分类号：F810　　学校代码：10697

密　级：公开　　学　　号：201931207

西北大学
Northwest University

专业学位硕士论文

SPECIALIZED MASTER'S DISSERTATION

基于 CNN-LSTM 的股票价格预测及量化选股研究

学科名称：金融学

专业学位类别：金融硕士

作者：李晨阳　　指导老师：徐璋勇 教授

西北大学学位评定委员会

二〇二一年六月

Reasearch on Stock Price Prediction and Quantitative Stock Selection Based on CNN-LSTM

A thesis submitted to

Northwest University

in partial fulfillment of the requirements

for the degree of Master

in Finance

By

Chenyang Li

Supervisor: Zhangyong Xu

June 2021

摘要

随着我国资本市场的快速发展和居民收入水平的逐步提升，股票市场吸引了越来越多的投资者参与其中，特别是在 2020 年新冠疫情影响下，为提振经济实行的较为宽松的货币政策令股票市场表现不俗，股市吸引了大量新增投资者。数据显示，截至 2020 年 12 月，中国 A 股投资者共计 1.77 亿户，户均市值达到 45 万，全年新增投资者 1802.11 万户，同比增加了 36.02%。股市的参与人数众多说明市场交易活跃，但是，个人投资者在交易中存在非理性投机倾向，盲目地追涨杀跌不利于股票市场的健康发展。倘若能够利用技术手段对股票价格走势进行预测，同时为投资者提供投资参考建议，那么此举将会促进资本市场的良性发展并提升投资者的收益水平，具有较强的理论和现实意义。

本文在股票价格趋势预测中引入深度学习神经网络算法，将两种原理不同的神经网络架构 CNN 和 LSTM 相结合来对股价涨跌进行预测。首先，对上证 50 指数成分股自 2011 年至 2020 年的日频原始交易数据进行数据预处理，采用滑窗滚动方式生成数据样本来训练神经网络模型，通过对不同输入变量下模型预测准确度之间相互比较来确定本文研究模型的输入变量，最终得到包括个股基本交易指标、上证 50 指数指标和个股技术指标在内的共计 20 个特征因子。其次，设置多个对照实验组，通过改变模型的内部参数组合，在研究不同参数对模型预测效果影响的同时挑选出预测准确度最佳的参数组合，并将该参数组合下的 CNN-LSTM 模型同 BP 模型、CNN 模型以及 LSTM 模型的预测效果进行对比。最后，对回测期间的市场状态进行牛熊市的划分，利用 CNN-LSTM 模型预测的涨跌概率作为选股条件构建量化选股策略并进行回测，并对牛熊市状态下的策略的绩效进行了分析。

本文实证研究表明：（1）CNN-LSTM 预测模型在上证 50 股指成分股涨跌预测中有较好的表现，相比其他神经网络架构的单模型，将两种架构相结合能够提升模型预测能力。（2）基于 CNN-LSTM 模型构建的量化选股策略能够在不同市场状态下都取得显著高于基准的收益，说明将模型输出涨跌概率视作选股因子，可以获取超额收益，涨跌概率可以作为量化选股的有效方法，对投资者在选股时具有投资参考价值。（3）我国股市尚未达到弱势有效市场。由于本文研究的模型输入变量中包含了个股交易数据以及相关

技术指标，依据此模型构建的交易策略能够在市场中获取超额收益，说明股票价格并未完全反映所有历史信息，通过对股价历史数据的挖掘和处理可以在市场中获利。

关键词：股价趋势预测，神经网络模型，上证 50 指数，量化选股

ABSTRACT

With the rapid development of China's capital market and the gradual improvement of residents' income, the stock market has attracted more and more investors to participate in it. Especially under the influence of the COVID-19 in 2020, the relatively loose monetary policy to boost the economy has made the stock market perform well, and the stock market has attracted a large number of new investors. Data show that as of December 2020, China A share investors total 177 million households, the average market value of the household reached 450,000, the new investors in the whole year 18021,100 households, an increase of 36.02%. The large number of participants in the stock market indicates that the market trading is active. However, individual investors have the tendency of irrational speculation in the trading, and blindly chasing after the rise and killing the fall is not conducive to the healthy development of the stock market. If we can use technical means to predict the trend of stock prices and provide investment suggestions for investors, then this will promote the sound development of the capital market and improve the return level of investors, which has a strong theoretical and practical significance.

In this paper, deep learning neural network algorithm is introduced into the trend prediction of stock price, and two neural network architectures with different principles, CNN and LSTM, are combined to predict the rise and fall of stock price. First of all, to the Shanghai 50 index from 2011 to 2020, the frequency of the original transaction data for data preprocessing, the sliding window scroll way to generate the data samples to train neural network model, and through the comparison on the accuracy of the prediction model under different input variables to determine the final input variables in this paper, we study model, are including basic trading stocks index, the Shanghai 50 index, index and technique index stocks, a total of 20 characteristic factors. Secondly, several control experimental groups were set. By changing the internal parameter combinations of the model, the influence of different parameters on the

prediction effect of the model was studied, and the parameter combination with the best prediction accuracy was selected, and the prediction effect of the CNN-LSTM model under the parameter combination was compared with that of BP model, CNN model and LSTM model. Finally, the market state during the backtest period is divided into bull and bear markets, and the rise and fall probability predicted by CNN-LSTM model is used as the condition for stock selection to construct quantitative stock selection strategy and conduct backtest. Moreover, the performance of the strategy under the bull and bear market conditions is analyzed.

The empirical conclusions of this paper prove that: (1) CNN-LSTM prediction model has a good performance in predicting the rise and fall of Shanghai Stock Index components. Compared with the single model with other neural network architectures, the combination of the two architectures can improve the prediction ability of the model. (2) The quantitative stock selection strategy based on CNN-LSTM model can achieve significantly higher returns than the benchmark under different market conditions, indicating that the model output probability of rise and fall can be regarded as a stock selection factor to obtain excess returns, and the probability of rise and fall can be used as an effective method of quantitative stock selection, which has investment reference value for investors in stock selection. (3) China's stock market has not yet reached the weak efficient market. Since the input variables of the model studied in this paper include individual stock trading data and related technical indicators, the trading strategy based on this model can obtain excess returns in the market, indicating that stock prices do not fully reflect all historical information, and profits can be made in the market by mining and processing the historical data of stock prices.

Keywords: Stock Price Trend Prediction, Neural Network Model, SSE 50 index, Quantitative Stock Selection

目录

摘要 I

ABSTRACT III

第一章 绪论 1

1.1 研究背景及意义 1

1.1.1 研究背景 1

1.1.2 研究意义 2

1.2 研究思路与方法 3

1.2.1 研究思路 3

1.2.2 研究方法 3

1.3 研究内容与框架 4

1.3.1 研究内容 4

1.3.2 研究框架 5

1.4 论文的主要贡献 7

第二章 文献综述 8

2.1 关于股票价格可预测性及预测方法的研究 8

2.1.1 关于股票可预测性的研究 8

2.1.2 关于股票价格预测方法的研究 9

2.2 关于利用深度学习预测时间序列的研究 10

2.2.1 关于深度学习预测股票价格的研究 10

2.2.2 关于深度学习预测其他时序问题的研究 12

2.3 文献评述 12

第三章 相关概念与理论基础 14

3.1 相关概念界定 14

3.1.1 机器学习的概念 14

3.1.2 深度学习的概念 14

3.2 神经网络基本理论 15

3.2.1 神经网络的基本单元 15
3.2.2 神经网络的网络结构 16
3.2.3 神经网络的训练过程 17
3.2.4 神经网络的特点及应用 19
3.3 深度学习的神经网络模型 20
3.3.1 CNN 神经网络算法原理 20
3.3.2 RNN 神经网络算法原理 24
3.3.3 LSTM 神经网络算法原理 25
3.4 股价预测中引入神经网络的可行性分析 27
3.5 本章小结 28
第四章 股价预测模型构建 29
4.1 实验平台及数据介绍 29
4.2 数据预处理 30
4.2.1 数据预处理方法 30
4.2.2 输入变量特征因子构建 30
4.2.3 样本标记及数据集划分 36
4.3 模型设计与构建 38
4.3.1 模型的评价指标 38
4.3.2 初始模型的网络结构设置 39
4.3.3 初始模型训练及预测效果展示 42
4.4 模型结构及参数优化 44
4.4.1 网络参数对模型的影响 45
4.4.2 优化参数对模型的影响 49
4.4.3 正则化参数对模型的影响 51
4.4.4 最终模型预测结果展示 52
4.5 CNN-LSTM 模型同其他神经网络模型对比 54
4.6 本章小结 56
第五章 量化选股策略构建 58

5.1 策略思想 58
5.2 策略评价指标 59
5.3 策略回测效果分析 60
5.3.1 牛熊市划分 60
5.3.2 策略牛熊市表现 61
5.4 本章小结 68
研究结论与展望 69
参考文献 71
致谢 75

第一章 绪论

1.1 研究背景及意义

1.1.1 研究背景

在 2016 年的世界围棋大赛中，谷歌 DeepMind 公司开发的 AlphaGo 击败了韩国棋手李世石，这一事件使得神经网络和机器学习受到人们的广泛关注，因为 AlphaGo 的架构正是基于多层神经网络搭建的深度学习模型。深度学习的核心——神经网络具有通用性强、可扩展性高等优点，这使得它成为解决大型、复杂机器学习任务的理想选择，比如对海量图片视频进行分类、支持语音识别服务和汽车驾驶辅助系统，而相关成果也已经部分实现了商用化，比如谷歌的 image 图像识别、苹果的 siri 和特斯拉的自动驾驶。

近年来，随着国内量化交易和计算机技术的发展，伴随着学科交叉现象的普及，越来越多的学者在开展研究时尝试引入人工智能、机器学习等一系列计算机学科理论方法，最终目的在于研发出新的能够预测股票价格并指导量化交易的方法。在此背景下，神经网络成为了一个主要的研究途径，按照网络架构可以将神经网络分为人工神经网络（ANN）、卷积神经网络（CNN）和循环神经网络（RNN）及长短期记忆神经网络（LSTM），其中，ANN 还包括了前馈神经网络，即 BP 神经网络。

在神经网络流行起来之前的很长一段时期内，人们在预测股价时大都热衷于采用 GARCH、ARIMA 等线性回归方法对股价进行建模，并且取得了较为不错的预测效果，传统的线性模型拥有诸多优点，比如易于使用，原理简明，对股价有较强的解释性，但是，由于现实中股票价格受到多个方面和因素的影响，并且时常发生的灰犀牛、黑天鹅事件也会影响到股市行情，这些都表明股票市场不仅具有线性特征，也具有非线性特征，用传统的线性模型研究股票价格和股票市场仍存在不妥之处。神经网络本身是一个非线性模型，网络结构的复杂堆栈让神经网络相较于线性模型能够更好的识别股票市场的非线性特征，从而更准确的对股票价格进行建模型及预测，随着神经网络深度的增加，模型对输入数据的特征组合和特征提取能力也在不断增强，因此，使用神经网络模型研究存在大量噪音的非线性金融时间序列具有一定的优势。

基于以上背景，本文将运用神经网络算法训练机器学习模型来对股票价格走势进行

预测，并构建相应的预测模型，从模型预测准确度指标出发，将 CNN-LSTM 模型和其他神经网络模型预测效果做比较，随后根据 CNN-LSTM 模型预测结果在上证 50 指数成分股中选股并构建投资组合，形成相应的交易策略，通过对预测模型的效果评估和交易策略的收益评价来开展实证研究，突出 CNN-LSTM 模型在股票市场中选股的可行性及适用性。

1.1.2 研究意义

（1）理论意义

在现有的使用非线性方法对股价预测的研究中，使用最多的是传统的机器学习方法，如朴素贝叶斯、随机森林、支持向量机等方法，这些机器学习方法不仅用在金融时间序列的研究中，同时也广泛用于对其他时间序列问题建模，是最常用的机器学习方法，作为机器学习的分支，基于深度学习的神经网络模型则主要用于对图像分类、音频分类以及文本分类，并且在金融时序预测中也以仅是以 BP 神经网络为主。在深度学习的几个常用的神经网络模型中，RNN、LSTM 相比 CNN 更适合对时间序列进行建模，因此大量研究都是围绕 RNN 及其变体 LSTM 开展的，将 CNN 与 LSTM 两种神经网络架构结合起来的研究不多，因此，本文将 CNN-LSTM 神经网络相结合对股价建模并进行趋势判断更具有创新性。此外，由于 CNN-LSTM 神经网络模型输入变量是二维结构，如何将一维时间序列数据二维化以及具体的使用效果同样值得探究。

（2）现实意义

本文将 CNN-LSTM 神经网络预测股价走势与构建量化交易策略相结合，在一定程度上丰富了传统的诸如因子选股等量化选股模型，研究结果在一定程度上证明了股票价格趋势是可以预测的。实证研究中从模型预测效果和策略回测结果两个方面对模型的有效性进行了双重确认，通过对不同神经网络模型之间的对比，突出 CNN-LSTM 模型在我国股票市场的适用性，此外，本文在策略回测中对我国股票市场的牛、熊市区间做了划分，分阶段分析了本文构建的模型在不同时期的预测效果及策略收益表现，对投资者在不同市场状态下的投资决策行为具有指导和借鉴意义。

1.2 研究思路与方法

1.2.1 研究思路

本文将理论和实践相结合，主要研究基于 CNN-LSTM 的股票价格变动趋势预测模型，即对股票上涨下跌的预测，并依据预测模型在不同数据集上的预测结果构建交易策略，通过对模型的预测效果和回测的收益来评价模型的有效性。本文的研究对象为上证 50 指数成分股，原始数据为上证 50 指数以及其成分股自 2011 年 1 月 1 日至 2020 年 12 月 31 日的交易量价数据，即每日个股开盘价、收盘价、最高价、最低价和成交量的“四价一量”信息。首先对数据预处理，将由于停牌、退市等原因导致的缺失数据剔除，然后在个股日频交易数据的基础上计算个股技术指标，将输入的数据维度从“个股四价一量”扩大为“个股四价一量+上证 50 指数四价一量+个股技术指标”，由于研究对象不是单只股票的单一时间序列，因此需要消除不同股票横截面数据的差异，按照归一化原则将输入数据转化为[0,1]之间的无量纲数，接下来以 20 天为滑窗，1 天为步长沿时间轴生成样本，将模型的输入数据结构转化为“20×T”的“图片”，其中 T 代表输入变量的特征因子，随后按照时间窗口滚动划分的形式将 2011-2020 年十年数据划分为 2011-2018、2012-2019、2013-2020 三个数据集分别训练神经网络模型。在此过程中利用 Python 编程语言以及 Keras 机器学习库进行模型的训练、测试以及模型内部超参数的调整优化，最后对 CNN-LSTM 和其他神经网络模型之间的预测效果进行比较分析，并依据模型的预测结果，进行选股及策略回测，从策略的实际收益效果上再次验证 CNN-LSTM 模型在量化选股方面的适用性和优越性。

1.2.2 研究方法

本文主要采用的研究方法有：

（1）实证研究法。本文实证所使用的理论方法是基于 CNN-LSTM 神经网络的深度学习算法，首先对包括神经网络模型原理及训练过程在内的基本理论做了简单介绍和梳理，然后基于上证 50 指数成分股构建股票涨跌预测模型，最后根据模型预测结果设计交易策略并进行策略回测。

（2）对比分析法。本文使用了比较分析的方法，在确定模型输入变量时，通过设置对照实验组确定了 20 个特征因子，在进行 CNN-LSTM 模型的参数优化时，通过设置对

照实验组来挑选出预测效果最好的模型，同时也比较了 CNN-LSTM 模型和 CNN 模型、LSTM 模型以及 BP 模型的在股价涨跌预测时的效果，突出 CNN-LSTM 模型的预测优异性，此外，还对依据 CNN-LSTM 模型构建的选股策略在牛、熊时期的表现进行了对比分析。

1.3 研究内容与框架

1.3.1 研究内容

第一章，绪论。本章首先介绍和梳理目前深度学习在各个领域的广泛应用，作为深度学习的核心，神经网络在处理非线性复杂问题上的优势使之成为一种新的预测股票价格趋势的研究方向，接下来介绍了本文的研究意义、研究思路，以及在研究中使用的方法，并对研究内容和框架做出总结，最后对本文的主要贡献进行了介绍。

第二章，文献综述。本章首先搜集近些年在股票价格预测方面的国内外研究文献，按照研究中使用的方法大致将其分为计量模型和机器学习模型，随后对机器学习模型中有关深度学习的神经网络在股票价格预测方面的研究做了详细梳理，同时也梳理了一些其他学科利用深度学习研究时序问题的实践，最后从股票价格可预测性和深度学习预测方法两个方面对已有研究做出总结归纳。

第三章，神经网络相关概念与理论基础。本章主要从概念和原理的角度，梳理和介绍了机器学习、深度学习的概念以及神经网络的算法原理，对神经网络模型训练过程中出现的拟合不足、过拟合等问题的解决措施进行了介绍，从理论方面、技术方面和前人研究三个角度分析了股票价格预测中引入神经网络的可行性。

第四章，股价预测模型构建。本章主要工作是基于 CNN-LSTM 神经网络构建股价涨跌预测模型，首先对构建模型使用的实验平台和实验数据进行介绍，然后对数据预处理并生成数据样本及标签，随后分数据集训练神经网络预测模型并进行超参数的调整优化，在此基础上确定了最终模型结构及参数。最后将 CNN-LSTM 架构的股价预测模型同 CNN、LSTM、BP 三种模型从预测效果的角度进行比较和评价。

第五章，股票量化选股策略。本章主要依据第四章构建的模型预测结果，在上证 50 指数成分股中选股并进行策略回测，首先确定策略思想和评价指标，然后将模型预测结果结合选股条件生成买卖信号在聚宽量化交易平台进行回测，最后利用非参数法对回测

期的市场状态做牛熊市划分，并分牛熊市对策略回测结果进行分析。

第六章，研究结论与展望。本章对全文研究过程进行回顾，对研究所得出的结论进行总结，对研究过程中存在的不足进行反思归纳，并对未来研究需要改进的方向做出展望。

1.3.2 研究框架

本文的研究框架如图 1-1 所示：

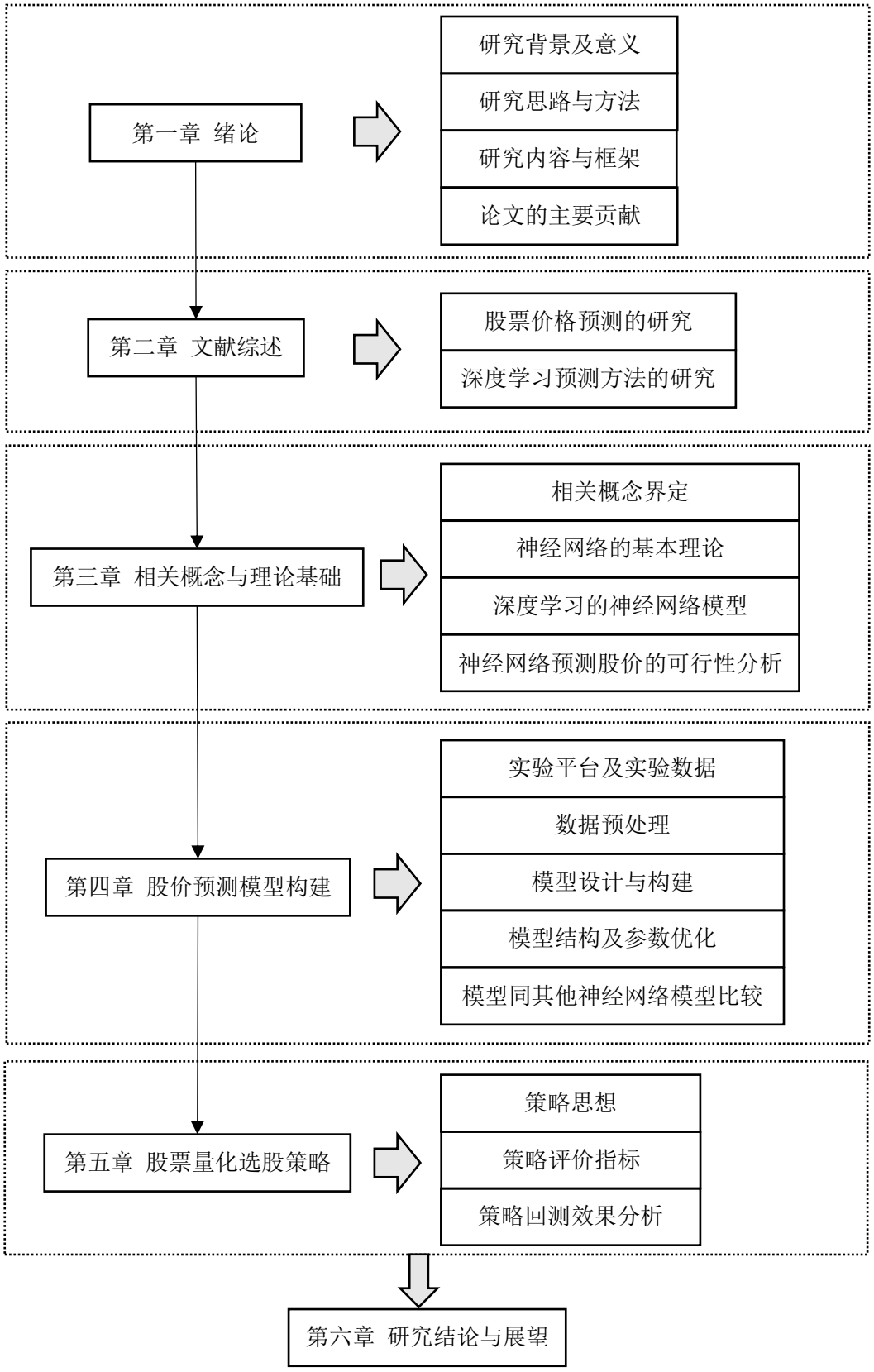


图 1-1 研究框架图

1.4 论文的主要贡献

本文的主要贡献主要有以下两点：

第一，将 CNN-LSTM 神经网络引入股价趋势预测中，并对预测模型做出改进。现有的股价涨跌预测模型的大都单独研究一种机器学习方法或将几种机器学习方法相比较，本文构建模型时将两种原理不同的神经网络模型相结合，有效利用了卷积神经网络的特征提取优势和长短期记忆神经网络的时序处理优势，相比于单模型，CNN-LSTM 双模型在一定程度上提升了股价涨跌的预测表现，并且基于 CNN-LSTM 模型构建的选股策略在牛市和熊市行情下都能获取较为不错的投资收益。

第二，在策略回测时将预测模型的输出作为选股的条件，丰富了量化选股因子。传统量化交易在确定选股标准时，多使用规模类因子、估值类因子、成长类因子、盈利类因子等标准在股票池中选取标的股票构建投资组合，这些因子能够在市场中获取较高的超额收益，本文构建的股价涨跌预测模型输出的是个股上涨下跌的概率，具体的个股涨跌概率值也可以视作选股标准，基于涨跌概率构建的选股交易策略在回测期间能够持续获取高于当年市场基准的收益，说明利用涨跌因子选股是有效的，可以作为选股标准在市场中获取超额收益。

第二章 文献综述

2.1 关于股票价格可预测性及预测方法的研究

2.1.1 关于股票可预测性的研究

尤金·法玛（Eugene Fama）在 1970 年提出了有效市场假说（EMH）[1]，法玛认为，市场在充分交易、无摩擦且一切市场信息都是充分有效传递的假设下，股票价格准确且充分地反映了所有信息，但是有效市场假说是以完美市场为前提的，该假设过于理想化，如交易无摩擦、充分竞争、无信息成本等都是很难实现的，因此，实际的股票市场并不是有效的，通过对信息的收集研究可以获取超额利润，此外，大量的实证研究结论表明，我国股票市场尚未达到有效市场状态，股票价格在一定程度上是可预测的。

周爱民（1999）对沪股、港股指数建立自回归模型，对主要指标进行协整性检验，指出技术分析指标能够在一定精度的前提下解释股指的变动，因此技术分析是有效的，并得出上海股市的弱有效性是在逐步增强的[2]。陈春晖和曾德明（2006）在对上证 180 指数进行预测分析时引入了分形理论，发现利用当天与前一天的多重分析普参数的变化可以一定概率预测第二天股价的涨跌，并且该结论在上证 A 股指数、上证基金指数、深证综合指数、深圳成分指数、深证 A 股指数、深证基金指数中都成立[3]。苏治、方明和李志刚（2008）基于 STAR 与 ANN 模型对上证 180 股指价格趋势进行预测并构建投资策略，得出结论为，相比文中其它模型，ANN 模型能更好地刻画上证 180 指数的非线性特征，有较好的样本外预测能力，同时发现我国证券市场未达到弱势有效[4]。柴宗泽和姚长辉（2009）利用一阶 VAR 方法研究沪、深两市的 A 股简单算术平均指数回报率的可预测性，分析结果表明上述指数的一个月和三个月回报率都有显著的可预测性，并且发现市场关于未来红利的信息和关于未来回报率的信息具有很强的正相关性[5]。杨光艺（2018）使用 IVX-WALD 方法对沪深 300 全收益超额收益率的可预测性进行了稳健性检验，证明与资金利率水平相关的债券到期收益率和短期收益率对股票市场具有可预测性，并且发现考虑情绪因子的变量在牛市中具有更好的预测效果[6]。

2.1.2 关于股票价格预测方法的研究

对于如何预测股价，选择何种方法来对股票价格序列进行建模，也有学者对股价预测方法开展了大量实证研究。传统的股票价格预测方法大都以统计学原理为依据，将股票历史价格信息视作时间序列，通过建立时序模型拟合股价变化来预测股票价格未来走势，如自回归差分移动平均模型（ARIMA）、自回归条件异方差模型（ARCH）和广义自回归条件异方差模型（GARCH）。吴玉霞和温欣（2016）对华泰证券 250 个交易日的股票收盘价进行建模，并且将拟合的有效模型用于价格的短期预测，证明 ARIMA 模型具有较好的适用性，但该模型在计算预测值时存在误差累积的缺陷，模型对原始数据的选取依赖度较高[7]。徐枫（2006）选取南方航空和东方航空作为研究对象，通过自回归过程得到两只股票价格预测的完整模型，研究发现 GARCH 模型在预测股票价格时存在短期记忆性[8]。许舒雅和梁晓莹（2019）将上述两种模型相结合对宇通客车的未来 20 日股价做出预测并进行了残差检验，实证结果表明 ARIMA-GARCH 模型在预测股票价格上有较好的表现[9]。石鸿雁、尤作军、陈忠菊（2014）利用小波分析对上证指数收盘价序列进行分解与重构，在低频和高频序列上分别构建 ARIMA 模型并将预测结果合并，结果显示引入小波分析能够提升 ARIMA 模型在预测上证指数中的精确性[10]。

然而基于统计学方法构建的预测模型大都是线性回归模型，通常只能捕捉股票价格信息的线性关系，而影响股价的因素丰富多样，徐忠兰、许永龙、赵亮（2004）运用回归分析方法对上证指数建模，分析归纳出了影响我国股价趋势变动的几个因素，包括宏观经济状况、对外贸易状况、国内市场环境和货币汇率政策等[11]。吴玉桐和梁静国（2008）认为股票市场反映了投资者的心理预期，股价和投资者心理预期之间存在相互作用的关系 [12]。由此可见，宏观货币政策、公司经营状况和投资者的行为都会对股票价格产生不同程度的影响，只通过线性回归模型来研究股价变动存在一定的局限性。此外，Shleifer 和 Summers（1990）也认为情绪会干扰投资者的投资决策[13]。

伴随着金融理论的发展和大数据的到来，与机器学习相关的模型逐步成为股票价格预测的主流分析方法。机器学习由 Arthur Samuel 于 1959 年首次提出[14]，其原理是通过某种映射把输入量转换为输出量，在此过程中试图寻找出股票历史价格变化的规律用于回归和分类，即对股票价格时间序列的预测和价格涨跌趋势的判断。目前在股价预测领域常用的机器学习模型有支持向量机（SVM）、支持向量回归（SVR）、随机森林、朴

素贝叶斯和人工神经网络（ANN）。杨新斌和黄晓娟（2010）认为股价时间序列具有复杂的非线性演化行为，采用基于支持向量机和留一法的前向浮动特征筛选算法建立了股价预测模型，实证结果表明 SVM 模型可用于股价中、长期预测和分析[15]。Kim（2003）选取了 12 个技术指标并利用 SVM 来对韩国 KOSPI 指数进行预测，通过实证验证了 SVM 在金融时序预测上较 BP 神经网络方法效果更好[16]。Huang（2012）基于遗传算法改进的 SVR 对台湾股票市场建立了选股模型，结果表明通过遗传算法进行特征选择和参数调优可以提升 SVR 模型的预测效果[17]。张潇和韦增欣（2018）基于价值成长投资策略选取了 16 个股票技术指标作为特征，利用随机森林模型预测股票涨跌并进行策略回测，得到了较高的超额收益[18]。吴微、陈维强和刘波（2001）利用 BP 神经网络模型对沪市综指涨跌做预测，预测准确度达到 70%，初步说明 ANN 模型在股票价格预测上的可行性和实用性[19]。Zhang 等人（2003）通过研究发现，神经网络模型在预测非线性时序问题上效果比 ARIMA 模型更好[20]。刘恒和侯越（2018）利用贝叶斯正则化算法改进了 BP 神经网络，提升了模型的泛化能力，一定程度上改善了 BP 神经网络易陷入局部最优值的问题[21]。常松和何建敏（2001）提出了一种将小波包分解和神经网络相结合的 WPNN 模型，并对上证综指价格进行预测，结果表明小波包分解能显著提升神经网络模型的预测精度[22]。崔建福和李兴绪（2004）将 GARCH 模型和 BP 神经网络模型进行比较，发现从非线性系统的角度对股价建模要优于从非平稳时间序列的角度出发进行建模[23]。

2.2 关于利用深度学习预测时间序列的研究

2.2.1 关于深度学习预测股票价格的研究

深度学习是机器学习的一种方法，最早由多伦多大学的 Geoffrey Hinton 和 Ruslan Salakhutdinov 提出[24]，“深度”意为含有多隐层节点的神经网络结构，是和浅层机器学习相对的概念，早期的机器学习模型如支持向量机、逻辑回归都可以理解为浅层机器学习模型，这类模型在结构上相当于只含一层或不含任何隐层节点。相比于浅层机器学习模型，深度学习可以更好的提取原始数据的特征结构，具有较好的泛化能力，深度神经网络是深度学习实现的途径，比较有代表性的深度神经网络有 CNN、RNN 和 LSTM。其中，LSTM 是 RNN 的一种变体，相比 RNN，LSTM 具有记忆性，能够检测数据中的

长期依赖关系，因此更适用于对股票价格时间序列进行预测。

在股票价格预测领域，已有大量学者对于深度学习的应用进行了研究。孙瑞奇（2016）对标普 500 指数、道琼斯指数和上证指数的价格趋势进行了预测，对比了 BP 神经网络、RNN 神经网络和 LSTM 神经网络的预测效果，并基于拟牛顿法原理调节神经网络模型中的学习率来避免模型陷入局部最优解的困境，实现了 LSTM 模型算法方面的优化改进[25]。周凌寒（2018）将投资者情绪纳入模型的输入数据中，在 LSTM 神经网络模型中加入了情感特征因素和基本面特征因素，证明了多种信息源能有效提升 LSTM 神经网络模型的预测准确性[26]。彭燕、刘宇红和张荣芬（2019）利用 LSTM 模型预测苹果公司的收盘价，通过确定合适的 LSTM 层数和前馈网络层中隐藏神经元个数，显著提升了模型预测的准确率[27]。曾安和聂文俊（2019）充分考虑了时间序列向前、向后两个方向的上下文关系，提出了一种基于深度双向 LSTM 的神经网络预测模型，并且在标普 500 数据集上降低了预测误差[28]。贺毅岳、李萍和韩进博（2020）在 LSTM 神经网络模型的数据预处理方法上，引入了自适应噪声完备集合经验模态分解（CEEMDAN），对股指进行分解、重构，结果表明，改进后的方法在预测股市指数上有更低的预测误差和滞后性，在构建量化择时策略领域有较好的应用前景[29]。Zhou（2019）构建了一个基于 EMD 和 FNN 的两阶段模型来预测上证综指、纳斯达克指数和标普 500 指数日收盘价,利用信号分解的 EMD 经验模态分解方法，将分解出来的信号分量（IMF）作为输入变量训练 FNN 模型，模型从预测效果和回测交易表现都很优秀[30]。CNN 神经网络从 20 世纪 80 年代起一直用于图像识别领域，近年来，随着计算机算力的提升，大数据时代到来带来了更多的可训练数据，CNN 神经网络也被应用到对股票价格时间序列的预测研究中。陈祥一（2018）基于卷积神经网络模型来预测沪深 300 指数的涨跌，并将 CNN 和逻辑回归、支持向量机、决策树模型比较分析，证明 CNN 模型的预测效果优于传统机器学习模型[31]。黄志辉（2019）通过增加一个技术指标维度将一维时间序列二维化，并采用窗口滚动方式划分数据集来训练模型，基于 CNN 构建的沪深 300 成分股量化选股策略具有较高的夏普比率和较低的回撤，证明了卷积神经网络在量化选股上有优良的适用性[32]。文宇（2018）结合了 CNN 和 LSTM 两种模型，对螺纹钢期货的价格变化做短期预测，结果表明 CNN-LSTM 神经网络能有效预测期货短时价格变动[33]。Kim（2019）将 LSTM 与 CNN 相结合，对标普 500 指数 ETF 的时间序列和 K 线烛形图像提取特征并预测股票

价格，实证发现将相同数据中的时间特征和图像特征组合可以有效减少预测误差[34]。

2.2.2 关于深度学习预测其他时序问题的研究

深度学习除了应用于股票价格、期货价格等金融时间序列预测领域，在其他时间序列分析预测中也有广泛应用和研究。罗文慧、董宝田和王泽胜（2017）将 CNN 模型用于交通流特征提取，通过对 G103 国道短时交通流的预测，证明了 CNN-SVR 模型能够在交通流波动较大情况下仍保持精准的预测效果[35]。陈亮、王震和王刚（2017）使用长短期记忆神经网络对某省电力公司电力负荷时序数据进行回归预测，研究发现 LSTM 模型可以有效预测短期电力负荷变化，有利于保障电力稳定供应[36]。李梅、李静和魏子健等人（2018）提出了运用 LSTM 模型预测短期客流量的方法，并基于上海轨道交通客流量数据开展研究，该方法相比传统 MLR 线性模型和 BP 神经网络具有更好的特征提取能力，提高了客流量预测的准确性[37]。罗向龙、李丹阳等（2018）提出一种将 K-最邻近（KNN）方法和 LSTM 模型相结合的方法，综合考虑了道路网中交通流的时空相关性和交通流序列的时间依赖关系，改进的方法将交通流的平均预测精度提高了 12.28%[38]。王庆荣、李彤伟和朱昌锋（2020）引入了小波分解方法对交通流数据进行去噪，降低了原始数据对模型的干扰，显著提升了短时交通流的预测准确率[39]。以上这些研究充分说明了深度学习在时间序列预测及趋势判断上有广泛的应用前景。

2.3 文献评述

通过梳理近些年来国内外对于股票价格可预测性以及相关预测方法、模型的应用研究，不难发现，在中长期上，股票价格可以通过基本面分析等相关理论进行价格和走势的预测，但更多的是一种站在价值投资角度的定性的趋势判断。在中短期内，股票价格也是可预测并且能够量化的，这使得对股票价格中短期进行预测更具有实际意义，而预测方法也经历了从早期的计量学方法发展到机器学习方法，再到当前备受关注的深度学习方法，以 CNN 和 LSTM 为典型代表。此外，大量研究围绕深度学习在股票价格预测领域所关注的问题都是如何通过现有手段来提升模型的预测效果以及模型的泛化能力。针对所关注的问题，专家和学者的研究主要集中在三个方面：

第一，对现有的预测模型进行算法方面的优化，包括优化模型层次架构以及参数调优，又或是将两种或是几种模型相结合，充分发挥不同模型的比较优势，提高模型在数

据集上的实际表现；第二，对模型输入数据的选择上，引入其他相关指标，如技术指标、情绪指标、政策指标、宏观经济指标，而不是仅仅用“四价一量”（开盘价、收盘价、最高价、最低价和成交量）的量价信息作为模型的输入数据；第三，对模型输入数据的预处理上，引入经验模态分解算法(EMD)和小波变化等信号去噪技术对输入数据进行降噪、分解和重构，过滤掉时间序列上的“噪声”，提高信噪比，这种方法主要用在对单一时间序列进行回归预测上，在对股票价格涨跌这种分类预测上使用较少。在具体的预测模型中，重指数轻个股是一个普遍现象，大部分模型利用深度学习在分类问题上的优势，对某一股票指数进行价格趋势的预测，少部分研究对大量个股的涨跌进行预测，此外，由于股票价格序列在牛、熊市时期的不同阶段有不同的特点，现有相关研究也很少对股市所处的时期做区分。

基于目前已有研究中所存在的问题和改进的方向，本文将结合 CNN 和 LSTM 两个模型的各自优势，在输入数据的选择上添加多个技术指标来对上证 50 指数成分股进行建模，具体模型结构借鉴已有模型的架构设计，并在网络结构部分做调整和优化，最终实现对股价变动趋势的预测和判断，即预测个股的涨跌，此外，根据模型预测结果在聚宽量化交易平台构建选股策略，通过回测收益进一步验证 CNN-LSTM 模型在实际投资中的适用性。

第三章 相关概念与理论基础

3.1 相关概念界定

3.1.1 机器学习的概念

卡内基梅隆大学的 Tom Mitchell 教授将机器学习定义为“如果一个程序可以在任务 T 上，随着经验 E 的增加，效果 P 也可以跟着增加，那么称这个程序可以从经验中学习并不断成长改进[40]。”

机器学习是一门能够让计算机从海量的数据中学习的计算机科学，根据机器学习是否需要在人类的监督下训练，可以将其划分为监督式学习、无监督式学习和强化学习三类。监督式学习是最常用的一种机器学习，在监督式学习时，提供给模型的训练数据必须是经过标记的，这些数据会告诉计算机在特定情况下的正确输出，并且希望模型在遇到新的样本时也能产生一个准确的预测结果。无监督式学习的训练数据是未经标记的，在训练过程中，机器会主动学习数据的特征并将数据分为若干类别，相当于形成了未知的标签，也正是这一特点让无监督式学习被广泛用于数据挖掘领域，人们希望借助机器学习模型的帮助在大量无标签数据中发现某些规律。强化学习与监督学习、无监督学习的区别最明显，在强化学习中，提供给模型的数据没有数量上的要求，因此无需大量的训练数据，强化学习的学习系统能够观察环境，做出选择，通过不断与变化的环境互动和试错，利用奖惩机制实现更新策略的过程。

3.1.2 深度学习的概念

深度学习是指通过一系列的非线性变换对数据样本的特征进行抽象提取的算法过程，该方法属于机器学习的一个分支。传统的机器学习方法在特征提取上需要人工参与，但在处理复杂问题时人工构建特征集合低效又耗时耗力，因此，得益于能够自动将简单的特征组合为多层复杂的特征，深度学习成为了一种更有效的机器学习方法。

深度学习也可理解为利用多隐藏层的神经网络进行机器学习，其基本原理类似于一个 n 层网络结构，数据从输入层进入网络，经过 n 层网络后输出结果，在这个过程中，每层网络的输出作为下一层的输入，并且每层网络都对上一层数据信息进行特征提取，

高层的网络表达由低层表达组合而成，通过这种方式实现对输入数据的分级表达。由于深度学习的神经网络隐藏层数量很多，从而能够实现无限逼近任何非线性函数，便于对非线性问题进行拟合研究。近年来，深度学习在很多研究领域突破了传统机器学习的瓶颈，不断推动着人工智能的发展。

3.2 神经网络基本理论

3.2.1 神经网络的基本单元

上一节对机器学习和深度学习的概念做了释义，本节将对深度学习的核心——神经网络的相关理论做介绍和梳理。神经网络是一种由神经单元通过互联的方式组合而成的结构，并且能够模拟生物的神经系统对外界做出反应和交互，因此，神经网络具有较强的适应性。

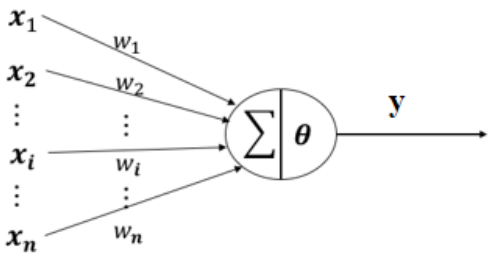


图 3-1 M-P 神经元模型

Warren McCulloch 和 Walter Pitts 受到生物神经元的启发，于 1943 年提出了一种能够模仿人类大脑工作原理的“M-P 神经元模型”[41]，该结构奠定了神经网络发展的基础，神经元也叫神经网络节点，同时也是组成神经网络的最基本的单元。

如图 3-1 所示，在神经网络中，上一层神经元和当前神经元之间是相互连接的。左侧x_i为某一时刻t来自第i个神经元的输入信号，每个输入信号乘以各自权重w_i后传递给当前神经元，当前神经元在收到所有上一层传入的信号后，会对所有信号求加权和，并将信号加权和同当前设定的阈值θ相比较，若超过阈值，则判定当前神经元处于激活状态，通过激活函数$f(x)$处理信号再输出给下一层神经元。以上表述可以将神经元的状态用公式表达为：

$$y(t) = f(\sum_{i=1}^{n} w_i\, x_i(t) - \theta) \qquad (3.1)$$

在上述的信号传递过程中，加入激活函数的目的是对模型做非线性化变换，将原本

的线性关系转化为非线性关系，常用的激活函数有 ReLU 函数、sigmoid 函数和 tanh 函数，三种函数图像如图 3-2 所示。

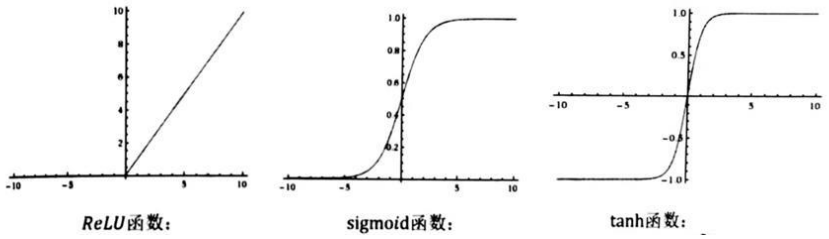


图 3-2 三种常用的激活函数

显然，这三种激活函数图像有很大差异因此各自特点也不同，ReLU 函数能够加快神经网络模型收敛速度，sigmoid 函数图像连续、平滑、严格单调，导数就是自身，是一个非常好的激活函数，但是当 x 趋于两侧时，sigmoid 函数导数趋向于 0，因此在模型训练过程中容易发生梯度消失，tanh 函数图像同样连续且平滑，相比 sigmoid 函数收敛速度更快，但同样存在梯度消失的问题。三种激活函数表达式如下所示。

$$\text{relu 函数：} f(x) = \max(x, 0) \tag{3.2}$$

$$\text{sigmoid 函数：} f(x) = \frac{1}{1+e^{-x}} \tag{3.3}$$

$$\text{tanh 函数：} f(x) = \frac{1-e^{-2x}}{1+e^{-2x}} \tag{3.4}$$

3.2.2 神经网络的网络结构

以神经元单元为基础，将多个神经元按照顺序进行堆叠连接，就可以组合成一个人工神经网络。如图 3-3 所示，左侧为一个输入层， 中间为两个隐藏层，右侧为一个输出层，图 3-3 展示了一个三层全连接神经网络，其中，输入层神经元用于接收外部输入数据，隐藏层的作用是从输入数据的特征中抽取更高维度的特征，完成特征变换、处理和传递，输出层将最终结果输出。

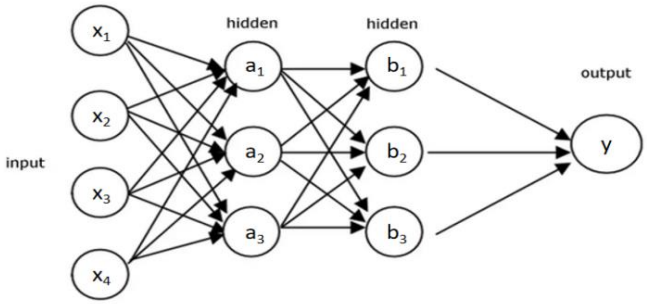


图 3-3 神经网络结构图

而如果想要对复杂信息进行处理和建模，那就要用到更多的神经元组成的层数更多的深度神经网络，进而实现对数据信息的深度学习。

3.2.3 神经网络的训练过程

神经网络在信号传递的过程中通过反复迭代来更新各项参数最终完成模型的训练，信号传递过程实质上是输入信号X到输出信号Y的映射过程，而映射函数的系数就是我们所要训练的神经网络参数W，神经网络模型的训练围绕得到最优的参数W展开。

神经网络的训练及优化过程由两部分构成，首先进行输入信号从前向后的传播，该过程也叫信号前传，然后进行误差梯度从后向前的反向传播，也叫误差反传，并不断重复这一过程，直至损失函数最小。以图 3-3 为例，输入信号的前向传播是指输入信号传输至第一个隐藏层的三个神经元a_i结构上，经过激活函数$h(x)$变化后作为第二层隐藏层的输入信号传输至第二个隐藏层的三个神经元b_i结构上，再经过激活函数变化后传输至输出层，由输出层神经元处理后产生输出信号Y的过程，在此过程中，数据信号信息通过网络向前传递，用公式可以表达为（为表达方便，此处暂不设偏置项）：

$$a_1 = W_{1,1}^{(1)}x_1 + W_{2,1}^{(1)}x_2 + W_{3,1}^{(1)}x_3 + W_{4,1}^{(1)}x_4 \tag{3.5}$$

$$a_2 = W_{1,2}^{(1)}x_1 + W_{2,2}^{(1)}x_2 + W_{3,2}^{(1)}x_3 + W_{4,2}^{(1)}x_4 \tag{3.6}$$

$$a_3 = W_{1,3}^{(1)}x_1 + W_{2,3}^{(1)}x_2 + W_{3,3}^{(1)}x_3 + W_{4,3}^{(1)}x_4 \tag{3.7}$$

$$b_1 = W_{1,1}^{(2)}h(a_1) + W_{2,1}^{(2)}h(a_2) + W_{3,1}^{(2)}h(a_3) \tag{3.8}$$

$$b_2 = W_{1,2}^{(2)}h(a_1) + W_{2,2}^{(2)}h(a_2) + W_{3,2}^{(2)}h(a_3) \tag{3.9}$$

$$b_3 = W_{1,3}^{(2)}h(a_1) + W_{2,3}^{(2)}h(a_2) + W_{3,3}^{(2)}h(a_3) \tag{3.10}$$

$$y = W_1^{(3)}h(b_1) + W_2^{(3)}h(b_2) + W_3^{(3)}h(b_3) \tag{3.11}$$

式中$W_{i,j}^{(n)}$上标表明神经网络的层数，下标表明连接节点的编号，输出层的 y 也可以表达为如下所示的矩阵乘法：

$$y = h\left(h\left([x_1 \quad x_2 \quad x_3 \quad x_4]\begin{bmatrix} W_{1,1}^{(1)} & W_{1,2}^{(1)} & W_{1,3}^{(1)} \\ W_{2,1}^{(1)} & W_{2,2}^{(1)} & W_{2,3}^{(1)} \\ W_{3,1}^{(1)} & W_{3,2}^{(1)} & W_{3,3}^{(1)} \\ W_{4,1}^{(1)} & W_{4,2}^{(1)} & W_{4,3}^{(1)} \end{bmatrix}\right)\begin{bmatrix} W_{1,1}^{(2)} & W_{1,2}^{(2)} & W_{1,3}^{(2)} \\ W_{2,1}^{(2)} & W_{2,2}^{(2)} & W_{2,3}^{(2)} \\ W_{3,1}^{(2)} & W_{3,2}^{(2)} & W_{3,3}^{(2)} \end{bmatrix}\right)\begin{bmatrix} W_1^{(3)} \\ W_2^{(3)} \\ W_3^{(3)} \end{bmatrix} \tag{3.12}$$

在神经网络输入信号的前向传播结束后，模型开始对误差进行反向传播，这一步也是训练神经网络的核心，首先，模型会记录下所有样本输出值和真实值的误差并计算一个损失函数$J(\theta)$， 通常在分类问题中，常使用交叉熵（Cross Entropy，CE）作为损失函数，而在回归问题中，则更多的将均方误差（Mean Square Error，MSE）作为损失函数。确定好损失函数$J(\theta)$后，利用链式梯度求导法，逐层求出损失函数对于模型各层参数的导数$\frac{\partial J(\theta)}{\partial W_{i,j}^{(n)}}$，然后以$\Delta W_{i,j}^{(n)} = -\eta \frac{\partial J(\theta)}{\partial W_{i,j}^{(n)}}$来更新权重参数。$\eta$代表了参数更新时的步长，也叫学习率，如果学习率太小，模型需要耗费较长的时间才能达到收敛，如果学习率太大，模型又有可能在最优值附近徘徊而无法收敛，在实际中设定好初始学习率后，通过从大到小试错来确定模型的学习率。

神经网络通过反向传播误差来调整模型内部参数，最终达到损失函数最小化，此时神经网络模型训练结束，接下来通过模型在样本外数据上的预测表现来评价模型效果。但在实际训练过程中，模型经常会出现因为网络层数和模型参数过多，在训练集上损失函数达到最小并且几乎能够完全拟合所有样本，但在样本外数据上的损失函数很大，并且预测效果也很差，这就是发生了过拟合现象，图 3-4 分别展示了欠拟合、适度拟合和过拟合的图示。

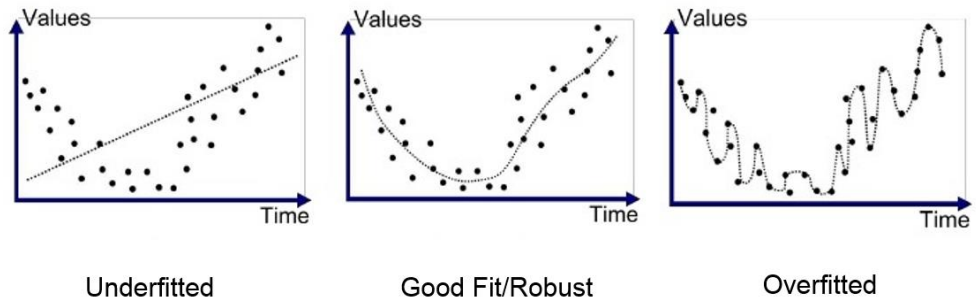


图 3-4 模型不同拟合程度示意图

模型欠拟合，说明模型的表达能力不强，可以通过继续训练和增加模型的深度及参数量来解决，一般的深度学习神经网络模型由于层数较多，模型表达能力较强，欠拟合的现象较少发生，而经常会出现的是模型过拟合，因此，防止并缓解模型的过拟合现象是研究过程中需要主要关注的问题。

模型出现过拟合现象，主要是因为神经网络模型的设计结构过于复杂，模型学习能力太强，在训练时把训练集上的很多噪音也学了进来。针对过拟合，主要的解决方案有三个，第一，数据集扩增，可以通过增加数据样本数量或在数据中加上随机噪音来实现，

在图像识别领域，人们通过将图片翻转、对称、变大缩小来增加数据量，这也是一种数据扩增的手段；第二，降低模型的复杂度，用简单的模型替代复杂的模型对数据进行训练，或者在神经网络中添加 Dropout 层，如图 3-5 所示，人为地让一些神经元之间的联系断开，不再参与每次训练时的参数更新，这种方法在 Hinton 在 2014 年的研究中被证明是一种非常好用解决过拟合现象的方法[42]；第三，提前停止训练，该方法是指模型在训练集上达到收敛之前就停止训练来防止过拟合，具体而言，在每一轮训练结束时，考察模型在验证集上的相关指标，包括模型预测准确率和损失函数，倘若指标不再向优化的方向变化，则认为此时模型已经无需继续训练，采取提前停止训练的措施。

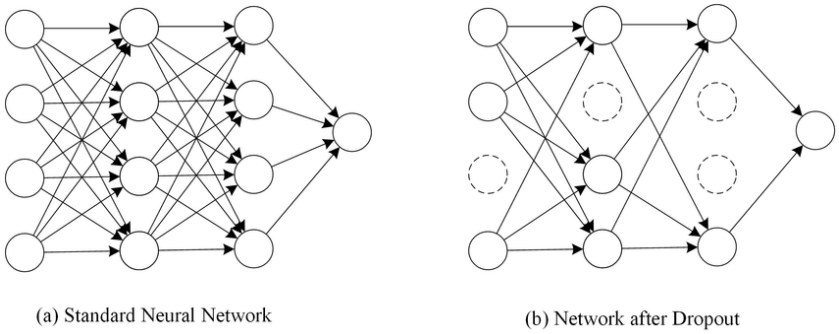


图 3-5 Dropout 原理示意图

3.2.4 神经网络的特点及应用

通过前几节对神经网络的原理结构和训练过程的梳理，可以对神经网络的特点做一个总结。第一，神经网络具有良好的自学习和自适应性。这是指模型在训练过程中能够根据新的输入训练样本自发调整内部参数，改变映射关系，这和人脑的运作机制较为相似。第二，神经网络的非线性。这主要由两方面造成，一方面，模型中的神经元处于激活还是抑制状态是非线性的，另一方面，由图 3-2 可知，激活函数自身通常是非线性函数，当神经元的输出经过激活函数作用后，整个神经网络模型就也拥有了非线性的特征，用非线性的神经网络对非线性的现实问题进行建模和分析效果更好，也更加合理。第三，模型的黑箱（Black Box）性质。这主要是针对深度神经网络而言的特点，层数较多且复杂的深度神经网络往往对现实问题建模效果非常好，但人们难以解释模型产生输出值的具体原因，模型在不同层上提取特征的原理和依据是什么，这些都不为使用者所了解，相比于可解释性较强的传统计量学模型，这一点限制了神经网络的普及和应用。比如，早已有学者对神经网络在金融机构的信用评级中应用开展了大量研究，但在实际中，金融机构在授信时基本不会根据神经网络模型的预测结果决定是否发放贷款，因为金融机

构无法向客户解释所依据神经网络得出结论的依据和原因。

神经网络模型和计量模型也有许多相似之处，计量经济学中一个简单的最小二乘模型如下所示：

$$y_i = \omega + \beta_i X_i + \varepsilon \quad (3.13)$$

神经网络模型的输入值对应计量模型的解释变量X_i，输出值对应计量模型的被解释变量y_i，（X_i，y_i）总是成对出现，分别表示一个样本和一个样本标签，如果被解释变量y_i是连续的，在机器学习中就叫做回归问题，比如预测具体的股票价格，如果被解释变量y_i是离散的，在机器学习中就属于分类问题，比如判断股票的涨跌趋势。此外神经网络还可用于聚类，聚类是指神经网络按照指定的类别个数来对一组数据进行分组，在这个过程中无需人为指定每组具体特征，模型会自动寻找数据之间的关联从而进行聚类，比如互联网公司根据按照手机使用习惯将客户群体划分为有明显的特征区别的细分群体，进而更精准的进行信息推送服务。

3.3 深度学习的神经网络模型

3.3.1 CNN 神经网络算法原理

卷积神经网络（CNN）是一种深度学习中常用的前馈神经网络，从 20 世纪 80 年代起就已被广泛用于图像识别。上世纪 60 年代 Hubel 和 Wiesel 通过对动物的大脑视觉皮层研究提出了“感受野”的概念[43]，他们认为视觉皮层神经元只对视野的局部区域内的视觉刺激做出反应，虽然不同神经元的感受野可能会有重叠，但整体上会平铺整个视觉区域。Fukushima 等人在此研究基础上于 1982 年提出了神经感知机 [44]，不同于全连接神经网络，神经感知机是第一个具有神经元间局部连接和层次结构性质的人工神经网络。Yann LeCun 等人在研究手写数字识别问题时提出了卷积神经网络模型（CNN），并于 1998 年将卷积层和池化层引入神经网络架构中，提出了 LeNet-5 架构[45]，该架构广泛用于手写数字识别领域，并且奠定了后续卷积神经网络发展的结构基础，成为最经典的 CNN 模型结构之一。

在计算机视觉领域，ILSVRC 大赛是一项备受瞩目的赛事，每年不断有新的卷积神经网络架构被提出并投入应用，除了 LeNet-5 架构，先后有 AlexNet、GoogLeNet、ResNet 等架构都在这个大赛上崭露头角，神经网络模型的深度也越来越深，ResNet 架构下的卷

积神经网络甚至达到了 152 层，实现了真正意义的“深度”学习。

本文构建模型的一部分结构参考了 LeNet-5 架构设计，因此以该模型架构为例，介绍卷积神经网络的算法原理。如图 3-6 所示，LeNet-5 卷积神经网络由一层输入层、两层卷积层、两层池化层、两层全连接层和一层输出层组成，通常卷积层和池化层成对出现，主要起到在样本上进行特征采样的作用，采样完毕后将特征传递给全连接层做更复杂的特征组合处理，后者连接至输出层产生一个输出结果。

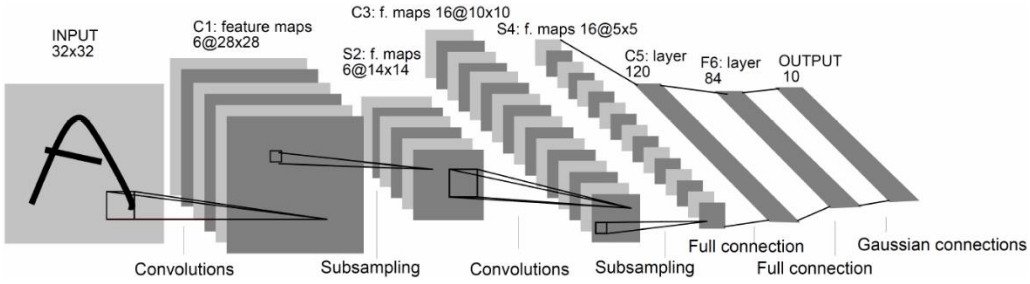


图 3-6 LeNet-5 卷积神经网络

卷积层是 CNN 神经网络的核心，数据传递到卷积层时会在该层进行卷积运算，离散的卷积运算用数学公式表达为：

$$c(n) = (f * g)(n) = \sum_{\tau=-\infty}^{\infty} f(\tau)g(n-\tau) \tag{3.14}$$

式中，f代表卷积层收到的原始样本图像，g代表卷积核加权函数，c代表卷积核特征采样后的样本图像，由于 CNN 神经网络多用于处理图像数据，图像数据可以理解为二维矩阵，每个元素即图像所在位置的像素大小，因此，可以借助如图 3-7、图 3-8 所示的样例来理解卷积层运算原理，假设输入图像 f 尺寸为 4×4，卷积核 g 大小为 3×3，移动步长为 1，图 3-7 不进行填充，图 3-8 进行全 0 填充，填充的目的是保持输出图像 c 和输入图像 f 尺寸一致，选择 ReLU 函数作为卷积层的激活函数。

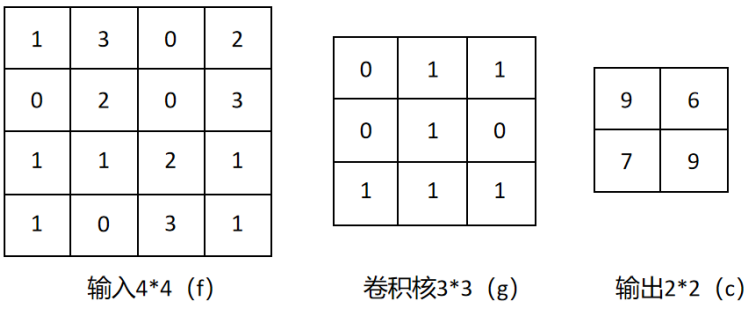


图 3-7 卷积层计算样例（不使用填充）

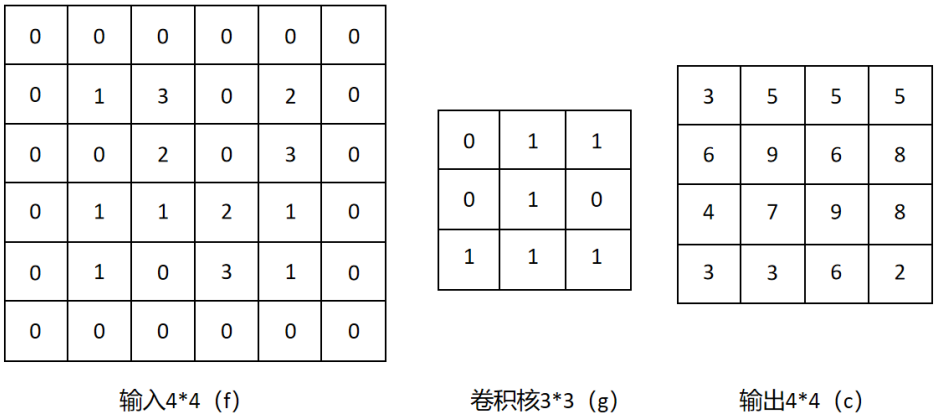


图 3-8 卷积层计算样例（全 0 填充）

输出图像 c 中第i行j列的元素可以由下式进行计算（行和列从 0 开始计数）：

$$c_{i,j} = ReLU(\sum_{x=0}^{2}\sum_{y=0}^{2} g_{x,y} f_{i+x,j+y} + b) \tag{3.15}$$

式中，$g_{x,y}$代表卷积核对应位置的权重参数，$f_{i,j}$代表输入图像对应位置的元素，b 代表卷积核的偏置项，默认为零。因此图 3-7 输出图像 c 中的各元素计算如下：（图 3-8 同理，在此不做展示）

$$c_{0,0} = ReLU(0*1+1*3+1*0+0*0+1*2+0*0+1*1+1*1+1*2+0) = 9$$

$$c_{0,1} = ReLU(0*3+1*0+1*2+0*2+1*0+0*3+1*1+1*2+1*1+0) = 6$$

$$c_{1,0} = ReLU(0*0+1*2+1*0+0*1+1*1+0*2+1*1+1*0+1*3+0) = 7$$

$$c_{1,1} = ReLU(0*2+1*0+1*3+0*1+1*2+0*1+1*0+1*3+1*1+0) = 9$$

卷积层工作时，卷积核会按照设定好的步长遍历输入图像的数据，每移动一个步长都会将所覆盖区域加权求和并产生一个结果，由此可见，卷积层具有特征提取的作用，并且不同局部区域共用一个卷积核，同一个卷积核的权重参数$g_{x,y}$相同，这是卷积神经网络的第一个特点——“权值共享”。在实际使用中，每个卷积核只负责对输入变量的一种特征进行提取，如果想要增强模型的特征提取能力，通常的做法是增加卷积核的数量，这样一来，模型可以对输入变量的不同特征进行全方位的识别与提取。此外，卷积核只能覆盖输入图像的一部分，每次只针对所覆盖部分做特征提取，因此相邻两层之间是局部连接而非全连接的，将卷积神经网络的这一特点称为“局部连接”，这也是卷积神经网络的第二个特点，图 3-9 展示了全连接和局部连接的效果对比。

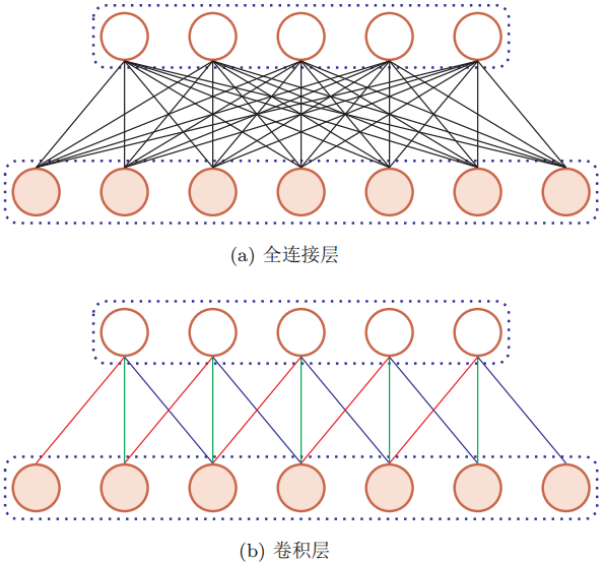


图 3-9 全连接和局部连接

池化层与卷积层成对出现，作用是对卷积后的图像进行下采样处理，在保留大部分特征信息的同时压缩图像尺寸，从而减少参数量，防止神经网络模型过拟合，在具体的池化方式的选择上，池化时常采用最大池化和平均池化两种方式。二者的主要区别如图 3-10 所示，以图 3-8 全 0 填充的输出图像 c 为例，设置池化层过滤器尺寸 2×2，移动步长为 2，采用最大池化得到 3-10(a)，采用平均池化得到 3-10(b)。

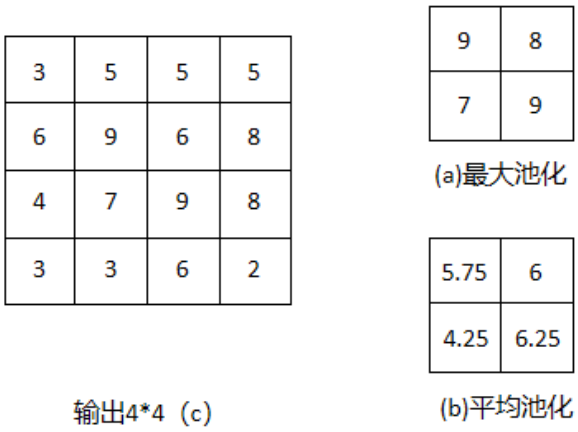


图 3-10 最大池化和平均池化

全连接层在 CNN 神经网络的尾部，也是较为重要的网络层，所谓全连接，就是说此层每个神经元都和前一层神经元互相连接。由图 3-6 可知，LeNet-5 架构中有两层全连接层，其中，第一层全连接层负责将高维特征矩阵一维化，同时对特征进行更复杂的非线性组合，第二层全连接层负责将学习到的特征映射到样本标签中，配合激活函数起到分类的作用，通常使用 Softmax 作为激活函数，输出值为每个类别的概率。

3.3.2 RNN 神经网络算法原理

RNN 神经网络起源于 1982 年 Saratha Sathasivam 提出的霍普菲尔德神经网络（Hopefield Networks）[46]。在以往的 CNN 神经网络或是 BP 神经网络中，网络结构都可以归纳为图 3-3 所示的形式，即模型依次由输入层、隐藏层和输出层组成，每层网络之间相互连接，但网络内部节点之间相互并未连接，数据信号只是沿着网络逐层传递。RNN 神经网络和前者本质的区别在于，在循环神经网络的隐藏层中，每层网络的神经元节点也是相互连接的，如图 3-11 所示，在 RNN 神经网络中，隐藏层除了接收输入层的输入信息，同时也会受到前一时刻隐藏层h_{t-1}的输入信息影响，这样一来，RNN 神经网络结构便能够刻画一个时间序列当前输出和之前历史信息的关系[47]。相较于 CNN 神经网络，RNN 神经网络更多地用于对时间序列数据进行处理和预测，这也是循环神经网络的结构特点所决定的。

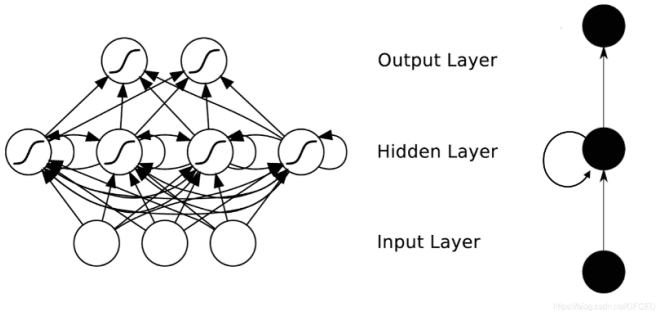


图 3-11 循环神经网络结构图

如果沿着时间维度将 RNN 神经网络展开，我们将会得到如图 3-12 所示的一系列相同的循环体构成的网络结构，图中，U 是输入层到隐藏层的权重系数矩阵，V 是隐藏层到输出层的权重系数矩阵，W 是隐藏层和前一时刻隐藏层的权重系数矩阵，由此看来，RNN 神经网络也具有类似 CNN 神经网络中权值共享的特征，唯一区别在于 CNN 神经网络是在空间位置上进行权值共享，而 RNN 神经网络是在时间维度上进行权值共享。

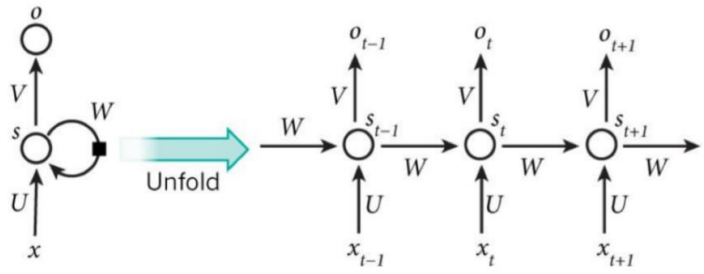


图 3-12 展开后的循环神经网络结构图

RNN 神经网络的隐藏层和输出层信息的计算公式如下所示，式中 f 和 g 都为激活函数：

$$s_t = f(Ux_t + Ws_{t-1}) \quad (3.16)$$

$$o_t = g(Vs_t) \quad (3.17)$$

$$o_t = g(Vs_t) = g\big(Vf(Ux_t + Ws_{t-1})\big) = g\big(Vf(Ux_t + Wf(Ux_{t-1} + Ws_{t-2}))\big) \quad (3.18)$$

将 3.16 代入 3.17 中得到 3.18，观察 3.18 可以发现 RNN 神经网络的输出o_t会受到之前每一期输入值$\{x_t，x_{t-1}，x_{t-2}，\ldots，x_1\}$的影响，因此，RNN 神经网络具有记忆储存能力，能够将往期数据信息作用于新的后续节点输出。

但在实际使用中，RNN 神经网络也并非是完美的，随着循环次数的增加，在计算梯度时由于梯度反向传播中的连乘效应会出现梯度爆炸和梯度消失的现象，表现为距离当前时间越远的数据反馈的梯度信号越不显著，模型只能学习到较近距离的数据特征，而无法学习到较远距离数据的依赖关系，因此无法有效处理长序列数据，在此背景下，RNN 的变体——LSTM 神经网络的出现解决了这个问题。

3.3.3 LSTM 神经网络算法原理

LSTM 神经网络结构是由 Sepp Hochreiter 和 Jürgen Schmidhuber 于 1997 年提出的一种 RNN 的变体结构[48]。作为 RNN 神经网络的变体，LSTM 神经网络最大的不同之处在于它借鉴了人类大脑的选择性输入和选择性遗忘机制，引入了遗忘门、输入门和输出门三个“门”结构，以及一个记忆单元来选择性接收传入神经网络的信息。其中，属于逻辑单元的“门”结构只负责在神经网络中其它部分与记忆单元相接的边缘地带完成权值的设定工作，而不会对其他神经元节点产生影响[49]，LSTM 神经网络结构如图 3-13 所示。

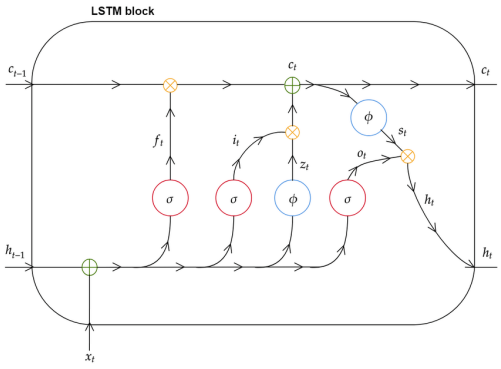


图 3-13 LSTM 神经网络结构图

在 LSTM 的网络结构中，遗忘门负责接收前一时刻信息至当前时点，也就是决定了前一时刻的记忆单元状态c_{t-1}有多少能被保留到当前时刻c_t，设置该逻辑单元的目的是遗忘掉之前的无用信息，遗忘门根据前一时刻的输出h_{t-1}和当前时刻的输入x_t计算出信息保留程度f_t：

$$f_t = sigmoid(W_f[h_{t-1}, x_t] + b_f) \tag{3.19}$$

经过$sigmoid$激活函数作用后，f_t的取值范围在 0-1 之间，f_t越接近 1，则保留的信息越多。在保留了部分之前的信息后，神经网络还需要从当前时刻的输入x_t中产生新的信息来更新记忆单元状态c_t。

输入门的作用是根据h_{t-1}和x_t来决定哪些信息可以加入到前一时刻记忆单元状态c_{t-1}中并更新为新的记忆单元状态c_t，首先，输入门根据前一时刻的输出h_{t-1}和当前时刻的输入x_t计算出i_t：

$$i_t = sigmoid(W_i[h_{t-1}, x_t] + b_i) \tag{3.20}$$

式中，W_i表示前一时刻隐层输出h_{t-1}和当前时刻输入x_t的参数矩阵，b_f代表偏置项。

再计算出z_t：

$$z_t = \tanh(W_g[h_{t-1}, x_t] + b_g) \tag{3.21}$$

更新后的当前时刻记忆单元状态c_t为：

$$c_t = f_t \cdot c_{t-1} + i_t \cdot z_t \tag{3.22}$$

输出门的作用是决定有多少当前时刻的信息能够被输出，根据前一时刻的输出h_{t-1}和当前时刻的输入x_t计算出信息输出程度o_t：

$$o_t = sigmoid(W_o[h_{t-1}, x_t] + b_o) \tag{3.23}$$

最后输出当前时刻的输出h_t:

$$h_t = o_t \cdot \tanh(c_t) \tag{3.24}$$

至此，LSTM 神经网络中循环体的前向传播就结束了，随后要进行的就是误差的反向传播，并在误差反传的过程中修正各项权重系数，核心思路仍是对每个门单元中经过的参数求偏导数，并按照梯度收敛的方向更行权值，在此不再赘述。

从以上推导可以看出，LSTM 神经网络的结构设计十分灵巧，非常适合对金融时间序列进行建模分析，同时，LSTM 神经网络的出现也解决了 RNN 神经网络所不能处理的问题，改善了模型训练过程中梯度消失及梯度爆炸问题。通过在每一层的神经元中加

入多个“门”结构逻辑单元，LSTM 神经网络使得误差在方向传播中可以直接通过“门”，改善了误差反传过程中梯度消失和发散的现象，以至于无论距离当前时间多远的数据都不会出现梯度完全消失的现象，此外，LSTM 神经网络的梯度在传播过程中较为稳定，不会像 RNN 神经网络梯度反向传播时发生相同权重矩阵连乘的现象。

3.4 股价预测中引入神经网络的可行性分析

本章前三个小节介绍了机器学习的定义、分类以及深度学习的特点，并且对神经网络的理论和算法原理做了梳理和总结，关于神经网络能否应用于股价预测研究，以及其依据是什么，有必要在此进行方法的可行性分析。

股票价格是非平稳且非线性的金融时间序列，传统对股价建模的 ARMA、ARIMA、ARCH、GARCH 等方法尽管在改进之后可以用于处理非平稳时间序列数据，但仍属于线性回归，由于股价的影响因素复杂多样，并且各因素对股价影响程度大小也不尽相同，对股价进行线性预测在一定程度上限制了模型的预测效果。

神经网络通过对人脑工作机制的模拟，以神经元之间权重连接的方式逐层建立起网络架构，并且每个神经元都有偏置项和激活函数对数据信息进行非线性化变换，这使得运用神经网络对股票价格建模能够克服金融时序的非平稳性和非线性，从而最大限度的逼近和拟合实际的股票价格，因此神经网络在预测股票价格的问题上优于传统的建模方法。此外，神经网络的各层神经元节点之间相互连接，可训练参数量通常可以达到上千上万个，更多的可训练参数可以更有效地发现数据之间的关系从而强化模型的表达能力。基于以上两点，利用神经网络预测股票价格在理论上是可行的。

然而，在数据量不足的情况下，使用更多的可训练参数去拟合股价问题很容易出现过拟合现象，表现为神经网络模型在样本内数据上的预测准确率很高，但在样本外数据上表现并不好，这使得模型对过去的股票价格有很好的解释能力，但对于未来的股票价格没有任何预测能力，此外，神经网络层数越深，节点越多，训练时所花费的时间也就越长。随着大数据时代的来临，各类数据的获取变得方便快捷，股票在每日的交易中也生成了海量的数据信息，充足的数据量保障了神经网络模型能够尽可能学习到股票价格的全部特征并在一定程度缓解了过拟合现象，在计算机的软硬件发展方面，谷歌（GOOGLE）公司在 2015 年将 Tensorflow 开放源代码，这将允许世界各地的研究人员

能够使用 Tensorflow 来搭建神经网络模型，此外英伟达（NVIDIA）公司也针对深度学习的研究人员专门开发出 CUDA 运算平台和 CUDNN 库来加速深度神经网络的计算，这项技术充分利用了计算机的显卡 GPU 性能，大幅缩减了训练神经网络所花费的时间。基于以上两点，利用神经网络预测股票价格在技术上是可行的。

最后，本文通过对相关研究文献的梳理和回顾，发现国内外已有学者在股价预测中引入了神经网络方法并取得了相比传统计量经济模型较为不错的预测效果，因此，基于理论角度、技术角度和现有研究成果角度三方面，本文认为在股价预测中引入神经网络是可行、有效的。

3.5 本章小结

本章首先介绍了机器学习的相关概念，主要介绍深度学习相比传统机器学习方法的优势，随后梳理了神经网络的相关算法原理，对神经网络模型在训练过程中最容易出现的问题——过拟合及其原因和解决方法做出归纳与总结。

接下来从算法原理的角度对本文研究中要用到的两个神经网络模型进行了解释，CNN 神经网络中各层之间局部相连，卷积核做卷积采样时权值共享，这两个特点让 CNN 神经网络相比全连接神经网络，有效缩减了模型中的参数量，提高了模型的训练速度，此外，卷积层和池化层够自动提取并组合输入图像的特征，这让 CNN 神经网络在图像识别这样的分类问题上表现出更好的适用效果。而 RNN 神经网络的记忆特性决定了它更适合对时间序列进行建模研究，但它也存在缺陷，即如果时间序列太长，模型只能学习到较近距离的数据特征，而无法学习到较远距离数据的依赖关系。RNN 神经网络的变体 LSTM 神经网络有效的解决了长期记忆失效的问题，通过引入了门结构和记忆单元，LSTM 神经网络可以有选择性的保留和遗忘往期的信息，并且在误差方向传播时，能够很好地解决由于梯度连乘效应导致的梯度消失和梯度爆炸问题，目前已经逐步取代 RNN 神经网络，广泛的应用于自然语言处理和金融时序预测等领域。

最后，本文从理论方面、技术方面和前人研究三个角度对股价预测中引入神经网络的可行性进行了分析概述。基于此，本文将在第四章结合 CNN 和 LSTM 神经网络模型各自的特点，学习上证 50 指数成分股数据集的特征，来对股票涨跌做判断。

第四章 股价预测模型构建

4.1 实验平台及数据介绍

本文构建的神经网络模型是基于 Window10（64 位）操作系统开发的，处理器配置为 Intel(R)Core(TM)i7-10870H CPU@2.20GHz 2.21GHz，内存 16G，显卡型号为 NVIDIA GeForce RTX2070，显存 8G，该显卡的算力支持 GPU 加速，可以大幅缩减模型训练时间，良好的实验平台性能保证了神经网络模型训练时的高效率和稳定性。此外，在代码编写部分使用 Jupyter Notebook 作为集成开发环境来完成。神经网络模型的搭建使用的是基于 Tensorflow.Keras 的 Sequential 序贯模型，此外，在因子构建部分使用到了 TA-Lib 库来获取个股的技术分析指标，模型效果评估及绘图部分使用到的模块库有 sklearn 库和 matplotlib 库。

本文实验数据来源为 JoinQuant 聚宽量化交易平台，原始数据为上证 50 指数成分股自 2011 年 1 月 1 日至 2020 年 12 月 31 日的日频股票交易数据，选取的标的为截至 2020 年 12 月 31 日的上证 50 指数成分股，股票交易数据包括开盘价、收盘价、最高价、最低价、交易量和交易额，具体的原始数据样本如表 4-1 所示：

表 4-1 上证 50 指数成分股日频交易原始数据样本

date	code	open	high	low	close	volume	money
2011/1/4	600000.XSHE	4.69	4.76	4.78	4.64	276480564	1307175000

上证 50 股票指数是由上海证券交易所发布的金融指标，该指数覆盖了上交所交易活跃的优质股票，是一个能够综合反映主流投资方向的市场指数。从最近一年公募基金的火爆程度来看，越来越多的机构投资者投资方向集中在绩优股领域，资金抱团的趋势明显，个股呈现出一种强者恒强的态势，而上证 50 指数中有大量白马股、蓝筹股，是较为不错的投资标的，因此，研究上证 50 指数成分股的股性并在上证 50 指数成分股中构建交易策略，能够取得较为稳健和理想的收益。此外，上证 50 指数成分股资质优良，少有停牌、退市等情况，能够保证实验数据样本的连续性和完整性，有利于模型的构建。

4.2 数据预处理

4.2.1 数据预处理方法

决定神经网络模型预测效果好坏的主要因素有两个，第一是模型网络结构设计及模型参数选择，本文将在 4.3、4.4 节模型构建及参数优化时讨论，第二是数据的预处理工作，在上证 50 指数成分股中，同一只股票的价格和交易量、交易额不在同一数量级上，并且不同股票的价格跨度范围也比较大，因此要对实验数据进行预处理。数据预处理包括对异常值和空值的处理，以及对数据的归一化处理，异常值是指偏离大多数样本的数据，是在数据集上不合理的值，一般认为距离样本平均值 3σ 以外的样本为异常值，对于异常值，常用的处理方法有利用平均值进行修正和直接剔除，空值是指缺失的样本数据，对于空值，常采用数据填充或直接剔除进行处理，由于本文研究的是金融时间序列数据，利用平均值填充有失数据的真实性，因此对于空值采用直接剔除的方式处理，数据归一化是把不同规模、不同单位的数据统一变换处理，让不同维度之间的数据在数值上变得可比，便于神经网络模型训练时进行特征提取。

本文使用的交易数据只包含了个股的量价信息，因此不做异常值处理，个股的空值即数据缺失现象是存在的，这主要是由于上证 50 股指成分股有可能在研究期内因为停牌或尚未选入指数，导致交易数据不能够完全覆盖 2011 年至 2020 年的全部区间，并产生了缺失数据，针对缺失值，本文采用直接剔除空值的处理方式。在做数据归一化时，本文希望将所有数据缩放至 0 到 1 之间，同时保持原有数据结构，因此采用 Max-Min 归一化方法对各特征因子做无量纲化处理，归一化的具体公式如 4.1 所示。

$$y_i = \frac{x_i - \min\limits_{1\le j\le n}\{x_j\}}{\max\limits_{1\le j\le n}\{x_j\} - \min\limits_{1\le j\le n}\{x_j\}} \tag{4.1}$$

式中，x_i 为原始数据，y_i 为标准化后的无量纲数据，j 为所在序列的第 j 个元素。

4.2.2 输入变量特征因子构建

神经网络模型的输入变量通常为可以直接或间接影响股价涨跌的因子，从短期来看，股价受交易相关的量价信息影响最大，并且在股票市场中，最容易获取的信息就是个股每日的开盘价、最高价、最低价、收盘价和成交量，“四价一量”可以说是当前市场状况最好的反映，因此大量的股价预测都是基于股票的基本交易指标开展的，大量研究直

接利用股票交易的“四价一量”数据作为神经网络的输入，然后借助神经网络自学习能力来提取数据中隐含的特征和非线性关系。

本文在构建输入变量特征因子时，在“四价一量”的基础上设置对照实验组，加入上证 50 指数的“四价一量”指标以及个股的技术分析指标，希望通过增加输入特征的维度，一方面，增加模型输入数据的信息丰富程度，另一方面，由于本文构建的 CNN-LSTM 模型输入数据的格式为类似于二维图片的样本数据而不是一维序列，太小的输入样本尺寸不利于卷积核进行特征提取，从尽量扩大图片尺寸的角度出发，构造出一共 20 个因子作为对照模型的输入变量。在以上三类指标中，个股的“四价一量”反映了股票交易的最直观状况，指数的“四价一量”反映了股票市场的整体趋势状况，个股的技术分析指标是根据股票量价信息统计出来的，用来反映包括股票近期的趋向和买卖情况在内的多个维度的特征。

本文选取的 10 个技术指标分别为 MA、CCI、RSI、W%R、ROC、MACD、ATR、ADX、OBV 和 MOM。其中，按照所属类别划分，属于均线型的指标有 MA、MACD，属于超买超卖型的指标有 CCI、RSI、W%R、ROC 和 MOM，属于趋势型的指标有 ADX，属于波动量型的指标有 ATR，属于成交量型的指标有 OBV，这 10 个技术也是个人和机构投资者做技术分析时较为常用的指标。

MA（Moving Average）为指数移动平均值，该指标的原理是对股票收盘价以指数式递减进行加权移动平均，用于判断价格未来走势的变动趋势，本文中使用 5 个交易日的收盘价来计算 MA，具体的计算公式如式 4.2 所示。

$$MA_t(5) = \frac{Close_t + \cdots + Close_{t-4}}{5} \tag{4.2}$$

CCI（Commodity Channel Index）为顺势指标，该指标由 Donald Lambert 于 20 世纪 80 年代提出，早期用于期货市场，后推广至股票市场，用于对趋势做分析判断,相比于 KDJ 和威廉指标这些超买超卖型指标，CCI 没有上下界的限制，因此有利于投资者对短期内暴涨暴跌的非常态行情更好地做研判。通常 CCI 值在 100 到-100 的范围是正常交易范围，超出此范围的则表明存在超买或超卖情况，具体的计算公式如式 4.3、式 4-4 所示。

$$CCI(14) = \frac{TP - MA(14)}{0.015 \times MD(14)} \tag{4.3}$$

$$TP(14) = \frac{High_{14}+Low_{14}+Close}{3} \quad (4.4)$$

式中，$MD(14)$为过去 14 个交易日的平均绝对误差，$High_{14}$为过去 14 个交易日的最高价，Low_{14}为过去 14 个交易日的最低价。

RSI（Relative Strength Index）为相对强弱指标，该指标是根据一段时期股价涨跌幅之和的比率构造的技术曲线，本文使用 12 个交易日的股票涨跌情况来计算 RSI，具体的计算公式如式 4.5 所示。

$$RSI(12) = 100 - \frac{100}{1+RS} \quad (4.5)$$

式中，RS 为相对强度，该值等于 12 个交易日内收盘价累计上涨数之和的平均值除以 12 个交易日内收盘价累计下跌数之和的平均值。RSI 的取值范围在 0-100，一般而言，当 RSI 比较大时说明此时该股票已经达到超买状态，未来股价很可能下跌调整，当 RSI 比较小时说明此时该股票处于超卖状态，未来股价由反弹回升的趋势。

W%R 为威廉指标，该指标由美国著名投资家 Larry Williams 于 1973 年在《How I made one million dollars last year trading commodities》中提出，利用过去一段时期的最高价、最低价、收盘价，计算当日收盘价所处过去一段时期内的价格区间相对百分位置，具体的计算公式如式 4.6 所示。

$$W\%R(14) = \frac{High_{14}-Close}{High_{14}-Low_{14}} \quad (4.6)$$

ROC（Price Rate of Change）为变动率指标，用于计算股票过去一段时期内收盘价变动的比例，该指标可以用于判断股票价格是否有反转的趋势，也属于一种辅助类指标，具体的计算公式如式 4.7 所示。

$$ROC(14) = \frac{Close_t-Close_{t-14}}{Close_{t-14}} \quad (4.7)$$

MACD（Moving Average Convergence Divergence）为异动移动平均线，该指标是从双指数移动平均线（DEMA）的基础上发展而来的，也是目前较为主流的技术指标，MACD 由快、慢均线的聚散状态反映市场的多空态势。本文在计算时选取快速线周期为 12 天，慢速线周期为 26 天，二者离差值的移动平均线周期为 9 天，具体的计算步骤如下：

第一步，计算$EMA_t(12)$和$EMA_t(26)$：

$$EMA_t(12) = \frac{2}{13} \times Close_t + \frac{11}{13} \times EMA_{t-1}(12) \quad (4.8)$$

$$EMA_t(26) = \frac{2}{27} \times Close_t + \frac{25}{27} \times EMA_{t-1}(26) \quad (4.9)$$

第二步，计算离差值DIF：

$$DIF = EMA_t(12) - EMA_t(26) \quad (4.10)$$

第三部，计算离差值的 9 日均线DEA：

$$DEA = DIF_t \times \frac{2}{10} + DEA_{t-1} \times \frac{8}{10} \quad (4.11)$$

第四步，计算$MACD$：

$$MACD = (DIF - DEA) \times 2 \quad (4.12)$$

ATR（Average True Ranger）为真实波动幅度均值，该指标用于衡量股票价格波动的剧烈程度，而不能直接反映股票价格变动趋势，具体的计算步骤如下：

第一步，计算当日最高价与最低价的波幅A：

$$A = High_t - Low_t \quad (4.13)$$

第二步，计算前一日收盘价与当日最高价的波幅B：

$$B = Close_{t-1} - High_t \quad (4.14)$$

第三步，计算前一日收盘价与当日最低价的波幅C：

$$C = Close_{t-1} - Low_t \quad (4.15)$$

第四步，计算真实波幅TR：

$$TR = max\{A, B, C\} \quad (4.16)$$

第五步，计算真实波幅 TR 的 14 日移动平均值，得到 ATR(14)。

ADX（Average Directional Indicator）为平均趋向指标，该指标是 DX 指标的移动平均，取值范围在 0-100，该值越大说明股价上涨或下跌趋势越强，具体的计算步骤如下：

第一步，计算动向变化（$+DM$和$-DM$）：

$$up = High_t - High_{t-1} \quad (4.17)$$

$$down = Low_{t-1} - Low_t \quad (4.18)$$

$$+\mathrm{DM} = \begin{cases} up, & up > \max(down, 0) \\ 0, & up \le \max(down, 0) \end{cases} \quad (4.19)$$

$$-\mathrm{DM} = \begin{cases} down, & down > \max(up, 0) \\ 0, & down \le \max(up, 0) \end{cases} \quad (4.20)$$

第二步，计算真实波幅 TR：计算公式见式 4.16。

第三步，计算动向指数（+DI 和-DI）：

$$+DI(14) = \frac{+DM(14)}{TR(14)} \times 100 \quad (4.21)$$

$$-DI(14) = \frac{-DM(14)}{TR(14)} \times 100 \quad (4.22)$$

第四步，计算 ADX：

$$DX = \frac{(+DI14) - (-DI14)}{(+DI14) + (-DI14)} \times 100 \quad (4.23)$$

第五步，计算 DX 的 14 日移动平均值，得到 ADX(14)。

OBV（On Balance Volume）为能量潮，也叫人气指标，由投资分析家 Joe Granville 所创，该指标注重考察股票交易量的变化和价格的变化的关联关系，通过成交量的变动推测股价的变动方向。在计算时，首先，将成交量划分为正值和负值来，每日将收盘价同前一日收盘价比较，若前者较大，则将当日成交量值记为正值，反之则记为负值，然后，根据当日收盘价和前一日收盘价的比较计算 OBV，如果二者相等，则 OBV 为零，否则 OBV 为往期成交量正负序列累加之和。

MOM（Momentum Index）为动量线指标，该指标用于衡量股价中短期波动状况，通过观察股价围绕中心线的周期性波动，从而判断股价的峰顶和谷底，具体的计算公式如式 4.24 所示。

$$MTM(n) = Close_t - Close_{t-n} \quad (4.24)$$

综上，将本文所构建的特征因子总结为表 4-3。

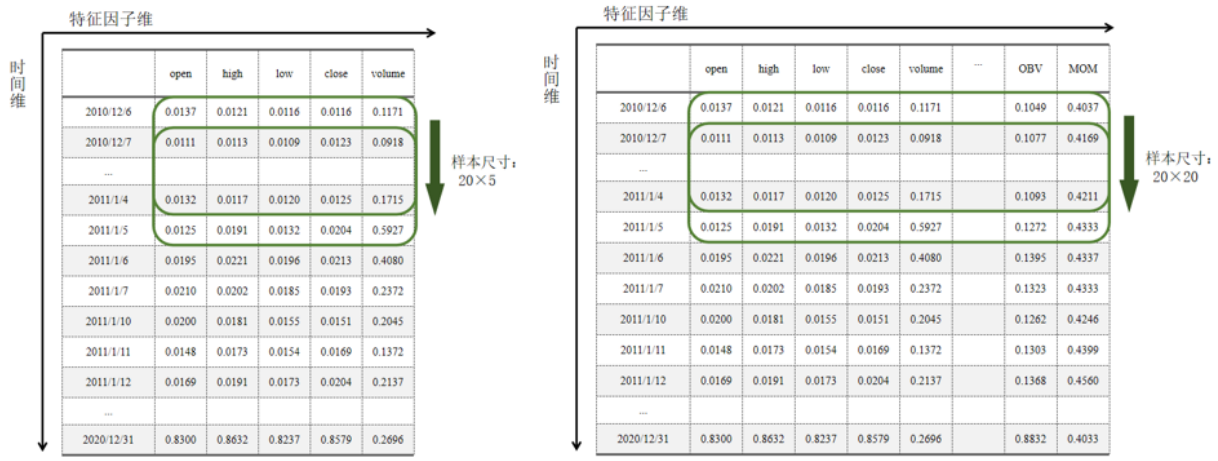
表 4-3 输入特征因子总结

因子类别	编号	指标名称	指标简称
T1: 个股基本交易指标	（1）	当日开盘价	open
	（2）	当日最高价	high
	（3）	当日最低价	low
	（4）	当日收盘价	close
	（5）	当日成交量	volume
T2: 上证 50 指数指标	（6）	当日开盘价	sz_open
	（7）	当日最高价	sz_high
	（8）	当日最低价	sz_low
	（9）	当日收盘价	sz_close

	（10）	当日成交量	sz_volume
T3: 个股技术指标	（11）	简单移动平均线	MA(5)
	（12）	顺势指标	CCI(14)
	（13）	相对强弱指标	RSI(12)
	（14）	威廉指标	W%R(14)
	（15）	变动率指标	ROC(14)
	（16）	平滑异同移动平均线	MACD
	（17）	真实波动幅度均值	ATR(14)
	（18）	平均趋向指标	ADX(14)
	（19）	能量潮指标	OBV
	（20）	动量线指标	MOM(10)

至此，模型输入变量的特征因子的选取工作已经完成。在实际操作时，首先在聚宽量化交易平台上获取上证 50 指数成分股在十年期间的“四价一量”数据，然后获取上证 50 指数十年期的“四价一量”数据，最后通过 TA-Lib 库计算上证 50 指数成分个股的相关技术指标，接下来将以上指标因子按照日期和个股进行数据对齐并归一化处理，按照图 4-1 所示的方式在数据集上滚动生成输入数据样本图片，图 4-1 的左侧为将成分股“四价一量”的个股基本交易指标 T1 作为模型输入变量的样本生成方式，图 4-1 右侧为增加了指数“四价一量”T2 和个股相关技术指标 T3 的对照实验组模型输入变量的样本生成方式，样本尺寸为“20×20”。

以 20 天为窗口，1 天为步长对每只股票进行遍历进行样本生成。特征构建完成后的数据结构为“N×20×T”，可以理解为 N 个“20×T”像素的特征图像，其中 20 为过去 20 个交易日，T 为输入变量特征因子的个数，样本的个数 N 即为上证 50 指数成分股在 2011 年 1 月 1 日至 2020 年 12 月 31 日一共生成的“20×T”样本的个数，去掉因停牌而产生的缺失样本后，最后得到的有效样本图片个数 N 共计 100307 个。



	open	high	low	close	volume
2010/12/6	0.0137	0.0121	0.0116	0.0116	0.1171
2010/12/7	0.0111	0.0113	0.0109	0.0123	0.0918
...					
2011/1/4	0.0132	0.0117	0.0120	0.0125	0.1715
2011/1/5	0.0125	0.0191	0.0132	0.0204	0.5927
2011/1/6	0.0195	0.0221	0.0196	0.0213	0.4080
2011/1/7	0.0210	0.0202	0.0185	0.0193	0.2372
2011/1/10	0.0200	0.0181	0.0155	0.0151	0.2045
2011/1/11	0.0148	0.0173	0.0154	0.0169	0.1372
2011/1/12	0.0169	0.0191	0.0173	0.0204	0.2137
...					
2020/12/31	0.8300	0.8632	0.8237	0.8579	0.2696

	open	high	low	close	volume	...	OBV	MOM
2010/12/6	0.0137	0.0121	0.0116	0.0116	0.1171		0.1049	0.4037
2010/12/7	0.0111	0.0113	0.0109	0.0123	0.0918		0.1077	0.4169
...								
2011/1/4	0.0132	0.0117	0.0120	0.0125	0.1715		0.1093	0.4211
2011/1/5	0.0125	0.0191	0.0132	0.0204	0.5927		0.1272	0.4333
2011/1/6	0.0195	0.0221	0.0196	0.0213	0.4080		0.1395	0.4337
2011/1/7	0.0210	0.0202	0.0185	0.0193	0.2372		0.1323	0.4333
2011/1/10	0.0200	0.0181	0.0155	0.0151	0.2045		0.1262	0.4246
2011/1/11	0.0148	0.0173	0.0154	0.0169	0.1372		0.1303	0.4399
2011/1/12	0.0169	0.0191	0.0173	0.0204	0.2137		0.1368	0.4560
...								
2020/12/31	0.8300	0.8632	0.8237	0.8579	0.2696		0.8832	0.4033

图 4-1 模型输入样本生成方式（以浦发银行为例）

4.2.3 样本标记及数据集划分

特征构建完成后，还需要对样本进行标记，该过程也叫做给样本打标签。本文预测的是股票价格的短期变动趋势，即根据个股过去 20 个交易日的输入特征因子判断 5 个交易日后的涨跌，因此按照股票在未来 5 个交易日后的收益率正负将标签分为上涨和下跌两个类别，具体的划分方式如式 4.25 所示。

$$\begin{cases} 上涨，\ r > 0 \\ 下跌，\ r \leq 0 \end{cases} \tag{4.25}$$

在对标签进行标记时，本文采取了独热编码的方式完成标记，将上涨样本记为[1,0]，下跌样本记为[0,1]。使用独热编码的意义在于我们希望神经网络模型能够输出每一个样本的涨跌概率值，便于第五章量化交易策略选出合适的股票池，此外，将标签转换为独热编码在一定程度上能够提高模型运算效率，经过标记的两类标签数量共有 100307 个，其中上涨样本 49676 个，下跌样本 50631 个，各类标签大致比例为 1:1，标签比例均衡。

在构建模型之前，需要先进行数据集划分工作，一般将数据集划分为两部分，数量较多的一部分样本作为训练集，用于让神经网络模型学习数据样本的特征，利用该部分数据样本进行拟合，并在训练集中划分一部分数据样本用于验证模型的样本内预测能力，另一部分数量较少的样本作为测试集，用于检验神经网络模型的样本外预测能力，即检验模型的泛化能力。如果模型在测试集预测准确度很高而测试集的表现很差，说明模型存在过拟合，需要进行模型结构和参数的调整优化，只有当模型在测试集和训练集的表

现都较好时，才能说明模型学习到了数据集上的特征，可以将此时的模型保存并用做后续的选股策略构建。

本文研究对象为上证 50 指数成分股，在剔除停牌等缺失样本后的有效样本仅有 100307 个，通常来说，要想在神经网络模型中取得较好的训练效果，需要数据样本量越多越好，此外，随着时间推移，越早的数据包含的信息量对当前数据的贡献会越来越小，模型的预测效果会衰减，因此，单将数据划分为一个训练集和一个测试集不能准确反映模型预测效果，为解决样本量不足，变相增加神经网络训练的样本量，同时消除策略构建部分的未来函数影响，在样本数据集划分时采用滑窗滚动的方式。按照上述的数据划分方式，本文将 2011 至 2017 年作为第一个训练集，2018 年作为第一个测试集，将 2012 至 2018 年作为第二个训练集，2019 年作为第二个测试集，将 2013 至 2019 年作为第三个训练集，2020 年作为第二个测试集，具体的数据集划分方式如图 4-2 所示。

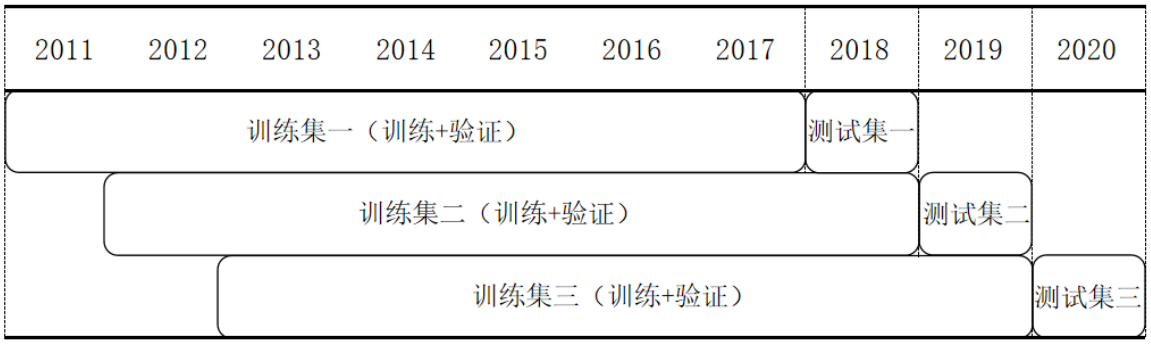


图 4-2 样本数据集划分方式

经过数据集划分后各子数据集的样本标签分布情况见表 4-4，各训练集样本涨跌的比例整体维持在 1:1，可以认为样本分布是均衡的。

表 4-4 各数据集样本标签分布统计

时间	标签类别	
	上涨	下跌
训练集一：2011-2017	32704	33430
测试集一：2018	4768	5955
训练集二：2012-2018	33617	34456
测试集二：2019	6495	5020
训练集三：2013-2019	35397	35114
测试集三：2020	5709	6226

4.3 模型设计与构建

4.3.1 模型的评价指标

本文构建的神经网络股价预测模型为对股票价格上涨下跌趋势的判断 ，属于二分类问题，并且涨跌样本分布均衡，不存在样本不平衡的问题，可以将准确度（Accuracy）作为模型主要的评价指标，同时将精确率（Precision）、召回率（Recall）和$F_\beta - score$作为辅助参考的评价指标，通过 sklearn 库中的函数可以直接调取查看模型在训练集和测试集上的评价指标。为了更清楚地解释这四个评价指标，可以借助混淆矩阵来辅助理解，混淆矩阵如表 4-5 所示，主对角线上的元素为上涨下跌样本正确分类的个数。

表 4-5 二分类混淆矩阵

混淆矩阵		预测值	
		上涨	下跌
实际值	上涨	T_U	$F_{U,D}$
	下跌	$F_{D,U}$	T_D

在表 4-5 所示的混淆矩阵中，假设上涨为正例，则下跌为反例，因此T_U、$F_{U,D}$、$F_{D,U}$和T_D分别代表了二分类中的真正例、假反例、假正例和真反例。

准确率也叫准确度，该指标描述了在所有样本中，正确分类的样本总数占据的比例，本例中准确率为主对角线上元素之和除以总样本数，计算方法如式 4.26 所示。

$$Accuracy = \frac{T_U+T_D}{T_U+F_{U,D}+F_{D,U}+T_D} \tag{4.26}$$

精确率也叫查准率，该指标描述了在所有预测为上涨的样本中，真正上涨的样本总数占据的比例，也就是真正例占预测为正例的比例，计算方法如式 4.27 所示。

$$Precision = \frac{T_U}{T_U+F_{D,U}} \tag{4.27}$$

召回率也叫查全率，该指标描述了在所有实际为上涨的样本中，真正上涨的样本总数占据的比例，也就是真正例占实际为正例的比例，计算方法如式 4.28 所示。

$$Recall = \frac{T_U}{T_U+F_{U,D}} \tag{4.28}$$

从式 4.27、4.28 中可以看出，查准率和查全率负相关，因此单看一个指标不能很好地判断分类准确性。F1-score 是精确率和查全率的调和平均，在评价模型分类效果时可

以参考F1-score指标，但在实际的一些分类问题中，对查准率和查全率的重视程度不同，比如本文研究的对股票涨跌分类问题，我们更重视在预测为上涨的样本中有多少是真正例，即更关注查准率，因此可以利用 F1 度量的一般形式——F_β作为模型的评价指标，F_β的计算方法如式 4.29 所示。

$$F_\beta = \frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R} \tag{4.29}$$

当β等于 1 时，上式就变为了 F1-score，此时对查准率和查全率是等权重关注的，当β大于 1 时，查全率更重要，当β小于 1 时，查准率更重要，本文将β值取 0.7，采用$F_{0.7}$代替 F1-score 指标。

4.3.2 初始模型的网络结构设置

本文在构建神经网络模型时部分借鉴了已有研究的模型架构，最终构建的 CNN-LSTM 模型的基本框架如图 4-3 所示。该模型由两部分组成，第一部分为 CNN 神经网络，第二部分为 LSTM 神经网络，两个神经网络架构之间通过 Lambda 层来连接，Lambda 层可以实现对数据结构进行调整，让 CNN 神经网络的输出数据经处理后可以输入至 LSTM 层中。

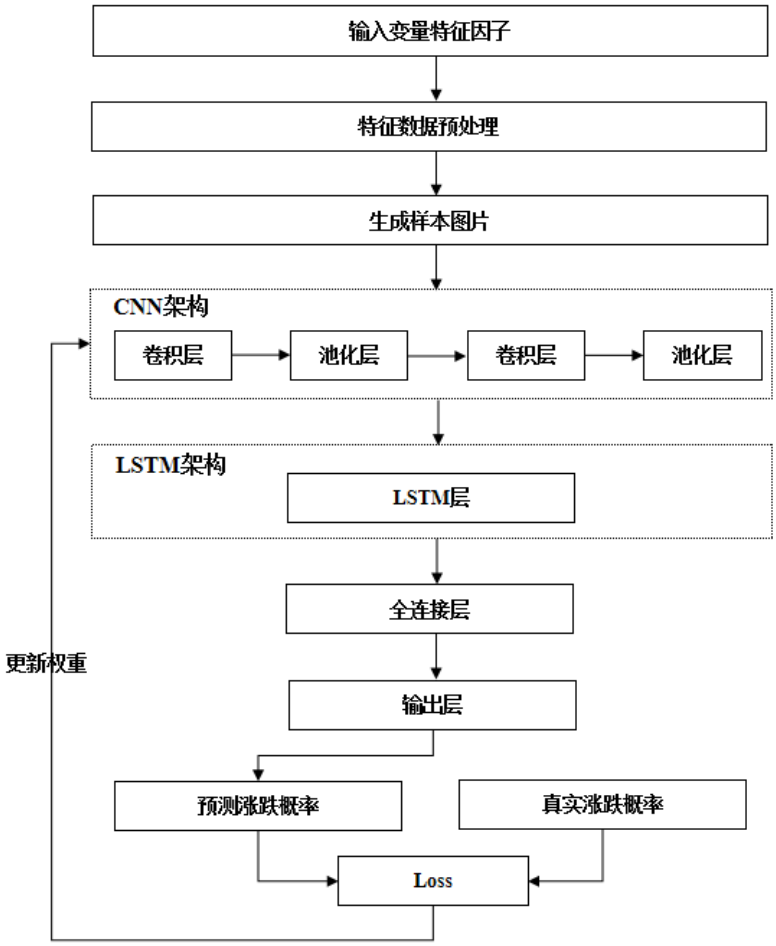


图 4-3 CNN-LSTM 预测模型基本框架图

卷积神经网络设计部分以经典的 CNN 神经网络构架 LeNet-5（图 3-6）为基础，由于传入 CNN 神经网络模型的样本结构为“20×T”的二维样本图片，卷积核在两个维度上提取特征，因此使用 Tensorflow.Keras 库中的二维卷积 Conv2D 函数来构建卷积层，池化层是卷积神经网络的核心要素，考虑到本文输入变量特征因子数量较多，因此在卷积层后面加入池化操作，LeNet-5 架构中有两次卷积池化操作，因此，本文同样设置两个卷积层和两个池化层，第二个池化层后面接一个卷积层，这个卷积层的卷积核为 1，目的是为 Lamba 层的输入做数据结构调整准备，在 Lamba 层之后接 LSTM 层，在 LSTM 层数设置上，Chen 等人的研究发现适当增加 LSTM 网络层数可以增强输入序列的特征提取[50]，因此先将 LSTM 层数暂定为 1 层，在 4.4.1 小节中具体探讨该参数对模型预测准确性的影响，在 LSTM 层之后再接两层全连接层，第一个全连接层用于对前面各层网络提取的特征做组合，第二个全连接层用于将特征信息转换为二分类中涨跌的概率。

将输入变量为 5 个特征因子的 CNN-LSTM 模型命名为 M0，将输入变量为 20 个特

征因子的 CNN-LSTM 模型命名为 M1，初始模型的的具体网络结构如表 4-6 所示。

表 4-6 初始模型网络参数一览表

层数	名称	网络参数	数据结构	激活函数
（1）	输入层	无	（N,20,T,1）	无
（2）	卷积层	kernel_num=16 kernel_size=(2,2)	（N,20,T,16）	ReLU
（3）	池化层	pool_size=(2,2)	（N,10,T/2,16）	无
（4）	卷积层	kernel_num=32 kernel_size=(2,2)	（N,10,T/2,32）	ReLU
（5）	池化层	pool_size=(2,2)	（N,5,T/4,32）	无
（6）	卷积层	kernel_num=1 kernel_size=(2,2)	（N,5,T/4,1）	ReLU
（7）	Lambda 层	无	（N,5,T/4）	无
（8）	LSTM 层	节点数：64	（N,64）	sigmoid/tanh
（9）	全连接层	节点数：64	（N,64）	sigmoid
（10）	Dropout 层	dropout rate = 0.5	（N,64）	无
（11）	全连接层	节点数：2	（N,2）	softmax

对模型的解释：

第一层为输入层，输出数据结构为（N,20,T,1），N 是指训练集的总体样本量，（20,T）代表用过去 20 日的 T 个因子来预测未来 5 日股价收益率变动情况，1 指通道数，在图像识别领域中，通道数是由红绿蓝（RGB）三通道构成的，本模型输入数据类似于手写识别 minist 数据集，因此通道数设置为 1。第二层为卷积层，有 16 个卷积核，每个卷积核的边长为 2，同时为了保持数据尺寸不变，采用填充处理，第三层为池化层，池化面积为 2×2，步长在时间维和因子维都为 2，池化的目的在于对卷积后的样本特征做进一步的下采样，第四层为卷积层，有 32 个卷积核，卷积核边长和第二层相同且仍为 2，第五层为池化层，其作用和第三层一样，用于对样本特征做下采样，第六层为卷积层，有 1 个卷积核，卷积核边长为 2，第七层为 Lambda 层，第八层为 LSTM 层，该层神经元节点设置为 64，第九层为全连接层，该层神经元节点同样设置为 64，第十层为 Dropout

层，丢弃率设置为 0.5，该层的作用在于缓解模型的过拟合现象，第十一层为输出层，采用 softmax 激活函数，选择交叉熵“categorical_crossentropy”作为模型的损失函数，尽管交叉熵常用于多分类问题，但由于本文已将标签转换为独热向量，输出的是上涨下跌二分类每个类别的概率，因此采用该交叉熵作为模型的损失函数，此外，在每层卷积层后面加入 BatchNormalization 层，一方面缓解过拟合现象，同时能够加快神经网络的训练和收敛的速度[51]。

4.3.3 初始模型训练及预测效果展示

初始模型网络结构搭建完成后，在训练模型之前还需对包括学习率、batch_size、epochs 和 validation_split 在内的参数进行设置。

学习率是指神经网络在误差反向传播时每次更新权重参数的步长，太小或太大都不利于模型收敛，本文选取 Adam 作为优化器，将学习率设定为 0.0001。batch_size 是指单个训练批次样本数量，也就是每一次抓取训练集上多少样本来更新权重参数，batch_size 设置的太小会导致梯度更新速度过慢，导致损失函数波动较大，设置的太大会导致神经网络陷入损失函数的局部最优解，难以找到真正的全局最优解，本文通过对模型的预调试，选取 batch_size 为 128。validation_split 是指用于验证神经网络模型损失函数的样本占训练集样本的比例，通常将数据集划分成训练集和测试集，直接用测试集来调试模型预测效果会造成人为的过拟合，因为这样做相当于针对测试集的表现来有目的的调试模型，会弱化模型的泛化能力，因此验证集不使用测试集数据，而是指定一部分训练集数据来验证模型的效果，本文将 validation_split 设置为 0.1，在指定验证集的同时需要设置 Shuffle，该值决定了对数据样本是否打乱顺序进行训练，将 Shuffle 设置为 True。epochs 为用全部样本对模型进行一轮训练，以训练集三数据为例，训练初始模型并观测损失函数和准确率的变动趋势，模型的在训练集和验证集的准确率如图 4-4 所示。观察图 4-4 可以发现模型在训练 40 个 epoch 左右达到较为平稳的状态，此时模型在验证集的准确率已经趋于平稳，但考虑到三个数据集都要进行训练，每个数据集达到收敛时的迭代次数不一定完全一致，因此本文将 epoch 设置为 100，较大的 epoch 也有利于模型权重参数尽可能迭代至最优。

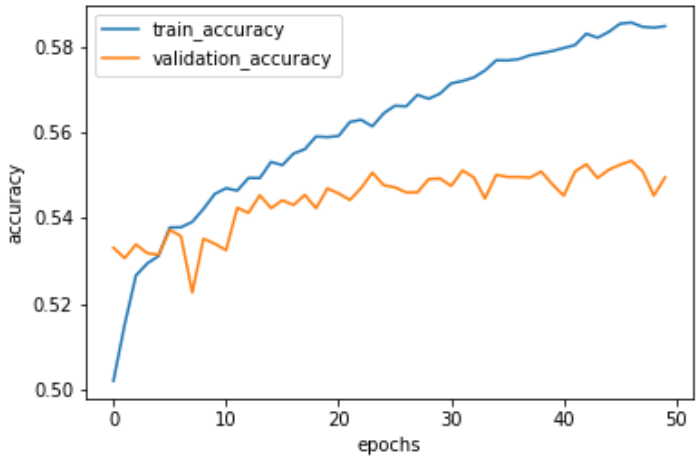


图 4-4 训练集三准确率变动趋势图

本文按照初始模型的架构来搭建股价涨跌预测模型，然后分别在三个训练集上训练，在三个训练集各自对应的测试集上完成对训练好的模型预测效果的评价。在模型训练期间，为了及时对最优的模型进行保存，在训练时调用了 Keras 库的 ModelCheckpoint 函数，该函数可以指定一个监控目标及条件，将监控目标设置为验证集损失函数“val_loss”，条件为“min”，只有当每轮训练后验证集的损失函数有下降才保存模型，经过对模型的训练，两组不同输入变量的神经网络初始模型在各数据集上测试年份的预测效果如表 4-7、表 4-8 所示。

表 4-7 M0 模型分数据集预测效果（实验组）

模型编号	数据集	准确率（%）	$F_{0.7}$
M0 输入变量 5 个特征因子	2011-2017	52.39	0.4824
	2012-2018	52.01	0.5511
	2013-2019	51.83	0.5070
	Avg.	52.11	0.5135

表 4-8 M1 模型分数据集预测效果（对照组）

模型编号	数据集	准确率（%）	$F_{0.7}$
M1 输入变量 20 个特征因子	2011-2017	55.32	0.4856
	2012-2018	55.82	0.6372
	2013-2019	54.29	0.4286
	Avg.	55.51	0.5171

从表 4-7、表 4-8 可以看出，将个股的“四价一量”基本交易指标作为实验组 M0 模型的输入变量已经可以得到 52.11%的涨跌预测准确率，大于随机猜测的 50%概率，初步说明个股的“四价一量”和股价涨跌有一定的相关性，可以根据个股基本交易指标对股价涨跌进行预测，而在输入变量中增加了上证 50 指数的“四价一量”指标和个股的技术指标后，对照组 M1 模型在 2018-2020 三年的测试集上的平均预测准确率为 55.51%，相比 M0 模型准确率提升了 3.4%，预测效果有所提升，说明在模型输入变量中加入上证 50 指数指标和个股技术指标可以增强模型预测能力，因此可以认为适当增加深度神经网络输入变量的特征因子数量能够改善模型的预测准确度，而 M0 和 M1 模型的$F_{0.7}$分数仅相差 0.36%，说明两个模型输出的涨跌预测结果在分布上大致相等，综上，从提升模型预测准确度的角度出发，本文选取个股基本交易指标 T1、上证 50 指数指标 T2 和个股技术指标 T3 这三类共 20 个因子作为模型的输入变量，将对照组 M1 模型设定为本文接下来研究的模型基准。

由于 M1 模型是基于已有的 CNN 神经网络架构（LeNet-5）和 LSTM 神经网络架构搭建的，模型在一些细节上还未做优化，为了获得更好的预测结果，需要继续在 M1 模型的基础上对各层神经网络结构及参数进行调整，通过对模型参数的调试来找出预测效果最佳的模型，并将此模型作为第五章构建策略时的最终依据，在对模型结构及参数调优的过程中研究不同参数对于模型预测效果的影响。

4.4 模型结构及参数优化

影响神经网络模型预测效果的参数可大致分为网络参数、优化参数和正则化参数三类。网络参数主要指神经网络隐藏层的参数设置，在 M1 模型中，第二层至第十层全部都属于神经网络的隐藏层，隐藏层的网络参数包括了每个卷积层中卷积核的大小及数量、LSTM 层的个数及网络节点数、全连接层中的网络节点数，优化参数主要是优化器的选择，正则化参数主要包括权重初始化方式和 dropout rate 大小的设定。本文在研究某个具体参数对模型预测效果的影响时，仅以准确率为各组对照模型的评价指标，采取控制变量法，只对所研究参数做改动，其他参数不做变动，下面就各参数对 M1 模型的预测效果影响分别讨论。

4.4.1 网络参数对模型的影响

初始模型 M1 是由 CNN 神经网络和 LSTM 神经网络组合构建的，CNN 神经网络的最大特征是卷积核在样本数据中做卷积采样，卷积核尺寸越大，网络感受野的范围也就越大，模型能够学习到更整体的特征，但更大的卷积核也增加了模型的参数量，在实际应用中偏向于用多层小卷积核来替代大卷积核。初始模型三个卷积层的卷积核个数分别为 16、32、1，这里借鉴了 VGGNet 架构设计，对前两个卷积层的卷积核数量做翻倍处理，而且卷积核的个数逐层递增有利于模型对样本特征进行更高维度的组合变换。

（1）卷积核大小对模型预测效果的影响

在图像识别领域，已有成熟架构在选取卷积核大小多采用边长为 3 的设置，但也有使用大卷积核的架构，比如 GoogleNet 架构在设置卷积核大小时采用了边长为 5 和边长为 7 的卷积核，卷积核大小直接决定了感受野的大小，感受野又会影响卷积核每次卷积时的计算范围，通常大卷积核可以从输入变量中获得更多的特征，但大的卷积核计算所耗费的时间也会随之增大，因此大卷积核不利于模型深度的增加。在实际操作中，卷积核多采用边长为 3 的设置方式，并且倾向于用多个小卷积核代替一个大卷积核。

本文在 M1 模型的基础上，将对照模型的所有卷积层中卷积核大小设置为(3×3)、(4×4)和(5×5)，对照模型命名为 M1_1、M1_2 和 M1_3，分别在三个训练集上再次训练神经网络模型，模型的预测效果如表 4-9 所示。

表 4-9 改变卷积核大小对模型预测效果影响

模型编号	数据集	准确率（%）
M1_1 (kernel_size=3×3)	2011-2017	55.53
	2012-2018	54.46
	2013-2019	54.23
	Avg.	54.74
M1_2 (kernel_size=4×4)	2011-2017	54.97
	2012-2018	55.11
	2013-2019	54.73
	Avg.	54.94
M1_3 (kernel_size=5×5)	2011-2017	55.76
	2012-2018	54.05
	2013-2019	54.48
	Avg.	54.76
M1	Avg.	55.51

从表 4-9 可以看出，增加了卷积核大小后，对照模型的平均预测准确率相较于初始模型 M1 都有略微下降，并且三个对照模型的平均预测准确率基本相同，由此可见在本文构建的模型中，改变卷积核大小对模型预测准确率的影响不大，考虑到试验平台的性能以及增加卷积核给模型带来的效用，决定沿用初始模型的卷积核数量。

（2）卷积核数量对模型预测效果的影响

卷积核负责提取输入样本的特征，通常来说，每个卷积核只对一种特征进行提取，增加卷积核数量可以增加模型对输入样本特征的识别能力，因此，本文在 M1 模型的基础上设置对照模型，将第二、第四层卷积核的个数分别改为(16,16)、(32,32)、(32,64)、(64,64)，并将其命名为 M1_4、M1_5、M1_6、M1_7，分别在三个训练集上再次训练神经网络模型，模型的预测效果如表 4-10 所示。

表 4-10 改变卷积核数量对模型预测效果影响

模型编号	数据集	准确率（%）
M1_4 kernel_num=(16,16)	2011-2017	53.81
	2012-2018	54.10
	2013-2019	53.58
	Avg.	53.83
M1_5 kernel_num=(32,32)	2011-2017	56.75
	2012-2018	55.49
	2013-2019	54.12
	Avg.	55.45
M1_6 kernel_num=(32,64)	2011-2017	55.89
	2012-2018	55.28
	2013-2019	54.66
	Avg.	55.28
M1_7 kernel_num=(64,64)	2011-2017	55.41
	2012-2018	54.62
	2013-2019	52.38
	Avg.	54.14
M1	Avg.	55.51

从表 4-10 可以看出，在减少卷积层卷积核数量后，对照模型 M1_4 的平均预测准确率较初始模型 M1 有所下降，这是因为减少卷积核数量后，模型的参数也相应减少了导致模型的表达能力被弱化，而增加卷积核数量后，对照模型 M1_5、M1_6 相比初始模型 M1 的平均预测准确率并无明显变化，而对照模型 M1_7 的平均预测准确率有所下降，究其原因很可能是由于输入样本包含的有效信息是有限的，过多的卷积核数量提取了重复的特征信息限制了模型的泛化能力，因此，在卷积核数量上保持不变，仍采用初始模型 M1 的参数设置。

（3）LSTM 层数对模型预测效果的影响

除了卷积核数量外，LSTM 层数量也同样可能会对模型预测效果产生影响，初始模

型的 LSTM 层数量为 1，而参考已有研究，LSTM 层数通常就是 1-2 层，因此增加一层 LSTM 层结构来探究 LSTM 层数量对模型预测效果的影响，神经元节点设置为 64，将该对照模型命名为 M1_6，模型的预测效果如表 4-11 所示。

表 4-11 改变 LSTM 层数对模型预测效果影响

模型编号	数据集	准确率（%）
M1_8 增加一层 LSTM 层	2011-2017	55.57
	2012-2018	55.12
	2013-2019	53.27
	Avg.	54.65
M1	Avg.	55.51

从表 4-11 可以看出，增加了一层 LSTM 层后，模型在数据集上的的平均预测准确率相比初始模型 M1 下降了 0.86%，尽管已有部分研究表明适当增加 LSTM 层能够提升模型预测能力，但在本文研究模型中，一层 LSTM 层就已经能够保证模型预测的效果，因此仍保持初始模型 M1 中的一层 LSTM 层结构不变。

（4）全连接层网络节点的模型预测效果的影响

全连接层位于神经网络的末端，该层每个神经元都与前一层神经网络上的网络节点相连接，因而这层神经网络能够汇集前面网络层提取的特征信息并进行更深度的处理。初始模型 M1 中，第一个全连接层的神经元个数为 64，第二个全连接层的神经元个数为 2，由于第二个全连接层负责将特征输出为两个类别的概率，网络节点个数固定为 2，因此，只对第一个全连接层神经元节点做调整。本节设置对照模型，将全连接层的神经元个数分别减小为 32 和增加至 128、256 来探寻不同的神经元个数对于模型训练效果的影响，将对照模型命名为 M1_9、 M1_10、 M1_11，模型的预测效果如表 4-12 所示。

表 4-12 改变全连接层神经元个数对模型预测效果影响

模型编号	数据集	准确率（%）
M1_9 dense_num=32	2011-2017	55.60
	2012-2018	54.66
	2013-2019	53.36
	Avg.	54.54
M1_10 dense_num=128	2011-2016	56.25
	2012-2017	56.42
	2013-2018	55.45
	Avg.	56.04
M1_11 dense_num=256	2011-2016	54.90
	2012-2017	53.42
	2013-2018	54.21
	Avg.	54.18
M1	Avg.	55.51

从表 4-12 中可以看出，M1_10 模型在三个数据集的平均预测准确率最高，相比初始模型 M1 提升了 0.57%，说明适当增加全连接层的神经元节点数能够增强该层网络对输入样本的特征提取及组合能力，并且能够提升模型的表达能力，M1_11 模型将全连接层神经元节点数继续增加到 256，模型平均预测准确率却出现了下降的现象，查看模型具体参数，发现 M1_11 模型的可训练参数为 37459 个，相比 M1_10 模型的可训练参数为 28883 个，提升了近 30%，更多的参数虽然提升了模型对训练集数据的拟合能力，但也限制了模型的样本外预测能力，表现在模型在测试集的预测准确率出现下降的现象，从提升模型预测准确率的角度出发，本文将全连接层的神经元节点数由 64 个替换为 128 个。

4.4.2 优化参数对模型的影响

在神经网络模型的训练过程中，模型通过误差反向传播来更新权重参数，并且通过不断多次迭代来逼近最优解，最终确定一组权重参数使得模型输出的预测值和真实值之间的损失函数达到最小化，当模型网络层数较多且参数数量较大时，这一过程可能会耗

费较长时间，为了加快训练速度，不同的优化器基于各自原理提供了不同的优化路径，选择正确的优化器可以大大缩减模型训练的时间，从而提高神经网络的训练效率。按照不同的原理可以将优化器分为梯度下降法、动量优化法和自适应学习率优化法，其中典型的代表有 SGD、AdaGrad、Adadelta、RMSProp 和初始模型 M1 中使用的 Adam，由于 SGD 优化器模型收敛的时间常常是其他优化器的数倍，模型难以收敛，考虑到训练时间，本文将设置对照模型，重点研究 AdaGrad、RMSProp 和 Adam 优化器对模型预测效果的影响，模型的预测效果如表 4-13 所示。

表 4-13 改变优化器对模型预测效果影响

模型编号	数据集	准确率（%）
M1_12 优化器：AdaGrad	2011-2017	55.48
	2012-2018	55.81
	2013-2019	52.58
	Avg.	54.62
M1_13 优化器：Adadelta	2011-2017	55.90
	2012-2018	54.26
	2013-2019	52.08
	Avg.	54.08
M1_14 优化器：RMSProp	2011-2017	55.39
	2012-2018	56.47
	2013-2019	54.19
	Avg.	55.35
M1	Avg.	55.51

从表 4-13 可以看出，在使用不同的优化器来训练神经网络模型时得到的模型预测效果有所差异，四种优化器中，使用 RMSProp 优化器的 M1_14 模型的预测效果与初始模型 M1 最为接近，其平均预测准确率仅比 M1 模型降低 0.16%，而使用 AdaGrad 的 M_12 模型与使用 Adadelta 优化器的 M_12 模型相比 M1 模型预测效果下降较为明显，此外，在使用 AdaGrad 和 Adadelta 训练神经网络模型时，模型收敛所耗费的时间相比 RMSProp 和 Adam 更长，因此，沿用初始模型 M1 的 Adam 优化器不变。

4.4.3 正则化参数对模型的影响

过拟合是深度神经网络模型中最常见的问题，表现在训练神经网络过程中，模型的相关监测指标不再继续优化，比如，模型在训练集上的 loss 不断下降，而在验证集和测试集的 loss 不断上升，说明模型在不断学习训练集的数据特征，但对于样本以外的数据拟合效果不佳，模型的泛化能力不强。正则化是一种缓解模型过拟合的手段，这种方法旨通过在目标函数后面加上一个正则项来约束神经网络的连接权重，除了正则化，丢弃率法是实际应用中使用最多的缓解过拟合的方法，通过指定一个丢弃比率 d，按照比例将每层神经元进行失活处理，只保留 1-d 这部分神经元来拟合训练集数据。Dropout 层一般设置在全连接层前后，因为全连接层上每个神经元都和上层节点相互连接，最容易产生多重共线性，从而造成过拟合。太小的 d 系数起不到限制过拟合的作用，太大的 d 系数虽然限制了过拟合，但是在相同训练轮数下会降低模型的准确度，也就是说达到相同的预测准确度需要训练更长时间，本文设置对照模型，将 dropout 分别设置为 0、0.25、0.75 来研究不同丢弃比率对模型预测效果的影响，模型的预测效果如表 4-14 所示。

表 4-14 改变 dropout 比率对模型预测效果影响

模型编号	数据集	准确率（%）
M1_15 d=0	2011-2017	53.75
	2012-2018	54.32
	2013-2019	52.66
	Avg.	53.58
M1_16 d=0.25	2011-2017	55.56
	2012-2018	54.82
	2013-2019	53.33
	Avg.	54.57
M1_17 d=0.75	2011-2017	56.09
	2012-2018	53.73
	2013-2019	53.99
	Avg.	54.60
M1	Avg.	55.51

将表 4-14 整理为如图 4-5 所示的柱状图，可以看出，随着 dropout 的增大，模型在数据集上的平均预测准确率呈现出先上升再下降的趋势，说明增加丢弃比率能在一定程度上提高模型在样本外即测试集上的准确度，但过于大的 dropout 会限制模型的权重更新效率，导致模型在相同的训练迭代次数下得到了较低的预测准确度，因此，仍沿用初始模型 M1 丢弃比率，将 d 设置为 0.5。

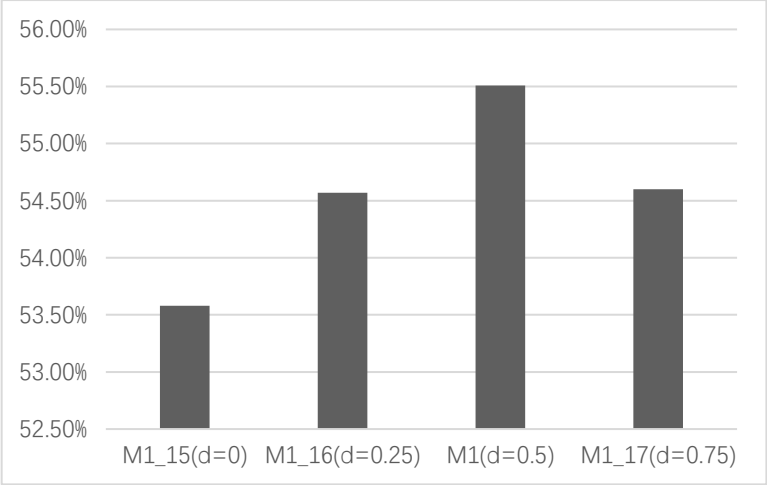


图 4-5 准确率变动趋势图

4.4.4 最终模型预测结果展示

通过本章前几个小节设置对照模型的实验分析工作，本文在初始模型 M1 的基础上以提升模型预测准确度为目标，按照控制变量的原则进行调参，最终对全连接层的神经元节点数做了调整，由 64 个增加至 128 个，而其他参数诸如卷积核大小、卷积核数量、LSTM 层数、优化器的选择对模型预测效果的提升作用有限，因此未做变更，仍采用初始模型 M1 的设置方式，调参后的最终模型参数如表 4-15 所示。

表 4-15 CNN-LSTM 最终模型要素表

模型编号	卷积核尺寸	卷积核数量	全连接层神经元个数	优化器	Dropout 比率	准确率
M1_10	2×2	（16,32,1）	128	Adam	0.5	56.04%

利用最终模型 M1_10 对三个数据集进行训练，各测试集的预测结果的混淆矩阵以及相应的评价指标总结如下。

表 4-16 测试集一（2018 年）预测结果混淆矩阵

混淆矩阵		预测值	
		上涨	下跌
实际值	上涨	2876	1692
	下跌	2999	3156

表 4-17 测试集一（2018 年）预测结果评价

标签类别	评价指标		
	精确率	召回率	$F_{0.7}-score$
上涨	0.4895	0.6296	0.5281
下跌	0.6510	0.5128	0.5980
Avg.	0.5703	0.5712	0.5631

表 4-18 测试集二（2019 年）预测结果混淆矩阵

混淆矩阵		预测值	
		上涨	下跌
实际值	上涨	3295	2833
	下跌	2186	3201

表 4-19 测试集二（2019 年）预测结果评价

标签类别	评价指标		
	精确率	召回率	$F_{0.7}-score$
上涨	0.6012	0.5377	0.5787
下跌	0.5305	0.5942	0.5499
Avg.	0.5659	0.5660	0.5643

表 4-20 测试集三（2020 年）预测结果混淆矩阵

混淆矩阵		预测值	
		上涨	下跌
实际值	上涨	2818	2891
	下跌	2426	3800

表 4-21 测试集三（2020 年）预测结果评价

标签类别	评价指标		
	精确率	召回率	$F_{0.7}-score$
上涨	0.5374	0.4936	0.5222
下跌	0.5679	0.6103	0.5812
Avg.	0.5527	0.5520	0.5517

从表 4-17、表 4-19 和表 4-21 对各年度预测结果的评价可以看出，在 2018 年和 2020 年的预测中，模型对下跌的预测精确率高于对上涨的预测精确率，而在 2019 年的预测中，模型对上涨的预测精确率高于对下跌的预测精确率，结合表 4-4，发现 2018 年和 2020 年下跌的样本多于上涨样本，而 2019 年上涨样本多于下跌样本，说明模型对不同类别的预测表现会受到不同类别样本数量及分布的影响，从整体的预测表现来看，最终模型 M1_10 在各测试集上的平均$F_{0.7}$分数也都分别到达了 0.5631、0.5643 和 0.5517，表明模型在综合考虑精确率和召回率后，仍能取得较为不错的表现，可以将经过调优的最终模型 M1_10 的预测结果作为接下来构建选股策略的依据。

4.5 CNN-LSTM 模型同其他神经网络模型对比

在第三章相关概念和理论基础中，本文对 CNN 神经网络、RNN 神经网络及 LSTM 神经网络的算法原理进行了梳理和比较，并且在本章的模型构建中同时用到了 CNN 和 LSTM 两种网络模型架构，那么相比单独仅使用一种神经网络架构的单模型，将 CNN-LSTM 相结合的模型预测效果是否有提升同样值得研究分析，因此，本节选取 CNN 神经网络、LSTM 神经网络以及 BP 神经网络三种机器学习中常用的神经网络模型作为对照实验组，通过不同模型之间预测效果的比较开展研究。

在构建 CNN 神经网络对照模型时，由于模型输入变量和 M1_10 模型完全相同，仍是 N 个 20×20 的样本图片，数据结构为（N,20,20,1），因此无需对输入变量进行数据结构上的调整，对照模型的网络结构参考了 LeNet-5 架构（图 3-6），由一层输入层，两层卷积层和池化层，一层 flatten 层，一层全连接层，一层 Dropout 层和一层输出层构成，并在每层卷积层后面同样加入了 BatchNormalization 层来防止过拟合，在模型参数设置方面，将两层卷积层中卷积核数量都设置为 16，其余如卷积核大小及全连接层节点数等

参数的设置同 M1_10 模型保持一致，将构建好的 CNN 神经网络模型用于预测各测试集年份股票涨跌，模型预测结果如表 4-22 所示。

表 4-22 CNN 神经网络模型预测结果

模型类别	数据集	准确率（%）
CNN	2011-2017	54.30
	2012-2018	54.46
	2013-2019	53.70
	Avg.	54.15

在构建 LSTM 神经网络对照模型时，模型输入变量的数据结构变为（样本数，时间维，特征维）的格式，相比 M1_10 模型少了一个维度，因此只需要通过 reshape 函数将原始输入数据重构为（N,20,20）即可，对照模型模型由一层输入层，一层 LSTM 层，一层全连接层，一层 Dropout 层和一层输出层构成，其中 LSTM 层和全连接层的神经元节点都设置为 64 个，将构建好的 LSTM 神经网络模型用于预测各测试集年份股票涨跌，模型预测结果如表 4-23 所示。

表 4-23 LSTM 神经网络模型预测结果

模型类别	数据集	准确率（%）
LSTM	2011-2017	52.29
	2012-2018	53.17
	2013-2019	53.94
	Avg.	53.13

BP 神经网络模型是基于全连接神经网络构建的，在构建 BP 神经网络对照模型时，模型输入变量的数据结构同 LSTM 神经网络模型一致，因此也需要对原始输入数据进行重构，对照模型由一层输入层、两个隐藏层和一层输出层构成，隐层的神经元节点分别设置为为 32 和 64，并使用 ReLU 函数实现去线性化，输出层的激活函数保持 Softmax 函数不变，将构建好的 LSTM 神经网络模型用于预测各测试集年份股票涨跌，模型预测结果如表 4-24 所示。

表 4-24 BP 神经网络模型预测结果

模型类别	数据集	准确率（%）
BP	2011-2017	54.52
	2012-2018	52.97
	2013-2019	52.86
	Avg.	53.45

通过对表 4-22、表 4-23、表 4-24 的比较，可以发现 CNN 神经网络模型在三组对照模型中的预测效果最好，准确率达到了 54.15%，说明卷积神经网络更适和对分类问题进行建模，但仍低于本文构建 M1_10 模型的预测准确率 1.89 个百分点，LSTM 神经网络模型和 BP 神经网络模型在上证 50 指数成分股涨跌预测中准确率分别为 53.13%和 53.45%，预测结果较为接近，但也都高于随机猜测的 50%概率，说明 LSTM 和 BP 神经网络同样可以用于股价涨跌的预测，综合比较本文构建的 M1_10 模型和对照实验组的三个神经网络模型，可以发现基于 CNN-LSTM 两种架构的双模型预测准确度最高，将两种架构相结合可以提升模型的预测效果。

4.6 本章小结

本章首先对构建股价预测模型用到的实验平台和实验数据来源进行了介绍，在数据预处理部分介绍了异常值、空值的处理方法以及数据归一化的方法，选取模型输入变量特征因子时，本文在“四价一量”的基本交易数据指标基础上，增加上证 50 指数指标、个股技术指标作为对照实验组，在对各因子做无量纲处理后，在研究区间上以 20 天为滑动窗口，1 天为步长，生成 N 张尺寸为 20×T 的样本作为模型的训练数据（T 为特征因子个数），并且独热编码的形式对所有样本按照个股未来 5 个交易日后的涨跌进行标记，对数据集进行划分时同样采用了滑窗的方式分别生成了“2011-2018”、“2012-2019”、“2012-2020”三个数据集，用各数据集上前七年数据训练和验证模型，用第八年数据测试模型。

随后本文基于经典的卷积神经网络 LeNet-5 架构建了初始模型，通过比较两组不同输入变量特征因子的模型预测准确率，确定了将对照实验组模型 M1 的 20 个因子作为模型输入变量。为了对现有模型 M1 进行结构和参数的优化以获得更好的预测效果，从

网络参数、优化参数和正则化参数三类共 6 种参数出发设置多个对照模型，通过修改各类参数来探究不同参数变化对模型预测效果的影响，最后确定了预测准确度最高的对照模型 M1_10，将该模型与 CNN 神经网络、LSTM 神经网络以及 BP 神经网络模型进行对比分析后，发现基于 CNN-LSTM 两种架构构建的 M1_10 模型对上证 50 指数成分股价格涨跌的预测表现在四组神经网络模型中最好，预测准确率明显高于其他三种神经网络模型。

第五章 量化选股策略构建

5.1 策略思想

个人投资者在股票二级市场交易时，大都采取的是主观交易策略，即通过个人对上市公司的经营状况、公司所处行业发展以及宏观经济预期来做出买卖决策。而量化交易区别于主观交易，将缜密的数学模型同先进的计算机技术相结合，能够避免传统股票交易中人在做投资决策时的非理性判断，量化交易策略则涉及到交易的细节，包括股票标的的选择、买卖时机的确定以及仓位的管理，不少私募证券投资基金公司发行的私募产品都是基于量化交易开发的，私募基金能够获取高出同期公募基金较大的收益，但相对而言，私募基金的风险也高于公募基金，因此只面向合格投资者发行。

本文研究的是 CNN-LSTM 神经网络在上证 50 指数成分股中对股价涨跌的预测效果，在上一章的研究中已经确定了最终模型 M1_10 的网络结构以及具体的模型参数，本章构建交易策略时将根据该模型在三个测试集上的预测结果进行选股，具体的策略思想如下：由于模型输出层的神经网络节点数为 2，并且激活函数采用了 softmax 函数，模型输出的是个股上涨和下跌的预测概率值，并且取值范围在 0-1 之间，因此在选股时可以根据模型输出的未来 5 个交易日后每只股票涨跌的概率做研判，基于此，提出两个选股方案。

方案一：每个调仓日对 50 只股票的上涨和下跌概率进行排序，选取其中上涨概率排名前 10%的股票，即排名靠前的 5 只股票构建组合做买入开仓操作，同时对下跌概率排名靠前的 5 只股票做卖出平仓操作，由于 A 股市场没有做空机制，因此，在仅在当前持仓股中有下跌概率排名靠前的股票时才执行卖出。

方案二：由于模型输出的是每只股票的涨跌概率，而涨或跌的输出值都在 0-1 之间，因此，可以设置一个阈值 θ，对上涨概率大于阈值 θ 的股票做买入开仓操作，对下跌概率大于阈值 θ 的股票做卖出平仓操作，同样仅在当前持仓股中有下跌概率大于阈值 θ 的股票时才执行卖出。

上述两种方案的区别在于，方案一每个调仓日买入卖出的股票个数是确定的，并且

在涨跌概率相近的情况下更容易选股，而方案二每个调仓日买入卖出的股票个数不确定，并且潜在的隐患是有可能因为阈值设置过于严苛导致没有符合条件的个股产生从而无法进行交易，本文将在接下来通过回测比较两种选股方案的收益效果。

在调仓频率及仓位管理的设置上，由于模型训练时使用的是股票的日频交易数据，并且是对个股未来 5 个交易日后的涨跌做预测，因此，将调仓频率设置为周调仓，在每周的第一个交易日进行调仓，每个调仓日使用可用资金的十分之一对待买入的股票做等现金买入操作，对待卖出的股票直接做清仓卖出操作。

在进行选股策略回测时，本文使用了聚宽量化交易平台完成回测，由于在回测环境中重新训练神经网络模型耗时较长，本文将 M1_10 模型在本地端对回测期间年份，也就是 2018 年至 2020 年的上证 50 指数成分股每日涨跌概率进行预测，并将模型输出的涨跌概率导入聚宽平台的策略环境并生成每日的买卖信号进行回测，为了更真实地模拟股票交易，在正式回测前还需对初始资金、交易手续费进行设定，将初始资金设置为 100 万元，交易手续费为买入时万分之三，卖出为万分之三加千分之一的印花税，每笔交易手续费最低为 5 元。

5.2 策略评价指标

聚宽量化交易平台的回测环境提供了多项策略评价指标，按照所属可以分为收益类指标和风险类指标两大类，其中收益类指标有策略收益（Total Return，以下简称 TR）、策略年化收益（Total Annual Return，以下简称 TAR）和超额收益（Alpha），风险类指标有夏普比率（Sharp Ratio，以下简称 SR）、信息比率（Information Ratio，以下简称 IR）和最大回撤（Max Drawdown，以下简称 MD）。

策略收益是策略在回测期间的总收益率，能够最直观地反映一个交易策略的收益能力。策略年化收益是指回测时期超过一年时，将策略总收益折算为一年期的收益率。超额收益是衡量策略收益的一个重要指标，具体指策略收益超出基准收益部分的收益水平，考虑到股票资产变动的复利效应，聚宽平台提供的超额收益为除法版超额收益，即“超额收益=（策略收益+100%）/（基准收益+100%）-1”，通常在回测前设定好一个基准，通过将策略收益与基准收益相比较，观察所构建策略获取超额收益的能力，本文在回测中将基准收益设置为上证 50。

倘若只根据策略的收益状况判断策略的好坏并不是十分全面，因为作为个人投资者可以只关注策略的收益，但机构投资者，比如大型公募基金和多数私募基金在基金募集成立前就已经设置好了基金份额的预警线和止损点，相比个人投资者，基金管理人更重视策略收益的稳健性，策略收益波动较大更容易触发基金份额预警线，进而引发投资者赎回持有份额，不利于基金管理人的长期管理，因此通过监测策略的风险指标，可以及时发现策略存在的问题并进行相应的优化调整。

夏普比率是一个反映承担单位风险获取超额收益的指标，该指标越高，说明策略在承担相同风险情况下，能够获得更多的超额收益，这也是一个非常重要的策略评价指标。信息比率是从主动投资管理角度出发衡量策略收益相对于基准收益的风险调整后的收益回报。最大回撤是交易策略重点关注的风险评价指标，具体指在回测期间内，从任意时点往后推算，策略净值下落至最低点时，策略收益率回撤幅度的最大值。

综上，本文将通过以上六个指标来对构建的策略进行评价、比较和分析。

5.3 策略回测效果分析

5.3.1 牛熊市划分

为了比较选股策略在不同市场状态下回测效果，需要对回测期间的市场状态进行区分，本文在实际操作时参考了 Pagan 和 Sossounov（2003）[52]以及何兴强和周开国（2006）[53]对牛、熊市周期的判别方法，通过非参数法寻找股市价格变化的波峰和波谷来判断牛、熊市周期，具体做法如下：

首先，假设股市月度价格水平为P_t，则月度对数价格$p_t = \log(P_t)$。如果p_t为窗口长度为 3 个月中的最大值，即满足式 5.1，则 t 为一个波峰，如果p_t为窗口长度为 3 个月中的最小值，即满足式 5.2，则 t 为一个波谷。

$$p_{t-3}, p_{t-2}, p_{t-1} < p_t > p_{t+1}, p_{t+2}, p_{t+3} \quad (5.1)$$

$$p_{t-3}, p_{t-2}, p_{t-1} > p_t < p_{t+1}, p_{t+2}, p_{t+3} \quad (5.2)$$

其次，要求波峰波谷交替出现，同时将连续波峰中的较低波峰和连续波谷中的较高波谷剔除，此外，考虑到短期的急涨急跌对行情的影响以及尽可能排除短期行情造成的虚假牛熊市，要求划分的牛熊周期满足：①若牛熊市的单程时长未满 4 个月，则价格逆转前后的涨跌幅度须大于 20%；②牛熊市周期长度应该大于半年；③波峰和波谷距离端

点应该大于 4 个月。

按照非参数法的条件和方法对 2018 年 1 月至 2020 年 12 月期间的上证 50 指数进行牛熊市行情判别，得到如下结论。

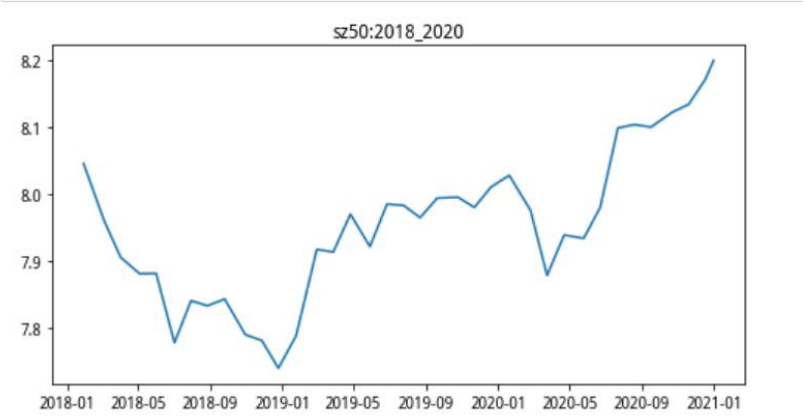


图 5-1 2018 年-2020 年回测期间上证 50 对数价格月度走势图

首先，调取如图 5-1 所示的上证 50 指数月度价格走势，根据何兴强等人的牛熊市判别依据，可以得出 2018 年 7 月、2018 年 12 月和 2020 年 3 月为波谷，2018 年 9 月和 2020 年 1 月为波峰，剔除掉不满足条件的波峰波谷后，得到如表 5-1 所示的市场状态划分结果。

表 5-1 市场状态划分结果

回测期间	牛熊市类别
2018.1-2018-12	熊市
2019.1-2020.12	牛市

由表 5-1 可见，本文构建交易策略回测期间包含了至少一个完整的牛熊市周期，在完整的牛熊市期间中构建选股交易策略能够检验策略的普适性和实用性，并且能够体现模型预测结果的实际意义和价值。

5.3.2 策略牛熊市表现

将 M1_10 模型在测试集一（2018 年）、测试集二（2019 年）和测试集三（2020 年）上的预测结果按照 5.1 节的选股策略构建思想转化为交易信号导入回测环境，并按照方案一的策略思想进行回测，生成的策略回测收益图如图 5-2 所示。

图 5-2 方案一选股策略回测收益图

由图 5-2 方案一策略回测收益图可以看出，交易策略在三年回测中最终获得了192.78%的策略收益，年化收益高达 44.47%，相比同期上证 50 指数基准获得了 130.04%的超额收益，此外，策略的胜率为 78.4%，盈亏比为 44.42，说明三年期间不但盈利次数较多，并且都是大赢小亏，策略整体收益较为理想。从策略风险指标来看，策略的夏普比率为 1.876，信息比率为 2.552，说明根据方案一构建的股票组合在承担相同风险下，可以获得更高的超额收益，策略的最大回撤为 14.14%，并且发生在 2020 年 3 月，其原因很大程度与 2020 年的新冠疫情相关，尽管 2020 年 2 月开市已经有过一轮下跌行情，但结合当时的市场背景，3 月的大幅回撤源于当时人们对于新冠疫情反弹的悲观预期。综合考虑策略的风险和收益，方案一的选股策略整体表现较为不错。

在对方案一策略分年度进行回测后，得到如图 5-3、图 5-4、图 5-5 所示的回测收益图。

图 5-3 方案一选股策略 2018 年回测收益图

图 5-4 方案一选股策略 2019 年回测收益图

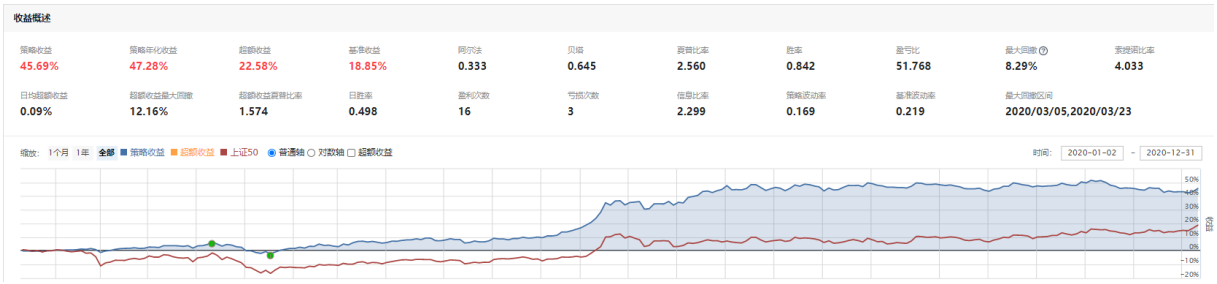


图 5-5 方案一选股策略 2020 年回测收益图

根据 5.3.1 节对回测期间的市场状态划分结果，2018 年为一个熊市周期，2019-2020 年为一个牛市周期，而图 5-3、图 5-4、图 5-5 的策略分年度回测收益图也印证了这一市场状态划分，上证 50 在 2018 年处于下跌行情中，而 2019-2020 年处于上涨行情中，赚钱效应比较明显。

为便于比较策略在不同市场状态的表现，将策略分年度回测收益风险指标整理为如表 5-2 所示的列表形式。

表 5-2 方案一的回测收益评价

回测期	市场状态	收益类指标			风险类指标		
		TR	TAR	Alpha	SR	IR	MD
2018 年	熊市	1.16%	1.19%	26.19%	-0.162	1.481	12.61%
2019 年	牛市	85.93%	88.79%	39.19%	3.352	2.410	14.73%
2020 年		45.69%	47.28%	22.58%	2.250	2.299	8.29%

通过观察表 5-2 可知，基于选股方案一构建的策略无论是在 2018 年熊市行情或是 2019-2020 年牛市行情中都跑赢了上证 50 基准，并分别获得了 26.19%、39.19%和 22.58%的超额收益，并且最大回撤也都控制在 15%以内，从交易胜率来看，策略仅在 2018 年交易中亏损次数超过了盈利次数，并且该年交易盈亏比也只有 1.782，远低于 2019 年和 2020 年的 103.415 和 51.768，出现这种情况也是受整体市场行情影响，在熊市中，交易亏损的概率大幅上升，亏钱效应明显，因此在实际投资中，应当避免在下跌行情中交易。综上所述，依据方案一构建的选股交易策略能在熊市中能够稳而不倒，做到熊市不亏损，并且在牛市中获取较大超额收益，一方面，说明方案一选股策略有效，构建的股票组合能够赚取比市场组合更多的超额收益，另一方面，从实际交易的角度证明了基于 CNN-LSTM 神经网络对股价进行涨跌的预测是可行并且可靠的。

对策略进行归因分析，可以得到如图 5-6 所示的策略在回测期间的股票持仓情况，通过观察可以发现，方案一构建的选股策略只是通过买入并持有的方式就获得了较高的收益，一方面，离不开回测期间整体市场行情的驱动，另一方面，也说明减少交易频次，选择有长期业绩增长驱动的优质股票就能获得较为不错的收益。将其整理为如表 5-3 所示的列表，可以看出回测期间的前十大持仓股主要集中在科技、消费、金融板块的白马股上，这些股票也都是有业绩增长支撑的优质股票。

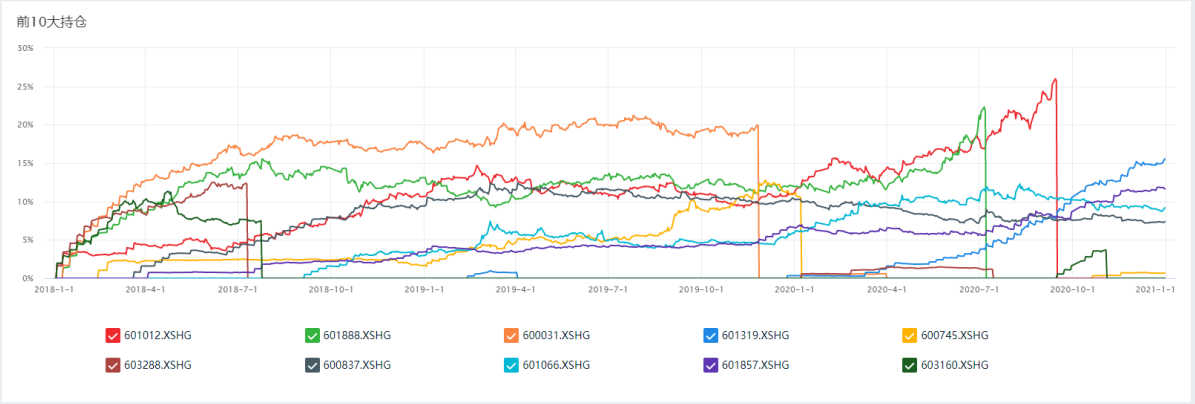


图 5-6 前十大持仓股持有天数分布图

表 5-3 方案一选股策略前十大持仓股

序号	股票代码	股票名称	最高仓位占比（%）
1	601012.XSHG	隆基股份	25.99
2	601888.XSHG	中国中免	22.31
3	600031.XSHG	三一重工	21.23
4	601319.XSHG	中国人保	15.55
5	600745.XSHG	闻泰科技	12.78
6	603288.XSHG	海天味业	12.5
7	600837.XSHG	海通证券	12.43
8	601066.XSHG	中信建投	12.29
9	601857.XSHG	中国石油	11.87
10	603160.XSHG	汇顶科技	11.34

方案二的策略思想是设定一个阈值，当模型输出的上涨下跌概率满足阈值条件时进行交易操作，通过对模型在三个训练集的涨跌输出概率进行分位点的统计，以及对模型

的预调试，最终确定了如表 5-4 所示的上涨下跌的阈值，将模型输出的上涨下跌概率 95%分位点平均数作为阈值，在每个调仓日，如果股票上涨概率大于 0.5854，则对符合条件的股票做开仓买入，如果股票下跌概率大于 0.6170，则对符合条件的股票做平仓卖出。

表 5-4 模型输出概率 95%分位点

测试集	上涨 95%分位点	下跌 95%分位点
2018 年	0.5962	0.5819
2019 年	0.5721	0.5532
2020 年	0.5878	0.7160
Avg.	0.5854	0.6170

按照方案二的策略思想进行回测，生成的策略回测收益图如图 5-7 所示。

图 5-7 方案二选股策略回测收益图

由图 5-7 方案二策略回测收益图可以看出，交易策略在三年回测中最终获得了 162.14%的策略收益，年化收益为 39.10%，相比同期上证 50 指数基准获得了 105.97%的超额收益，和方案一相比，方案二的收益出现明显下降，并且最大回撤由 14.14%上升至 19.28%，虽然策略胜率由 0.784 上升至 0.833，但盈亏比却从 44.42 下降至 26.4，从策略风险指标来看，方案二的夏普比率为 1.471，相比方案一的 1.876 也有所下降，此外，方案二的交易次数较方案一也减少了，说明阈值设置过于严苛限制了交易次数，综上所述，方案二策略的整体表现不如方案一。

在对方案二策略分年度进行回测后，得到如图 5-8、图 5-9、图 5-10 所示的回测收益图。

图 5-8 方案二选股策略 2018 年回测收益图

图 5-9 方案二选股策略 2019 年回测收益图

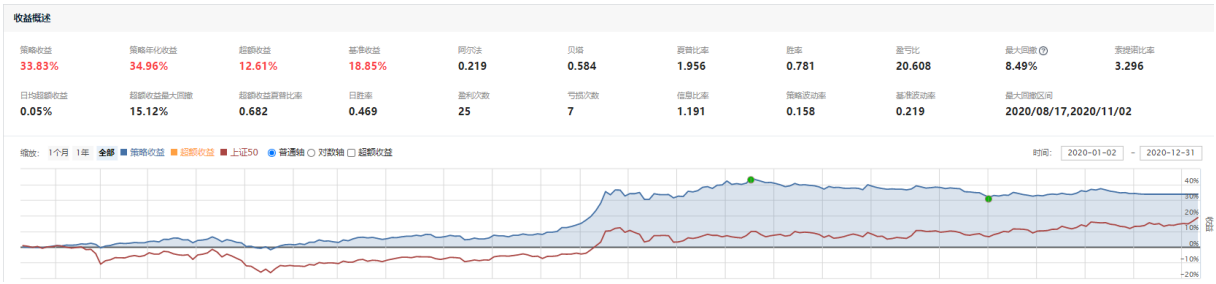


图 5-10 方案二选股策略 2020 年回测收益图

为便于比较策略在不同市场状态的表现，将策略分年度回测收益风险指标整理为如表 5-5 所示的列表形式。

表 5-5 对方案二的回测收益评价

回测期	市场状态	收益类指标			风险类指标		
		TR	TAR	Alpha	SR	IR	MD
2018 年	熊市	-1.97%	-2.03%	22.28%	-0.305	0.947	19.28%
2019 年	牛市	85.81%	88.66%	39.10%	3.307	2.197	12.82%
2020 年		33.83%	34.96%	12.61%	1.956	1.191	8.49%

通过观察表 5-5 可知，基于选股方案二构建的策略同样能够在 2018 年熊市行情和 2019-2020 年牛市行情中跑赢上证 50 基准，并分别获得了 22.28%、39.10%和 12.61%的超额收益，但相比方案一，方案二未能在 2018 年熊市中帮助投资者获取正的收益，且

熊市中的最大回撤也由 12.61%上升至 19.28%，而在 2019-2020 年牛市中，虽然方案二在 2019 年表现和方案一较为接近，但在 2020 年无论是收益能力还是风险指标都逊色于方案一构建的选股策略，尽管方案二策略表现不如方案一，但总体来看，方案二能够在牛市中获得较大收益，这说明行情和趋势对于投资组合收益是十分重要的。

对策略进行归因分析，可以得到如图 5-11 所示的策略在回测期间的股票持仓情况以及表 5-4 所示的前十大持仓股列表，通过观察可以发现，方案二构建的选股策略挑选的股票组合相比方案一集中度更高，排名最靠前的个股汇顶科技仓位最高时达到了 38.24%，高于方案一排名第一的隆基股份的 25.99%近十个百分点，并且在 2018 年熊市中，汇顶科技的仓位在前半年一直在三至四成之间，这使得方案二构建的策略遭受了巨大的亏损，导致熊市中方案二整体收益表现不如方案一，说明在熊市行情中要合理的控制仓位，此外，通过对比表 5-6 和表 5-3 可以发现，方案一和方案二在选股上有近 80%是重合的，说明基于 CNN-LSTM 的股价涨跌预测模型在上证 50 指数成分股中的选股范围比较集中。

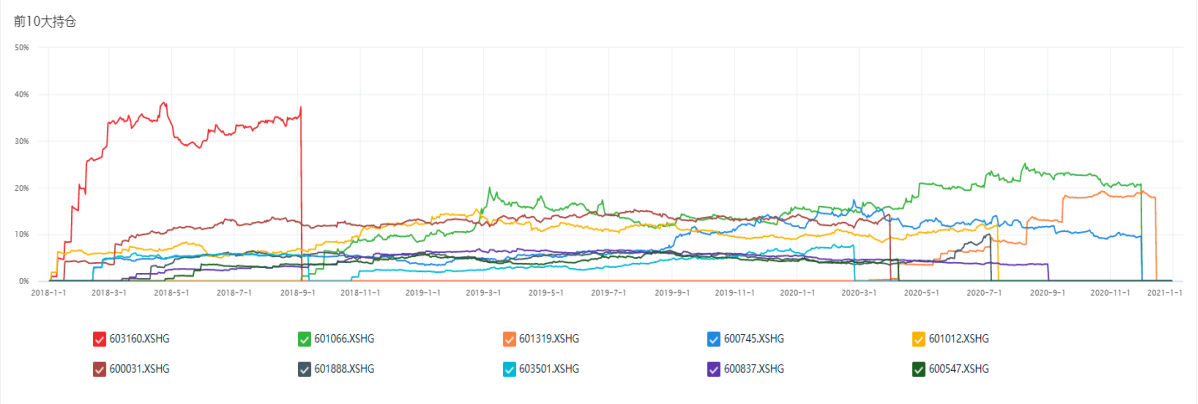


图 5-11 前十大持仓股持有天数分布图

表 5-6 方案二选股策略前十大持仓股

序号	股票代码	股票名称	最高仓位占比（%）
1	603160.XSHG	汇顶科技	38.24
2	601066.XSHG	中信建投	25.22
3	601319.XSHG	中国人保	19.31
4	600745.XSHG	闻泰科技	17.41
5	601012.XSHG	隆基股份	15.42

6	600031.XSHG	三一重工	15.30
7	601888.XSHG	中国中免	10.05
8	603501.XSHG	韦尔股份	7.82
9	600837.XSHG	海通证券	6.94
10	600547.XSHG	山东黄金	6.56

通过对构建的两种选股策略的回测对比，本节分牛熊市从策略收益和风险方面对策略进行了全面评价，并从策略的胜率、交易次数等多个角度进行了分析，方案一和方案二的回测结果表明，无论是根据上涨下跌概率排序选股或是通过设置阈值选股，两种选股思想都能够在回测中获得较为理想的收益，同时能够将策略风险和回撤控制在合理范围。分牛、熊市来看，本文构建的两个选股策略都能够在牛熊市获取显著高于基准的超额收益，并且，方案一构建的交易策略甚至可以做到在 2018 年大熊市中仍有正收益，因此，可以得出基于 CNN-LSTM 的股价预测模型在上证 50 指数成分股中构建量化选股策略是有效、可行的结论。

5.4 本章小结

本章首先介绍了基于 CNN-LSTM 选股模型的策略思想，根据 M1_10 模型输出的个股涨跌概率，提出了两个方案来构建选股策略，接着对策略的评价指标进行分类并介绍了各指标的含义，随后参考和借鉴了前人的研究经验，通过寻找上证 50 对数价格月度走势中的波峰波谷，对回测期间的牛熊市场状态进行划分，并通过回测展示了两种方案的选股策略在 2017-2020 年三年的回测收益表现。在策略回测效果分析时，从不同市场状态的视角对策略进行了多角度的分析，并且比较了两种选股方案的收益差异，以及形成差异的原因，对两种方案所选出股票所属行业等特点进行了总结。

本章通过构建选股策略并进行回测，再次印证了 CNN-LSTM 神经网络股价预测模型在股票市场的有效性，依据该模型构建的选股策略具有应用前景和实际意义，能够给投资者提供部分投资参考与建议。

研究结论与展望

本文首先介绍了神经网络模型预测股价的研究背景及意义，然后，对现有相关研究进行梳理与总结，随后，从机器学习中的深度学习出发，对神经网络的基本理论进行介绍，接着，对本文研究中使用的 CNN 神经网络、LSTM 神经网络进行算法原理上的说明，并且对股价预测中引入神经网络的可行性进行了简要分析。在股价涨跌预测模型搭建时，本文通过借鉴和参考已有模型来设计初始模型 M1，并在初始模型的基础上调参，在此过程中探寻了不同神经网络结构及网络参数对于模型预测效果的影响并在实验平台允许的性能范围内确定了最终模型 M1_10。接着本文基于 CNN-LSTM 股价预测模型设计了两个选股方案，并形成了相应的选股策略，在回测之前，先利用非参数法对回测期内的市场周期做划分，将回测期划分为 2018 年熊市和 2019-2020 年牛市并在聚宽量化交易平台进行回测，策略回测总收益最高可达 192.75%，并且取得 130.04%的超额收益，依据方案一构建的交易策略能在 2018 年大熊市中仍获取正收益。

本文的研究结论为：

（1）使用 CNN-LSTM 神经网络预测股价涨跌是可行的。本文从模型预测效果和策略回测表现两方面，证明了 CNN-LSTM 神经网络可以用于预测股价涨跌，并且基于 CNN-LSTM 神经网络的股价预测模型能够对上证 50 指数成分股做出较为准确的涨跌趋势判断，相比 CNN、LSTM 以及 BP 神经网络单模型，CNN-LSTM 双模型在预测上证 50 指数成分股涨跌的表现更好，预测准确度更高，说明将两种模型相结合能够在一定程度上提升模型的分类能力；

（2）利用 CNN-LSTM 双模型在上证 50 指数成分股中构建选股策略是有效的。将模型输出的涨跌概率视为选股条件，一定程度上丰富了传统的选股因子，构建的选股策略在牛熊市都有较为不错的收益表现，证明该模型对于投资者具有一定的投资借鉴意义和使用参考价值；

（3）我国股票市场不属于弱势有效市场。由于本文构建的 CNN-LSTM 股价涨跌预测模型的输入变量包含了股票基本交易数据及个股技术指标，通过神经网络对此类价格信息的特征提取，进而构建的选股策略可以在市场中获取超额收益，说明股票价格并未

完全反映所有历史信息，股价的未来走势和历史信息是相关的。

本文存在的不足和需要改进的地方主要集中在以下两方面：

第一，在于神经网络模型训练部分。受限于实验平台的性能，本文用手动调参替代了网格搜索法（Grid Search）的调参方式，有可能错过了最优的参数组合，找到的只是相对表现较好的参数组合。此外，在数据样本生成时，本文只是在“四价一量”的基础上加入了上证 50 指数的“四价一量”信息和 10 个股票技术分析指标，相较于样本尺寸最小的“32×32”的 LeNet-5 架构，本文构建模型的输入样本尺寸仅为“20×20”，这在一定程度上限制了神经网络层的数量以及卷积层的特征提取，使得深度神经网络不能较好地发挥作用。在后续研究中可以考虑加入情绪因素、上市公司财务数据、宏观政策因素等相关的因子继续增大样本尺寸，更好地发挥卷积层的特征提取优势，从而提升模型预测效果。

第二，在于选股策略构建部分。本文构建的神经网络涨跌预测模型所使用的训练集为前七年的股票相关数据信息，随后将接下来一整年的股票相关数据信息作为测试集并预测该年每日个股涨跌，以一年为步长更新数据集并在此基础上构建选股策略，这样做相当于每年更新一次涨跌预测模型，时间跨度稍长，由于股票的价格具有较强的动态性，每年更新模型会弱化模型的预测效果进而影响到所构建的选股策略回测收益表现。今后的研究可以考虑缩短模型更新的周期，每半年或每季度更新一次涨跌预测模型，让模型更好的适应当前市场状态，同时提升选股策略的时效性和收益性。

参考文献

[1] Fama E F . The Behavior of Stock Market Price[J]. The Journal of Business, 1965, 38(1):34-105.

[2] 周爱民.股市可预测性与技术指标协整性的模型检验[J].数理统计与管理,1999(01):6-11.

[3] 陈春晖,曾德明.我国股市可预测性的统计研究[J].统计与决策,2006(15):6-7.

[4] 苏治,方明,李志刚.STAR 与 ANN 模型:证券价格非线性动态特征及可预测性研究[J].中国管理科学,2008(05):9-16.

[5] 柴宗泽,姚长辉.股票回报的可预测性及方差分解[J].经济科学,2009(02):108-116.

[6] 杨光艺.中国股市可预测性的稳健性检验[J].金融发展研究,2018(12):3-9.

[7] 吴玉霞,温欣.基于 ARIMA 模型的短期股票价格预测[J].统计与决策,2016(23):83-86.

[8] 徐枫.股票价格预测的 GARCH 模型[J].统计与决策,2006(18):107-109.

[9] 许舒雅,梁晓莹.基于 ARIMA-GARCH 模型的股票价格预测研究[J].河南教育学院学报(自然科学版),2019,28(04):20-24.

[10] 石鸿雁,尤作军,陈忠菊.基于小波分析的 ARIMA 模型对上证指数的分析与预测[J].数学的实践与认识,2014,44(23):66-72.

[11] 徐忠兰,许永龙,赵亮.股票价格影响因素研究[J].天津师范大学学报(自然科学版),2004(02):61-63+72.

[12] 吴玉桐,梁静国.股票价格的影响因素研究[J].现代管理科学,2008(07):111-112.

[13] Shleifer A , Summers L H , Long J B D , et al. Noise Trader Risk in Financial Markets[J]. Journal of Political Economy, 1990, 98(4):703-738.

[14] Samuel A L . Some Studies in Machine Learning Using the Game of Checkers[J]. IBM Journal of Research and Development, 1959, 3(3):211-229.

[15] 杨新斌,黄晓娟.基于支持向量机的股票价格预测研究[J].计算机仿真,2010,27(09):302-305.

[16] Kim K J . Financial Time Series Forecasting Using Support Vector Machines[J]. Neurocomputing, 2003, 55(1/2):307-319.

[17] Huang C F . A Hybrid Stock Selection Model Using Genetic Algorithms and Support Vector Regression[J]. Applied Soft Computing Journal, 2012, 12(2):807-818.

[18] 张潇,韦增欣.随机森林在股票趋势预测中的应用[J].中国管理信息化,2018,21(03):120-123.

[19] 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):9-15.

[20] Peter G , Zhang. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model[J]. Neurocomputing, 2003,50(17):159-175.

[21] 刘恒,侯越.贝叶斯神经网络在股票时间序列预测中的应用[J].计算机工程与应用,2019,55(12):225-229+244.

[22] 常松,何建敏.基于小波包和神经网络的股票价格预测模型[J].中国管理科学,2001(05):9-16.

[23] 崔建福,李兴绪.股票价格预测:GARCH 模型与 BP 神经网络模型的比较[J].统计与决策,2004(06):21-22.

[24] Geoffrey E , Hinton, Simon, et al. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation,2006,18(7).

[25] 孙瑞奇.基于 LSTM 神经网络的美股股指价格趋势预测模型的研究[D]. 北京: 首都经济贸易大学,2016.

[26] 周凌寒.基于 LSTM 和投资者情绪的股票行情预测研究[D]. 武汉: 华中师范大学,2018.

[27] 彭燕,刘宇红,张荣芬.基于 LSTM 的股票价格预测建模与分析[J].计算机工程与应用,2019,55(11):209-212.

[28] 曾安,聂文俊.基于深度双向 LSTM 的股票推荐系统[J].计算机科学,2019,46(10):84-89.

[29] 贺毅岳,李萍,韩进博.基于 CEEMDAN-LSTM 的股票市场指数预测建模研究[J].统计与信息论坛,2020,35(06):34-45.

[30] Zhou F , Zhou H M , Yang Z , et al. EMD2FNN: A Strategy Combining Empirical Mode Decomposition and Factorization Machine Based Neural Network for Stock Market Trend Prediction[J]. Expert Systems with Applications, 2019, 115(JAN.):136-151.

[31] 陈祥一.基于卷积神经网络的沪深 300 指数预测[D]. 北京: 北京邮电大学,2018.

[32] 黄志辉.基于卷积神经网络的量化选股模型研究[D]. 杭州: 浙江大学, 2019.

[33] 文宇.基于 CNN-LSTM 网络分析金融二级市场数据[J].电子设计工程,2018,26(17):75-79+84.

[34] Kim T , Kim H Y , Hernandez Montoya A R . Forecasting Stock Prices with a Feature Fusion LSTM-CNN Model Using Different Representations of the Same Data[J]. PLoS ONE, 2019, 14(2).

[35] 罗文慧,董宝田,王泽胜.基于 CNN-SVR 混合深度学习模型的短时交通流预测[J].交通运输系统工

程与信息,2017,17(05):68-74.

[36] 陈亮,王震,王刚.深度学习框架下 LSTM 网络在短期电力负荷预测中的应用[J].电力信息与通信技术,2017,15(05):8-11.

[37] 李梅,李静,魏子健,王思达,陈赖谨.基于深度学习长短期记忆网络结构的地铁站短时客流量预测[J]. 城市轨道交通研究,2018,21(11):42-46+77.

[38] 罗向龙,李丹阳,杨彧,张生瑞.基于 KNN-LSTM 的短时交通流预测[J].北京工业大学学报,2018,44(12):1521-1527.

[39] 王庆荣,李彤伟,朱昌锋.基于小波去噪和 LSTM 模型的短时交通流预测(英文)[J/OL].Journal of Measurement Science and Instrumentation:1-14[2021-03-23].

[40] Mitchell T M, Carbonell J G, Michalski R S. Machine Learning[M].McGraw-Hill, 2003.

[41] Warren S. McCulloch, Walter Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity[J]. The Bulletin of Mathematical Biophysics, 1943, 5(4):115-133.

[42] Srivastava N , Hinton G , Krizhevsky A , et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[43] Hubel D H, Wiesel T N. Receptive Fields and Functional Architecture of Monkey Striate Cortex.[J]. The Journal of Physiology,1968,195(1).

[44] Fukushima Kunihiko,Miyake Sei. Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position[J]. Pergamon,1982,15(6).

[45] LeCun Y, Boser B, Denker J , et al. Handwritten Digit Recognition with a Backpropogation Network. World of Computer Science & Information Technology Journal. 1990;2:299-304.

[46] Saratha S , Wan A . Logic Learning in Hopfield Networks[J]. Modern Applied Science, 2008, 2(3):57.

[47] Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[48] Yu D , Deng L . Recurrent Neural Networks and Related Models[M]. 2015.

[49] Graves A , Jürgen Schmidhuber. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks, 2005, 18(5–6):602-610.

[50] Chen K , Zhou Y , Dai F . A LSTM-based Method for Stock Returns Prediction: A Case Study of China Stock Market[C]// IEEE International Conference on Big Data. IEEE, 2015:2823-2824.

[51] Ioffe S , Szegedy C . Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015.

[52] Pagan A R , Sossounov K A . A Simple Framework for Analysing Bull and Bear Markets[J]. Journal of Applied Econometrics, 2003, 18(1):23-46.

[53] 何兴强,周开国.牛、熊市周期和股市间的周期协同性[J].管理世界,2006(04):35-40.

致谢

夏尽秋至，伴随着金黄的落叶，我步入美丽的西北大学校园，开启了两年的研究生求学之旅，同来自五湖四海的朋友同学习、共进步。如今回首往昔，两年的光阴如梭似箭，在此期间的学习生活中，我不但收获了知识，也收获了友情，更为重要的是提升了自己的学习研究能力，让我面对不熟悉的领域仍能有效地融会贯通，在今后时刻保持学习能力，丰富自己的知识储备，这一切离不开培育我的母校和各位老师。

首先，感谢我的研究生导师徐璋勇教授。在入学之前，徐老师就向我们推荐了数本金融学著作，这些书从逻辑和历史角度帮助我建立起了对金融学的认识和理解，入学后，每周例行的组会更是让我接触到了大量的优秀学术期刊。在硕士毕业论文的选题、开题及撰写定稿期间，老师也对我的论文提供了大量宝贵意见，这为我论文的顺利完成提供了重要保障，在此过程中，徐老师专业的学术素养和严谨的工作态度为我树立了优秀的榜样。

其次，感谢经济管理学院各位老师的辛勤付出。在研究生一年级的课程学习中，各位老师用负责的教学态度和生动的教学方式为我们授课，同时积极为我们联系校外企业交流和学习的机会，让我们得以在两年间接触到期货公司、券商和银行的学习培训，丰富和开拓了我的视野。

再次，感谢我的同学也是好友乔泽东、李京泽、焦禹铭、艾泽鑫和尹朝鹏。他们的出现让我的研究生生活更加丰富多彩，而且每个人身上都有值得我学习的地方，在此也衷心祝福他们在日后的生活中一帆风顺，事业进步。还要特别感谢我的父母及家人，他们在我本科毕业工作两年后仍支持我重返校园学习，并不断鼓励关怀着我，家永远是爱的港湾。

最后，向参加本次论文评审的各位老师献上我诚挚的谢意！