2022 届硕士专业学位研究生学位论文

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：51195100039

華東師範大學

East China Normal University

硕士专业学位论文

Master's Degree Thesis (Professional)

论文题目：基于技术面数据的自监督学习股市涨跌预测方法

院 系 名 称:	**数据科学与工程学院**
专业学位类别:	**工程硕士**
专业学位领域:	**电子信息（计算机技术）**
研 究 方 向:	**AI 量化投资**
指 导 教 师:	**钱卫宁 教授**
学 位 申 请 人:	**应泽林**

2021 年 11 月

Thesis (Professional) for Master's Degree in 2022

University Code: 10269

Student ID: 51195100039

East China Normal University

Title: **Predict Stock Trends with Self-supervised Learning Based on Technical Data**

Department:	School of Data Science and Engineering
Category:	Master of Engineering
Domain:	Electronic Information (Computer Technology)
Research Direction:	AI Quantitative Investment
Supervisor:	Prof. QIAN Weining
Candidate:	Zelin Ying

November, 2021

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于技术面数据的自监督学习股市涨跌预测方法》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名:_____ 日期:2021 年 11 月 18 日

华东师范大学学位论文著作权使用声明

《基于技术面数据的自监督学习股市涨跌预测方法》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的研究成果归华东师范大学所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和相关机构如国家图书馆、中信所和“知网”送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

（ ）1. 经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文 *，于 年月日解密，解密后适用上述授权。

（ ）2. 不保密，

导师签名:_____ 本人签名:_____

2021 年 11 月 18 日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

应泽林硕士学位论文导师指导小组声明

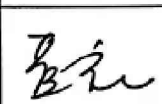
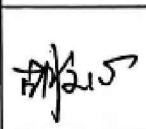
本人已认真阅读此篇学位论文。我认为，该论文已达到申请硕士学位的要求，可启动学位申请程序。

签名： 

钱卫宁，导师，电子信息（计算机技术）

2021 年 11 月 15 日

应泽林 硕士学位论文开题小组成员名单

姓名	职称	单位	签字
董启文*	研究员	华东师范大学 数据科学与工程学院	
黄定江	教授	华东师范大学 数据科学与工程学院	
胡文心	高级工程师	华东师范大学 数据科学与工程学院	

*硕士学位论文开题小组组长

应泽林 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
田秀霞	教授	上海电力大学	**主席**
黄定江	教授	华东师范大学	
董启文	研究员	华东师范大学	
李翔	研究员	华东师范大学	
王延昊	副教授	华东师范大学	

摘　要

金融股票市场是各类投资者热衷参与的一项重要经济活动，对于经济的发展有着重要的影响。投资者期望在金融股票市场上进行投资从而获取盈利，因此如何更有效地进行股票未来走势的预测便成了投资者所关心的一个任务。传统的市场趋势预测模型通常基于手工因子或特征，严重依赖于昂贵的专业知识，此外，很难发现股票时间序列数据中包含的隐藏特征，而这些特征将有助于预测股市趋势。本文提出了一个以自监督学习序列编码模型 S3E 为预训练模型的股票市场趋势预测框架 SMART。具体地来说，该模型将股票技术数据序列进行编码表征，表征通过多个自监督学习辅助任务进行进一步联合训练。通过多任务联合学习来训练 S3E 模型中的编码器，对于股票序列数据进行编码表征，随后基于 LSTM 和前馈神经网络进行股市趋势预测。本文在中国 A 股市场和美国纳斯达克市场上进行了大量的实验，实验结果表明本文所设计的模型对股票市场趋势预测是非常有效的。根据 SMART 框架的预测结果，本文基于模型输出结果所实现的投资收益明显优于其他方法。

基于自监督学习的股市预测算法模型研究完成后，如何去高效地管理与维护算法模型，使其实际地落地到工业场景上，也就是算法的落地过程，也是本文所考虑的一个问题。针对于算法落地的场景，本文设计了一个量化模型管理平台，来自动化地进行模型的滚动训练和预测任务等，同时对于每个模型各方面的评估和预测等都做了可视化展示，方便用户对于模型的业务评测效果有一个整体的把控。

总而言之，本文的核心贡献如下所示：

- 本文将自监督学习技术引入到金融量化投资领域，提出了一套基于技术面数据的自监督学习框架 SMART，设计了三种业务相关的自监督学习辅助任务，通过多任务联合训练的方式来建模金融市场股市涨跌的预测任务。
- 本文在多个数据集上进行了充分的实验，同时对比了现有的各种主流模型，评估了 SMART 框架及其变体和其他各种主流模型在股票涨跌分类任务的 accuracy 和 F1-score 效果，同时在业务指标上评估了 SMART 框架所能达到的累计收益率、信息系数和夏普率，充足有效的实验证明了本文所提出的方法框架的有效性。
- 为了有效地管理各个量化模型，本文设计了一套量化模型管理平台，给投资者一个直观立体的模型效果展示，包括模型每天的预测情况、量化策略的累计收益情况等，方便投资者比较各个模型之间的相关差别。

本文所提出的 SMART 金融股市预测框架，在多个数据集上取得了领先的效果，并且能够在金融市场上实现较高的年化收益率，领先于其他方法，说明自监督学习技术在股票金融序列数据编码中有着不错的效果，为自监督学习技术在金融量化投资领域上的应用奠定了一定基础。

关键词：股市趋势预测；自监督学习；量化交易；Transformer；多任务学习

ABSTRACT

Financial stock market is an important economic activity that all kinds of investors are keen to participate in, which has an important impact on economic development. Investors expect to invest in the financial stock market to obtain profits. Therefore, how to predict the future trend of stocks more effectively has become a task concerned by investors. The traditional market trend prediction model is usually based on manual factors or features, which seriously depends on expensive professional knowledge. In addition, it is difficult to find the hidden features contained in the stock time series data, which will help to predict the stock market trend. This paper proposes a stock market trend prediction framework SMART, which takes the self supervised learning sequence coding model S3E as the pre training model. Specifically, the model encodes and represents the stock technology data sequence, and the representation is further jointly trained through multiple self supervised learning auxiliary tasks. The encoder in S3E model is trained through multi task joint learning to encode and characterize the stock sequence data, and then the stock market trend is predicted based on LSTM and feed forward neural networks. Extensive experiments on both China A-Shares and NASDAQ stock datasets demonstrate that the features learned from stocks' daily data sequences are effective for stock market trend prediction. According to the prediction results of SMART framework, the investment return based on the model output is obviously better than other methods.

After the research on the stock market prediction algorithm model based on self supervised learning is completed, how to efficiently manage and maintain the algorithm model and make it actually land on the industrial scene, that is, the landing process of the algorithm, is also a problem considered in this paper. For the scenario of algorithm landing, this paper designs a quantitative model management platform to automatically carry out the rolling training and prediction tasks of the model. At the same time, it makes

a visual display of the evaluation and prediction of each model, so as to facilitate users to have an overall control over the business evaluation effect of the model.

In conclusion, the core contributions of this paper are as follows:

- This paper introduces self supervised learning technology into the field of financial quantitative investment, proposes a set of self supervised learning framework smart based on technical data, designs three business-related self supervised learning auxiliary tasks, and models the prediction task of stock market rise and fall in the financial market through multi task joint training.

- This paper has conducted sufficient experiments on multiple data sets, compared various existing mainstream models, evaluated the accuracy and F1-score effects of SMART framework and its variants and other mainstream models in the task of stock rise and fall classification, and evaluated the cumulative rate of return, information coefficient and sharp rate that SMART framework can achieve in terms of business indicators, Sufficient and effective experiments show the effectiveness of the proposed method framework.

- In order to effectively manage each quantitative model, this paper designs a set of quantitative model management platform to give investors an intuitive and three-dimensional model effect display, including the daily prediction of the model and the cumulative income of the quantitative strategy, so as to facilitate investors to compare the relevant differences between various models.

The SMART financial stock market prediction framework proposed in this paper has achieved leading results in multiple data sets, and can achieve high annualized rate of return in the financial market, which is ahead of other methods, indicating that the self supervised learning technology has a good effect in the data encoding of stock financial series, It lays a foundation for the application of self supervised learning technology in the field of financial quantitative investment.

Keywords: *Stock Trends Prediction; Self-supervised learning; Quantitative Investment; Transformer; Multi-task Learning*

目录

插图

表格

第一章　绪论

1.1　研究背景与意义

股票交易是一种重要的投融资金融活动 [1, 2]。基本上，所有的投资者对股价都很敏感 [3]，期望以较低的价格买入股票，然后以较高的价格卖出。因此，预测股市走势，判断未来几个交易日股价是涨是跌，是投资者及时调整投资策略所关心的一项有意义的任务，对个人投资者和国民经济都有着巨大的影响。在股票金融市场上，每个人都有着自己偏好的投资策略并践行之 [4, 5]，从大方向上来说，可以分为主动投资 [6] 和被动投资两大投资策略 [7]，如图1.1所示。主动投资指的是投资者会更加主动地在金融市场上进行主动选股和主动择时的操作，主动选股指的是在所有金融市场的股票中选取投资者认为在将来有一定上涨空间的股票形成投资组合进行买卖操作，主动择时指的是对于同一个股票，投资者在期望的价格低点买入，待股票价格上涨达到高点后进行卖出，从而赚取价差从而盈利。在主动投资策略中，又可分为传统投资 [8] 和量化投资 [9] 两方面，传统投资往往是投资者根据自己积累的相关投资经验，对上市公司的盈利情况、发展情况等基本面情况进行一个基本面的分析，同时也会结合上市公司的股票价格走势的相关情况进行一个技术面的分析 [10, 11]，最后得出对应的买卖决策从而进行交易；量化投资往往是借助计算机程序化的手段在金融市场上进行盈利，其中包括量化选股、量化择时、统计套利、算法交易等多方面，都可以借助计算机的手段来达到一个盈利的目的。在被动投资策略中，又可分为指数基金和指数复制，指数基金指的是以特定指数中的成分股作为投资的对象来构建投资组合，而指数复制则完全按照指数中成分股的权重比例进行配置来跟踪指数的收益情况。

在金融股票投资领域，投资策略所实现的收益率主要来源为两部分：alpha 收

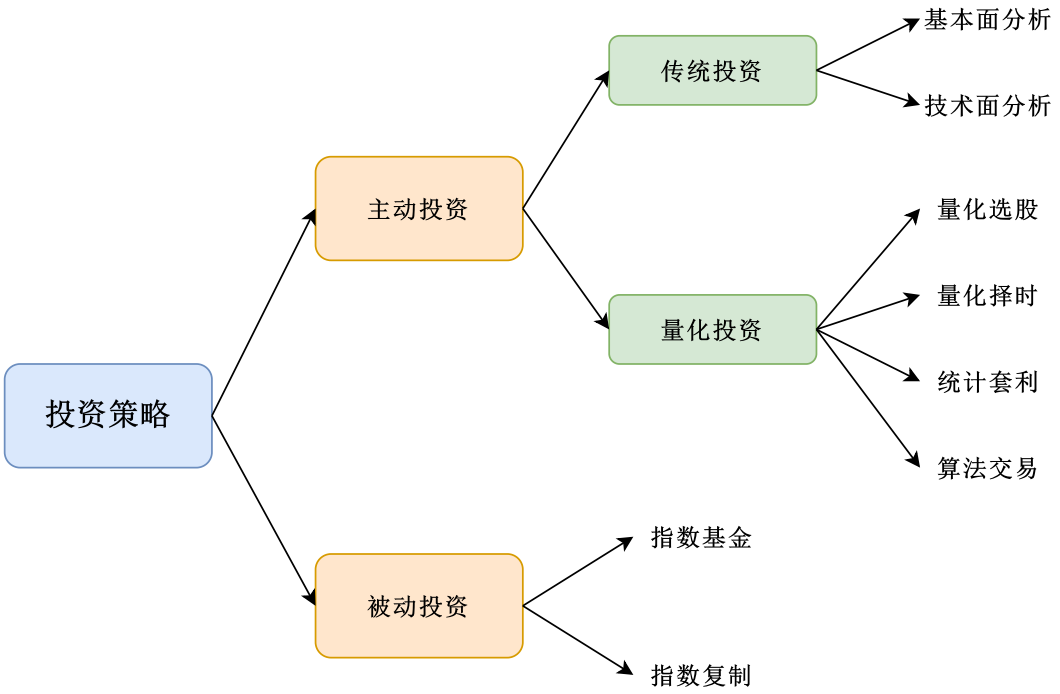


图 1.1: 投资策略分类

益 [12] 和 beta 收益 [13]，即：

$$Return\ Rate = alpha + beta \times Market\ Return \tag{1.1}$$

其中 beta 收益表示市场波动情况所带来的收益率，也被称之为系统性风险。当市场行情好的时候，大部分股票都处于上涨的情况，市场波动会给投资者带来一定程度的市场正收益率；而当市场行情差的时候，大部分股票都处于下跌的状态，市场波动会给投资者带来一定程度的市场负收益率。除此之外，还有一部分 alpha 收益率，指的是投资策略超过市场收益率的那部分收益情况，即投资者根据自己的投资经验进行买卖交易从而获得超过市场收益的部分，量化交易的核心目标也就是最大化 alpha 收益，捕获超过市场收益率的那部分超额收益。

量化投资, 就是一种采用数量化技术与计算机自动程序化执行的方法, 来完成买卖交易指令的投资方式, 并由此方法来获得稳健利润。量化投资在国外已有三十余年的发展历程, 由于长期投资业绩稳健, 使得国际量化交易的市场规模和份额也

在不断扩大, 并日益受到投资人的青睐。与此同时, 飞速发展的信息时代让许多新兴投资概念在世界范围内流传迅速, 使得越来越多的交易与投资人接触并认识到了量化投资的相关概念, 但量化投资作为一种概念, 也不是很新鲜, 一些国内的相关投资人也有所耳闻。但是，真正的量化交易基金在国内还是比较罕见的 [14]。同时，机器学习 [15, 16] 和深度学习 [17-19] 等人工智能技术 [20] 的不断发展，也对量化投资起到了一定的促进作用。

量化交易从历史的海量数据入手，有着严格的交易逻辑规则，当收益达到一定程度后会进行严格的止盈操作，同时当亏损达到一定程度后也会进行严格的止损操作，不会像传统投资一样受到人为主观以及情绪的影响。和传统投资相比，量化投资的优势主要如下：

- 系统性：多角度多层次地观察金融市场的海量数据，从中捕捉潜在的盈利机会。
- 纪律性：严格按照程序指令执行，克服人为主观交易中缺乏纪律性的特点。
- 准确性：能够在海量的数据中准确地捕捉到错误定价的投资标的，进行交易从而从中获利。
- 分散化：从历史数据的特征中筛选出有较大概率获利的投资组合从中盈利。
- 及时性：能够迅速及时地跟随市场变化，对市场的变化作出及时反应。

因此，相较于传统的人为投资，构建一个有效的量化投资模型，依据模型预测结果让程序自动化地去交易，克服在投资行为中人性的弱点，是一个巨大的挑战 [21, 22]。

1.2 本文工作与贡献

本文工作总体上可分为两块内容，一部分是自监督 [23, 24] 多任务联合学习 [25, 26] 的算法研究设计，本文提出了一套基于自监督学习技术的股市预测框架

SMART（**S**tock **Mar**ket **T**rend），基于历史的技术面交易数据构建深度学习模型，来对股票未来的趋势做出一个涨跌的得分预测；另一部分是算法研究的工程落地，对于研究完成的模型，本文研发了一个量化模型管理平台，来管理不同人员研发的各种各样的量化模型。

对于算法研究设计的相关内容，核心目标是基于自监督学习技术，本文设计了三种业务相关的自监督辅助任务，采用多任务联合的方式进行训练，得到一个预训练的模型，随后抽取出预训练模型的 Encoder 部分，来编码技术面数据得到技术面数据表征，将技术面表征作为特征输入到时序模型中，最后输出一个 0 到 1 的概率值，也称之为股票的得分值，分值越高的股票，说明模型预测该股票未来有着更大的概率上涨，分值越低的股票，说明模型预测该股票未来有着更大的概率下跌，对于大概率上涨的股票，期望进行买入的操作，而对于大概率下跌的股票，期望进行卖出的操作。

对于算法研究工程落地相关内容，在量化模型管理平台上我们可以看到每个量化模型的预测结果，可视化地展示持仓和交易情况，以及在每个时间点所能达到的收益率情况，能够简单便捷地比较模型效果之间的差异，给投资者一个直观立体的投资绩效展示。

针对金融市场的股市预测问题，本文的核心贡献主要如下：

- **本文将自监督技术引入到金融量化投资领域，针对股票涨跌预测的场景，提出了一套基于自监督技术的框架 SMART**。该框架由预训练的 S3E 序列编码模型和长短期记忆神经网络（LSTM）模型组合而成：本文针对 S3E 序列编码器设计了三种业务相关的自监督辅助任务，通过多任务联合训练的方式来预训练 S3E 序列编码器，训练完成后抽取出其中的 Encoder 部分来对技术面数据进行编码得到每日技术面表征；对于得到的每日技术面表征向量，将其输入到长短期记忆神经网络中进行股市涨跌预测的分类任务，以此建模金融市场股市的预测任务。

- **本文在多个数据集上进行了实验，充分有效的实验证明了本文所提出框架的**

有效性。本文评估了所提出的 SMART 框架在各个指标上的效果，同时对比了现有的各种主流模型，比较了 SMART 框架及其变体和其他各种主流模型在股票涨跌分类任务的 accuracy 和 F1-score 效果，同时在业务指标上评估了 SMART 方法所能达到的累计收益率、信息系数和夏普率，本文所提出的 SMART 框架在各个指标上都处于领先地位。

- **本文设计了一套量化模型管理平台，方便用户有效地管理各个量化模型，给投资者一个直观立体的模型效果展示**。用户可以在量化模型管理平台上查看、维护以及管理各个量化模型，同时量化模型管理平台提供模型的可视化展示功能，包括模型每天的预测情况、量化模型的累计收益情况等，方便投资者比较各个模型之间的相关差别。

1.3 本文架构

本文一共分为六个章节。

第一章是绪论章，主要介绍了量化投资的相关背景和意义，以及本文的研究内容与论文组织架构。

第二章是相关技术章，在第一章介绍了量化投资的背景和意义上，主要介绍了相关的量化投资模型，主要包括传统的多因子模型、机器学习方法和深度学习方法。

第三章是基于自监督多任务联合学习的 SMART 框架的研究，主要介绍了 SMART 框架的各个流程及其中的 S3E 序列编码预训练模型的结构，包括三种自监督辅助任务的设计及模型的相关损失函数。

第四章是实验对比和结果分析，在第三章 SMART 框架设计的基础上，设计多种实验去评估本文所提出的 SMART 框架的有效性。本章介绍了实验的设置及相关的评价指标，对比了一些主流的图表征模型和序列编码模型，从传统分类指标和业务指标多角度地评测了本文所设计的 SMART 框架，同时也进行了消融实验对比 SMART 框架及其对应的变体模型。

第五章是量化模型管理平台的设计与实现，在第三章和第四章设计并且验证了所提出的量化模型的有效性之后，设计实现量化模型管理平台对模型进行管理维护和可视化操作。本章主要介绍了量化模型管理平台如何进行技术选型及设计，同时基于一些主流的开发框架进行量化模型管理平台的实现。

第六章是总结章，主要总结了本文的所有主要内容、难点以及解决方式，并对未来的发展前景进行一定的展望。

第二章　量化投资模型概述

量化投资没有准确的定义，从广义上讲，可以认为所有借助数学模型 [27] 和计算机 [28] 实现的投资方法都可以称为量化投资。量化投资模型也有着比较悠远的发展历史，随着时间的推演大致可以分为传统的多因子模型 [29]，机器学习模型 [30] 和深度学习模型 [31] 三大类量化投资模型。目前国内常见的量化投资模型还是以传统的股票多因子模型为主，很多量化研究的团队致力于挖掘有效的因子，从而依据因子进行择优选股，但近些年来随着机器学习、深度学习等 AI 技术的兴起，人工智能开始慢慢渗透到每个行业，也包括金融量化领域，于是一些头部量化私募公司开始尝试使用机器学习、深度学习等技术在股票金融市场上进行建模，预测股票未来的涨跌走势，从相关私募公司的 AI 量化模型盈利情况来看，优秀的 AI 量化模型能够实现年化 30% 以上的 alpha 超额收益，由此可见 AI 技术在金融股票上建模取得了比较优越的收益效果。

2.1　传统多因子模型

传统的量化投资模型往往都是基于多因子进行构建的 [32]。因子 [33] 也被称为特征或者信号，指的是利用统计的方式或者历史经验积累所总结的规律，广义区分为基本面因子和技术面因子，比如净利润、资产负债率等就属于基本面因子，衡量的是一个上市公司的基本发展情况，而大单净流入量、大单比例等就属于技术面因子，衡量的是上市公司的股票在交易日进行交易时的一些投资者交易行为特征。不同因子在不同方面反映了一家上市公司的情况，有效的因子能够比较好地反映下一期股票的收益率，从而通过买卖股票进行盈利。如果拥有更多更有效的因子，基于这些因子研发模型就能取得比较好的效果。

多因子选股模型，以一系列因子为筛选股票的准则，对达到一定因子值的股票进行购买，反之则进行出售。不同的因子从不同的角度对股票的各个状态进行一定

的描述，模型的优势是可以综合多方面的信息，最后给出一个选股结果。选取的因子不同以及综合各个因子方式的差异也会形成不同的模型, 因此一般会选择打分法或者回归法作为综合因子的方式, 且打分法比较普遍。

打分法，顾名思义就是用每一个因子对股票进行打分，对于所得到的各个因子的分数，以一定的权重进行加权计算，最后得到一个总体分数，依据总体分数对股票进行一定选择，构建成一个投资组合，在股票市场上进行模拟交易的回测，最后根据模拟交易回测的结果得到模型最后所能达到的投资组合收益率，以此来评判模型优劣。打分法的一个优点是相对稳定，受极值影响较小，但缺点是打分法中的各个因子的权重分配需要主观定义，这也是打分法实际使用中的难点。同时，这些因子的有效性也是一个重要问题，随着市场格局的变化, 一些原有的影响因子可能会逐渐失效, 需要探索新的因子。在这种条件下，需要对选用的因子及模型本身做持续性的再评价和不断的改进策略来适应多变的市场环境，同时还需要考虑到交易时发生的交易成本等相关因素。

回归法是一个用于检测投资因子有效性的方法, 把 T 期的因子值和 T+1 期的股票收益率进行回归, 得到的回归系数即为 T 期的投资因子收益率。在回归法的分析中, 必须对数据进行相应的预处理, 并进行标准化、去极值和缺失值填充的处理, 对经过预处理后的数值进行最小二乘求得标准化回归系数, 以回归系数来衡量因子的有效性。

Fama-French 三因子模型 [34] 是一个非常著名的多因子模型，包括三个重要的因子, 分别是账面市值比因子（HML）, 市场投资组合（Rm - Rf）, 市值因子（SMB）。基于三个因子，通过回归模型的方式确定相关系数，之后便可基于当期的三个因子来预测下一期股票的涨跌情况。

2.2 机器学习方法

机器学习指的是利用一些算法从已有的数据中挖掘一定的特征来构建模型，以此来对新数据做出一定的预测判断，一般可以分为有监督学习 [35]、半监督学习

[36]、无监督学习 [37]、强化学习 [38] 等，而金融股市预测的场景往往采用有监督学习的机器学习方法为主。

传统的多因子模型主要基于构建的有效因子进行预测，这些因子往往都是基于专家经验进行构建，但是由于传统多因子模型方法比较简单，所以准确率不是很高。随着机器学习方法的兴起，越来越多的领域开始应用机器学习方法来解决一些问题，金融研究者也开始尝试采取机器学习的方法对股市趋势进行预测 [39, 40]，比如基于构建的有效因子，采用逻辑回归 [41]、支持向量机 [42]、XGBoost[43] 等方法，将因子作为特征输入到模型中，将对应股票下一天的涨跌作为标签，进行二分类任务的训练学习，当模型训练完成后，对下一个交易日的股票涨跌走势趋势进行预测，继而根据模型的预测结果进行对应的买卖操作。

机器学习可以使用的地方，一是构造因子，还有一个是给因子配权重。这里的构造因子，是在已有的数据库里，用机器学习的方法尽可能的挖掘信息。比如在传统的 EP（市盈率）和 DP（市净率的倒数）的基础上，探索 EP*DP，或者 EP 开 DP 次方等是否存在有效的信息。换句话说，就是利用机器的快速的大批量的处理能力，来尽可能的挖掘有用信息。但这类方法看似在理，却和传统的投资逻辑不符。传统的逻辑是，知道要从什么角度去投资，而计算机只是实现的工具。换句话说，像诸如 EP 或者 DP 这样的因子，构造的背后有着很强的理论依据。而机器学习寻找因子，却是在汪洋里面找寻未知的岛屿。即使找到了，也不知道这是数据拟合的结果，还是真的有效。再者，即使在组外测试效果好，在真实的交易中，真的要使用这个因子，也需要很大的信心。

2.2.1 逻辑回归

逻辑回归模型通常被用来解决分类任务而不是回归任务，且一般以二分类任务为主，它在线性回归的基础上套用了逻辑函数从而实现二分类任务的目标，比如一些垃圾邮件的判别、是否患病的检测等应用场景都可以采用逻辑回归的模型进

行建模处理，逻辑回归模型较为简单，有着比较好的可解释性，其计算方式如下：

$$P(y=1|x)=\frac{1}{1+e^{-w^Tx+b}} \tag{2.1}$$

其中 $P(y=1|x)$ 表示根据给定样本的特征，计算得出为正样本的概率，w 和 b 表示逻辑回归模型的参数，分别表示模型的矩阵变换的参数和偏置项，e 表示自然对数，其中所用到的激活函数也就是 sigmoid 函数：$\sigma(x)=\frac{1}{1+e^{-x}}$，通过激活函数对线性变换后的结果进行激活，增强了模型的非线性表达能力。

股票涨跌的二分类预测问题，属于一个天然的二分类任务，比较适合逻辑回归模型的应用场景。在当前时间 t 时，往前看前 k 天的时间窗口，计算每天每种因子的值，对于缺失的因子值进行填充处理，然后进行去极值和标准化等数据预处理的操作，汇总后作为逻辑回归模型的输入，而任务的标签则是下一个交易日对应股票的涨跌类别，如果股票上涨则为正样本，反之为负样本。模型参数求解过程主要是基于极大似然估计的思想，其对数似然函数为 $F(\boldsymbol{w})=\sum_{n=1}^{N}(y_n ln(p)+(1-y_n)ln(1-p))$，模型以负对数似然函数作为损失函数，优化过程则是通过梯度下降法让模型的参数以一定的步伐不断地向负梯度方向进行更新，来求得损失函数最小的的模型参数。

2.2.2 支持向量机

机器学习中的支持向量机是一种数据科学算法，属于监督学习的范畴，它分析数据集的趋势和特征，解决与分类和回归相关的问题。支持向量机是基于 VC 理论的学习框架（Vapnik-Chervonenkis 理论），每个训练数据点被标记为两个类别中的一个，然后迭代地构建一个区域，该区域将空间中的数据点分为两组，使得该区域中的数据点在边界上以最大宽度或间隙很好地分开。

支持向量机想要的就是找到各类样本点到超平面的距离最远，也就是找到最大间隔超平面。对于任意一个超平面，都可以用下面这个线性方程来描述：

$$w^Tx+b=0 \tag{2.2}$$

对于 n 维空间的一个点 $x=(x_1,x_2...x_n)$，其到超平面 $w^Tx+b=0$ 的间隔为：

$$d=\frac{|w^Tx+b|}{||w||} \tag{2.3}$$

其中 $||w||=\sqrt{w_1^2+w_2^2+...+w_n^2}$ 表示参数 w 的二范数，对于处于支持向量上的点 $x_{sv}=(x_1,x_2...x_n)$，其满足 $|w^Tx_{sv}+b|=1$，因此对于属于支持向量的正负样本，其到分离超平面的距离都为 $\frac{1}{||w||}$，所以最大化间隔超平面的距离又可表示为：

$$max\ \frac{2}{||w||} \quad s.t.\ y_i(w^Tx_i+b)\geq 1 \quad for\ i=1,2,...,n \tag{2.4}$$

由于 $\frac{2}{||w||}$ 不属于凸函数，在使用梯度下降法优化参数时容易收敛到局部最优点，因此我们将其转化为等价求解问题：

$$min\ \frac{1}{2}||w||^2 \quad s.t.\ y_i(w^Tx_i+b)\geq 1 \quad for\ i=1,2,...,n \tag{2.5}$$

对于存在的不等式约束条件，通常采用拉格朗日乘子法将其纳入到优化函数目标中，然后通过一系列数学手段求出支持向量机模型的参数。

对于金融股票市场的涨跌预测任务，支持向量机也同样适用，在当前时间 t 时，往前看前 k 天的时间窗口，计算每天每种因子的值，对于缺失的因子值进行填充处理，然后进行去极值和标准化等数据预处理的操作，最终作为支持向量机模型的输入，而任务的标签则是下一个交易日对应股票的涨跌类别，如果股票上涨则为正样本，反之为负样本。模型参数求解过程主要是最大化支持向量到超平面的距离，通过数学的方法优化目标函数从而求得最优的模型参数值。

2.2.3 XGBoost

XGBoost 是一种数学味深厚的经典计算模型，也因为其模型具有相当好的有效性，被广泛地运用于机器学习的各个实际使用情景中。XGBoost 从梯度提升决策树 (GBDT) 中发展而来，为一个以 CART 回归树为基学习器的提升树模式，在 GBDT 中以一阶负梯度替代残余误差的基础之上，进一步增加了二阶梯度来提升精度，同

时也增加了正则化项来减少过拟合情况的出现。

更具体地来说，XGBoost 模型具有加法性质，由 k 个基模型形成，假设第 t 次需要迭代训练的树模型是 $f_t(x)$ ，则有：

$$\hat{y}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2.6}$$

其中 $\hat{y}^{(t)}$ 表示第 t 次迭代后样本 i 的预测结果，$\hat{y}_i^{(t-1)}$ 表示前 $t-1$ 颗树的预测结果，$f_t(x_i)$ 表示第 t 颗树的模型。通过这种迭代的方式，每次迭代训练一个树模型，拟合的是之前所有的树模型预测结果的累计和与目标标签值的残差。同时在训练过程中，为了缓解过拟合情况的发生，通常采取收缩系数（Shrinkage）和子采样（Sub-Sampling）的方法，收缩系数主要的目的是避免一次迭代的时候让模型把残差直接学习到位，而是每次只学习残差的一部分，从而多衍生出几个树模型，更多的树模型更能够缓解过拟合情况的发生；子采样的方法主要是指在训练过程中不直接使用全部样本，而是采样一部分样本用来做训练，能更好地防止过拟合。

从目标函数的角度上来说，模型的偏差和方差共同决定模型的预测精度，损失函数表示经验化风险，也就是模型的偏差，优化模型除了优化模型的损失函数外，往往也会在目标函数中加入一定的正则化项，来控制模型的结构化风险，也就是方差，来缓解过拟合的情况，所以目标函数由两部分组成，一部分是模型的损失函数 L ，另一部分是控制模型复杂度的正则化项 Ω ，其定义如下：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \tag{2.7}$$

其中，$\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ 表示所有样本的损失，$\sum_{i=1}^{t} \Omega(f_i)$ 表示将全部 t 颗树的复杂度进行求和，以此为整体的目标函数进行优化，保证模型偏差小的同时，也缓解模型过于复杂带来的方差问题。

对于金融股票市场的涨跌预测任务，XGBoost 可以尝试应用在这个场景中，在当前时间 t 时，往前看前 k 天的时间窗口，计算每天每种因子的值，对于缺失的因子值进行填充处理，然后进行去极值和标准化等数据预处理的操作，最终作为

XGBoost 模型的输入，而任务的标签则是下一个交易日对应股票的涨跌类别，如果股票上涨则为正样本，反之为负样本。模型参数求解过程主要是通过优化模型的目标函数，使其最小化来求得最优的模型参数值。

2.3 深度学习方法

深度学习方法一般以神经网络为主，来学习数据内部潜在的数据特征，最终目标是让计算机学习人类处理信息和分析问题的能力，在语音 [44] 和图像识别 [45] 方面取得了非常显著的效果，远超了传统的方法，因此不断地被各个领域的研究者所用来解决遇到的实际问题。

在人工智能量化投资领域，一些前沿机构开始采取深度学习的方法进行建模，尝试去挖掘相关的历史数据与未来的股票涨跌间潜在的关联关系。总的来说可以分为三大方面：一是基于股票序列数据的编码模型，基于编码后的隐状态进行股票未来涨跌的预测；二是基于图数据结构的编码模型，将公司与公司之间的关联关系纳入到图数据结构中来，再基于图模型算法进行编码形成图表征，最后基于图表征进行股票未来涨跌的预测；三是事件驱动的股票预测模型，主要思想是基于发生的金融新闻事件，对事件进行表征，基于表征的事件进行股票未来涨跌的预测。

2.3.1 循环神经网络

循环神经网络 [46]（Recurrent Neural Network, RNN）往往用来建模序列数据，其相关模型结构如图2.1所示，对于输入的序列数据 $X = \{X_1, X_2, ..., X_T\}$ ，在每一个时刻 t，RNN 的循环单元有如下表示，具体的计算公式为；

$$\begin{aligned} s_t &= tanh(Ws_{t-1}, Ux_t) \\ \hat{y}_t &= softmax(Vs_t) \end{aligned} \tag{2.8}$$

其中 s_t 表示 RNN 模型在 t 时刻的隐状态，x_t 表示 RNN 模型在 t 时刻的序列的输入，W 、U 和 V 表示参数矩阵，$tanh$ 和 $softmax$ 表示激活函数，$\hat{y}_t$ 表示 RNN 模型在 t 时刻的输出。

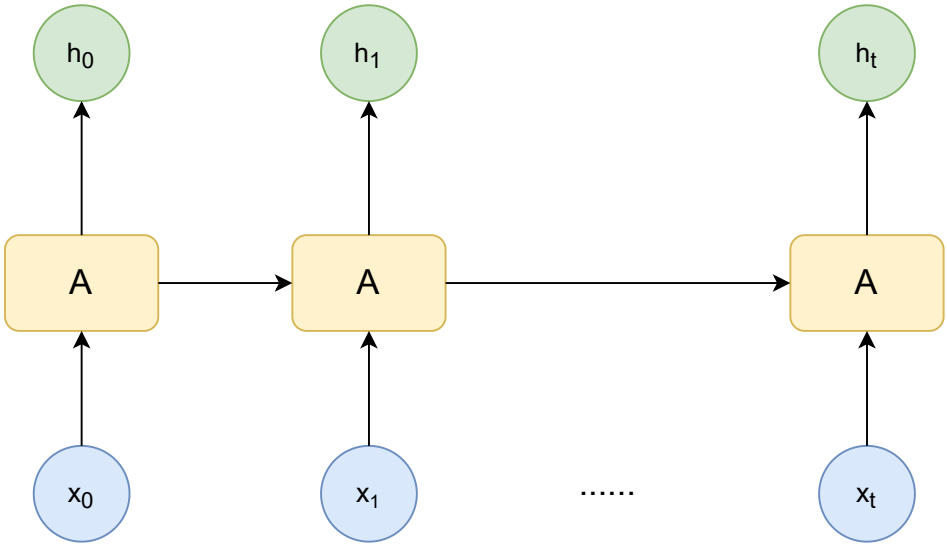


图 2.1: RNN 模型结构

在金融股票的涨跌预测应用场景中，由于股票数据的天然序列特性，就比较适合使用 RNN 进行序列建模。通过对专家经验构建的因子，或者是股票市场每日的技术面交易数据（开盘价、收盘价、最高价、最低价、成交量和成交金额），以此作为序列数据输入到 RNN 模型中，在最后一个时间步的隐状态接上一个分类器，最后模型输出涨跌的概率值。模型的损失函数采用交叉熵损失，即：

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)] \tag{2.9}$$

其中 y_i 为样本 i 的涨跌标签，若为涨则标签为 1，若为跌则标签为 0，p_i 为样本 i 预测为涨的概率值。模型的优化过程则是通过反向传播算法让模型的参数以一定的步伐不断地向负梯度方向进行更新，从而求得使得损失函数最小的的模型参数。

2.3.2 长短期记忆网络

长短期记忆网络（Long Short-Term Memory，LSTM）也通常是用来建模序列数据，其诞生主要是为了解决传统 RNN 存在的梯度消失和梯度爆炸的问题，特别是当序列长度比较长时，长距离序列编码的效果较差。因此为了解决这个存在的

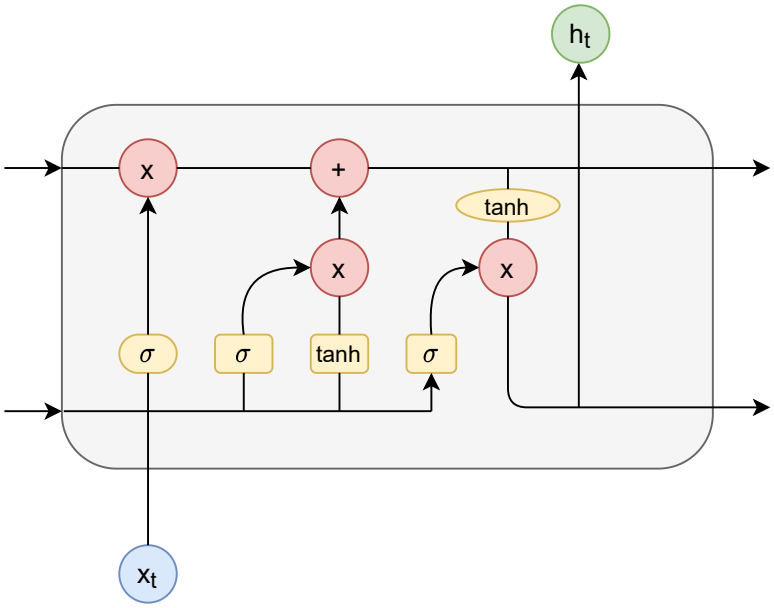


图 2.2: LSTM 模型结构

问题，LSTM 采用了三种门控机制，即遗忘门、输入门和输出门来缓解长距离序列的梯度爆炸和梯度消失现象，相关模型结构如图2.2所示，计算方式为：

$$
\begin{aligned}
&\text{遗忘门：} f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \\
&\text{输入门：} i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \\
&\tilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_C) \\
&C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \\
&\text{输出门：} o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \\
&h_t = o_t * tanh(C_t)
\end{aligned}
\tag{2.10}
$$

其中，W_f, W_i, W_c, W_o 和 b_f, b_i, b_C, b_o 都为模型需要学习的参数，h_t 表示 LSTM 模型在 t 时刻的隐状态，x_t 表示 LSTM 模型在 t 时刻的输入，f_t, i_t, o_t 分别表示遗忘比例、输入比例和输出比例，在当前时间步 t 时，计算出需要更新的细胞状态 $\tilde{C}_t$，然后将前一个时间步的细胞状态乘以遗忘比例，再加上当前时间步的更新状态乘以输入比例，最后得到当前时间步的细胞状态。对于当前时间步 t 的输出，则是将当前细胞状态乘以一个输出比例，来得到输出结果。在最后一个时间步的隐状

态接上一个分类器，最后模型输出涨跌概率。

对于金融股票涨跌预测的场景，长短期记忆神经网络应用非常广泛，主要可以分为两种应用场景：第一种场景是基于专家经验构建的因子，输入到长短期记忆神经网络中去，让模型去学习到当期的因子值和下一期股票涨跌的潜在关系，从而进行买卖操作；第二种场景是直接基于股票市场每日的技术面交易数据，也就是开盘价、收盘价、最高价、最低价、成交量和成交金额，将这些数据直接作为特征输入到长短期记忆神经网络中进行训练然后预测，最后再依据预测结果进行实际交易。由于模型的复杂程度较高，基于长短期记忆神经网络模型的预测效果往往比传统的多因子模型和机器学习模型效果更好，而且由于门控机制，不容易产生梯度爆炸和梯度消失的情况，使得 LSTM 模型对于序列数据的编码效果较好。模型的优化过程则是通过反向传播算法让模型的参数以一定的步伐不断地向负梯度方向进行更新，从而求得使得损失函数最小的的模型参数。

2.3.3 图表征股票预测模型

股票的涨跌趋势不仅仅与它自身的历史数据相关，还往往离不开相关股票间的关联关系，因为一个公司对应股票的价格涨跌往往会一定程度地受到其关联公司股票价格涨跌的影响，如何去把股票间的关联关系考虑到模型建模过程中来，是图表征股票预测模型所要研究的内容。图表征的股票预测模型往往是基于图数据结构，将公司及其它们之间的关联关系表示在一个图数据结构中，一家企业在图中的以一个节点来表示，两家企业间的持股关系以节点和节点之间的边来表示，边的权重的含义为持股比例，以此来构建一个图数据结构。对于构建完成的图数据结构，采用随机游走的算法在图中得到若干个节点序列，然后将节点序列当做一个句子，将节点当做一个单词，使用词向量的训练方式（比如 skip-gram）进行训练来得到节点的表征向量。在得到节点的表征向量之后，便可以利用 cos 相似度来计算节点与节点之间的相似性，在要预测目标企业的涨跌时，利用节点间的 cos 相似度选择出与目标企业最为接近的 K 家公司，拼接后并以此作为特征向量输入至 LSTM

中, 进行对股价涨跌走势的预测。

2.3.4 事件驱动的股票预测模型

股票的价格波动走势不仅仅与金融市场的股票交易序列数据相关，还与其他各种外源数据紧密关联，比如金融新闻的事件往往也会显著地影响股票价格的走势：对于一家上市公司的利好新闻发布，则会引入越来越多的投资者进行买入，从而推动股票价格上涨；对于一家上市公司的利空新闻发布，则会引入越来越多的投资者进行卖出，从而推动股票价格下跌。因此如何去处理好金融新闻事件对于股票价格走势的影响，便是一些事件驱动的股票预测模型所研究的内容。

神经张量网络是一种事件驱动的深度学习模型。它首先提取金融新闻事件，然后将其表征为低维稠密向量，然后使用深度卷积神经网络对事件的短期和长期影响进行建模。对于一个事件，将其表示为 E = (O1, P, O2)，其中 O1 表示主语，P 表示谓语，O2 表示宾语。神经张量网络的输入为向量，所以首先将 O1, P, O2 当成单独的词，然后进行 word embedding（事先用 skip-gram 算法针对金融语料库进行词嵌入训练得到模型）。如果 O1,P,O2 中含有多个词，那么可以采取将多个词向量取平均的方式来得到最终的词向量表示。获得事件表示后，采用长、中、短期事件向量相结合的方法：使用过去 30 天内的所有相关事件。如果某一天发生多个事件，则对该天的事件进行平均，以获得该天的事件表示。然后根据时间序列排列好 30 天的事件，设置滑动窗口，并在同一窗口中卷积事件。在卷积运算之后，所有卷积层的输出向量被合并以获得最终的长期事件向量。通过使用类似的操作获得中期事件向量。中期时间定义的时间是过去一周。短期事件不需要卷积池，昨天的事件直接聚合并平均。最后，可以得到长期、中期和短期的三个事件向量。将这三个向量拼接成一个向量，发送给后续的单隐层前馈神经网络，最后输出分类结果。

第三章　自监督学习股市涨跌预测方法

针对金融股市预测问题，本文提出了一种自监督多任务模型的框架，称之为SMART（**S**tock **Mar**ket **T**rend），工作流程图如下图3.1所示，首先构建股票序列对作为输入（步骤①），通过三个自监督辅助任务来预训练一个股票序列的表征模型S3E（**S**elf-supervised **S**tock **S**equence **E**ncoding model），用来对股票序列数据编码得到编码后的表征（步骤②）。在那之后，为了去预测股票在对应的 t+1 天的走势，将前 k 天的股票序列数据喂入到序列编码器中得到股票每天的技术面表征向量（步骤③和步骤④）。最后将股票每天技术面表征数据作为 LSTM 的序列输入（步骤⑤和步骤⑥），再将 LSTM 得到的最后一层隐向量作为前馈神经网络的输入（步骤⑦），最终得到的输出为下一个交易日股票上涨的概率得分（步骤⑧）。接下来本文首先介绍相关的符号定义，然后阐述 SMART 框架的具体实现过程。

3.1　符号定义

本文所提出的 SMART 框架是基于日频的技术面交易数据实现的，每一段序列由 k 天连续的日频技术面数据构成，相关符号定义如表3.1所示。定义一个股票 s 第 t 天的技术面交易数据为: $x_t^s = \{r_{t-k+1}^s, ..., r_{t-i}^s, ..., r_t^s\}$, 其中 r_{t-i}^s 表示第前 i 天的技术面交易数据记录。进一步定义 $r_{t-i}^s = (p^o, p^c, p^h, p^l, t^v, t^o)$ 为第 $t-i$ 天的日频技术面数据，分别代表股票的开盘价格、股票的收盘价格、股票的最高价格、股票的

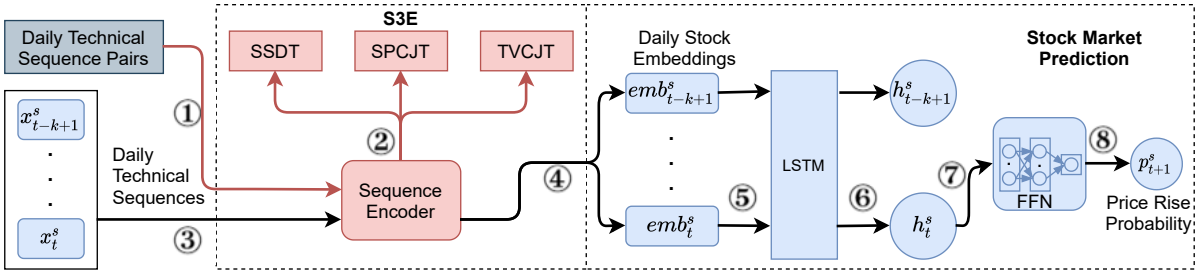


图 3.1: SMART 框架运行流程图

符号	对应含义
p^o	股票的开盘价格
p^c	股票的收盘价格
p^h	股票的最高价
p^l	股票的最低价
t^v	股票的成交价
t^o	股票的成交金额
r_t^s	股票 s 在第 t 天的日频技术面数据，$r_t^s = (p^o, p^c, p^h, p^l, t^v, t^o)$
x_t^s	股票 s 在第 t-k+1 天至第 t 天的长度为 k 的序列数据，即 $x_t^s = (r_{t-k+1}^s, ..., r_t^s)$

表 3.1: 相关符号定义

最低价格、股票的成交量、股票的成交金额。对于每一个 x_t^s，将其作为一个锚序列，并将其关联正样本序列和负样本序列，形成序列对的形式作为模型的输入。其中正样本指的是紧接着锚序列之后的那一段序列，因为它们之间有着更多相似的特征，所以被称为正样本；负样本包括时序负样本和对比负样本，时序负样本指的是和锚序列属于同一个股票但是在时间上差距比较远的样本，对比负样本指的是和锚序列属于不同股票但是在时间上一致的样本，锚序列与正样本之间数据特征会比锚序列和负样本之间的数据特征更为相似。

3.2 数据预处理

本文所采用的数据为日频的技术面交易数据，也就是每天每个股票都有其对应的技术面交易数据（开盘价、收盘价、最高价、最低价、成交金额、成交量），是通过 Wind 金融服务商处购买获得。由于股票有着除息除权的特性，除息是指因公司股东派发股利导致每股所代表的企业实际价值（每股净资产）下降而形成的消除行为，这一事实发生后，需要从股票市场价格中消除；除权是指由于公司股本增

加，每股所代表的企业实际价值（每股净资产）减少。对于该事实，需要将这些因素从股市价格中消除。除息除权一般分为两种，一种是通过股票红利进行派发股票，一种是通过现金红利派发现金，主要有四个重要日期：

- 股利宣布日：公司董事会宣布分红的日子。
- 股权登记日：参与当期股利的股东的计算确认日。只有在该日持有股份的股东才能参与分红。
- 除权除息日：股权登记的下一工作日，本日及本日之后买入的股东不享有本期股利。
- 派发日；实际派发股利股息的日期。

除息除权的存在导致了除息除权前后的价格不存在可比性，比如一个股票进行每持有一股派送一股的除息除权，除息除权前是 10 块钱，除息除权后股价变为 5 块钱，但本质上资产总量并没有发生变化，股票价格却缩水了 50%，因此为了避免这个问题，每个股票在每天都会有一个复权因子，进行复权操作，将成交量和价格调整到可比较的水平，来达到股价走势的连续一致性。一般可以通过前复权或者后复权的方式将相关数据进行一定的处理，来保证价格等数据之间可以互相比较：前复权是以当前价格作为基准进行复权操作，可以比较清楚的观察成本分布情况，比如相对的最高价、最低价、成本密集相关区域，以及目前股价所处的位置是在相对高位还是相对低位；后复权指的是维持上市首日的价格不变，根据处理股利分配数据后的价格，最后一天的价格将不会是实际交易价格，但可以看到股票的实际价值和股东的实际收益率的增加，如果进行价值投资，建议使用后复利权，这样计算的收益率相对正确，查询更直观。对于本文所用到的相关股票数据，均统一采取了后复权的方法对股票的价格和成交量数据进行了一定的处理，来保证了价格和成交量数据间的可比性。

将股票数据进行复权后，还需要进行一系列的数据预处理操作：对于某些天停牌或者没有交易的股票，采取上一个有效交易日的数据进行填充，填充完成后对

相关特征数据进行 z-score 标准化处理，z-score 也被称为 standard score，用来评估样本点到总体均值的距离，z-score 标准化处理是基于原始数据的均值和方差，通过将原始数据减去均值然后再除以方差来得到变换后的值，对于数据整体处理后可让数据聚集在 0 附近，标准差为 1，在这种数据分布下模型进行训练时收敛的效果能够更快更好，其具体的计算公式为：

$$x^* = \frac{x - \mu}{\sigma} \tag{3.1}$$

其中 x 为数据的原始值，μ 为原始数据的均值，σ 为原始数据的标准差，x^* 为变换后的数据值。

当完成股票数据的预处理操作后，便可以开始构建数据集。按照时间的先后关系，依次划分为训练集、验证集和测试集。采用长度为 k（比如 k = 30）的滑动窗口按照时间轴进行滑动，形成若干段股票序列数据，从而构建完成数据集。

3.3 序列数据编码

3.3.1 Attention 机制

Attention 机制又被称为注意力机制，主要从人类认知世界的角度演变而来。注意力是一种复杂的认知功能，感知的一个重要特性是人类不倾向于同时处理整个信息，取而代之的是，人类的注意力机制往往会在需要的时间和地点有选择地关注信息的一部分，而同时忽略其他可感知的信息。例如，人类利用视觉感知事物时，通常不会从头到尾看所有场景，而是根据需要去观察特定部分。当人类发现一个场景的某个部分经常有他们想要观察的东西时，他们将学会在类似场景再次出现时关注该部分，并将更多注意力集中在有用的部分。

人类的注意机制根据其产生方式可分为两类。第一类是自下而上的无意识注意，称为基于显著性的注意，由外部刺激驱动。例如，人们在谈话中更容易听到响亮的声音。它类似于深度学习中的最大池和选通机制，它将更合适的值（即更大的值）传递到下一步。第二类是自上而下的有意识注意，称为集中注意。集中注意力

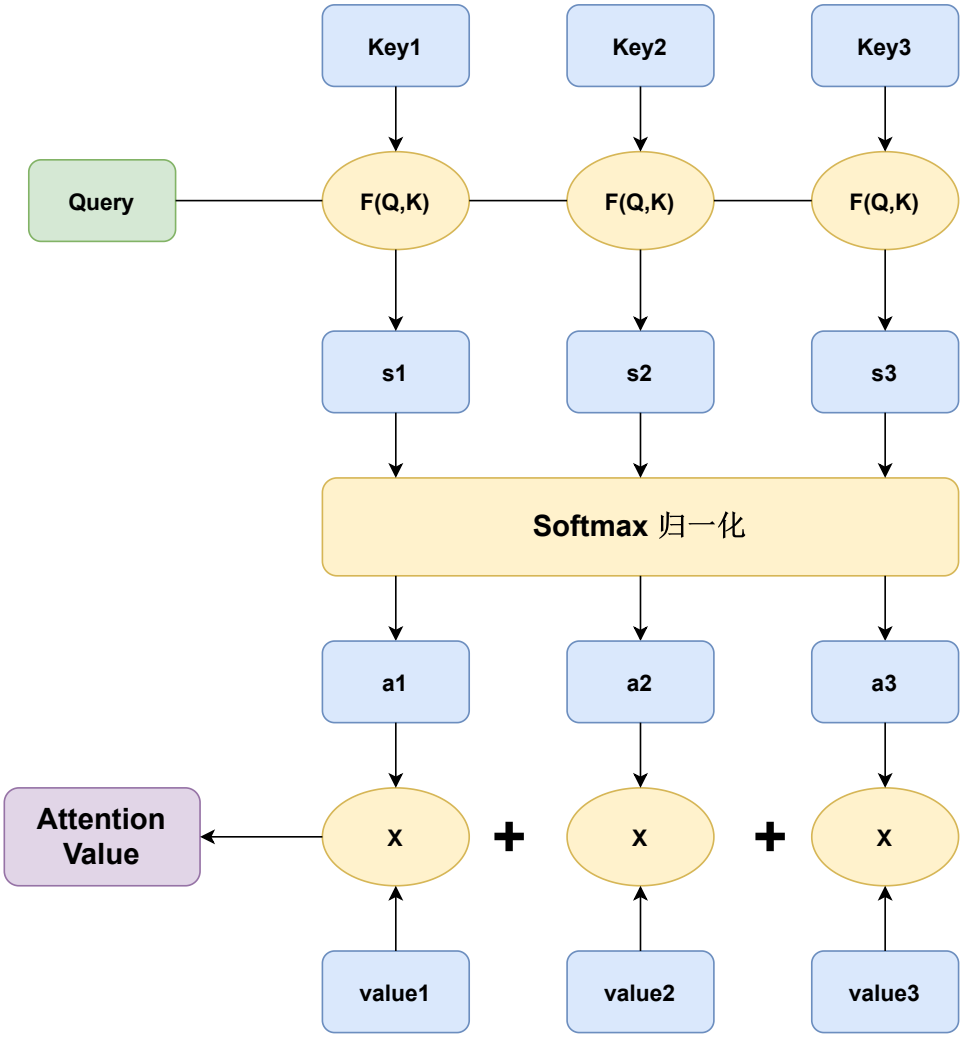


图 3.2: Attention 机制

是指具有预定目的并依赖于特定任务的注意力。它使人类能够有意识地、积极地将注意力集中在某个物体上。深度学习中的大多数注意机制都是根据特定的任务设计的，因此大多数都是集中注意力。

总而言之，注意力机制的核心是一种权重的分配方案，专注于解决信息量过大的问题，可以用有限的计算资源来处理更加重要的信息，其通用的结构如图3.2所示。最通用的 Attention 公式可以表示为；

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.2}$$

其中 Q,K 表示的是计算 Attention 权重的特征向量，V 表示的是输入特征向量，

它们都是由输入特征得到的，$Attention(Q,K,V)$ 的核心就是指根据 Q 和 K 求出关注程度然后再以对应权重乘以 V。其具体的流程可以表示为；

1. 对于两个向量 q 和 k，通过 Attention 打分函数 $a(q,k)$ 来计算两个向量间的相似度得分，即：$e_i = a(q,k)$ 。
2. 对于得到的相似度得分值，通过 softmax 函数将分值进行归一化，得到归一化后的概率值 α_i，即：$\alpha_i = \frac{e_i}{\sum_i e_i}$ 。
3. 对于归一化后的概率值 α_i，将其与 v 向量加权求和得到编码后的向量 c，即：$c = \sum_i \alpha_i v_i$ 。

对于流程 1 中 Attention 打分函数，一般可以分为以下三种：

$$score(q,k) = \begin{cases} q^T k & dot \\ q^T W k & general \\ v^T tanh(W[q;k]) & concat \end{cases}$$

其中 dot 指的是两个向量 q,k 间求内积来计算 Attention 的得分值，general 指的是通过两个向量 q,k 和一个参数矩阵 W 来计算 Attention 的得分值，concat 指的是先将两个向量 q,k 拼接后，在乘以参数矩阵 W，通过 tanh 激活函数激活后，再乘以向量 v^T 来计算 Attention 的得分值。

随着研究的发展，Attention 机制也衍生出了很多变体，比如 hard Attention 和 soft Attention，常用的 Attention 一般以 soft Attention 为主，保证各个时间步上 Attention 的权重都有一定的概率值，而 hard Attention 将概率最大的那个时间步上的权重置为 1，其他时间步的权重置为 0，强行让模型只关注到影响最大的那个时间步；以及 global Attention 和 local Attention，普通的 Attention 一般以 global Attention 为主，让模型关注到全局每一个时间步的权重影响，而 local Attention 让模型只关

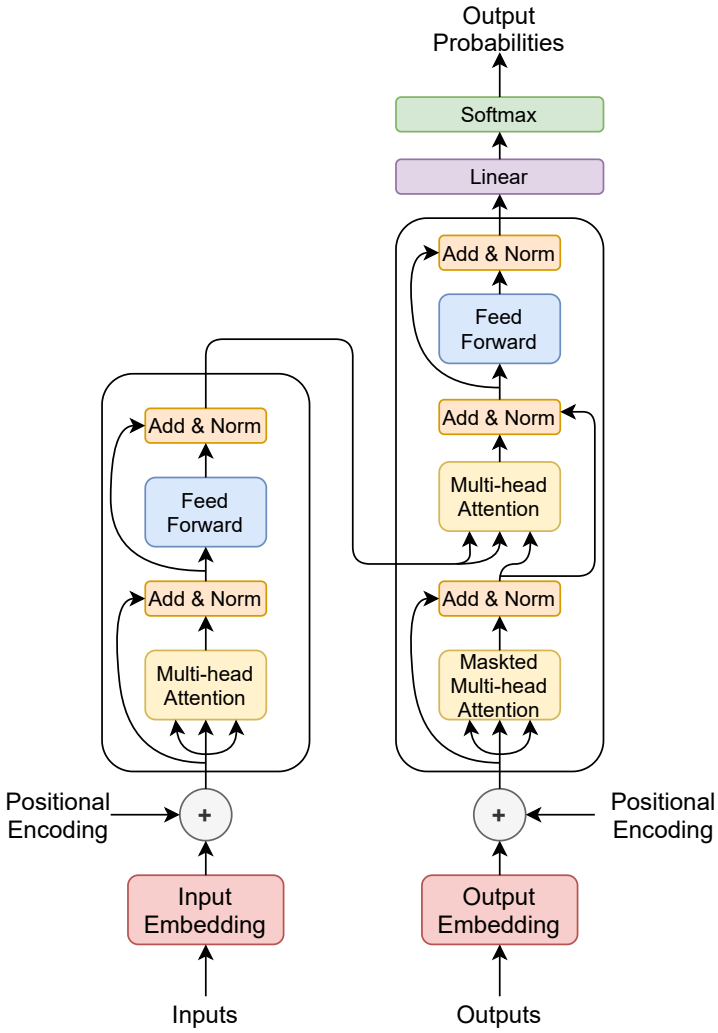


图 3.3: Transformer 模型结构

注到局部一定范围时间步的权重影响，能够降低计算的复杂度。除此之外，还有很多一些 Attention 机制的变体，比如 self-Attention 以自身的 Q, K, V 保持一致来加强本身的表征，multi-head Attention 采用多头注意力机制加快并行速度，同时能够捕获不同子空间内的数据特征等。

3.3.2 Transformer 结构

Transformer 模型结构主要是基于 Attention 机制设计的，如图3.3所示，由于其在自然语言、计算机视觉、强化学习等一系列领域都有着不错的效果，因此受到了人们广泛的关注。从总体上来说，Transformer 架构可以分为四个部分，分别是输

入部分、输出部分、编码器部分和解码器部分。

更具体地来说，对于输入部分，一般包含原文本嵌入层和对应的位置编码器，以及目标文本的嵌入层和对应的位置编码器；对于输出部分，一般包含线性层和softmax 层，用以表示输出的概率；对于编码器部分，一般由 N 个编码器堆叠而成，每个编码器通过两个子层结构连接组成，第一个子层结构为多头自注意力机制、归一化层和残差连接层，而第二个子层结构为前馈神经网络层、归一化层和残差连接层；对于解码器部分，一般也由 N 个解码器堆叠而成，每个解码器通过三个子层结构连接组成，第一个子层结构为多头自注意力机制、归一化层和残差连接层，第二个子层结构为多头注意力机制、归一化层和残差连接层，第三个子层结构为前馈全连接层、归一化层和残差连接层。从整体上来说，Transformer 是一个完全依赖于 Attention 机制，摒弃了传统的 RNN、CNN 模型结构，而且在多个领域的多项任务上都取得了非常不错的效果，尤其是 Transformer 的 encoder 结构，在多项任务中大放异彩，被应用在很多场景上。

3.3.3 S3E 序列编码器

S3E 模型包含了一个序列编码器部分和三种自监督辅助任务。序列编码器主要是用来编码输入的“锚——样本”序列对，形成低维稠密的每日股票技术面表征，模型结构如图3.4所示。更具体的来说，序列编码器由 Transformer 层和 Attention 层组成，对于一个输入的“锚——样本”序列对，锚序列和样本序列被输入到 Transformer 组件中，分别得到序列的隐状态，然后我们再基于 Attention 机制去学习到之前不同天对于当前时间的不同权重程度的影响，得到注意力机制加权后的锚序列和样本序列的表征，整个过程描述如下所示：

$$
\begin{aligned}
y_{t_1}^{s_1} &= \{z_{t_1-k+1}^{s_1}, \cdots, z_{t_1}^{s_1}\} = \text{Transformers}(x_{t_1}^{s_1}) \\
y_{t_2}^{s_2} &= \{z_{t_2-k+1}^{s_2}, \cdots, z_{t_2}^{s_2}\} = \text{Transformers}(x_{t_2}^{s_2}) \\
c_{t_1}^{s_1} &= \text{Attention}(y_{t_1}^{s_1}) \qquad c_{t_2}^{s_2} = \text{Attention}(y_{t_2}^{s_2})
\end{aligned}
\tag{3.3}
$$

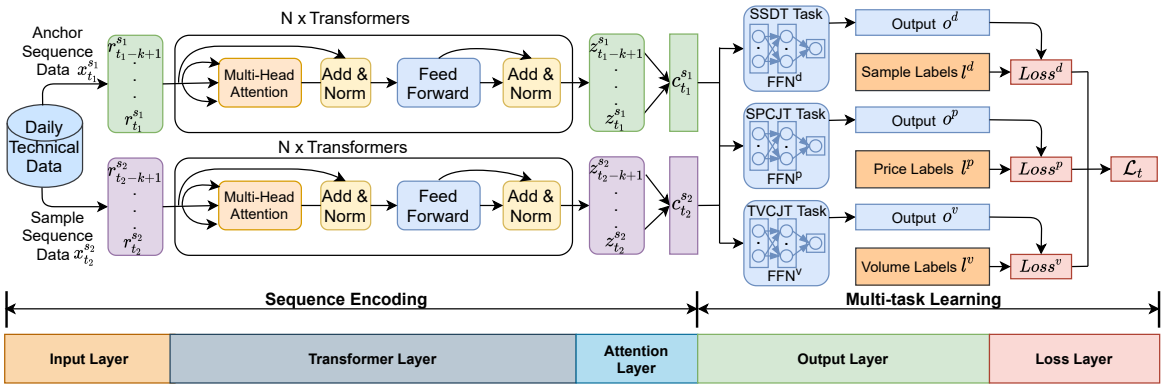


图 3.4: S3E 序列编码预训练模型

其中 $y_{t_1}^{s_1}, y_{t_2}^{s_2} \in \mathbb{R}^{k \times d}$ 是“锚——样本”序列对经过 Transformer 编码后的隐状态，$c_{t_1}^{s_1}, c_{t_2}^{s_2} \in \mathbb{R}^{d}$ 是锚序列 $x_{t_1}^{s_1}$ 和样本序列 $x_{t_2}^{s_2}$ 经过 Attention 机制加权求和后的技术面表征。

3.4 多任务联合学习

我们设计了三种自监督辅助任务来帮助训练 S3E 股票序列编码模型，这些自监督辅助任务分别是正负样本判别辅助任务（SSDT Task），价格变化同向性辅助任务（SPCJT Task），成交量变化同向性辅助任务（TVCJT Task）。更具体地来说，正负样本判别的辅助任务是帮助模型去学习到不同股票之间内在的独特特征；价格变化同向性辅助任务是帮助模型去学习到不同股票间的价格变化的关联关系；成交量变化同向性辅助任务是帮助模型去学习到不同股票间的成交量变化的关联关系。为了让模型综合地学习到多个任务中潜在的数据模式，我们在框架中采取多任务联合训练的方式进行训练，接下来我们具体的介绍每个自监督辅助任务以及最后的学习目标函数。

3.4.1 正负样本判别任务

在正负样本判别任务中，我们期望模型能够去判别“锚——样本”序列对中的样本是关于锚序列的正样本还是负样本，其中正样本是锚序列后面紧接着的一段序列，负样本又可分为对比负样本和时序负样本，对比负样本指的是和锚序列的

序列时间一样但采样自不同的股票的序列，时序负样本指的是和锚序列来自于同一个股票但时间差距较长的序列。正样本能够帮助模型识别与锚序列相邻的序列数据特征，对比负样本帮助模型学习到不同股票之间内在特征的差异性，时序负样本帮助模型学习到同一个股票内在特征随时间的变化性。

对于一对“锚——样本”序列对 $sp=(x_{t_1}^{s_1},x_{t_2}^{s_2})$，它的正负样本判别任务标签 l_{sp}^{d}，定义如下：

$$l_{sp}^{d}=\begin{cases}1 & x_{t_2}^{s_2}\in \text{正样本}\\ 0 & x_{t_2}^{s_2}\in \text{负样本}\end{cases} \tag{3.4}$$

如果 $x_{t_2}^{s_2}$ 是锚序列 $x_{t_1}^{s_1}$ 的正样本，则将标签设置为 1，否则将标签设置为 0。对于给定的输入序列对 $sp=(x_{t_1}^{s_1},x_{t_2}^{s_2})$, 编码后得到表征序列对 $pc=(c_{t_1}^{s_1},c_{t_2}^{s_2})$，喂入到前馈神经网络中去计算 $x_{t_2}^{s_2}$ 是 $x_{t_1}^{s_1}$ 正样本的概率，即 $o_{sp}^{d}=\mathrm{FFN}^{d}(pc)$。基于计算得到的概率值和真实的标签值，采取交叉熵损失作为正负样本判别自监督辅助任务的损失函数：

$$\begin{aligned} loss_{sp}^{SSDT}&=-[l_{sp}^{d}*\log(o_{sp}^{d})+(1-l_{sp}^{d})*\log(1-o_{sp}^{d})]\\ loss_{(s_1,t_1)}^{SSDT}&=\frac{1}{N_1}\sum_{sp\in AP_{t_1}^{s_1}} loss_{sp}^{SSDT}\\ Loss^{d}&=\frac{1}{D\times L}\sum_{s_1}\sum_{t_1} loss_{(s_1,t_1)}^{SSDT}\end{aligned} \tag{3.5}$$

其中 N_1 是“锚——样本”序列对的个数，D 是股票 s_1 对应数据集的总天数, L 是股票市场上所有股票的数目, $loss_{sp}^{SSDT}$ 是“锚——样本”序列对 sp 对应的正负样本判别任务的损失, $loss_{(s_1,t_1)}^{SSDT}$ 是股票 s_1 在第 t_1 所对应的损失，$Loss^{d}$ 是总体的正负样本判别任务的损失。

3.4.2 价格变化同向性任务

在价格变化同向性任务中，期望模型能够去学习到股票序列之间的价格变化同向性特征，对于一对“锚——样本”序列对 $sp=(x_{t_1}^{s_1},x_{t_2}^{s_2})$，对应的价格变化同向

性标签定义如下：

$$l_{sp}^{p}=\begin{cases}1 & \Delta p_{t_1}^{s_1}\times\Delta p_{t_2}^{s_2}\geq 0\\ 0 & \Delta p_{t_1}^{s_1}\times\Delta p_{t_2}^{s_2}<0\end{cases} \tag{3.6}$$

其中 $\Delta p_{t_1}^{s_1}=p_{t_1+1}^{s_1}-p_{t_1}^{s_1}$ 表示股票 s_1 在 (t_1+1) 时刻和 t_1 时刻的收盘价价差，$\Delta p_{t_2}^{s_2}=p_{t_2+1}^{s_2}-p_{t_2}^{s_2}$ 表示股票 s_2 在 (t_2+1) 时刻和 t_2 时刻的价差。如果 l_{sp}^{p} 等于 1，表明序列 $x_{t_1}^{s_1}$ 和序列 $x_{t_2}^{s_2}$ 有着价格变化同向性的数据特征，在接下来的交易日中有着更大的概率出现价格同涨同跌的走势。

对于给定的“锚——样本”序列对 sp，通过编码后得到两段序列的表征向量 $pc=(c_{t_1}^{s_1},c_{t_2}^{s_2})$，然后将表征向量喂入到前馈神经网络中去计算序列 $x_{t_2}^{s_2}$ 和 $x_{t_1}^{s_1}$ 在下一个交易日价格同涨同跌的概率 o_{sp}^{p}，即 $o_{sp}^{p}=\mathrm{FFN}^{p}(pc)$，其对应的损失函数如下所示：

$$\begin{aligned}loss_{sp}^{SPCJT}&=-[l_{sp}^{p}*\log(o_{sp}^{p})+(1-l_{sp}^{p})*\log(1-o_{sp}^{p})]\\ loss_{(s_1,t_1)}^{SPCJT}&=\frac{1}{N_1}\sum_{sp\in AP_{t_1}^{s_1}}loss_{sp}^{SPCJT}\\ Loss^{p}&=\frac{1}{D\times L}\sum_{s_1}\sum_{t_1}loss_{(s_1,t_1)}^{SPCJT}\end{aligned} \tag{3.7}$$

其中 $loss_{sp}^{SPCJT}$ 是序列对 sp 的价格变化同向性任务损失，$loss_{(s_1,t_1)}^{SPCJT}$ 是股票 s_1 在 t_1 时刻的损失，$Loss^{p}$ 是总体的价格变化同向性任务的损失。

3.4.3 成交量变化同向性任务

在成交量变化同向性任务中，期望模型能够去学习到股票序列之间的成交量变化同向性特征，对于一对“锚——样本”序列对 $sp=(x_{t_1}^{s_1},x_{t_2}^{s_2})$，对应的成交量变化同向性任务标签定义如下：

$$l_{sp}^{v}=\begin{cases}1 & \Delta v_{t_1}^{s_1}\times\Delta v_{t_2}^{s_2}\geq 0\\ 0 & \Delta v_{t_1}^{s_1}\times\Delta v_{t_2}^{s_2}<0\end{cases} \tag{3.8}$$

其中 $\Delta v_{t_1}^{s_1} = v_{t_1+1}^{s_1} - v_{t_1}^{s_1}$ 表示股票 s_1 在 (t_1+1) 时刻和 t_1 时刻成交量差，$\Delta v_{t_2}^{s_2} = v_{t_2+1}^{s_2} - v_{t_2}^{s_2}$ 表示股票 s_2 在 (t_2+1) 时刻和 t_2 时刻的成交量差。如果 l_{sp}^{v} 等于 1，表明序列 $x_{t_1}^{s_1}$ 和序列 $x_{t_2}^{s_2}$ 有着成交量变化同向性的数据特征，在接下来的交易日中有着更大的概率出现成交量同涨同跌的走势。

对于给定的“锚——样本”序列对 sp，通过编码后得到两段序列的表征向量 $pc = (c_{t_1}^{s_1}, c_{t_2}^{s_2})$，然后将表征向量喂入到前馈神经网络中去计算序列 $x_{t_2}^{s_2}$ 和 $x_{t_1}^{s_1}$ 在下一个交易日成交量同涨同跌的概率 o_{sp}^{v}，即 $o_{sp}^{v} = \mathrm{FFN}^{v}(pc)$，其对应的损失函数如下所示：

$$
\begin{aligned}
loss_{sp}^{TVCJT} &= -[l_{sp}^{v} * \log(o_{sp}^{v}) + (1 - l_{sp}^{v}) * \log(1 - o_{sp}^{v})] \\
loss_{(s_1,t_1)}^{TVCJT} &= \frac{1}{N_1} \sum_{sp \in AP_{t_1}^{s_1}} loss_{sp}^{TVCJT} \\
Loss^{v} &= \frac{1}{D \times L} \sum_{s_1} \sum_{t_1} loss_{(s_1,t_1)}^{TVCJT}
\end{aligned}
\tag{3.9}
$$

其中 $loss_{sp}^{TVCJT}$ 是序列对 sp 的成交量变化同向性任务损失，$loss_{(s_1,t_1)}^{TVCJT}$ 是股票 s_1 在 t_1 时刻的损失，$Loss^{v}$ 是总体的成交量变化同向性任务的损失。

3.4.4 目标函数

对于所设计的三种自监督辅助任务，其对应的任务损失分别命名为 $Loss^{d}$，$Loss^{p}$，和 $Loss^{v}$，进一步定义多任务联合训练的损失为：

$$
\mathscr{L}_t = \alpha * Loss^{d} + \beta * Loss^{p} + \gamma * Loss^{v} \tag{3.10}
$$

其中，α, β, γ 是模型用来调整三种自监督辅助任务间相对重要性的参数，我们使用 Adam 优化器来优化损失函数，神经网络的参数通过反向传播算法进行更新。

3.5 股市预测

基于 S3E 序列编码预训练模型，接下来就可以进行股市预测任务。为了预测一个股票在第 $(t+1)$ 天的价格走势，本文使用了距当前时间前 k 天的股票序列 $\{r_{t-k+1},\cdots,r_t\}$，将这些序列输入 S3E 模型的序列编码器中去得到每天的表征向量，对于得到的每天的表征向量 $E=\{emb_{t-k+1},\cdots,emb_t\}$，再将其输入到 LSTM 模型中进行编码，即：

$$H=\{h_{t-k+1},h_{t-k+2},\cdots,h_t\}=LSTM(E) \tag{3.11}$$

对于编码后的隐状态 H，取最后一个时间步的隐状态 h_t 喂入到前馈神经网络中去进行二分类预测：$p=\mathrm{FFN}_l(h_t)$，其中 p 是股票在下一个交易日第 $(t+1)$ 天上涨的概率。损失函数定义如下：

$$\begin{aligned} loss_i^s &= -[l_i^s * \log(p_i) + (1-l_i^s) * \log(1-p_i)] \\ Loss_l &= \frac{1}{D\times L}\sum_s\sum_i loss_i^s \end{aligned} \tag{3.12}$$

其中 $loss_i^s$ 是股票 s 在第 i 天的损失，$l_i^s\in\{0,1\}$ 是股票 s 在第 i 天的涨跌走势。

第四章　实验对比与结果分析

4.1　实验设置

本文在中国 A 股市场和纳斯达克市场两个数据集上进行了充分的实验来评估 SMART 股市预测框架的效果，分别收集了两个金融市场 2015 年 1 月 1 日至 2019 年 12 月 31 日的相关数据，每天的技术面交易数据包括开盘价、收盘价、最高价、最低价、成交量、成交金额。通过一个大小为 k 的滑动窗口沿着时间轴逐天滑动，以此来构建每日技术面交易序列数据，同时按照时间的先后关系切分成训练集和测试集，即 2015 年 1 月 1 日至 2018 年 12 月 31 日的数据作为训练集，2019 年 1 月 1 日至 2019 年 12 月 31 日的数据作为测试集，同时对股票的价格涨跌标签做了一定的处理，如果股票的价格相较于上一个交易日涨幅超过 0.5%，则标签被设置为 1，如果股票的价格相较于上一个交易日跌幅超过 0.5%，则标签被设置为 0。对于那些涨跌幅属于 $[-0.5\%, 0.5\%]$ 的股票，属于正常的波动范围，并没有太多的数据特征来反应下一个交易日的涨跌情况，将其排除在外，防止这些有噪声的数据影响到模型的训练效果。预测所有股票在下一个交易日的涨跌走势，采取 accuracy 和 F1-score 作为评价指标。

SMART 框架中的超参数设置如表4.1所示：序列长度 k 设置为 20, S3E 中序列编码器的 Transformer 的 block 个数 N 设置为 1，多任务联合训练的权重参数 α, β 和 γ 全部设置为 1，LSTM 隐层神经元的个数 hidden_num 设置为 256，模型的学习率 lr 初始值设置为 0.001。

4.2　评价指标

股票的涨跌分类是一个二分类问题，对于二分类问题，可以以正确率（Accuracy）和错误率（Error）作为一种评价指标，正确率表示所有预测的样本中预测正确的样本个数占所有预测样本个数的比例，错误率表示所有预测的样本中预测错

超参数	参数值
序列长度 k	20
S3E 中 Transformer 的 Encoder 个数 N	1
多任务联合学习权重 α, β, γ	1
LSTM 隐层神经元个数 $hidden\_num$	256
学习率 lr	0.001

表 4.1: 模型超参数设置

误的样本个数占所有预测样本个数的比例。虽然正确率是一个很直观的评价指标，但是在一些场景下正确率这个评价指标并不一定十分有效，比如在某些样本标签极不均衡的情况下，模型只要全部预测为标签多的那类样本便可以取得比较高的正确率，但此时模型的效果其实是比较差的，为了更好地衡量模型效果，本文还引入了 F1-score 作为评价指标，F1-score 相比于正确率和错误率来说是个更加有效的评价指标，特别是在一些样本不均衡的情况下能够有效地评价模型。

4.3 基线模型

为了衡量本文所提出的框架方法的效果，本文选取了多个图表征算法模型和序列编码算法模型来进行比较。

所选取的算法模型包括：

- **DeepWalk + LSTM**[47]：利用爬虫技术在企查查上爬取上市公司相关信息及其有关联的公司，以此形成一张图数据结构，图上的每一个节点代表一家上市公司，存在边代表上市公司和上市公司间存在着关联关系。DeepWalk 算法的核心是将图上的节点进行向量表征，采取随机游走算法的思想，规定好游走的路径长度，随机选取一个开始的节点进行游走，得到一段节点序列，然后借鉴自然语言处理（Nature Language Process）里的思想，将一段序列当做

一个句子，将序列里的每一个节点当做一个单词，采用 word2vec 的方法进行训练，最后得到节点的表征。在得到节点表征之后，再将其输入到 LSTM 模型中进行编码，最后基于编码后的隐状态进行股市未来涨跌趋势的预测任务。

- **node2vec + LSTM**[48]：node2vec 与 DeepWalk 的思路相当相近, 同样是采用了上市公司的图数据结构, 也采用了随机游走的算法思想, 但相对于 DeepWalk, 在随机游走的权重上也做出了一定程度的创新, 使得图表征的结果更加满足了网络的同质性与结构性。这里," 同质性" 指的是距离相近的节点, 它们的表征会更加相似," 结构性" 指的是在地图数据结构中结构距离相同的节点, 那么它们的表现就应该尽可能的接近。因此，node2vec 主要通过节点间的跳转概率，通过参数来控制游走回上一个节点还是游走到远方节点生成节点序列, 之后和 deepwalk 类似采用 word2vec 的方法对已经得到的节点序列进行训练, 最后得到节点表征。在得到节点表征之后，再将其输入到 LSTM 模型中进行编码，最后基于编码后的隐状态进行股市未来涨跌趋势的预测任务。

- **LINE + LSTM**[49]：LINE 的理念不同于 Deepwalk 和 node2vec，它主要是根据分布相似性学习图节点的表示，主要是指一阶邻近关系和二阶邻近关系，如果两个节点有连通的直边，则称为一阶邻近关系；如果两个节点不直接相连，但可以通过中间节点连接，这称为二阶邻近关系。最后生成的图节点表征向量, 它结合了一阶邻近和二阶邻近, 一旦获得了节点的表示，就将其引入 LSTM 模型进行编码。最后，根据编码后的隐藏状态进行未来股市涨跌趋势的预测。

- **GCN**[50]：图卷积神经网络，是一个应用在图数据结构上的特征提取器。基于构建的上市公司的图数据结构，采取图卷积神经网络的方法对节点进行表征，去学习到图中节点内部的相互关联关系，然后基于节点表征去进行股市未来涨跌趋势的预测任务。

- **GCN + LSTM**：对于股市序列模型先采用 LSTM 模型进行编码，基于 LSTM 编码后的隐状态，将其输入到图卷积神经网络中去，基于编码后的隐状态让图卷积神经网络去学习到股票与股票之间的相关性，得到上市公司的节点表征，最后基于节点表征去进行股市未来涨跌趋势的预测任务。

- **LR**[41]：逻辑回归是一个特别经典的算法，因为其简单、可并行、可解释性强的特点被广泛的应用，特别是在二分类任务中。对于股市预测二分类任务场景，将股票的技术面数据作为特征输入到逻辑回归模型中，预测下一个交易日的股票涨跌情况。

- **CNN**[51]：卷积神经网络具有一定程度上的表征学习能力，在一些任务中可以采用卷积神经网络对数据进行特征编码，对于股市预测的场景，通过卷积神经网络对股票序列数据进行表征，对于得到的表征向量，去进行股市未来涨跌趋势的预测任务。

- **LSTM**[52]：LSTM 对于时间序列数据有着较好的编码效果，因此被广泛地应用在时间序列数据编码的场景中。在股市预测的场景中，正好需要对股票的时序数据进行编码，因此 LSTM 在这个场景下非常合适，对于股票每日的技术面交易数据序列，利用 LSTM 进行编码后，基于编码的隐状态去预测下一个交易日的股票涨跌情况。

- **Attentive LSTM（ALSTM）**[53]：ALSTM 在传统的 LSTM 网络上做了相关创新，对于输入的特征，先经过一层前馈神经网络，再经过 LSTM 层，对于 LSTM 层编码的隐状态，ALSTM 并不直接基于隐状态进行预测，而是通过时序上的 Attention 机制来进行加权求和，最后基于加权求和后的向量去预测下一个交易日的股票涨跌情况。

- **Basic Transformer（B-TF）**[54]：Transformer 主要都是基于 Attention 机制的思想，摒弃了以往深度学习中使用的 CNN、RNN 等结构，当 Transformer 被

提出后在多个领域的多个任务中都验证了其编码的有效性。因此，在本文的股票涨跌预测任务中，考虑采用 Transformer 来编码股票的技术面数据，基于编码后的表征向量去预测下一个交易日的股票涨跌情况。

- **Hierarchical Multi-Scale Gaussian Transformer（HMG-TF）**[55]：层次化多粒度高斯 Transformer 是在传统的 Transformer 上做了相关改进，采用高斯先验增强 Transformer 的局部性，具有从金融序列中挖掘极长期相关性的优势，对于金融时序数据的编码效果较好，对于编码后的隐状态，进行预测下一个交易日的股票涨跌情况。

除此之外，本文还衡量了 SMART 框架的两种变体，分别是 SMART-noCB 和 SMART-noTB。SMART-noCB 表示负采样的样本中排除了对比负样本，仅仅只包含时序负样本；SMART-noTB 表示负采样的样本中排除了时序负样本，仅仅只包含对比负样本；而 SMART 表示负采样的样本中既包含对比负样本，又包含时序负样本。

对于图表征的算法，包括 DeepWalk + LSTM，node2vec + LSTM，LINE + LSTM，GCN 和 GCN + LSTM，将上市公司当做图中的节点，将上市公司和上市公司之间的关联关系当做节点间的边，然后采用图表征的算法得到节点的表征向量，对于一个要预测的目标公司，我们选取 top N 个最关联的相关公司，将它们的特征融合起来一起去预测目标公司下一个交易日的涨跌情况。对于序列编码算法模型，包括 LR，CNN，LSTM，ALSTM，B-TF，HMG-TF 和 SMART，将时序数据喂入到模型进行编码后得到编码后的向量，以此作为序列数据的表征，使用编码后的序列表征去进行股票下一个交易日的涨跌预测任务。

4.4 实验结果分析

4.4.1 分类评价指标

本文将所提出的方法与基线模型进行了对比，以此来衡量模型的优劣，首先比较分类任务的传统指标，主要是正确率（Accuracy）和 F1-score。本文首先评估

Methods		Accuracy(%)/F1-score(%)	
		China A-Shares	NASDAQ
Graph Embedding Methods	DeepWalk+LSTM	52.15/45.22	51.89/44.76
	node2vec+LSTM	52.88/46.34	52.30/45.23
	LINE+LSTM	53.19/46.69	52.67/45.30
	GCN	55.19/48.36	54.28/47.11
	GCN+LSTM	57.98/51.35	56.03/50.84
Sequence Encoding Methods	LR	52.30/39.15	52.03/37.86
	CNN	52.77/41.98	52.45/39.76
	LSTM	53.28/47.44	53.97/47.40
	ALSTM	55.27/49.96	54.01/47.56
	B-TF	57.22/50.32	54.93/49.57
	HMG-TF	57.88/51.13	56.01/50.89
	SMART	**58.96/52.77**	**57.64/51.62**
	SMART-noCB	58.44/52.35	56.96/50.43
	SMART-noTB	58.49/52.18	56.71/50.27

表 4.2: SMART 框架和一些主流模型在 2019 年的正确率和 F1-score

了所提出的 SMART 框架方法和其他基线模型方法的正确率和 F1-score，实验相关结果如下表4.2所示。在基于图表征的方法中，DeepWalk + LSTM 方法在中国 A 股市场上达到 52.15% 的正确率和 45.22% 的 F1-score，在美国 NASDAQ 市场上达到 51.89% 的正确率和 44.76% 的 F1-score；node2vec + LSTM 方法在中国 A 股市场上达到 52.88% 的正确率和 46.33% 的 F1-score，在美国 NASDAQ 市场上达到 52.30% 的正确率和 45.23% 的 F1-score；LINE + LSTM 在中国 A 股市场上达到 53.19% 的正确率和 46.69% 的 F1-score，在美国 NASDAQ 市场上达到 52.67% 的正确率和 45.30% 的 F1-score；GCN 方法在中国 A 股市场上达到 55.19% 的正确率和 48.36% 的 F1-score，在美国 NASDAQ 市场上达到 54.28% 的正确率和 47.11% 的 F1-score；GCN + LSTM 方法在中国 A 股市场上达到 57.98% 的正确率和 51.35% 的 F1-score，

在美国 NASDAQ 市场上达到 56.03% 的正确率和 50.84% 的 F1-score。在基于序列编码的模型方法中，LR 方法在中国 A 股市场上达到 52.30% 的正确率和 39.15% 的 F1-score，在美国 NASDAQ 市场上达到 52.03% 的正确率和 37.86% 的 F1-score；CNN 方法在中国 A 股市场上达到 52.77% 的正确率和 41.98% 的 F1-score，在美国 NASDAQ 市场上达到 52.45% 的正确率和 39.76% 的 F1-score；LSTM 方法在中国 A 股市场上达到 53.28% 的正确率和 47.44% 的 F1-score，在美国 NASDAQ 市场上达到 53.97% 的正确率和 47.40% 的 F1-score；ALSTM 方法在中国 A 股市场上达到 55.27% 的正确率和 49.96% 的 F1-score，在美国 NASDAQ 市场上达到 54.01% 的正确率和 47.56% 的 F1-score；B-TF 方法在中国 A 股市场上达到 57.22% 的正确率和 50.32% 的 F1-score，在美国 NASDAQ 市场上达到 54.93% 的正确率和 49.57% 的 F1-score；HMG-TF 方法在中国 A 股市场上达到 57.88% 的正确率和 51.13% 的 F1-score，在美国 NASDAQ 市场上达到 56.01% 的正确率和 50.89% 的 F1-score。总的来说，在图表征算法中，GCN＋LSTM 的表现最好，而在序列编码模型中，HMG-TF 的表现最好。

本文所提出的 SMART 框架方法及其变体，整体的表现较好。SMART 方法在中国 A 股市场上达到 58.96% 的正确率和 52.77% 的 F1-score，在美国 NASDAQ 市场上达到 57.64% 的正确率和 51.62% 的 F1-score；SMART-noCB 方法在中国 A 股市场上达到 58.44% 的正确率和 52.35% 的 F1-score，在美国 NASDAQ 市场上达到 56.96% 的正确率和 50.43% 的 F1-score；SMART-noTB 方法在中国 A 股市场上达到 58.49% 的正确率和 52.18% 的 F1-score，在美国 NASDAQ 市场上达到 56.71% 的正确率和 50.27% 的 F1-score。

此外，一件需要强调的事情是股市预测任务是具有挑战的，并不同于其他一些分类预测任务有着特别高的正确率，在分类传统指标上一些细微的提升能够带来显著的超额收益率，比如 0.5% 的正确率的提高能够带来超过 10% 的超额收益，本文所提出的 SMART 方法相比于已有的 SOTA（State of the art）算法模型提高了超过 1% 的正确率，这将会带来显著的累计收益率的提升，本文将在之后的实际应

Metrics	Ablation Settings			
	SMART	S-noSSDT	S-noSPCJT	S-noTVCJT
Accuracy(%)	**58.96**	57.31(-1.65)	55.40 (-3.56)	56.82(-2.14)
F1-score(%)	**52.77**	50.48(-2.29)	49.71 (-3.06)	50.83 (-1.94)

表 4.3: 多任务消融实验的正确率和 F1-score 的结果

用中进行展示。

4.4.2 消融实验

为了验证本文所设计的正负样本判别任务、价格变化同向性任务、成交量变化同向性任务一共三种自监督辅助任务的有效性，本文依次将三种自监督任务中的其中一种排除在外，得到 SMART 的三种变体。对于排除正负样本判别任务的 SMART 变体，将其命名为 S-noSSDT；对于排除价格变化同向性任务的 SMART 变体，将其命名为 S-noSPCJT；对于排除成交量变化同向性任务的 SMART 变体，将其命名为 S-noTVCJT。对于 SMART 的各种变体，将其单独进行训练，然后提取出 Encoder 部分对技术面数据进行编码，对于编码后的表征向量进行股市预测任务，相应的实验结果如表4.3所示。从实验结果来看，每种自监督辅助任务都在发挥着一定的作用，增强了序列编码的有效性从而提高了最后金融股市预测的效果，排除任一自监督辅助任务都会导致 SMART 框架方法的效果下降，特别是排除价格变化同向性辅助任务后，SMART 方法的效果下降最为明显，因为价格变化同向性辅助任务主要包含了股票之间价格关联性的影响，相比于其他两种自监督辅助任务，与最后股票的涨跌走势预测更为相关。

更进一步的，为了衡量序列长度 k 对 SMART 方法的效果影响，本文选取了不同的序列长度 k 进行实验，实验结果如表4.4所示。从实验结果可以看出，SMART 方法的效果随着序列长度 k 从 5 到 20 逐渐提高，因为随着序列长度的增加，包含

Sequence Length	China A-Shares		NASDAQ	
	Accuracy(%)	F1-score(%)	Accuracy(%)	F1-score(%)
k = 5	58.82	52.32	57.48	51.36
k = 10	58.87	52.56	57.55	51.48
k = 20	58.96	52.77	57.64	51.62

表 4.4: 不同序列编码长度对于 SMART 框架的影响

了更多数据信息，能够让模型学习到更丰富的隐状态，从而提高最后预测的效果，但是序列长度也不能过长，过长的序列长度使得模型更加复杂难以训练，容易出现潜在的梯度消失和梯度爆炸的问题，还可能让模型关注不到重点从而使预测效果变差，因此需要选择一个合适的序列长度 k 来保证模型有着比较好的效果。

4.4.3 业务指标

传统的分类任务评价指标从一个角度衡量了模型的效果，但与之相比实际的业务更加关注模型的业务指标效果，即按照模型给出的输出结果进行交易，最后能够达到的实际收益率水平，是量化模型最重要的一项评价指标。

为了评估模型的业务指标，本文研发了一套量化回测系统，用来对模型的输出股票进行买卖操作，最后会形成一个总的收益率，更具体的来说，按照以下流程进行:

1. 在第 t 期股市收盘后，基于技术面数据利用 SMART 方法对金融市场的每个股票输出一个预测得分，预测得分越高说明模型越看好这个股票，认为对应股票在下一个交易日的上涨概率更大；预测得分越低说明模型越不看好这个股票，认为对应股票在下一个交易日的上涨概率更低。

2. 在第 t+1 期股市开盘后，将股市里的所有股票按得分从高到低进行排序，然

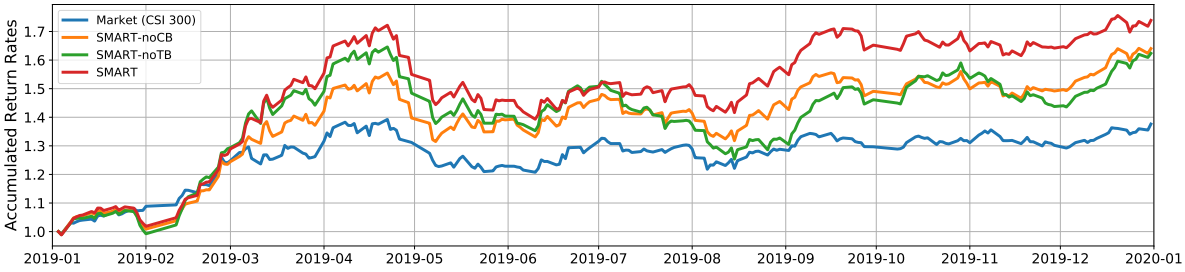


图 4.1: SMART 框架及其变体的累计收益率情况

后选取 Top K 个股票形成一个投资组合，对于已形成的投资组合里的股票，如果上一期已经存在对应的持仓，则对应股票不进行买卖操作，如果上一期不存在对应的持仓，则卖出持仓股票中得分低的股票，买入投资组合里得分高的股票，从而实现一整个调仓换股的操作。

3. 每当下一个调仓周期到来时，重复执行步骤 1 和步骤 2，定义回测开始时间的资金净值为 1，随着不断地调仓换股，持仓的投资组合每天都会产生对应的收益率，在净值的基础上叠加上收益率，最后会形成一条净值曲线，反映在整个模拟交易回测时间段上的净值情况。

值得一提的是，如果单纯地按照步骤 2 的方式去模拟交易执行，常常会因为频繁的交易导致手续费过高从而削减了收益，也会因为模型过度的偏好某一个行业的股票导致投资组合里相同行业的股票比例过高从而带来行业风险。因此，本文做了些模拟交易上的优化调整，通过限制换手率的上限来缓解频繁交易带来的手续费的冲击成本；对于行业偏好风险，按照行业进行分组，然后轮询行业去选择排名靠前的股票，以此来保证行业的多样化选股。基于上述优化后的模拟交易步骤，本文以沪深 300 指数（CSI 300）作为对照基准（benchmark），评估了 SMART 方法及其两个变体 SMART-noCB 和 SMART-noTB，相关的净值曲线实验结果如图4.1所示。就 SMART 方法的净值曲线和沪深 300 的净值曲线比较而言，从图中可以看出虽然 SMART 方法在 2019 年 2 月份稍微落后于沪深 300 指数，但在接下来的时间中收益逐渐反超沪深 300 指数，占据了绝对领先的地位；而从 SMART 的净值曲线与其变体 SMART-noCB 和 SMART-noTB 的比较结果来看，它们都有着比较相似的

走势趋向，SMART-noCB 和 SMART-noTB 容易受到不同时间阶段的影响而出现一些波动，而 SMART 的表现更加稳定，最后能够达到更高的累计收益率。

在金融量化领域，衡量量化模型的优劣除了从形成的净值曲线的角度外，往往也会衡量信息系数（Information Coefficient）和夏普率（Sharp Ratio）。

信息系数，也称为 IC 值，表示所选股票因子值与下一期股票收益率之间的横截面相关系数。这里的因子值在量化交易的场景中指的就是模型预测出的股票得分，相关系数一般可以分为皮尔逊相关系数（Normal IC）和斯皮尔曼相关系数（Rank IC）。皮尔逊相关系数表示某时点某因子在全部股票的因子暴露值与其下期回报的截面相关系数，其计算方式如下；

$$Normal\ IC = corr(f_{t-1}, r_t) \tag{4.1}$$

其中 f_{t-1} 为股票 t-1 期的因子值，r_t 为股票 t 期的收益率。但更常用的是斯皮尔曼相关系数，也就是 Rank IC，其计算方式如下：

$$Rank\ IC = corr(order_{t-1}^{f}, order_{t}^{r}) \tag{4.2}$$

其中 $order_{t-1}^{f}$ 为 t-1 期各股票的因子值排名，$order_{t}^{r}$ 为 t 期各股票收益率排名。

IC 值可用于评估当期的因子值对下一期股票收益率的预测能力。总的来说，如果该因子的绝对值越高，说明该因子对下一期股票的收益率具有良好的预测效果；如果该因子的 IC 值为正，则表明该因子的值与下一期股票收益率正相关；如果该因子的 IC 值为负，则表明该因子的值与下一期股票收益率负相关，即因子值越低，下一期的股票收益率越高。IC 的理论最大值为 1，对于多期的因子 IC 均值来说，当因子的绝对值大于 0.04 时，便可以认为是一个有效的阿尔法因子，能够带来一定程度的超额收益，有着不错的因子选股能力。

当然，金融市场上的收益和风险永远都是共生共存的，所以一个优秀的量化模型不仅仅把目光放在收益率上面，同时也需要考虑到相关的风险状况，夏普利

率就是一个综合衡量收益和风险的评价指标，其计算公式为：

$$Sharpe\ Ratio = \frac{E(R_p) - R_f}{\sigma_p} \tag{4.3}$$

其中 $E(R_p)$ 表示投资组合的预期收益率，R_f 表示无风险利率，σ_p 表示投资组合的标准差，也就是衡量投资组合的风险。其中，分子 $E(R_p) - R_f$ 也就是指超额收益率，表示投资组合超出无风险收益率的那一部分，分母表示投资组合的风险，两者相比也就表示每承担一单位的风险，能够达到多少程度的超额收益率，因此夏普利率越大则说明对应投资组合的效果越好。

因此，基于上述所提的量化领域的业务评价指标，本文在所提出的 SMART 方法及其变体 SMART-noCB 和 SMART-noTB 上做了评估，同时为了更好的比较，本文也选取了金融量化领域一些常用的因子，分别为：

- H2C(highest / close price)：股票当天的最高价除以股票当天的收盘价，用以衡量当天股票的上涨走势的最大强度。
- ROA(return on assets)：资产回报率，以税后净利润除以总资产表示，衡量单位资产创造的净利润。
- ROE(return on equity)：净资产收益率，净利润和股东平均权益的百分比，反映股东权益的收益水平，衡量公司自有资本的运用效率。
- MV(market value)：股票的市值，用来衡量公司资产规模，等于上市公司发行的普通股数量乘以股票价格。

相关实验的对比结果如表4.5所示，从表中可以看出，SMART 方法不仅在累计收益率上表现较好，达到了 73.89% 的累计收益率，在因子 IC 值和夏普率上都有着不错的表现，分别达到了 0.062 的 IC 值和 0.405 的夏普率，显著地高于其他方法。其次是 SMART 的两种变体，SMART-noCB 达到了 64.03% 的累计收益率，0.056 的 IC 值和 0.293 的夏普率，SMART-noTB 达到了 62.36% 的累计收益率，0.059 的 IC

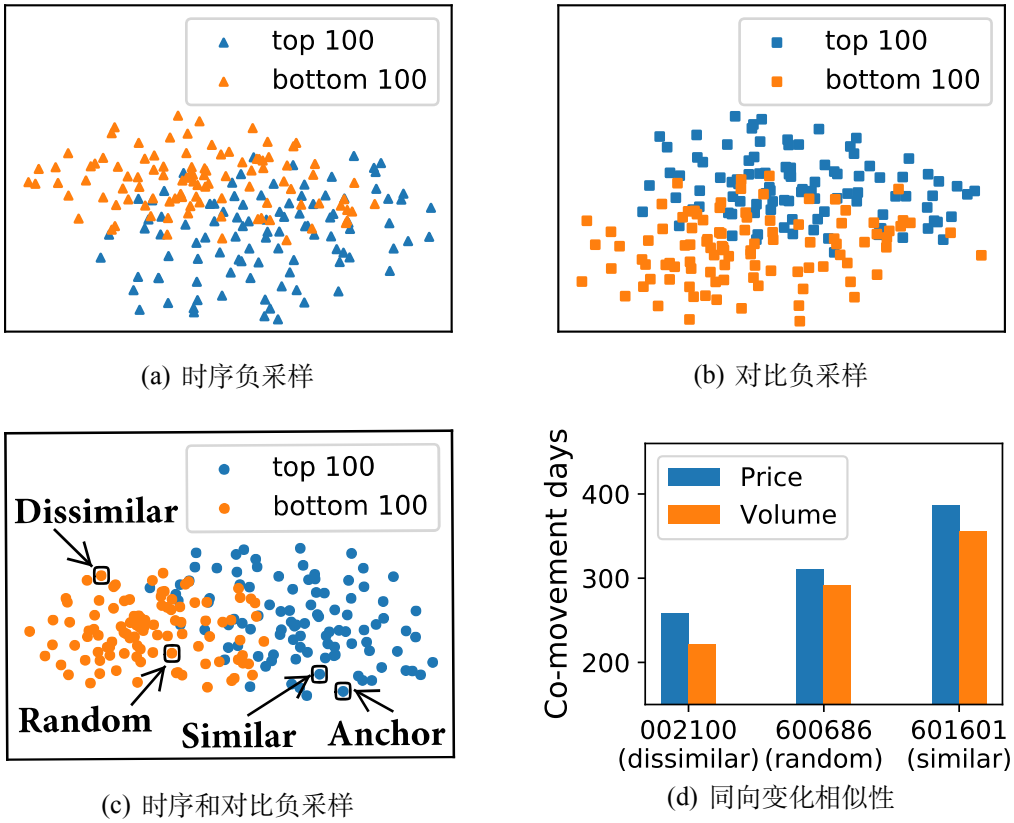
Input Factors	Accumulated Returns	IC Values	Sharpe Ratios	Market Return
H2C	26.56%	0.015	0.105	
ROA	32.83%	0.031	0.129	
ROE	34.27%	0.034	0.156	
MV	44.15%	0.042	0.228	33.47%
SMART-noCB	64.03%	0.056	0.293	
SMART-noTB	62.36%	0.059	0.345	
SMART	**73.89%**	**0.062**	**0.405**	

表 4.5: 不同因子在 2019 年的业务评价指标

值和 0.345 的夏普率，相比于完备的 SMART 方法虽然效果略逊一筹，但还是显著领先于其他方法。这充分地说明了我们所提出的 SMART 方法在股市预测任务中的有效性，以 SMART 方法构建的量化模型能够带来较高的累计收益率、IC 值和夏普率。

4.5 案例研究

为了更好地说明本文所提出的 SMART 方法在编码股票技术面数据的效果，本文将编码后的每日股票表征进行了取平均的操作，从而得到了一个股票对应的整体表征向量，然后基于 T-SNE 的方法进行降维处理，进行二维平面的表征可视化展示。在构建“锚——样本”序列对时，同时采取了两种负采样方法，基于采样的样本让模型进行学习，为了更好地说明两种负采样方法都起到了一定的作用，本文分别对仅使用对比负采样、仅使用时序负采样、同时使用对比负采样和时序负采样的方法所得到的二维平面表征进行了一定的对比，结果如下图4.2(a)、4.2(b)、4.2(c)所示。



(a) 时序负采样　(b) 对比负采样

(c) 时序和对比负采样　(d) 同向变化相似性

图 4.2: 不同股票的 S3E 序列表征可视化

从实验结果可见，负采样方法中的对比负采样和时序负采样都发挥了一定的效果，能够将排名靠前的股票和排名靠后的股票有效地区分开来，对比负采样所起的效果会略好于时序负采样，综合对比负采样和时序负采样的方法效果最好，能最有效地将排名靠前和排名靠后的股票区分开来。同时，在综合了对比负采样和时序负采样的方法中，选取一个股票作为“锚”，可以发现与当前“锚”股票相似的股票在表征空间上与“锚”股票更为接近，而与当前“锚”股票不相似的股票在表征空间上与“锚”股票距离更远，随机采样的股票则一般处于它们中间的位置，这在一定程度上说明了本文模型所编码的股票表征是有效的。从另一个角度来看，由于本文模型编码了股票之间的相关性，即价格和成交量的同涨同跌性，本文也选取了一些具体的股票进行观察，统计了股票和股票之间价格、成交量上的同涨同跌的情况，如下图4.2(d)所示，选择股票代码为“601318”的股票作为“锚”股票，选取与“锚”股票表征空间相近的股票作为相似股票（股票代码为“601601”），

选取随机采样的一个股票（股票代码为“600686”），以及选取与“锚”股票表征空间相远的股票（股票代码为“002100”），从结果可以观察得到编码表征相似的股票在价格和成交量上的同涨同跌性会比随机采样的股票或者不相似的股票更为明显，有着更多的趋同走势，这同样从一个角度上说明了股票之间的涨跌关联性已经被纳入到表征空间中。

第五章　量化模型管理平台

基于机器学习、深度学习的量化模型研发完成后，如何去高效地维护和管理好研发完成的模型，以及如何让模型适应股票市场风格的改变，使模型始终保持在一个较新的状态，是量化模型管理平台的难点和挑战，也是本文所关注的一个问题。对于算法的工业落地场景，本文提出了一个量化模型管理平台，架构如图5.1所示，主要可分为前端 UI、展示层、业务层、数据层和数据库。该平台主要是帮助使用者能够便捷地管理模型，同时能够可视化地观察模型所预测的每日持仓以及每日股票的得分情况等，直观地展示量化模型投资组合的收益情况，而对于股市市场风格的改变，采取定时任务的方式对模型进行滚动训练，使模型学习最新的市场数据特征。

量化模型管理平台主要分为两大块内容，一块是模型的管理维护，包含查询模型、创建模型、删除模型、修改模型等相关功能，方便使用者对于模型进行一定的操作；另一块是模型输出结果及绩效的可视化展示等，方便使用者直观地了解模型的效果及相应的评测情况，以及模型的定时任务滚动训练等。

5.1　技术选型

对于研发一套平台管理系统而言，有很多种可供选择的方案，比如基于 python web 进行开发或者是基于 Java Web 进行开发，python 的好处是语言简单，并且开发快、部署快，但是 python 语言运行速度较慢，不同版本间的 python 解释器之间可能存在不兼容的问题，而 JAVA 语言虽然相比于 python 语言更加繁琐和复杂，但由于运行速度较快，且代码的规范完整性较好，因此被广泛地应用在服务器开发、web 开发的各类场景中。综合多方面的考虑，结合量化模型管理平台的实际场景，本文最后决定采取 JAVA web 进行量化模型管理平台后端的开发，基于 JavaScript（JS）脚本语言进行前端的开发。而在实际的开发场景中，往往会采取一些前沿的

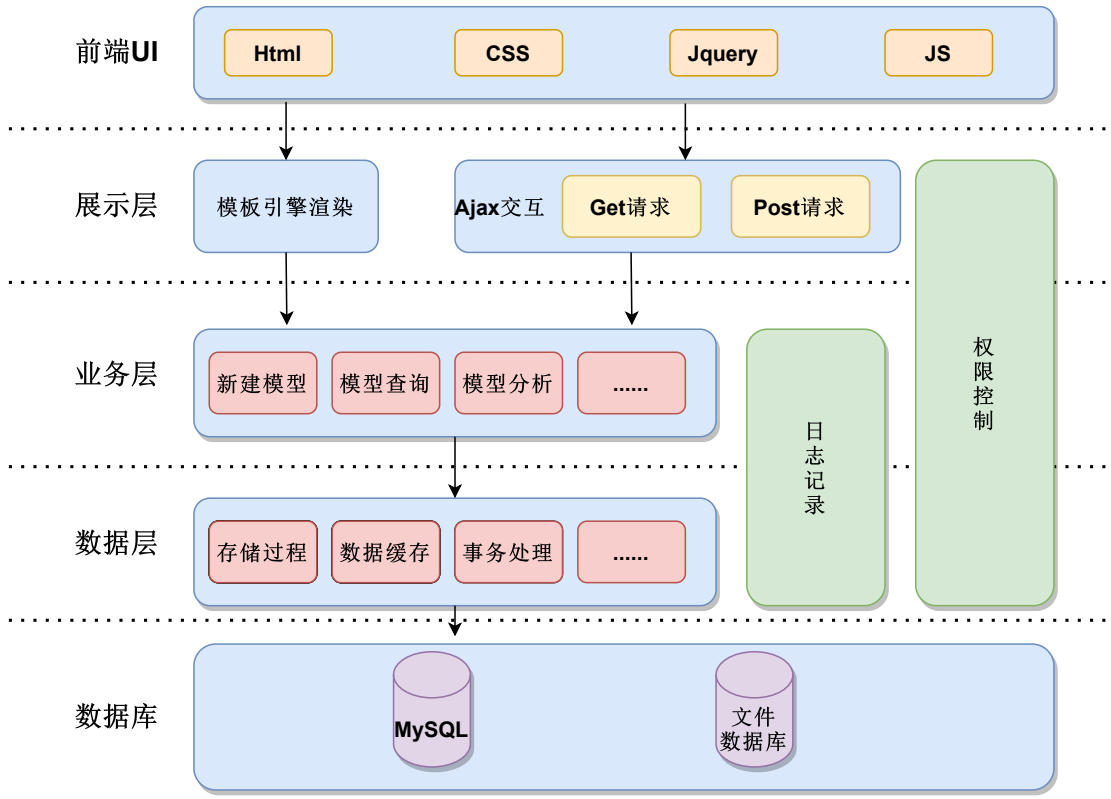


图 5.1: 系统架构图

主流框架进行开发，一方面加快开发的速度，在不考虑一些公共问题的情况下，开发人员可以专注于实现业务。另一方面，框架结构相对统一，便于学习和维护，方便需求变动带来的代码变动修改。因此，本文在前端开发上基于 vue 前端框架，在后端开发上基于 spring 后端框架进行开发，能够快速搭建便捷有效的系统，同时有着优美的页面和良好的用户体验。

5.2 数据库设计

一套有效并且高可用的平台系统往往离不开数据库，数据库作为一套系统最底层的数据支撑，是系统向外面提供服务最基本的保障。数据库用来持久化数据信息，包含各种需要存储的相关数据对象，目前主流的数据库主要分为两类，第一类是适合做事务的行存储关系型数据库，比如 MySQL、Oracle 等，它们是结构化的数据集，由行和列组合而成的二维表数据结构，适合存储结构化的数据；第二类

字段	含义
model_id	模型的 id，用来唯一标识一个模型。
model_name	模型的名字，由用户自己指定，用来区分模型。
algorithm	模型所使用的算法，比如 XGBoost、LSTM 等。
label_type	模型预测的 label 类别，主要包括收益回报率、夏普率。
label_handle_method	模型 label 的处理方式，主要包括标准化和中性化.
feature_handle_method	特征的处理方式，主要包括标准化和中性化。
model_root_path	模型的根目录，用来表明模型的根目录地址。
model_start_dt	模型的开始日期，用来表明模型的训练的开始日期。
train_period	模型的训练频度，用来表明每隔多久进行模型的训练。
model_status	模型的是否启用状态。
is_private	模型的是否公开状态。
is_valid	模型的是否有效状态。
create_time	以时间戳的形式来表明模型的创建时间。

表 5.1: 表字段及其含义

是键值对形式构建的非关系型数据库，比如 MongoDB、Redis 等，适合存储非结构化的数据。对于不同的应用场景，应当选取合适场景的数据库，才能够方便数据库最大程度地发挥功效，而对于量化模型管理平台的应用场景，主流的关系型数据库存储更加合适，因此本文主要选择 MySQL 作为数据库进行存储。量化模型管理平台的数据库主要分为两大部分，一部分是 MySQL 数据库，另一部分是文件数据库：MySQL 数据库主要用来存储模型的元信息，包括模型的 ID、模型的名称、模型使用的算法等等；文件数据库主要是以文件的形式（比如 csv）来存储模型的预测结果和模型评价分析的相关结果，由于相关结果较大，因此采用了文件存储的

文件数据	含义
RankIC	模型当期预测的得分值与下一期收益率的相关系数。
夏普率	模型投资组合收益的超额收益除以标准差。
年化收益率	模型投资组合进行交易所能实现的年化收益率。
波动率	模型投资组合收益率的标准差。
最大回撤率	净值从最高点到最低点的最大跌幅。
数据预测	模型输出的股票得分预测。

表 5.2: 文件数据存储相关含义

形式。

更具体的来说，本文设计了“model_info”模型元信息表在 mysql 中来存储模型相关的信息，包含了各个相关的字段，其对应字段的含义和解释如表5.1所示。

而对于每个具体的模型而言，都有着它的根目录，根目录里包含了各种各样的结果，比如模型分析的相关数据、模型预测的相关数据等，主要如表5.2所示。因为相关的数据内容不小，如果采取 mysql 存储的话数据量太大容易导致数据库的各种性能变慢，因此我们采取 csv 文件存储的方式。当需要展示对应模型的相应结果时，后端请求函数通过读取模型根目录下对应的文件数据，然后返回给前端页面进行展示。

5.3 功能模块

量化模型管理平台主要分为两大块内容，一是模型的管理和维护，二是可视化展示。模型的管理和维护提供给用户便捷管理维护模型的功能；可视化展示让用户了解到模型全方面立体的各种衡量评价指标，同时定时任务的设计使得模型能够及时进行训练学习到最新的金融市场上的数据特征，整个平台的功能模块如图5.2所示，接下来对于每个部分展开详细的介绍。

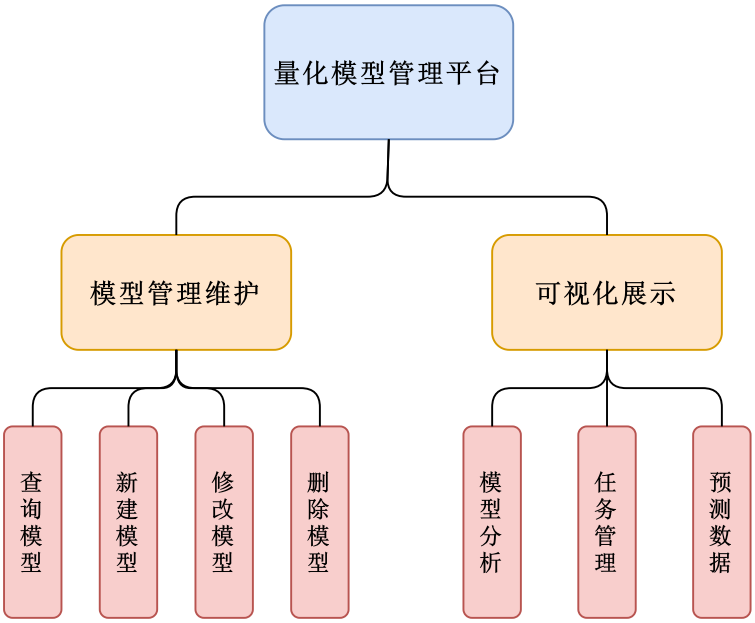


图 5.2: 功能模块图

5.3.1 模型管理维护

模型的管理和维护主要是让用户能够便捷有效地管理和维护模型，当用户想要添加一个新模型时，可以通过点击“新建模型”的按钮，然后配置模型名称、开始时间、根目录等参数信息来完成新建模型的功能，当模型新建完成后，便会加入到已有的模型列表中；当用户想要查询模型列表时，可以通过点击“模型库”来获取所有存在且生效的模型概要信息；当用户想要修改模型相关信息时，可以在“模型信息”处选择所需要修改的相关信息进行修改，修改完成后点击“修改”按钮进行保存；当用户想要删除一个模型时，可以点击“删除”按钮来删除对应不想要的模型。每个功能所对应的具体操作信息和流程，在接下来的内容中展开介绍。

对于新建模型的功能，需要用户输入若干个相关字段信息，从而完成模型的创建，具体包括：

- 模型名称：对应模型所起的名称。
- 开始时间：模型训练的开始时间。
- 根目录：对应模型所在位置的根目录。

- 所在服务器：对应模型所在的服务器位置及所属的用户。
- 训练频度：由于市场风格会随时间推移而发生一定的变化，所以需要通过不断地滚动训练来将最新的市场数据特征纳入进来，因此训练频度就是指多久时间重新进行模型的训练。
- 预测频度：预测频度与量化模型的调仓周期有关，如果所定的调仓周期为 1 天，那么模型的预测频度就应当为 1 天进行一次预测；如果所定的调仓周期为 1 周，那么模型的预测频度就应当为 1 周进行一次预测。
- 模型类型：模型类型可分为择股模型和择时模型，择股模型指的是在股票池里所有的候选股票中，选取一定量的股票形成一个投资组合，在股票金融市场上进行交易从而获取盈利；择时模型指的是针对同一个股票，在不同时间点进行买卖的操作，从而赚取价差进行盈利。
- 算法：算法指的是模型所采取的具体的算法，比如 XGBoost，GRU 等。
- 特征处理：特征处理指的是对于输入的特征数据进行相关的操作，比如市值中性化和行业中性化等，中性化的方式是对因子暴露值和市值、行业进行线性回归，最后用回归之后剩下的残差替代因子值，因此这个计算出来的残差肯定是和市值、行业无关的。
- 标签处理：对应标签的处理方式同样可以采取市值中性化和行业中性化的方法，来剔除掉标签中市值和行业带来的相关性的影响。
- 预测标签：预测标签主要可分为回报率和夏普率。回报率指的是将股票的收益率作为模型训练任务的标签，而夏普率指的是将夏普率作为模型训练任务的标签。
- 预测周期：预测周期指的是以多长时间周期来构建预测标签，比如预测周期为 1 天表示以下一天的回报率或者夏普率作为任务的标签进行预测，预测周期为 7 天表示以之后一周的回报率或者夏普率作为任务的标签进行预测。

模型名称	创建人	所在服务器	算法	特征处理	label处理	预测label	预测周期
stockrnn-basic-model-20210323	yingjia...	kd06_...					5日
GRUAVG10-5D-0.0.1	yingjia...	kd06_...					5日
AlphaNet-10D-0.0.2	yingjia...	kd06_...	ALPH...	标准化	标准化	RETU...	10日
XGB-5D-0.0.1	yingjia...	kd06_...	XGB	行业市...	行业市...	RETU...	5日
Aibs-t5-model	admin	kd06_...					
Aibs-t2-model	admin	kd06_...	GRU	行业市...	无中性化	RETU...	1日
Aibs-t3-model	admin	kd06_...	GRU	行业市...	无中性化	RETU...	1日
NOON-LR-1D-0.0.1	yongm...	kd06_...	XGB	行业中...	标准化	RETU...	1日
NOON-ALPHA-MODEL-5D-0.0.1	admin	kd06_...	ALPH...	无中性化	无中性化	RETU...	5日

图 5.3: 模型列表

- 版本：该模型所对应的版本号，版本号越高表示模型越新，方便区分旧模型和新模型之间的差别。
- 私密性：该模型是否是公开的，如果设置了私密性，则表示该模型不公开，普通用户没有权限查看该模型；如果不设置私密性，则表示该模型公开，任意用户都有权限查看该模型。

当在对应的重要字段填充完相关信息后，点击“确定”按钮即可完成模型的创建。根据前端页面的请求地址，在后端编写对应的请求处理函数“CreateModel”，来处理接受到的请求，进行一定的业务逻辑处理后，将相关信息更新到数据库进行持久化，然后将执行后的结果返回给前端从而完成前后端的交互行为，前端页面根据后端返回的结果进行相关的页面渲染操作。

对于查询模型列表的功能，通过点击“模型库”按钮，前端发送对应请求到后

端，后端有相应的处理函数 “ListModel” 来处理请求，从数据库中查询获取模型的相关信息后返回给前端，前端依据后端返回来的数据进行页面的渲染，从而呈现给用户，如图5.3所示。

对于修改模型信息的相关功能，点击 “模型信息” 可以看到之前所创建的模型的相关信息，由于某些字段创建后不允许修改，所以在前端的页面显示上属于灰色不可改的样式，而对于一些可以修改的字段，在对应字段上修改信息后，点击 “修改” 按钮，则会将所要修改的字段信息作为参数，发送到对应的后端修改请求地址，后端会有对应的 “UpdateModel” 处理函数来处理前端传过来的修改请求，根据修改请求的参数来处理修改信息的业务逻辑，修改完成后将数据保存在数据库中，同时将结果返回给前端进行页面的渲染。

对于删除模型的相关功能，点击 “删除” 按钮，前端将模型的 id 发送到后端对应的删除请求地址，后端会有对应的 “DeleteModel” 处理函数来处理前端传过来的删除请求，根据删除请求的 id 参数来删除数据库中对应 id 的模型，同时将结果返回给前端完成交互。

5.3.2 可视化展示

量化模型管理平台除了需要提供给用户便捷的模型管理功能外，还需要提供给用户关于模型的各方面多角度的可视化展示功能，让用户能够方便地了解模型的各个方面。总的来说，可以分为模型分析、任务管理、预测数据三大块主要内容，接下来依次介绍每块内容的各个详细方面。

对于模型分析的功能模块，有着各种各样的评价指标，其中包括 RankIC、夏普率、年化收益率、年化波动率和最大回撤率等等，一般通过分层模拟回测的形式来得到每层的股票组合的各类评价指标，主要如图5.4所示。其中 RankIC 指标用来指示当前模型的预测的得分值与下一期股票涨跌的相关性程度，如果 RankIC 值越高，说明当前模型的预测分值与下一期股票的涨跌相关程度越高，也就是模型效果越好，如果 RankIC 值越低，说明当前模型的预测分值与下一期股票的涨跌相关程

RankIC指标

模型名称	RankIC均值	RankIC标准差	IC_IR	IC>0占比
stockrnn-basic-model-…	0.06	0.04	1.65	0.96

alpha_statistics

名称	夏普率	年化收益	年化波动率	最大回撤率	Calmar
layer_0_net_value…	0.72	9.74%	13.60%	-13.54%	71.94
layer_1_net_value…	1.74	25.09%	14.42%	-10.15%	247.22
layer_2_net_value…	0.86	11.86%	13.83%	-12.80%	92.63
layer_3_net_value…	0.37	4.97%	13.37%	-13.53%	36.76
layer_4_net_value…	-0.13	-1.68%	13.12%	-15.46%	-10.85
layer_5_net_value…	-1.35	-18.78%	13.91%	-23.67%	-79.33
hedge_return_cu…	14.64	53.16%	3.63%	-0.62%	8632.73

图 5.4: RankIC、年化收益、夏普率、波动率等评价分析指标

度越低，模型效果越差；夏普率是综合了收益和风险的一个指标，用来衡量每承担一单位的风险能获得多少超额收益，越高的夏普率表明模型的整体收益风险效果会更好，越低的夏普率表明模型的收益风险整体效果较差；年化收益率是将投资组合的收益进行年化后的表示，也就是指一单位的初始资金，在一年之后能够实现多少收益率，越高的年化收益率表示该模型所能实现的收益情况越好，越低的年化收益率表示该模型所能实现的收益情况越差；年化波动率表示的是模型预测所形成的投资组合的收益率的波动情况，用来衡量该模型的收益情况是否比较稳定，如果年化波动率越低，说明该模型表现比较稳定，如果年化波动率越高，说明该模型的表现不稳定，容易产生较大的风险；最大回撤率是指在选定周期内任一历史时点往后推，资金净值走到最低点时的收益率回撤幅度的最大值，一般用来衡量投资者所能承受的最大回撤心理预期，如果模型所形成的投资组合最大回撤率较大，那么对于承担风险能力较小的投资者来说就不适合投资，一般来说模型所形成的投资组合的最大回撤率越小，说明该投资组合的风险越小，模型的表现会更好。以模型库中的其中一个模型为例，其各个评价指标相关的展示如图，实现过程为前端发送请求到后端对应的“EvalModel”处理函数，后端根据对应的请求找到

	模型名称	任务类型	任务模式	开始时间	结束时间	任务组用时	进展	任务状态
◉	stockrnn-ba	模型预...	自动执行	2021-0...	2021-0...	00:38:11	4 / 8	成功
○	stockrnn-ba	模型预...	自动执行	2021-0...	2021-0...	00:04:07	4 / 8	成功
○	stockrnn-ba	模型预...	自动执行	2021-0...	2021-0...	00:02:40	4 / 8	成功
○	stockrnn-ba	模型预...	自动执行	2021-0...	2021-0...	00:50:15	4 / 8	成功
○	stockrnn-ba	模型预...	自动执行	2021-0...	2021-0...	00:08:44	4 / 8	成功

图 5.5: 模型任务管理

模型的根目录，将根目录下的相关数据读取出来然后返回给前端进行页面的渲染，可视化的模型分析展示页面给用户一个立体模型分析感受，让用户能够了解模型的各方面评测结果。

对于任务管理的功能模块，主要是对模型训练以及预测任务进行相应的调度管理，主要如图5.5所示。每当新的一天过去时，新的数据会被自动地加入到模型中进行训练，而不需要人为手动地进行相应的操作；同理在每天股市收盘后，预测也不需要人为手动地去执行下一个交易日股票得分的预测，而是通过设置定时任务自动化地在收盘后某一时间点进行下一个交易日的股票得分预测，自动化地流程简化了模型的训练和预测过程。其具体的流程为前端会发送对应的请求，包含对应模型的 id 和定时任务调度的相关参数信息，后端收到对应请求后会有对应的“TaskSchedule”处理函数来处理相应请求，把定时任务的相关参数信息加入到服务器系统的 crontab 定时任务中来，然后每当调度时间触发后，系统自动地执行相关的训练或者预测任务。

对于预测数据的功能模块，主要是对于每一天各个股票的得分预测，然后可视化地展示成一个表格的形式，包括预测的日期、预测的排名、预测的个股以及对应的股票代码和得分情况，如图5.6所示。其具体的流程为前端发送对应模型 id 和相关参数的请求到后端，后端会有对应的“PredictResult”处理函数，根据请求找到对应模型的根目录，然后将相应的预测结果加载出来返回给前端，前端根据相

日期	排名	个股	代码	score
2021-09-22	1	康龙化成	300759	0.8566195964813232
2021-09-22	2	富临精工	300432	0.7979437112808228
2021-09-22	3	*ST 全新	000007	0.7647554278373718
2021-09-22	4	酒 鬼 酒	000799	0.7620120048522949
2021-09-22	5	中伟股份	300919	0.7619705796241759
2021-09-22	6	兴发集团	600141	0.7171013951301575
2021-09-22	7	震裕科技	300953	0.7127037644386292
2021-09-22	8	拓 尔 思	300229	0.7071682810783386

图 5.6: 模型对于股票的得分预测

应的结果进行页面的渲染展示，用户可以在此页面直观的看到模型对于各个股票的得分预测情况以及相对排名，更加的方便快捷和灵活。

总的来说，量化模型管理平台提供给用户便捷管理模型的功能，用户可以在该平台上创建注册对应的模型，平台会按照创建的相关信息进行模型的训练。当模型训练完成后，用户可以在平台上观察模型对应的评价指标来衡量模型的优劣，全面立体的可视化展示页面让用户对模型有个全方面的了解。同时，为了应对市场风格随时间可能发生潜在的变化，通过采取定时任务进行滚动训练，让模型学习到最新的数据特征，进一步提高了模型的效果。量化模型管理平台在使用过程中用户体验较好，模型训练能够定时运行，也能定时地给出模型的预测结果，对于训练发生异常或者出错的模型，也可以通过查看日志来定位问题，从而进行修复。

第六章　总结与展望

6.1　本文工作总结

本文所研究的问题主要是金融股市下的股票涨跌预测问题，如何在该场景中借助机器学习、深度学习的技术手段，从历史的交易数据中挖掘潜在的数据特征，从而对股票未来的涨跌走势做出一定程度的预测，最后来根据模型的预测结果，在股票金融市场上进行一定的买卖交易获取盈利。由于最近自监督技术在语音、图像和其他相关领域都有了一定的突破，通过设计一些与目标任务相关联的自监督辅助任务，可以帮助模型更好地捕获相关的数据特征，从而提升目标任务的效果。因此本文率先借鉴了语音、图像等领域的自监督学习任务设计的思想，将其应用在金融量化领域，来进行金融股市的预测任务。更具体地来说，本文设计了一个基于技术面数据的 SMART 框架，采取了自监督学习的技术，使用了多任务联合训练的方法，来完成金融股市预测的任务。SMART 框架通过采样“锚——样本”序列对的形式构建数据集，其中采样的样本包括正样本和负样本，同时关联设计了三种自监督辅助任务，分别是正负样本判别任务、价格变化同向性任务和成交量变化同向性任务，基于所设计的三种自监督辅助任务，来进行 S3E 序列编码预训练模型的多任务的联合训练。待训练完成后，抽取 S3E 序列编码预训练模型中的编码部分，来对股票的序列数据进行编码表征，从而得到股票每天的表征向量，继而将股票每天的表征向量输入到 LSTM + Attention 模型中，在 LSTM 模型的最后一个时间步的隐状态接上一个分类器，来输出该股票在下一个交易日的涨跌概率得分值，分值越高说明模型认为该股票在下一个交易日上涨的概率越大，分值越低说明模型认为该股票在下一个交易日上涨的概率越小。

本文在设计了 SMART 金融股市预测框架后，对 SMART 框架所能实现的效果进行了多方面的评测，其中包括传统的分类评价指标和业务指标，同时对比了一些相关的主流模型的效果。在分类评价指标上，主要是针对股票涨跌的分类任务

进行一定的评测，SMART 自监督多任务金融股市预测框架以 58.96% 的 Accuracy 和 52.77% 的 F1-score 在中国 A 股市场上领先于其他模型，以 57.64% 的 Accuracy 和 51.62% 的 F1-score 在美国纳斯达克市场上领先于其他模型；在业务评价指标上，主要是针对模型所预测的股票得分来构建投资组合，评估模型所预测的股票得分和下一期股票的实际收益情况的相关程度，以及评估对应投资组合在实际金融市场的模拟交易回测上能够实现多少收益率，以此来评判模型在业务指标上的相对优劣情况，本文所设计的 SMART 金融股市预测框架在 2019 年的模拟交易回测上实现了 73.89% 的累计收益率，领先于其他的一些主流模型。由此可见本文所设计的模型在多方面多角度上都处于一定的领先地位。

在本文所设计的算法模型研究完成后，如何将其应用到实际的工业生产环境中，也是一个值得深思熟虑的问题。本文提出了一套量化模型管理平台来管理模型，对于每一个研发完成的量化模型，进行自动化地滚动训练和定时预测任务，以及对于模型的各方面的业务评估，比如年化收益率、夏普率等进行可视化地展示，包括对于股票每一期的得分预测以及持仓情况也在页面上进行优美的展示，让用户能够全方面立体地了解到模型各方面的评估情况。

总体来说，本文的核心贡献主要包括：

- 本文将自监督技术引入到金融量化投资领域，针对股票涨跌预测的场景，提出了一个 SMART 金融股市预测框架，其采取了自监督学习的技术，使用了多任务联合训练的方法，完成金融股市涨跌预测的任务。

- 本文在多个数据集上进行了实验，充分有效的实验证明了本文所提出 SMART 金融股市预测框架的有效性，同时对比了现有的各种主流模型，实验结果表明 SMART 框架及其变体在股票涨跌分类任务的传统指标上和业务指标上都领先于其他方法。

- 本文设计了一套量化模型管理平台，来管理各个量化模型，包括查看、维护以及管理各个量化模型等，同时提供模型的可视化展示功能，包括模型每天

的预测情况、量化模型的累计收益情况等。

6.2 未来发展展望

随着科学技术的进步与发展，越来越多的应用场景开始引入机器学习、深度学习相关技术，来辅助人们进行决策或其他操作，在金融量化投资领域也毫不例外。传统的人为投资虽然目前仍然占据了金融股票投资领域的大部分市场，但基于人工智能的量化投资算法最近几年发展迅速，基于算法交易的资金规模也日渐扩大，逐渐成为一个未来的走势，因此如何去研发一个有效的人工智能量化模型便显得十分重要。本文是从日频技术面的交易数据作为输入数据进行模型的构建，但除了历史交易数据对股票未来涨跌走势的影响外，还有其他很多方方面面的因素都会对股票未来的涨跌走势有一定的影响。比如，上市公司的金融新闻事件会对上市公司的股价产生影响，利好新闻推动股票价格上涨，利空新闻推动股票价格下跌；宏观经济形式会对股市产生比较大的影响，良好乐观的宏观经济形式容易产生牛市一路上涨，萧条悲观的宏观经济形式容易产生熊市一路下跌；投资者的交易行为情绪也会对股市产生一定的影响，可以通过自然语言处理的技术在一些留言评论中进行捕获并分析情绪，乐观积极的交易情绪容易推动股市进一步上涨，而悲观消极的交易情绪容易推动股市进一步下跌，情绪可以用来控制仓位比例，乐观情绪时加大仓位，悲观情绪时减少仓位，从而更好地在金融市场上进行盈利。多种不同源的数据打开了 AI 量化模型的思路，如何有效地处理好各个来源的数据，以及如何综合整体地从各个方面考虑到对股票未来涨跌的影响，是一个非常有意义的研究问题。同时，对于所研究出来的模型，是否能够在牛市和熊市中持续带来稳定的收益，也是一个重要的考量。未来 AI 量化模型的目标必然是以多源数据作为输入，结合市场多方面的影响，来作出对股票未来涨跌走势的判断，随着市场风格改变，模型能够及时地进行学习调整，同时有着较好的鲁棒性，能够持续地在金融市场盈利，才是 AI 量化模型的最终发展目标。

随着 AI 技术在金融量化领域的发展，未来必定有更多的资金是基于算法进行

交易，这也就不可避免地会带来一些金融市场的监管问题。过多的资金处于算法交易容易引起股票市场的一些动乱，个别股票会存在严重的资金操纵行为，从而扰乱其他一些正常的交易，这也就违背了金融市场的初衷——“让金钱流向它最有价值的地方”。一个理想的状态应当是算法交易资金与常规人为投资资金处于一个动态平衡且合适的比例，共同在金融市场上进行生长，促进金融市场的繁荣!

参考文献

[1] DE SANTIS R A. Unobservable systematic risk, economic activity and stock market[J]. Journal of Banking & Finance, 2018, 97: 51-69.

[2] TUDOR C. Investors' Trading Activity and Information Asymmetry: Evidence from the Romanian Stock Market[J]. Risks, 2021, 9(8): 149.

[3] COFFIE E, DUEDAHL S, PROSKE F. Sensitivity Analysis with respect to a Stochastic Stock Price Model with Rough Volatility via a Bismut-Elworthy-Li Formula for Singular SDEs[J]. ArXiv preprint arXiv:2107.06022, 2021.

[4] CHHIMWAL B, BAPAT V, GAURAV S. Investors' preferences and the factors affecting investment in the Indian stock market: an industry view[J]. Managerial Finance, 2020.

[5] WEITZEL U, LAUDI M, SMEETS P. Do Financial Advisors Exploit Responsible Investment Preferences?[J]., 2021.

[6] KAROUI A, PATEL S. What Drives Active Share? Active Stock Selection or Active Stock Weights[J]. Journal of Investment Management (JOIM), 2020.

[7] CHEN X, SCHOLTENS B. The urge to act: A comparison of active and passive socially responsible investment funds in the United States[J]. Corporate Social Responsibility and Environmental Management, 2018, 25(6): 1154-1173.

[8] DAVYDOV D, TIKKANEN J, ÄIJÖ J. Magic Formula vs. traditional value investment strategies in the finnish stock market[J]. Nordic Journal of Business, 2016, 65(3-4): 38-54.

[9] BIRRU J, GOKKAYA S, LIU X. Capital market anomalies and quantitative research[J]. Fisher College of Business Working Paper, 2019(2018-03): 007.

[10] EDWARDS R D, MAGEE J, BASSETTI W C. Technical analysis of stock trends[M]. [S.l.]: CRC press, 2018.

[11] NAZÁRIO R T F, e SILVA J L, SOBREIRO V A, et al. A literature review of technical analysis on stock markets[J]. The Quarterly Review of Economics and Finance, 2017, 66: 115-126.

[12] FU C. Alpha Beta Risk and Stock Returns—A Decomposition Analysis of Idiosyncratic Volatility with Conditional Models[J]. Risks, 2018, 6(4): 124.

[13] SALIM D F, WASPADA I, UTAMA W, et al. Optimal Portfolios With Smart Beta, Alpha, Diversification, And Var On Horizon Indonesia’s Stock Exchange[J]. European Journal of Molecular & Clinical Medicine, 2020, 7(2): 5371-5381.

[14] 赵佳艺. 量化投资发展及我国现状分析[J]. 现代商贸工业, 2019, 000(008): 116-117.

[15] OBTHONG M, TANTISANTIWONG N, JEAMWATTHANACHAI W, et al. A survey on machine learning for stock price prediction: algorithms and techniques[J]., 2020.

[16] SHARMA A, BHURIYA D, SINGH U. Survey of stock market prediction using machine learning approach[C]//2017 international conference of electronics, communication and aerospace technology (ICECA): vol. 2. [S.l. : s.n.], 2017: 506-509.

[17] HU Z, ZHAO Y, KHUSHI M. A survey of forex and stock price prediction using deep learning[J]. Applied System Innovation, 2021, 4(1): 9.

[18] MOSAVI A, BATHLA Y, VARKONYI-KOCZY A. Predicting the future using web knowledge: state of the art survey[C]//International conference on global research and education. [S.l. : s.n.], 2017: 341-349.

[19] CHUN J, AHN J, KIM Y, et al. Using deep learning to develop a stock price prediction model based on individual investor emotions[J]. Journal of Behavioral Finance, 2020: 1-10.

[20] RAY R, KHANDELWAL P, BARANIDHARAN B. A survey on stock market prediction using artificial intelligence techniques[C]//2018 International Conference on Smart Systems and Inventive Technology (ICSSIT). [S.l. : s.n.], 2018: 594-598.

[21] RECCHION C H. Chicago Quantitative Alliance Investment Challenge: Strategy and Reflection[J]., 2017.

[22] FINNEY M A. The challenge of quantitative risk analysis for wildland fire[J]. Forest Ecology and Management, 2005, 211(1-2): 97-108.

[23] ZHAI X, OLIVER A, KOLESNIKOV A, et al. S4l: Self-supervised semi-supervised learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l. : s.n.], 2019: 1476-1485.

[24] TUNG H Y F, TUNG H W, YUMER E, et al. Self-supervised learning of motion capture[J]. ArXiv preprint arXiv:1712.01337, 2017.

[25] CARUANA R. Multitask learning[J]. Machine learning, 1997, 28(1): 41-75.

[26] REI M. Semi-supervised multitask learning for sequence labeling[J]. ArXiv preprint arXiv:1704.07156, 2017.

[27] IVANYUK V. Formulating the concept of an investment strategy adaptable to changes in the market situation[J]. Economies, 2021, 9(3): 95.

[28] IBIDAPO I, ADEBIYI A, OKESOLA O. Soft computing techniques for stock market prediction: A literature survey[J]. Covenant Journal of Informatics & Communication Technology, 2017, 5(2): 1-28.

[29] NG V, ENGLE R F, ROTHSCHILD M. A multi-dynamic-factor model for stock returns[J]. Journal of Econometrics, 1992, 52(1-2): 245-266.

[30] DAI Y, ZHANG Y. Machine learning in stock price trend forecasting[J]. Stanford University Stanford, 2013.

[31] NIKOU M, MANSOURFAR G, BAGHERZADEH J. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms[J]. Intelligent Systems in Accounting, Finance and Management, 2019, 26(4): 164-174.

[32] ZHANG R, LIN Z, CHEN S, et al. Multi-factor Stock Selection Model Based on Kernel Support Vector Machine[J]. J. Math. Res, 2018, 10(9).

[33] YU L, HU X W, GUO K. The Study of the Development of Chinese Stock Market Based on Factor Analysis[J]. Procedia Computer Science, 2015, 55: 422-430.

[34] PETKOVA R. Do the Fama–French factors proxy for innovations in predictive variables?[J]. The Journal of Finance, 2006, 61(2): 581-612.

[35] JORDAN M I, RUMELHART D E. Forward Models: Supervised Learning with a Distal Teacher[J]. Cognitive Science, 2010, 16(3): 307-354.

[36] ZHU X, GHAHRAMANI Z, LAFFERTY J D. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions[C]//Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA. [S.l. : s.n.], 2003.

[37] LI C, BISWAS G. Unsupervised learning with mixed numeric and nominal data[J]. Knowledge & Data Engineering IEEE Transactions on, 2002, 14(4): 673-690.

[38] XIONG Z, LIU X Y, SHAN Z, et al. Practical Deep Reinforcement Learning Approach for Stock Trading[J]. Papers, 2018.

[39] PAHWA N, KHALFAY N, SONI V, et al. Stock prediction using machine learning a review paper[J]. International Journal of Computer Applications, 2017, 163(5): 36-43.

[40] VADLAMUDI S. Stock Market Prediction using Machine Learning: A Systematic Literature Review[J]. American Journal of Trade and Policy, 2017, 4(3): 123-128.

[41] BILBREY JR J K, RILEY N F, SAMS C L. Short-term prediction of exchange traded funds (ETFs) using logistic regression generated client risk profiles[J]. Journal of Finance and Accountancy, 2013, 14: 1.

[42] KARA Y, BOYACIOGLU M A, BAYKAN Ö K. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange[J]. Expert systems with Applications, 2011, 38(5): 5311-5319.

[43] YUN K K, YOON S W, WON D. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process[J]. Expert Systems with Applications, 2021, 186: 115716.

[44] HINTON G, DENG L, YU D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[45] MVFNN: Multi-Vision Fusion Neural Network for Fake News Picture Detection[M]. [S.l.]: Computer Animation, 2020.

[46] YU D, DENG L. Recurrent Neural Networks and Related Models[M]. [S.l. : s.n.], 2015: 237-266.

[47] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]//KDD. [S.l. : s.n.], 2014: 701-710.

[48] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks[C]//KDD. [S.l. : s.n.], 2016: 855-864.

[49] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding[C]//WWW. [S.l. : s.n.], 2015: 1067-1077.

[50] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//ICLR. [S.l. : s.n.], 2017: 1-14.

[51] EAPEN J, BEIN D, VERMA A. Novel deep learning model with CNN and bi-directional LSTM for improved stock market index prediction[C]//2019 IEEE 9th annual computing and communication workshop and conference (CCWC). [S.l. : s.n.], 2019: 0264-0270.

[52] PAWAR K, JALEM R S, TIWARI V. Stock market price prediction using LSTM RNN[G]//Emerging Trends in Expert Applications and Security. [S.l.]: Springer, 2019: 493-503.

[53] FENG F, CHEN H, HE X, et al. Enhancing Stock Movement Prediction with Adversarial Training[J]. IJCAI, 2019.

[54] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in neural information processing systems. [S.l. : s.n.], 2017: 5998-6008.

[55] DING Q, WU S, SUN H, et al. Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction.[C]//IJCAI. [S.l. : s.n.], 2020: 4640-4646.

致　　谢

时间好快，转眼间就即将告别研究生生涯了，回想起来还是挺感慨的。

还记得2018年的那个冬天，保完研后提前一年来到了上海这边，进入了罗轶凤老师的实验室，参与了上海瞰点科技有限公司项目组。感谢罗老师这几年来对于我学习和生活上的关照，帮助我修改论文，虽然由于一些变故没能将我送到毕业，但还是非常感谢罗老师对于我学术上的指导和生活上的关心；也感谢上海瞰点科技有限公司，提供给我一个不错的实习机会，让我接触工业上的实际场景，从事我所热爱的AI量化投资工作，锻炼自己的代码能力，同时培养做事认真负责的态度。

同时，也感谢学院各位老师对我的指导，感谢那些帮我修改过论文的老师们，感谢帮助过我的学长学姐、同学和学弟学妹们，我们在华东师范大学数据学院相遇，本身就是一个很好的缘分了，感恩研究生生涯遇到的每一个人！

其实对于自己的研究生生涯并不满意，将大部分的时间和精力都花费在学习科研上了，最后也并没有取得什么显著的成果，小论文被拒了四次，现在仍然投稿中，可能这就是命吧，我不认命，但，我也认命了，特别感谢陈岑老师在这种关键时刻帮我改论文，希望这一次能有个好结果吧！

对于未来，很幸运自己拿了不少公司的offer，最后各方面权衡下选择签了字节跳动，希望自己不要被研究生生涯的一些挫折影响心态，仍然能够保持积极向上，乐观奋斗的状态，不要让自己，也不要让身边的老师、同学、朋友们失望。未来我想去接触好多新鲜的东西，想学唱歌，想学吉他，想学摄影，想去健身，也想找个女朋友分享生活中的点滴。

我很珍惜现在的生活，是非常非常珍惜，因为我知道一路走来有多么不容易。小时候，在小乡村里，接受着落后甚至是思想不正确的教育，而我只有通过自己的努力一步一步往上爬；小时候，父母吵架、离婚，连父母的亲情有时候对我来说都是一种奢侈；小时候，由于家里穷，甚至被亲戚看不起..... 其实这也正常吧，在落

后的小乡村，这些都再正常不过了。有时候都不知道自己是如何坚持下去的，大概就是想远离那种让我不舒服的环境吧！现在生活比以前好多了，感谢父亲、母亲、爷爷、奶奶和亲戚们的信任、支持和帮助，家庭永远是我最坚强的后盾！

读了将近二十年书，也不知道读没读明白了，即将告别象牙塔，希望自己以后能有个光明的未来吧，感谢遇见的每个人、每件事带给我的成长，感谢每一个在生命中照亮过我的人，也希望自己能够成为一束微光，在别人有需要的时候照亮他们前行，因为淋过雨，所以更想为别人撑伞吧！

天行健，君子以自强不息；地势坤，君子以厚德载物！

应泽林

二零二一年十一月

攻读硕士学位期间科研情况

■ 已申请的专利

申请号：202010240856.1

申请日：2021 年 03 月 31 日

申请人：华东师范大学

发明创造名称：一种基于自监督的股市预测方法

■ 攻读学位期间参加的科研项目

华东师范大学-瞰点科技金融大数据联合实验室，2020-2022.

■ 获得荣誉

[1] 2021 年华为软件精英挑战赛上合赛区冠军

[2] 第十七届“华为杯”中国研究生数学建模大赛三等奖

[3] 2020、2021 年度华东师范大学优秀学生

[4] 2020 年度华东师范大学优秀党员

[5] 华东师范大学数据学院企业奖学金