.01416	0.09422	0.97443	0.81816	81.82%

表 5.5 DWT-ARIMA(1,1,0)-GSXGB 模型预测指标的比较

	RMSE	MAE	P ²	AUC	Accuracy
ARIMA(1,1,0) model					
1. CNPC	0.02936	0.13106	0.86824	0.64402	64.65%
2. PING AN	1.32716	0.89993	0.84348	0.65671	65.66%
3. KMCL	141.964	9.47687	0.95082	0.72222	71.72%
4. SAIC Motor	0.19049	0.34145	0.94694	0.69721	69.70%
5. HT-Food	1.44728	0.93487	0.89790	0.69693	69.70%
6. CSCEC	0.07837	0.20890	0.94329	0.68714	68.69%
7. S.F	0.14896	0.29985	0.98270	0.73874	73.74%
8. Fuyao Group	0.86317	0.69817	0.86691	0.72714	72.73%
9. CASTECHINC	0.07393	0.20649	0.90697	0.70649	70.71%
10. CITIC	0.07055	0.20318	0.87263	0.66632	66.67%
XGBoost model					
1. CNPC	0.00539	0.05300	0.97581	0.73261	73.74%
2. PING AN	0.54991	0.57517	0.93514	0.72258	71.72%
3. KMCL	81.3232	7.08542	0.97182	0.64814	63.64%
4. SAIC Motor	0.13253	0.29011	0.96308	0.75490	75.76%
5. HT-Food	2.27602	1.14520	0.83944	0.56408	56.57%
6. CSCEC	0.05220	0.18098	0.96222	0.75714	75.76%
7. S.F	0.32150	0.41549	0.96266	0.56362	55.56%
8. Fuyao Group	0.22360	0.37932	0.96552	0.65714	65.66%
9. CASTECHINC	0.02889	0.13483	0.96363	0.71384	71.72%
10. CITIC	0.03054	0.13427	0.94486	0.72775	72.73%
GSXGB model					
1. CNPC	0.00424	0.04464	0.98094	0.73874	73.74%
2. PING AN	0.58371	0.59601	0.93115	0.72606	72.73%
3. KMCL	79.2032	6.68972	0.97256	0.69074	68.69%
4. SAIC Motor	0.12607	0.28218	0.96488	0.73874	73.74%
5. HT-Food	2.48793	1.22223	0.82449	0.57551	57.58%
6. CSCEC	0.04312	0.15927	0.96880	0.79836	79.80%
7. S.F	0.28475	0.41577	0.96693	0.66837	66.67%
8. Fuyao Group	0.20067	0.35596	0.96905	0.67632	67.68%
9. CASTECHINC	0.03154	0.14312	0.96030	0.66789	66.67%
10. CITIC	0.02727	0.12423	0.95075	0.75693	75.76%
DWT + ARIMA(1,1,0) + XGBoost model					
1. CNPC	0.00240	0.03733	0.98919	0.84656	84.85%
2. PING AN	0.33199	0.45930	0.96084	0.82835	82.83%
3. KMCL	36.3691	4.89102	0.98740	0.77777	77.78%
4. SAIC Motor	0.05828	0.19787	0.98376	0.77925	77.78%
5. HT-Food	0.48167	0.56790	0.96602	0.80795	80.81%
6. CSCEC	0.03395	0.14335	0.98266	0.77734	77.78%
7. CSCEC	0.13170	0.25053	0.98470	0.83694	83.84%
7. S.F	0.13170	0.25053	0.98470	0.83694	83.84%
8. Fuyao Group	0.25357	0.41413	0.96090	0.73734	73.74%
9. CASTECHINC	0.02633	0.13310	0.96686	0.83823	83.84%
10. CITIC	0.02482	0.12644	0.95518	0.80775	80.81%
DWT + ARIMA(1,1,0) + GSXGB model					
1. CNPC	0.00183	0.03186	0.99177	0.87847	87.88%
2. PING AN	0.32868	0.45127	0.96123	0.84758	84.85%
3. KMCL	17.6599	3.42838	0.99388	0.81851	81.82%
4. SAIC Motor	0.03579	0.15802	0.99003	0.81014	80.81%
5. HT-Food	0.09147	0.24848	0.96602	0.92897	92.93%
6. CSCEC	0.02666	0.13397	0.99354	0.83836	83.84%
7. CSCEC	0.06124	0.18284	0.99288	0.86579	86.87%
7. S.F	0.04759	0.17606	0.99266	0.87836	87.88%
8. Fuyao Group	0.01682	0.10477	0.97882	0.87928	87.88%
9. CASTECHINC	0.01418	0.09421	0.97439	0.81816	81.82%
10. CITIC					

表 5.6 DWT-ARIMA(2,1,1)-GSXGB 模型预测指标的比较

	RMSE	MAE	P ²	AUC	Accuracy
ARIMA(2,1,1) model					
1. CNPC	0.93132	0.92350	-3.1782	0.54582	54.55%
2. PING AN	635.427	24.6705	-73.939	0.48506	48.48%
3. KMCL	186842.	430.860	-63.722	0.46481	46.46%
4. SAIC Motor	154.924	12.3286	-42.152	0.53518	53.54%
5. HT-Food	1048.60	31.6009	-72.971	0.43428	43.43%
6. CSCEC	2.27300	1.20419	-0.64451	0.54551	54.55%
7. S.F	237.885	15.1482	-26.622	0.46481	46.46%
8. Fuyao Group	170.689	12.1630	-25.317	0.54551	54.55%
9. CASTECHINC	7.31028	2.34663	-8.1988	0.52512	52.53%
10. CITIC	1.68569	0.99995	-1.8971	0.59545	59.55%
XGBoost model					
1. CNPC	0.00539	0.05300	0.97581	0.73261	73.74%
2. PING AN	0.54991	0.57517	0.93514	0.72258	71.72%
3. KMCL	81.3232	7.08542	0.97182	0.64814	63.64%
4. SAIC Motor	0.13253	0.29011	0.96308	0.75490	75.76%
5. HT-Food	2.27602	1.14520	0.83944	0.56408	56.57%
6. CSCEC	0.05220	0.18098	0.96222	0.75714	75.76%
7. S.F	0.32150	0.41549	0.96266	0.56362	55.56%
8. Fuyao Group	0.22360	0.37932	0.96552	0.65714	65.66%
9. CASTECHINC	0.02889	0.13483	0.96363	0.71384	71.72%
10. CITIC	0.03054	0.13427	0.94486	0.72775	72.73%
GSXGB model					
1. CNPC	0.00412	0.04660	0.98148	0.76759	76.77%
2. PING AN	0.54595	0.58225	0.93141	0.70785	70.71%
3. KMCL	81.6389	7.09883	0.97172	0.64444	63.64%
4. SAIC Motor	0.12769	0.28507	0.96443	0.72810	72.73%
5. HT-Food	2.41562	1.20023	0.82959	0.55510	55.56%
6. CSCEC	0.04659	0.16267	0.96628	0.77795	77.78%
7. S.F	0.33958	0.45438	0.96056	0.61067	60.61%
8. Fuyao Group	0.25533	0.39315	0.96063	0.68673	68.69%
9. CASTECHINC	0.03196	0.14804	0.95977	0.67769	67.68%
10. CITIC	0.02854	0.12601	0.94847	0.73714	73.74%
DWT + ARIMA(2,1,1) + XGBoost model					
1. CNPC	0.00246	0.03829	0.98892	0.85720	85.86%
2. PING AN	36.6624	4.86288	0.98730	0.78888	78.79%
3. KMCL	0.06293	0.20289	0.98247	0.78989	78.79%
4. SAIC Motor	0.56006	0.62002	0.96049	0.80816	80.81%
5. HT-Food	0.02680	0.13398	0.98060	0.83836	83.84%
6. CSCEC	0.13884	0.25705	0.98387	0.86783	86.87%
7. S.F	0.24130	0.40903	0.96279	0.73734	73.74%
8. Fuyao Group	0.02633	0.13459	0.96686	0.82843	82.83%
9. CASTECHINC	0.02404	0.12178	0.95866	0.76388	76.40%
10. CITIC					
DWT + ARIMA(2,1,1) + GSXGB model					
1. CNPC	0.00183	0.03165	0.99177	0.87745	87.88%
2. PING AN	20.5723	3.66786	0.99287	0.79814	79.80%
3. KMCL	0.04056	0.16687	0.98870	0.79950	79.80%
4. SAIC Motor	0.31242	0.47413	0.97796	0.83836	83.84%
5. HT-Food	0.01634	0.10402	0.98817	0.84836	84.85%
6. CSCEC	0.09404	0.21539	0.98908	0.89668	89.90%
7. S.F	0.11392	0.28894	0.98243	0.79775	79.80%
8. Fuyao Group	0.01690	0.10632	0.97873	0.86948	86.87%
9. CASTECHINC	0.01424	0.09191	0.97551	0.83156	83.15%
10. CITIC					

表 5.7 DWT-ARIMA(3,1,0)-GSXGB 模型预测指标的比较

	RMSE	MAE	P ²	AUC	Accuracy
ARIMA(3,1,0) model					
1. CNPC	0.03094	0.13428	0.86115	0.65364	65.66%
2. PING AN	3.99522	1.62830	0.52882	0.56403	56.57%
3. KMCL	384.751	16.0459	0.86672	0.70185	69.70%
4. SAIC Motor	0.78712	0.71621	0.78075	0.57467	57.58%
5. HT-Food	3.77732	1.48301	0.73353	0.43428	63.64%
6. CSCEC	0.39802	0.48888	0.71202	0.56571	56.57%
7. S.F	0.53290	0.58146	0.93812	0.62581	62.63%
8. Fuyao Group	2.77757	1.30040	0.57174	0.64653	64.65%
9. CASTECHINC	0.36459	0.47406	0.54120	0.56617	56.57%
10. CITIC	0.28792	0.40646	0.48023	0.56591	56.57%
XGBoost model					
1. CNPC	0.00539	0.05300	0.97581	0.73261	73.74%
2. PING AN	0.54991	0.57517	0.93514	0.72258	71.72%
3. KMCL	81.3232	7.08542	0.97182	0.64814	63.64%
4. SAIC Motor	0.13253	0.29011	0.96308	0.75490	75.76%
5. HT-Food	2.27602	1.14520	0.83944	0.56408	56.57%
6. CSCEC	0.05220	0.18098	0.96222	0.75714	75.76%
7. S.F	0.32150	0.41549	0.96266	0.56362	55.56%
8. Fuyao Group	0.22360	0.37932	0.96552	0.65714	65.66%
9. CASTECHINC	0.02889	0.13483	0.96363	0.71384	71.72%
10. CITIC	0.03054	0.13427	0.94486	0.72775	72.73%
GSXGB model					
1. CNPC	0.00395	0.04506	0.98227	0.75695	75.76%
2. PING AN	0.54794	0.57639	0.93537	0.72708	72.73%
3. KMCL	78.1873	6.85354	0.97291	0.64444	63.64%
4. SAIC Motor	0.12625	0.28177	0.96483	0.73874	73.74%
5. HT-Food	2.39310	1.19128	0.83118	0.57510	57.58%
6. CSCEC	0.04835	0.16927	0.96501	0.77816	77.78%
7. S.F	0.29923	0.41952	0.96525	0.62990	62.63%
8. Fuyao Group	0.23090	0.37282	0.96439	0.70673	70.71%
9. CASTECHINC	0.02589	0.12936	0.96741	0.74693	74.75%
10. CITIC	0.02859	0.12542	0.94837	0.77734	77.78%
DWT + ARIMA(3,1,0) + XGBoost model					
1. CNPC	0.00252	0.03761	0.98868	0.84656	84.85%
2. PING AN	0.39389	0.48846	0.95354	0.81669	81.82%
3. KMCL	36.4617	4.88869	0.98736	0.78888	78.79%
4. SAIC Motor	0.06011	0.19773	0.98325	0.77925	77.78%
5. HT-Food	0.48241	0.57075	0.96596	0.81795	81.82%
6. CSCEC	0.02718	0.13362	0.98033	0.83836	83.84%
7. S.F	0.14030	0.26006	0.98370	0.85822	85.86%
8. Fuyao Group	0.30944	0.46130	0.95228	0.71734	71.72%
9. CASTECHINC	0.02645	0.13466	0.96670	0.81862	81.82%
10. CITIC	0.02487	0.12645	0.95509	0.80775	80.81%
DWT + ARIMA(3,1,0) + GSXGB model					
1. CNPC	0.00188	0.03207	0.99153	0.86783	86.87%
2. PING AN	0.26000	0.40326	0.96933	0.87745	87.88%
3. KMCL	20.3013	3.67230	0.99296	0.79999	79.80%
4. SAIC Motor	0.03717	0.16133	0.98964	0.81014	80.81%
5. HT-Food	0.23419	0.41648	0.98347	0.83816	83.84%
6. CSCEC	0.01671	0.10384	0.98790	0.85836	85.86%
7. S.F	0.09321	0.21630	0.98917	0.89668	89.90%
8. Fuyao Group	0.16920	0.34245	0.97391	0.76816	76.77%
9. CASTECHINC	0.01693	0.10634	0.97869	0.84987	84.85%
10. CITIC	0.01421	0.09410	0.97433	0.82816	82.83%

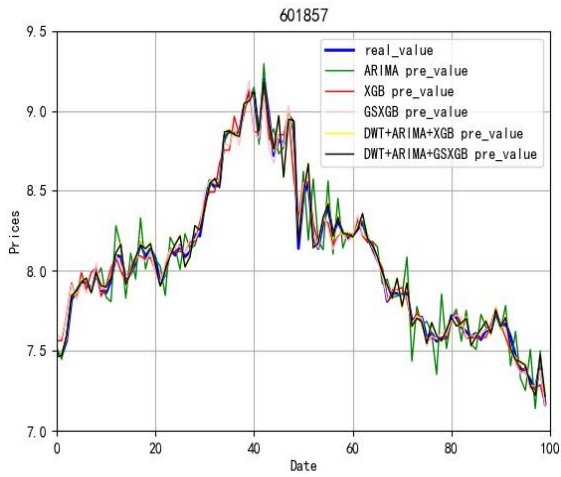


图 5.2 601857 股票预测对比图.

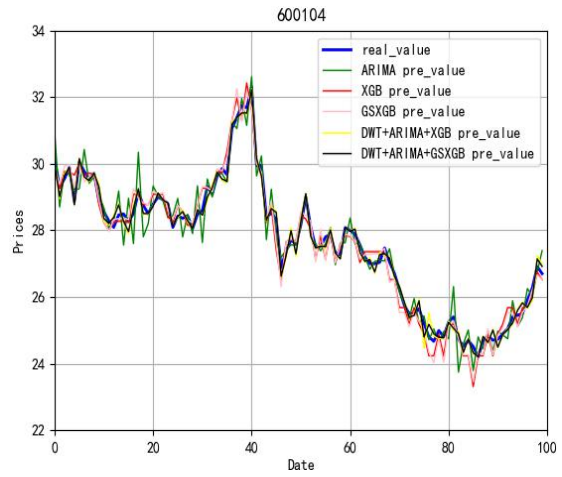


图 5.3 600104 股票预测对比图.

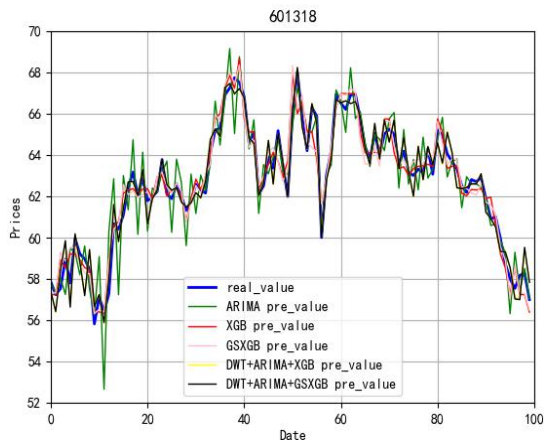


图 5.4 601318 股票预测对比图.

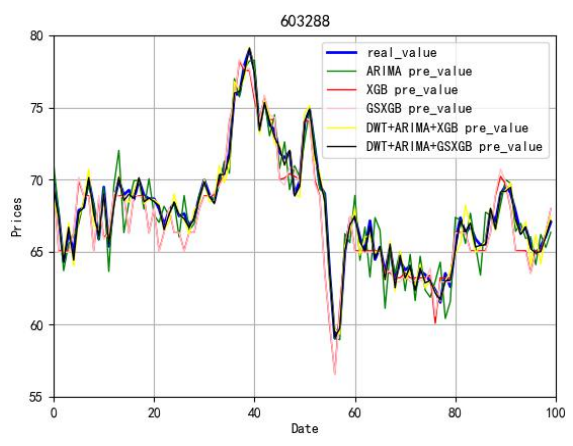


图 5.5 603288 股票预测对比图.

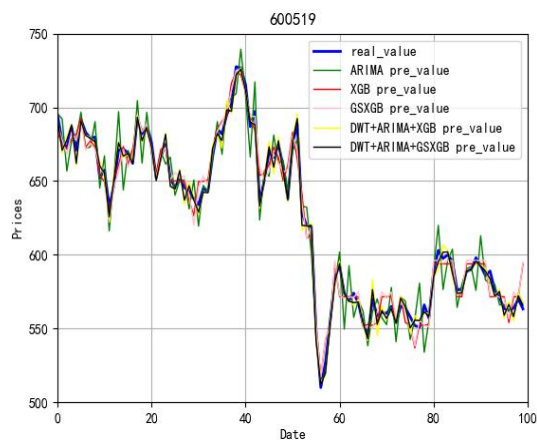


图 5.6 600519 股票预测对比图.

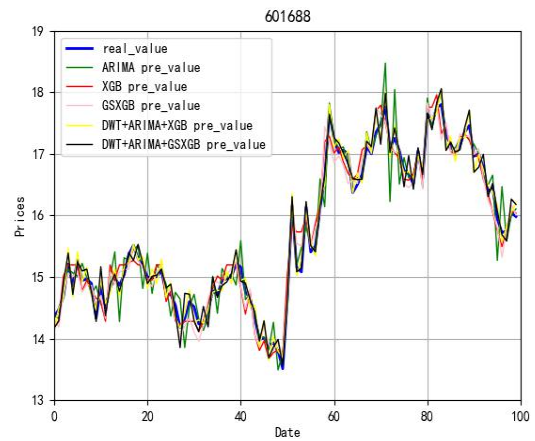


图 5.7 601688 股票预测对比图.

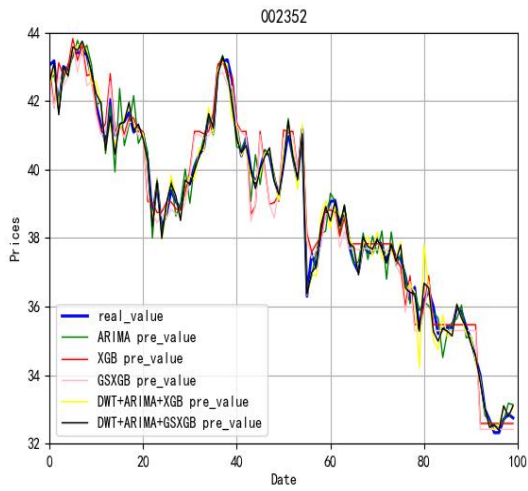


图 5.8 002352 股票预测对比图.

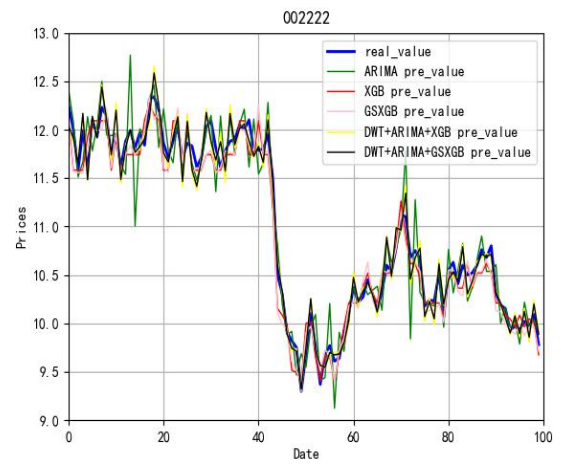


图 5.9 002222 股票预测对比图.

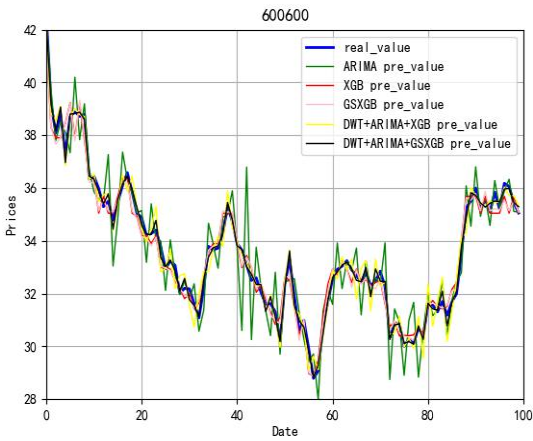


图 5.10 600600 股票预测对比图.

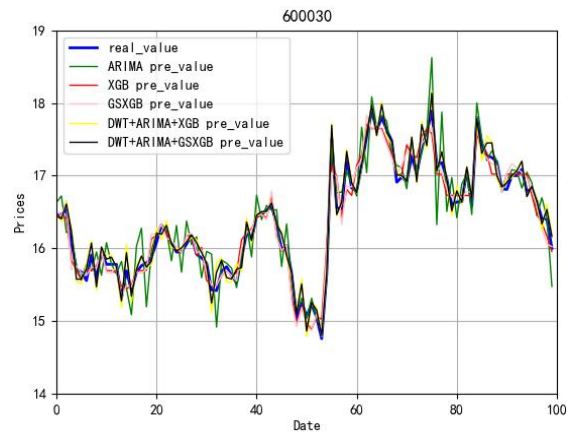


图 5.11 600030 股票预测对比图.

5.4 本章小结

股票历史数据集是一个金融时间序列数据集，可以分解为近似部分数据集和误差部分数据集。ARIMA 在时间序列预测的许多领域占据主导地位，该模型基本上是一种面向数据的方法，适用于数据本身的结构。然而，任何误差部分数据集都会限制该模型的预测性能。XGBoost 中的 CART 树能解释模型中的非线性关系和变量之间相互依赖的现象，在非线性数据预测中具有良好的性能。由于 DWT 分解的误差部分具有非线性特性，因此，本章采用 XGBoost 模型进行股票数据集中的误差部分进行预测。本章提出的混合模型（DWT-ARIMA-GSXGB）采用离散小波变换将数据拆分为近似和误差两部分，ARIMA 模型来处理近似部分数据，网格搜索改进的 XGBoost 模型（GSXGB）来处理误差部分数据。根据 10 组股票数据集的实验对比，发现 DWT-ARIMA-GSXGB 模型的误差均小于 ARIMA、XGBoost、GSXGB 和 DWT-ARIMA-XGBoost 四种预测模型。仿真结果表明，该融合模型极大地改善了单个 ARIMA 模型或单个 XGBoost 模型在预测股票价格方面的预测性能。从理论上和经验上看，杂交两种不同的模型可以减少预测误差。但是，模型参数的多样性，使得两种最佳个体模型的简单组合不一定能产生最佳结果。因此，混合模型的最佳参数的结构化选择具有重要的研究意义。

总结与展望

随着信息科技的不断进步，机器学习领域得到了长足的发展。关于如何利用机器学习方法来解决金融领域的相关问题，专家和学者一直都在深入研究。股票预测方法也属于金融领域其中的一个部分。由于人民生活水平日益提高。在解决温饱问题之余，有了可供投资的余财。越来越多的人将目光转向股市投资，为股市发展提供了资金条件。然而在纷繁复杂的股票市场，如何寻找最优股成为亟待解决的问题。这不仅是投资者单方面的困惑，也是股票价格预测领域中学者们所关心的重点。因此，对股票市场预测系统的设计和实施不仅具有深刻的理论意义，而且具有非常重要的使用价值。

本文探究了 XGBoost 与 ARIMA 的混合模型在股票预测中的应用。然后在 XGBoost 与 ARIMA 模型的基础上，结合机器学习相关理论，提出相应的模型结构改进和优化，并做出相应的模型对比，从而提升股票预测精度。下面将会对本文的工作做简要的总结。

(1) 本文提出网格搜索算法优化的 XGBoost 金融预测模型(GS-XGBoost)。首先，根据网格搜索算法的思想，先设定将要选择的参数组合区间，基于 Xgboost 算法，在参数寻优的过程中，结合网格搜索算法的思想，不断地训练模型，通过评价函数对每个参数组合得到的分类结果进行评价，最终得到最优参数组合，最后将最优参数组合代入 Xgboost 算法，从而使预测性能得到提升。

(2) 本文提出离散小波变换与优化的 XGBoost 算法结合的股价预测模型(DWT-GS-XGBoost)。综合考虑了 DWT 与 XGBoost 模型的优点，其中为了降低股票数据集中的噪声，采用在去噪方面表现良好的离散小波变换进行数据降噪处理与分解，然后使用网格搜索优化的 XGBoost 模型对降噪处理后的股票数据集进行训练和预测，并与 GS-XGBoost 模型预测结果进行对比分析。通过实验预测结果表明 DWT-GS-XGBoost 模型预测效果优于 GS-XGBoost 模型。

(3) 本文提出了一种离散小波变换、ARIMA 和优化的 XGBoost 的混合模型(DWT-ARIMA-GSXGB)来解决股票价格预测问题。其中混合模型采用离散小波变换将数据集拆分为近似部分和误差部分，ARIMA 模型处理近似部分数据，网格搜索改进的 XGBoost 模型处理误差部分数据。实验结果表明，DWT-ARIMA-GSXGB 股价预测模型具有较好的拟合能力和泛化能力，极大地改善了单个 ARIMA 模型或单个 XGBoost 模型在预测股票价格方面的预测性能。

虽然本文在股票预测中取得了一些实验进展，但由于时间和能力有限，还存在一些问题没有解决。考虑到现实生活中股票价格波动是由多因素引起，且存在

人为操控等特点，还需要对模型做进一步的探索与研究。下一步的研究重点是：

（1）考虑到网络舆情对股市的影响，后期将对网络舆情进行特征提取，结合现有的股票属性，进行分析。从而使股票波动的预测精度得到提升，进一步为投资者在股市中的投资提供更牢靠的参考。

（2）其次是利用传统的统计方法与机器学习算法相结合，寻找出一种能够找到股票变动正常的方法，从而使在进行股票涨跌判断时对于超出正常趋势的股票可以做出筛选，然后采用机器学习算法进行预测分析，使模型在股票预测时，可以达到更高的预测精度。

（3）由于近几年深度学习的发展，基于深度学习研究股票预测的方法也日益增多，把传统的股票预测方法与深度学习方法进行优势互补，应用到更加复杂的金融领域，将作为以后主要的研究方向。

参考文献

- [1] 苑小康. 基于随机系统理论的股票定价模型的研究[D].北京交通大学,2018.
- [2] Fengmei Yang,Zhiwen Chen,Jingjing Li,Ling Tang. A novel hybrid stock selection method with stock prediction[J]. Applied Soft Computing Journal,2019,80.
- [3] Artificial Neural Networks; New Artificial Neural Networks Findings from M. Paluch and Co-Researchers Described (Hybrid Models Combining Technical and Fractal Analysis with ANN for Short-Term Prediction of Close Values on the Warsaw Stock Exchange)[J]. Computers, Networks & Communications,2019.
- [4] 杜纯文. 我国金融结构对经济增长的边际效应演化分析[D].厦门大学,2018.
- [5] Yiu Kuen Tse, S. H Tung. Forecasting Volatility in the Singapore Stock Market[J]. Asia Pacific Journal of Management, 1992, 9(1).
- [6] 郑丕谔, 马艳华. 基于 RBF 神经网络的股市建模与预测[J]. 天津大学学报, 2000(04):70-73.
- [7] 吴微, 陈维强, 刘波. 用 BP 神经网络预测股票市场涨跌[J]. 大连理工大学学报, 2001(01):12-18.
- [8] Kim K J . Financial time series forecasting using support vector machines[J]. Neurocomputing, 2003, 55(1-2):307-319.
- [9] 张燕平, 张铃, 吴涛, et al. 基于覆盖的构造性学习算法 SLA 及在股票预测中的应用[J]. 计算机研究与发展, 2004(第 6 期):979-984.
- [10] Otengabayie E F, Frimpong Magnus J. Modelling and Forecasting Volatility of Returns on the Ghana Stock Exchange[J]. American Journal of Applied Sciences, 2006, 3(10).
- [11] Miao K, Chen F, Zhao Z G. Stock Price Forecast Based on Bacterial Colony RBF Neural Network[J]. Journal of Qingdao University, 2007.
- [12] Sui X, Hu Q, Yu D, et al. A Hybrid Method for Forecasting Stock Market Trend Using Soft-Thresholding De-noise Model and SVM[C]// Rough Sets, Fuzzy Sets, Data Mining & Granular Computing, International Conference, Rsfdgrc, Toronto, Canada, May. 2007.
- [13] Zhe Liao,Jun Wang. Forecasting model of global stock index by stochastic time effective neural network[J]. Expert Systems With Applications,2009,37(1).
- [14] Ticknor J L. A Bayesian regularized artificial neural network for stock market

- forecasting[J]. Expert Systems with Applications, 2013, 40(14):5501-5506.
- [15] Laboissiere L A, Fernandes R A S, Lage G G. Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks[J]. Applied Soft Computing, 2015, 35:66-74.
- [16] Ariyo A A , Adewumi A O , Ayo C K . Stock Price Prediction Using the ARIMA Model[C]// Uksim-amss International Conference on Computer Modelling & Simulation. IEEE, 2015.
- [17] Chen M Y, Chen B T. A hybrid fuzzy time series model based on granular computing for stock price forecasting[J]. Information Sciences, 2015, 294(2):227-241.
- [18] Tsantekidis A, Passalis N, Tefas A, et al. Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks[C]// IEEE Conference on Business Informatics. 2017.
- [19] Shuheng Wang, Guohao Li, Yifan Bao. A novel improved fuzzy support vector machine based stock price trend forecast model[J]. Papers, 2018.
- [20] 曹红辉, 杨欣, 申慧. 股票市场非线性随机游走检验[J]. 中央财经大学学报, 2003(4):24-28.
- [21] 朱扬光. 金融危机传染检验方法与联动行为分析[D]. 中国科学技术大学, 2017.
- [22] Bates J M, Granger C W J. The Combination of Forecasts[J]. Journal of the Operational Research Society, 1969, 20(4):451-468.
- [23] Clemen R. Combining forecasts: a review and annotated bibliography with discussion. International Journal of Forecasting 1989;5:559–608.
- [24] Menezes Lilian M, Bunn Derek W, Taylor James W. Review of guidelines for the use of combined forecasts. European Journal of Operational Research 2000;120:190–204.
- [25] Lam KF, Mui HW, Yuen HK. A note on minimizing absolute percentage error in combined forecasts. Computer & Operations Research 2001;28:1141–7.
- [26] Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 2003, 50(1):159-175.
- [27] Zhang Y , Shan R , Wang H , et al. A new wavelet-neural network-ARIMA shares index combination forecast model[C]// International Conference on Automatic Control & Artificial Intelligence. IET, 2013.
- [28] Shi S, Liu W, Jin M. Stock Price Forecasting Based on a Combined Model of ARMA and BP Neural Network and Markov Model[J]. International Journal of

- Information Processing & Management, 2013, 4(3):215-221.
- [29] Ye, Tian. [IEEE 2017 3rd International Conference on Information Management (ICIM) - Chengdu, China (2017.4.21-2017.4.23)] 2017 3rd International Conference on Information Management (ICIM) - Stock forecasting method based on wavelet analysis and ARIMA-SVR model[J]. 2017:102-106.
- [30] Zhu H, Wei H, Jing Y, et al. Soil organic carbon prediction based on scale-specific relationships with environmental factors by discrete wavelet transform[J]. Geoderma, 2018, 330:9-18.
- [31] Kim Taewook, Kim Ha Young. Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data.[J]. PloS one, 2019, 14(2).
- [32] Tsantekidis A, Passalis N, Tefas A, et al. Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks[C]// IEEE Conference on Business Informatics. 2017.
- [33] Wang C, Bai X. Boosting Learning Algorithm for Stock Price Forecasting[J]. IOP Conference Series: Materials Science and Engineering, 2018, 322:052053.
- [34] Kei Nakagawa, Tomoki Ito, Masaya Abe. Deep Recurrent Factor Model: Interpretable Non-Linear and Time-Varying Multi-Factor Model[J]. Papers, 2019.
- [35] A. Okay Akyuz, Mitat Uysal, Berna Atak Bulbul. Ensemble approach for time series analysis in demand forecasting: Ensemble learning[C]// 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, 2017.
- [36] Xueheng Qiu, Ponnuthurai Nagaratnam Suganthan, Gehan A. J. Amaratunga. Fusion of multiple indicators with ensemble incremental learning techniques for stock price forecasting[J]. Journal of Banking and Financial Technology, 2019, 3(1):33-42.
- [37] Chen T, Tong H, Benesty M. xgboost: Extreme Gradient Boosting[J]. 2016.
- [38] Omkar Giraka, Vasantha Kumar Selvaraj. Short-term prediction of intersection turning volume using seasonal ARIMA model[J]. Transportation Letters The International Journal of Transportation Research, 2019:1-8.
- [39] Jing Bi, Libo Zhang, Haitao Yuan. Hybrid task prediction based on wavelet decomposition and ARIMA model in cloud data center[C]// 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2018.

- [40] Rudra P. Pradhan. Development of stock market and economic growth: the G-20 evidence[J]. Eurasian Economic Review, 2018, 8(2):161-181.
- [41] Simplicio A. Asongu, Jacinta C. Nwachukwu. Political Regimes and Stock Market Performance in Africa[J]. Working Papers, 2018, 16(3):240–249.
- [42] Jammazi, Rania, Ferrer, Román, Jareño, Francisco. Main driving factors of the interest rate-stock market Granger causality[J]. International Review of Financial Analysis, 2017, 52:1-40.
- [43] Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin, & Victor Chang. An innovative neural network approach for stock market prediction[J]. Journal of Supercomputing, 2018(1):1-21.
- [44] Konstantinos Pliakos, Pierre Geurts, Celine Vens. Global multi-output decision trees for interaction prediction[J]. Machine Learning, 2018, 107(7).
- [45] Ming-Chi Lee. Using support vector machine with a hybrid feature selection method to the stock trend prediction[J]. Expert Systems with Applications, 2009, 36(8):10896-10904.
- [46] Abidatul Izzah, Yuita Arum Sari, Ratna Widyastuti, & Toga Aldila Cinderatama. Mobile app for stock prediction using Improved Multiple Linear Regression[C]// 2017 International Conference on Sustainable Information Engineering and Technology (SIET). IEEE, 2017.
- [47] Ritika Singh, Shashi Srivastava. Stock prediction using deep learning[J]. Multimedia Tools & Applications, 2016:1-16.
- [48] A. Douglas Harris. The Impact of Hot Issue Markets and Noise Traders on Stock Exchange Listing Standards[J]. University of Toronto Law Journal, 56(3):223-280.
- [49] Yannick Timmer. Cyclical investment behavior across financial institutions[J]. Journal of Financial Economics, 2018, 129:págs. 268-286.
- [50] Wei Y, Ying S, Fan Y, et al. The cellular automaton model of investment behavior in the stock market.[J]. 2003, 325(3):507-516.
- [51] Freund, Yoav, Iyer, Raj, Schapire, Robert E., Singer, Yoram, & Dietterich, Thomas G. An Efficient Boosting Algorithm for Combining Preferences[J]. Journal of Machine Learning Research, 2004, 4(6):170--178.
- [52] 曹正凤. 随机森林算法优化研究[D].首都经济贸易大学,2014.
- [53] 曹莹, 苗启广, 刘家辰, 高琳. AdaBoost 算法研究进展与展望 [J]. 自动化学报, 2013, 39(06):745-758.
- [54] Liao Z, Huang Y, Yue X, et al. In Silico Prediction of Gamma-Aminobutyric Acid

- Type-A Receptors Using Novel Machine-Learning-Based SVM and GBDT Approaches[J]. BioMed Research International,2016,(2016-8-8), 2016, 2016(6):1-12.
- [55] Chyon-Hwa Yeh. Classification and regression trees (CART)[J]. 12(1):95-96.
- [56] 何晓旭. 时间序列数据挖掘若干关键问题研究[D].中国科学技术大学,2014.
- [57] Riswan Efendia, Nureize Arbaiy, Mustafa Mat Deris. A New Procedure in Stock Market Forecasting Based On Fuzzy Random Auto-Regression Time Series Model[J]. Information Sciences, 2018, 441.
- [58] Wei, Liang-Ying, Cheng, Ching-Hsue, Wu, Hsin-Hung. A hybrid ANFIS based on n-period moving average model to forecast TAIEX stock[J]. Applied Soft Computing Journal, 19:86-92.
- [59] Shuzhen Shi, Wenlong Liu, Minglu Jin. Stock price forecasting using a hybrid ARMA and BP neural network and Markov model[C]// Communication Technology (ICCT), 2012 IEEE 14th International Conference on. IEEE, 2012.
- [60] Pankratz A . Forecasting with Univariate Box-Jenkins Models: Concepts and Cases[M]. 2008.
- [61] Meek S A, Hipke A, Guelachvili G, et al. Doppler-free Fourier transform spectroscopy[J]. Optics Letters, 2018, 43:162-165.
- [62] Demiralp T, Ademoglu A, Schürmann M, et al. Detection of P300 waves in single trials by the wavelet transform (WT).[J]. Brain & Language, 1999, 66(1):108.
- [63] Hiemstra C, Jones J D. Testing for Linear and Nonlinear Granger Causality in the Stock Price - Volume Relation[J]. Journal of Finance, 1994, 49(5):1639-1664.
- [64] Angelos Kanas. Nonlinearity in the stock price–dividend relation[J]. Journal of International Money & Finance, 2005, 24(4):583-606.
- [65] Tsai C S, Hsieh C T, Huang S J. Enhancement of damage-detection of wind turbine blades via CWT-based approaches[J]. IEEE Transactions on Energy Conversion, 2006, 21(3):776-781.
- [66] 徐晓明. SVM 参数寻优及其在分类中的应用[D]. 大连海事大学, 2014.
- [67] Tsantekidis A, Passalis N, Tefas A, et al. Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks[C]// IEEE Conference on Business Informatics. 2017.
- [68] Jaiwang G, Jeatrakul P. A forecast model for stock trading using support vector machine[C]// Computer Science & Engineering Conference. 2017.
- [69] Guoying Z , Ping C . [IEEE 2017 IEEE International Conference on Smart Cloud (SmartCloud) -New York, NY (2017.11.3-2017.11.5)] 2017 IEEE International

- Conference on Smart Cloud (SmartCloud) - Forecast of Yearly Stock Returns Based on Adaboost Integration Algorithm[J]. 2017:263-267.
- [70] Xiao Z , Zengxin W , Tianshan Y . The Application of GBDT Combination Model in Stock Forecasting[J]. Journal of Hainan Normal University(Natural Science), 2018.
- [71] 王燕,郭元凯.改进的 XGBoost 模型在股票预测中的应用[J].计算机工程与应用,2019,55(20):202-207.
- [72] Maggioni M, Katkovnik V, Egiazarian K, et al. Nonlocal transform-domain filter for volumetric data denoising and reconstruction[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2013, 22(1):119-133.
- [73] Boashash B. Note on the use of the Wigner distribution for time-frequency signal analysis[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 2002, 36(9):1518-1521.
- [74] Diffie W, Hellman M E. Special Feature Exhaustive Cryptanalysis of the NBS Data Encryption Standard[J]. Computer, 2006, 10(6):74-84.
- [75] Mcqueen G, Thorley S. Are Stock Returns Predictable? A Test Using Markov Chains[J]. Journal of Finance, 1991, 46(1):239-263.
- [76] Zhang G, Zhang X, Feng H. Forecasting financial time series using a methodology based on autoregressive integrated moving average and Taylor expansion[J]. Expert Systems, 2016, 33(5):501-516.
- [77] Baldauf B, Santoni G J. Stock price volatility: Some evidence from an ARCH model[J]. Journal of Futures Markets, 2010, 11(2):191-200.

致 谢

不知不觉当中，我已经在兰州理工大学求学近三个年头，在兰州理工大学攻读硕士学位期间，我无论在学术研究上，还是在生活能力上都有了非常大的进步。在这里，我学会了很多，也成长了很多。在论文完成之际，我想要真诚的向曾经帮助过我的人道一声感谢！

首先，我要特别感谢我的母校兰州理工大学，以足够大的包容性接受了我这个来自专科学校的学生，让我有机会再一次步入高校的大门继续自己未完成的求学梦想，让我对生活以及教育有了更进一步的认识。其次，要感谢我的导师王燕教授，在老师的谆谆教导之下，自己从一个学术小白不断的成长为一个对于软件工程一数据挖掘方面有见地的学生，在科研过程中对自己启发最大的是作任何事情都要有自己的观点和看法，不能人云亦云。在三年的研究生学习生涯中，从对科研的一知半解到发现自己感兴趣的方向，以及从研究中发现不断尝试解决问题的过程中，在王燕老师的帮助与指导下，自己不仅学会了独立思考，而且也掌握了更多解决问题的办法。王燕老师以其自身渊博的知识，严谨的治学态度，敏锐的学术洞察力，不断影响着我的学术成长。在我论文投稿期间，王老师也给予了我很多建设性的建议和意见，完成了学校的论文发表要求。老师对于我的影响远不止文中所表达的这些，我会继续努力，不负老师的教诲。

再次，感谢计通院各位任课老师在课堂上的传道、受业、解惑，感谢 C207 实验室的每一位成员，感谢他们在学习生活中带给我的帮助，能够和各位同门在实验室进行科研探讨，于我而言也是人生中一段不可磨灭的美好回忆。感谢三年来在生活中给予我理解与包容的舍友们，让我能够有一个温馨的宿舍环境来缓解自己每天的疲惫。感谢我的女朋友李洋洋同学，在我读研究生这三年来，一直不离不弃的帮助与鼓励，让我能够全身心投入到学习与科研中。

最后，我要满怀愧疚之情感谢我的父母。无论我求学多久，都一如既往的选择支持我，回想自己二十多年的求学经历，不管我在生活和学习上遇到什么困难，他们都选择站在我的背后和我一起面对，自己曾以为的“靠山”也随着岁月的流逝慢慢变老，看着父母日渐苍老的容颜，自己内心满是亏欠，只有更加刻苦学习，努力工作，才能无愧于父母的无私付出，无愧于自己的执着与努力。此外，我还要特别感谢各位百忙中对我论文进行评审的专家和老师，感谢你们的辛勤工作和指导帮助！

附录 A 攻读硕士学位期间发表的学术论文

- [1] 王燕,郭元凯.改进的 XGBoost 模型在股票预测中的应用[J].计算机工程与应用.
(已录用)
- [2] Yan Wang, Yuankai Guo.Forecasting Method of Stock Market Volatility in Time
Series Data Based on Mixed Model of ARIMA and XGBoost[J].China
Communications. (已录用)