西南财经大学

Southwestern University of Finance and Economics

课程论文

学年学期：2020-2021 学年第二学期

课程名称：数据科学实战

论文题目：基于数据驱动的量化交易

学生学号：41811025、41827208、41827239

学生姓名：宋佳[1]、谭雷[2]、赵晨曦[2]

学　　院：1 经济信息工程学院 2 统计学院

年级专业：1 2018 级信息管理与信息系统（金融智能与信息管理实验班）

2 2018 级数据科学与大数据技术

评语：

得　分：

评阅教师签字：　　年　月

摘　要

量化交易是指在公开的资本市场上，通过对一系列资产进行有计划的买卖操作，实现规避风险和资产增值目的的投资行为。近几十年来，在国外成熟的金融市场，以数据挖掘为基础的量化投资方法和程序化交易方式已经成为主流。其中量化选股是对股票进行系统分析，量化其相关技术指标后根据相关选股策略选出收益率最高、风险最低的股票组合的行为。

本文主要基于多源异构数据的分析方法，对国内股指数据、研报数据与新闻数据进行研究，最终以 A 股钢铁行业、化学原料行业以及种植业为例，实现了对未来股价的初步预测工作，预测结果准确度相对较高。通过多渠道收集多源异构数据对数据进行预处理。结构化数据部分，进行指标因子的构造，例如 OBV、CCI 等共计 40 个特征，分别对这些特征进行因子有效性检验，均通过测试。非结构化数据部分，对研报数据与新闻数据提取情感值与信息熵，以及构造的文本语句特征等，用来评估行情舆论对股市的影响。然后分别使用 LSTM、CNN 与 CNN-LSTM 模型进行求解，用五天的历史数据来预测第六天的股价或者涨跌情况。

最终我们实现了基于数据驱动的量化交易，对收集数据、处理数据、模型求解、模型评估以及修正等理论方法有进一步的学习理解，创新的实现了 CNN 与 LSTM 模型的融合，模型能够一定程度上进行选股与模拟交易策略，在量化交易方面做出了进一步贡献。

关键词：量化交易，多源异构数据，因子分析，CNN，LSTM，交易策略；

目录

一 绪论

1.1 研究背景及意义

1.1.1 量化交易的产生及意义

量化交易是指在公开的资本市场上，通过对一系列资产进行有计划的买卖操作，实现规避风险和资产增值目的的投资行为。

量化交易利用人们从事相关业务得到的经验，以建模的手段生成相关数学模型，并利用计算机技术从大量历史数据中选取能够获取超额收益的事件，对其加以分析然后制定策略，以减少人为情绪等各种主观因素带来的消极影响，从而达到收益最大化。

近几十年来，在国外成熟的金融市场，以数据挖掘为基础的量化投资方法和程序化交易方式已经成为主流。其中量化选股是对股票进行系统分析，量化其相关技术指标后根据相关选股策略选出收益率最高、风险最低的股票组合的行为。

传统的根据金融技术指标的基本面分析看似全面却不一定准确，对于结构复杂的大数据更是难以快速正确处理。由于人为因素影响，某些信息会被主观的进行放大或者忽略。而量化交易的工作能够处理海量的信息并且不受人为因素的影响，如量化选股模型能够在对海量数据进行分析后，得出最有可能获得较高收益承担较低风险的股票组合。本文对量化方面的研究可以弥补传统投资策略中忽略的影响因素，处理大数据并排除人为因素影响，并基于多元异构数据构建有效因子，提取出新的技术指标，带入模型进行对股票价格进行预测，从而给出投资策略。

1.1.2 多源异构数据的融合

量化投资最大的特点是以概率取胜，即不断地从历史数据中挖掘有望重复的规律并加以利用，即过去和当前的现象可能表明未来活动的发展趋向，所以对历史数据进行有效分析是量化交易的关键技术之一。

根据金融市场传统的有效市场假说理论，股票的价格基本反映目前可用的信息，同时指出股票价格总是由理性的投资者驱动，大致反映出工期预期未来收益

的合理现值。因此，股票价格受新信息的影响很小，遵循随机的变化路径。但是随着信息技术的发展，越来越多的投资者会关注于股票市场相关的信息，并且不断的做出改变，以减的不一致使得股票的实际价格于内在价值产生差异，最终产生价格的波动。

近年来行为金融学领域的实证研究表明，股票走势可以在一定程度上预测。Chan 在研究公司新闻发布对股票的影响中发现股票面对公司负面新闻时表现不佳，会出现比较大的波动，而面对好消息时则表现出比较小的波动。Vega 研究了私人信息和公共新闻对股票的影响，实证结果表明，投资者对资产真正价值的了解越多，对该信息的认同程度越高，交易的异常收益波动越小。这证明了信息对于股票市场是存在影响的。

现阶段研究中股票市场预测主要是依靠三方面的信息：基本面信息、技术指标信息以及网络媒体信息。基本面信息主要包括公司的总体经营情况、财务报告、管理能力以及宏观经济一些指标信息。Cheung 研究了随时间变化公司规模与股票价格波动之间的关系。技术指标信息主要包括反应当天的交易情况数据，例如每日收盘价、最高价和最低价等等。前两种信息是定量信息，即结构化数据，获取相对而言较为容易。本文中使用的国内层面的结构化数据取自于 Wind 与 CSMAR 数据库，包括个股数据，行业数据，上证、深圳板块数据等，国外层面的数据来源于纳斯达克。

由于技术的进步，信息交互方式产生了变化，由单向传播变为了双向多元传播，人们对股票相关信息的态度和看法有了更多的渠道进行表达，网络媒体的重要性越来越大。Li 提出了媒体感知量化交易框架，发现公共情绪会根据公司的特征对股票走势产生不同的影响。Nguyen 提出基于方面的情感分析方法，通过大规模的实验研究了社交媒体对股票走势的影响。此种属于定性信息，即我们能够爬取的非结构化数据，主要是来自于新浪财经的研报和新闻内容，分析人们发表在社交媒体上的对股票的态度，以及分析新闻对公共情绪的影响。

非结构化数据与结构化数据互补，多种来源的信息共同影响股票的走势。

1.1.3 因子构建

国内基金市场上采取的量化选股模型沿用的国外已经使用过的交易策略思想，不少实证研究表明大部分国外的优秀策略中的有效因子在国内表现一般甚至

不能较好的融入中国 A 股的市场环境。在此背景下，基于多源异构数据构建适应我国股市不同行情的有效因子以实现交易策略的本土化，推动金融市场发展很有必要。

本文通过尝试寻找不同市场行情中有效因子的表现规律，进行因子构建。多因子选股模型在国内的有关领域，但并未得到业界公认的适合中国股市的模型和权威的结论。不同研究中确定的有效因子存在较大差异，且不同文献对于因子权重的判定方法不同，常见的多因子选股模型主要采用回归法和打分法两种方式确定因子系数。考虑到不同因子对股票收益率的影响不同，本文采用回归法和打分法分别进行量化交易因子的选择。根据我们爬取到的 2020 年 1 月 1 日至 2020 年 12 月 31 日的多源异构数据，构建不同因子，分别进行因子有效性检测，保留有效因子以在下一步建模工作中使用。

二 量化交易综述

2.1 量化交易概念

量化交易兴起于 1980 年代初，迄今只有短短几十年的历史。这是因为量化交易是基于市场规律和大量历史数据，用计算机编程实现一套全自动化的交易流程，其发展与计算机技术的普及和发展密不可分。

量化交易与传统的投资策略类似，都需要投资者的经验作为指导，而量化交易更偏重基于数据做出决策。通过量化模型来实现全自动交易，避免了投资者的主观意愿影响，降低不理性决策带来的危害。除此之外，在行为金融学领域已有文献证明股价并不是无迹可寻的。根据有效市场假说理论，股价反映了当前市场已知的信息，并受理性的投资者驱动，一定程度上反映了企业预期未来收益的合理现值。随着信息社会的发展，投资者会从各处寻找与市场相关信息，并不断改变自己的意见。此时意见的不一致使得股价的实际价格和内在价值产生差异，从而使股价发生波动。比如一家公司的负面新闻被披露出来时，股价会发生较大的波动，而面对好消息，股价的波动则较小。在投资者挖掘企业披露出来的各类数据时，由于各行业在会计准则上有较大差异，因此财务数据难以运用于一个大型的全自动交易系统上。自然语言处理的技术发展也使得人们开始使用新闻、股吧评论等文本数据。财经新闻等报告不仅包含了市场现状，还蕴含着造成其现状的潜在原因。与公司财务方面的报道也决定着其对投资者的吸引力。而人们无法亲自阅读每一条新闻报告，因此利用基于计算机的量化交易技术可以帮助人们了解真实全面的市场情况，从而做出理性的判断。

量化交易的大致流程如下：

Figure 1 量化交易流程示意图

第一步是数据收集过程。数据可以从外部来源或数据供应商中收集，一般分为结构化数据和非结构化数据。第二步，为了获得规范的可输入模型的数据形态，采用各种数据清洗和预处理方式来处理数据。第三步，建模过程的作用主要是建立能够准确预测股价的模型、并进行统计分析和优化模型。第四步，将分析结果可视化，成为投资决策的标准。最后两个阶段，建模和分析通常用于评估趋势、确定策略、回溯测试和评估投资组合表现的迭代过程。

在整个量化交易的过程中，最核心的部分是第三步：建立模型预测股价。股价的预测对制定交易策略以获得可观的收益起着重要的作用，预测结果是构建和优化投资组合的先决条件。然而股票市场的不确定性使得股价预测成为了一项机遇与挑战并存的任务。在接下来的部分，我们将重点介绍股价预测的国内外研究现状。

2.2 量化交易国内外研究现状

量化交易最早可追溯到 19 世纪初的法国的数学家 Louis Bachelier(1900)，发表了第一篇关于期权定价的学位论文，名为 “Théorie de la spéculation (投机交易理论)”，它是公认意义上的现代金融学的里程碑。随后，巴菲特的老师、享有“华尔街教父”美誉之称的 Benjamin Graham(1934)，在《Security analysis》(证券分析)一书中通过理论分析，提出通过研究发掘公司情况来决定是否购买股票。随后到了 1952 年，Harry Markowitz(1952)[3] 所著的《portfolio theory》（投资组合理论），真正意义上提出了量化交易第一个理论模型——均值-方差模型，提出了以风险为核心的投资量化策略。这标志着现代资产组合理论的开端，也标志着量

化交易的开端。

后来，金融学家在此基础上不断完善和优化，提出了不同的理论假说用来预测股票价格走势，较为著名的有资本资产定价模型(CAMP)、有效市场假说（EMH）、随机游走理论(Random Walk Theory)和行为金融学（BF）等，这些理论是金融资产定价和预测股票市场波动的代表理论。

19 世纪 60 年代，William Sharpe，John Lintner 和 Jan Mossin 等人共同建立和完善了资本资产定价模型。

资本资产定价模型的基本公式为：

$$E(r_i) = r_f + \beta_{im}(E(r_m) - r_f)$$

其中，$E(r_i)$表示资产的预期回报率，r_f表示无风险利率，β_{im}表示衡量系统性风险的变量，$E(r_m) - r_f$表示市场风险溢价。该模型揭示了投资有效组合的收益和风险之间的均衡关系，认为投资组合的收益率仅与该组合超过市场投资组合收益的大小有关，而敏感度用 β 衡量。

有效市场假说是经济学家 Eugene Fama(1970)[4]提出的。有效市场假说认为，当前市场本身包含了关于它的所有信息，如果任何新的信息被收集，它就会被市场吸收，并反映在它的价格上。它间接得出的结论是，投资者无法通过任何其他手段预测股票市场。同时，有效市场假说理论进一步将有效市场分为弱有效市场、半强有效市场和强有效市场。

M.F.M Osborne(1956) 受到布朗运动启发，提出随机漫步模型。随机漫步模型指出，未来的股票价格不能由历史和当前的股票价格预测，因为它是高度波动的，彼此独立，没有任何规律可循。随机漫步模型为量化交易的发展奠定了基础。而上述理论仅立足于金融市场，没有结合实际过程中人的行为对市场带来的影响，不足以解释人们的风险决策行为。

Daniel Kahneman(1979) 建立了人类进行金融决策的心理学理论，把心理学和金融学相结合，奠定了行为金融学的基础。行为金融学揭示了金融市场的非理性决策行为及其内在规律，认为股票价格不仅由企业的基本信息和内在价值决定，而且在很大程度上也受到投资者主体行为的影响，即在金融市场中，投资者的心理和行为对价格决策及其走势有重要影响。

但这些模型不总是有效的，其理论不仅受股票市场波动影响较大，而且时间有效性也较短。例如 Douglas Hodgson 在该模型应用研究发现，资本资产定价模

型对 1969 年之后的数据失效，试验结果表明β不显著，需要有新的模型对其进行更新与修正。后来 Stephen Ross(1976)提出了套利定价理论(APT)，在原先的资本资产定价模型基础上提出了一种新的资产定价模型，认为收益率需要由多个因子共同确定与解释。多因子决策模型标志着量化交易理论进入全新领域。

19 世纪 80 年代以来，量化交易均由多因子模型展开，其中最经典的是 Fama-French 三因子模型和五因子模型。

Fama 和 French(1997)所提出的 Fama-French 三因子模型的公式如下：

$$E(R_{it})-R_{ft}=\beta_i[E(R_{mt})-R_{ft}]+s_iE(SMB_t)+h_iE(HMI_t)$$

其中 R_{ft} 表示时间 t 的无风险收益率; R_{mt} 表示时间 t 的市场收益率; R_{it} 表示资产 i 在时间 t 的收益率; $E(R_{mt})-R_{ft}$ 是市场风险溢价， SMB_t 为时间 t 的市值因子的模拟组合收益率， HMI_t 为时间 t 的账面市值比因子的模拟组合收益率。 β_i，s_i 和 h_i 分别是三个因子的系数。

其回归模型如下所示：

$$R_{it}=\alpha_i+R_{ft}+\beta_i[E(R_{mt})-R_{ft}]+s_iE(SMB_t)+h_iE(HMI_t)+\varepsilon_i$$

其中 ε_i 为回归模型随机噪音项。三因子模型将股票价格的变动和个股的收益解释为市场资产组合因子、市值因子和账面市值比因子共同作用所得到的，模型对股市收益与风险因素进行了有效解释，并且研究表明其能够有效代替其它的风险因子的作用。

后来，很多金融学者对 Fama-French 三因子模型进行实证分析，发现有些股票的 α 显著不为 0，这说明三因子模型中的三个因子并不能解释所有超额收益，其中所包含的其他未被解释的因子不能被忽略。基于此，Fama 和 French(2015) [10] 又提出的 Fama-French 五因子模型，其公式如下：

$$E(R_{it})-R_{ft}=\beta_i[E(R_{mt})-R_{ft}]+s_iE(SMB_t)+h_iE(HMI_t)+r_iE(RMW_t)+c_iE(CMA_t)$$

其中三因子部分保持不变， RMW_t 为时间 t 的盈利水平因子的模拟组合收益率， CMA_t 为时间 t 的投资水平因子的模拟组合收益率。 r_i 和 c_i 是新增的两个因子的系数。

其回归模型如下所示：

$$R_{it} = \alpha_i + R_{ft} + \beta_i[E(R_{mt}) - R_{ft}] + s_i E(SMB_t) + h_i E(HMI_t) + r_i E(RMW_t) + c_i E(CMA_t) + \varepsilon_i$$

其中ε_i为回归模型随机噪音项。Fama-French 五因子模型对 Fama-French 三因子模型中不能很好解释的盈利水平风险和投资水平风险进行了补充解释，使得模型更加完善。

至此，量化交易基本理论体系基本完善，后续研究均由上述理论为基础，结合多因子模型与机器学习、深度学习等前沿方法展开。对于此，我们基于金融市场理论，研究探讨现阶段国内外基于数据驱动的量化交易模型的理论研究与目前进展。

国外文献方面，Ruchira Ray，Prakhar Khandelwal 与 B. Baranidharan(2018)研究了机器学习技术在股票市场预测中的实现。文献表明，支持向量机和人工神经网络是目前股票市场预测应用最广泛的机器学习技术。支持向量机的低方差行为可以较为精确捕捉股票市场的行为；由于股票市场具有高度波动和非线性模式，人工神经网络可以提高股票价格预测的准确性。Weiwei Jiang(2020)对不同的数据源、各种神经网络结构和常用的评价指标进行了分类，而且还对其实现和再现性进行了研究，总结归纳了深度学习在股票市场预测中应用的最新进展。Van-Dai Ta , CHUAN-MING Liu 和 Direselign Addis Tadesse(2020)提出了一种基于历史数据的 LSTM 来预测股票波动，基于 LSTM 预测模型构建的投资组合优于线性回归、支持向量机等机器学习方法构建的投资组合预测模型。Gourav Kumar，Sanjeev Jain 和 Uday Pratap Singh(2020)对基于计算智能方法的股票市场预测的现有文献进行了最新综述，对前处理、降维和预测未来趋势或预测未来股价的文章进行了研究和讨论，认为计算智能方法在股票市场领域已显示出良好的效果。Yang Liu, Qi Liu, Hongke Zhao,Zhen Pan, Chuanren Liu(2020)我们提出了针对量化交易问题的自适应的 iRDPG 交易模型，采用了模仿学习技术来平衡对贸易代理的探索和开发，模型的准确度高，市场适应能力强。Chunchun Chen，Pu Zhang，Yuan Liu 和 Jun Liu(2020)使用卷积神经网络(CNN)技术对金融投资进行量化，通过深度学习技术进行量化投资研究，预测股票市场的涨跌准确度较高。RuCheng, Qing Li(2021)提出了基于属性驱动的图注意网络，来捕获股票的动量溢出效应，并基于张量的融合模块来捕获不同信号之间的相互作用，该模型对具有特征交互

作用的市场信息空间进行建模可以进一步提高对股票的预测。Yujie Fang, Juan Chen 和 Zhengxuan Xue(2019)采用随机森林模型、LSTM 模型和 SVR 模型预测 50 支 ETF 的价格，实验结果表明，基于深度学习的量化投资策略比传统的量化投资策略具有更高的收益率，收益率曲线更稳定，抗跌落性能更好。

国内文献方面，潘莉(2011)等人基于因子模型，研究 A 股回报率的规律,探索构建适用于 A 股市场的因子模型，其研究发现股票市值、市盈率对回报率的影响显著,杠杆率对回报率的影响前期较强,近期减弱，其余因素无显著影响。孔翔宇(2016)等人深度挖掘了财经新闻主题内容与股市市场的相关性，结合自动文本分析技术与机器学习技术，提出了一种基于理解当日新闻主题分布来分析中国股市涨跌的预测模型，其实验结果表明算法能较准确得预测当日股市涨跌，而建立在其上的投资策略也取得了很好的效果。耿立校(2021)等人提出了一种基于多源异构数据的 LSTM 模型，通过对融合资本市场交易数据、技术指标数据、投资者情绪三种源数据的量化来预测股票指数的走势，其实际研究结果显示 LSTM 模型的预测准确率比传统模型更为优秀，数据源的增加也对模型准确率的提升有较大贡献，验证了多源异构数据分析方法的可行性和有效性。黄创霞(2020)等人运用情感分析技术，在情感倾向点互信息算法的基础上，引入“拉普拉斯修正”和“情绪分类阈值”，提出了一种改进的个体投资者情绪度量的情感倾向点互信息算法，研究表明：改进情感倾向点互信息算法的情感识别精度更高，并且算法发现积极情绪是股票收益率的格兰杰原因。同时，当投资者处于积极状态时，会热衷于使用表情符号表达情绪，这有助于投资者更好的利用网络论坛信息进行投资决策。赵一鸣(2021)等人提出了一种新型的基于多知识图谱构建中文文本语义图的方法，实现了实体层面和概念层面两个层次的中文文本语义化表示，可应用于新闻类文本分类、文本分析等自然语言处理任务。王成龙(2021)等人研究发现，投资者情绪可以影响到股票价格，将量化投资与投资者情绪结合起来，根据能反映投资者情绪的指标设计量化投资策略，发现该策略能够获得有显著收益差异的股票投资组合，从而获得高于市场平均水平的利润。马长峰(2020)等人总结了应用文本分析研究金融问题的一般步骤，理清了应用文本分析进行会计和金融研究的脉络，并构建全新指标评估文本分析。杨妥(2020)等人采用 LSTM 模型对新闻内容信息进行情感分析，再将分析得到的情感分类结果与股票的技术指标相结合作为特征，利用 BP 神经网络进行预测，最终得到与一般模型相比有明显提升的

预测效果。高雅(2020)等人引入注意力机制的思想。通过改进 TF-IDF 算法和 Bi-LSTM 神经网络构建注意力模型，筛选出对情感极性相对更有影响的子文本，通过子文本进行文本情感极性分类判断，实验证实该模型比一般的深度学习模型效果更好。唐振鹏(2020)等人在极值理论中引入行为金融学，结合标值自激发点过程刻画股指收益率极端值序列的集聚性、短期相依性，并将传统的超阈值模型所描述的齐次泊松过程拓展为非齐次泊松过程，探讨投资者情绪对极端收益率的冲击。实证结果表明，沪深股市在短期内股指连续暴跌现象时有发生，投资者极度负面情绪会加剧股市的剧烈动荡。贺康(2020)等人以 2007～2017 年沪深两市 A 股上市公司年报为研究样本，实证检验年报文本信息复杂性对资产误定价的影响，并考察其作用机制。其研究发现，年报文本信息复杂性与资产误定价之间显著正相关，同时当机构投资者持股比例越高、新闻媒体报道越多时，年报文本信息复杂性对资产误定价的影响会被削弱。

综上所述，国内外量化交易均基于多因子模型展开，通常结合机器学习与深度学习方法，预测股市价格走向，并均取得较好结果；此外，对于影响股市的非客观因素，也有较为全面、详尽的分析，例如公司年报、研报与新闻，以及行为金融学分支下研究的投资者情绪影响。这为我们后续开展多源异构数据分析来预测股市价格提供了可靠的理论支撑，证实了从股指数据、文本数据与情感数据等多源数据来进行量化交易模型构建的可行性与有效性。

2.3 量化交易数据平台

我们的研究对象为全部 A 股，时间范围为 2020 年 1 月 1 日至 2020 年 12 月 30 日。

所收集的结构化数据为每日每只股票的开盘价、收盘价、最高价、最低价与交易量。数据来源为国泰安数据库。因子数据来源于同花顺 Mind Go 量化交易平台。

非结构化数据主要来源于新浪财经的股票研报。新浪财经平台将整个股票市场分成了 104 个行业，部分行业代码及名称示例如下：

Table 1　部分行业代码及名称示例

sw2_110100	**种植业**	**sw2_210300**	**其他采掘**
sw2_110200	渔业	sw2_210400	采掘服务
sw2_110300	林业	sw2_220100	石油化工
sw2_110400	饲料	sw2_220200	化学原料
sw2_110500	农产品加工	sw2_220300	化学制品
sw2_110600	农业综合	sw2_220400	化学纤维
sw2_110700	畜禽养殖	sw2_220500	塑料
sw2_110800	动物保健	sw2_220600	橡胶
sw2_210100	石油开采	sw2_230100	钢铁
sw2_210200	煤炭开采	sw2_240200	金属非金属新材料

我们利用 scrapy 框架编写网页爬虫程序，分别对 104 个行业进行爬取，可以获取到每个行业下所有股票自上市以来的所有相关研报，且能获得有关这个行业的宏观研报。最终所有数据存在 MongoDB 中。爬虫程序的具体逻辑如下：

1. 爬取所有行业的名称及行业代码，用于生成初始链接；
2. 进入初始链接（通常为某个行业研报网站的第一页）；
3. 获取这个网页中每条研报的标题等信息，并点击进入单条研报爬取正文；
4. 根据所获得的数据通过正则表达式等处理生成不同的数据项；

3. 根据研报的分类存入不同的数据表。

后续对数据进行空值处理、去重等操作后，获得了两张数据表，分别存着针对个股的研报与针对行业宏观环境的研报。其中个股研报 company 表中总数据量为 310855 条，有以下字段：

Table 2 个股研报字段名及含义

字段名	含义
cate	研报的分类，如“公司”、“创业板”
content	研报的正文内容
date	研报发布日期
ind	所属行业
ins	研究机构
person	研究员
stock_name	股票名称
stock_num	股票代码
title	研报标题

行业研报 industry 表中总数据量为 74630 条，有以下字段：

Table 3 行业研报字段名及含义

字段名	含义
cate	研报的分类，如“策略”、“行业”、“晨报”
content	研报的正文内容
date	研报发布日期
ind	所属行业
ins	研究机构
person	研究员

title	研报标题

2.4 量化交易研究方法

2.4.1 文本挖掘技术

由于我们所收集的数据都不带标注，且网上暂无已标注的股票研报数据库，因此我们无法使用机器学习或深度学习的模型来训练得到情感值或其他特征。在阅读大量有关文本量化的文献后，我们提出了这样的非结构化数据量化手段：

$$\underbrace{\text{情感}\quad \pm \quad a \quad \times \quad b}_{\text{研报的真实作用}} \quad \text{可读性} \times \text{信息价值}$$

整个研报的真实作用由以上公式来衡量。其中，研报中所展现出来的情感作为正负号，若为积极或中性的情感，则为正，若为消极情感，则为负。可读性反映了研报的信息能否顺利传达给读者，因此作为一个范围为 0 至 1 的系数乘在研报真实的信息价值前。以上三个方面我们都通过自然语言处理技术挖掘出一些文字上的浅层特征来衡量。

情感分析 基于金融情感词典来计算文本的情感值。

可读性 从词、句、段、语法树四个层面构建了如下特征，全面衡量一篇研报的可读性。

Table 4 文本浅层特征

	平均字笔画数	法树	语法分析树的高度之和
	至少四个字组成的词数		高度不低于 16 的语法分析树的个数
	单字与词组比例		语法分析树的节点的个数之和
	超过九笔字比例		语法分析树的名词短语的个数之和
	名词比例		语法分析树的动词短语的个数之和
	连接词与虚词比例		语法分析树的形容词短语的个数之和

	词频统计
	平均句长
	非文本类信息比例

	语法分析树的平均高度
	高度不低于 16 的语法分析树的比例
	语法分析树的平均节点数
	单词的平均节点数
	语法分析树的平均名词短语个数
	语法分析树的平均动词短语个数
	语法分析树的平均形容词短语个数

信息价值 用信息熵作为研报的信息价值。

任何信息都存在冗余，冗余大小与信息中每个符号（数字、字母或单词）的出现概率或者说不确定性有关，并且香农借鉴了热力学的概念，把信息中排除了冗余后的平均信息量称为“信息熵”，对于任意一个随机变量 X，它的信息熵定义

$$H(X) = -\sum_{x \epsilon X} P(x) log P(x)$$

如下：

无监督的新闻评估方法可以从新闻内信息的冗余程度来评估，我们可以认为文本冗余度越高，新闻的质量越差。反之，新闻的冗余度越低，新闻的质量越好。所以我们可以用文本的熵作为文本质量的评估标准。具体代码逻辑如下：

1.文本预处理，去除特殊字符等；

2.使用 jieba 分词对文本进行分词并统计每个单词的个数；

3.计算每个词出现的概率，也就是上述公式中的 P(x)；

4.利用上述公式计算文本的熵。

2.4.2 深度学习模型——基于注意力机制的 LSTM

深度神经网络自 xx 年被引入股价预测中。传统的 BP 神经网络的输入输出只能为一条数据，而股票市场上的数据多为时间序列数据，数据之间的时序性无法被 BP 神经网络捕捉学习到。递归神经网络不同于传统的 BP 神经网络的结构只在层与层之间建立了连接，它在同一层的不同神经元之间也建立了连接，这样的结构使得 RNN 可以处理序列变化的数据。

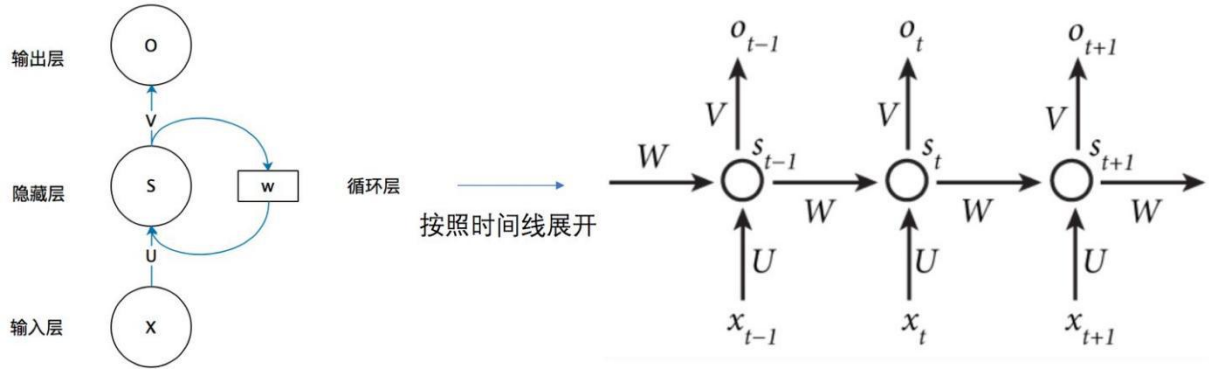


Figure 2 RNN 模型原理示意图

在优化 RNN 的过程中，由于时间序列的长期依赖情况，一些参数在优化的过程中是一系列小于 1 的数字相乘，使得出现梯度消失的现象。除此之外，若网络参数的值太大，也可能是一系列大于 1 的数字相乘从而出现梯度爆炸。解决这个问题的办法就是 LSTM。

LSTM（Long Short-Term Memory，长短期记忆网络）在 1997 年首先被 Sepp Hochreiter 和 Jurgen Schmidhuber 提出，它在 RNN 的基础上在每个神经元内加入了三个门来控制信息的流入、存储和流出。

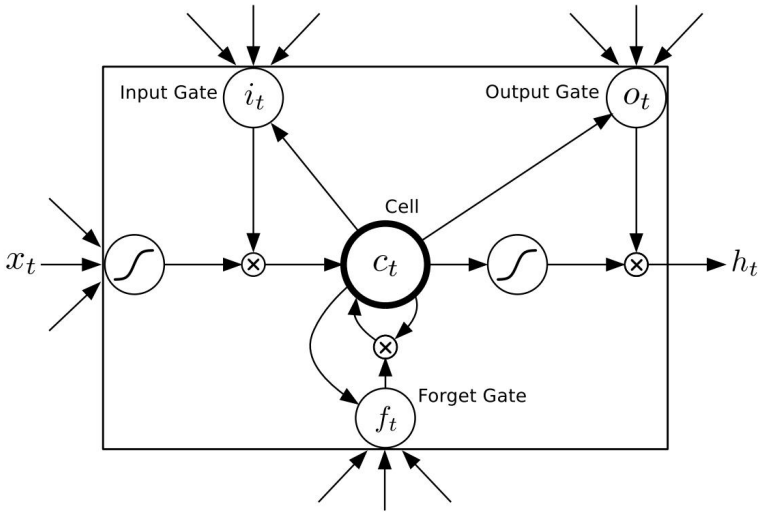


Figure 3 LSTM 原理示意图

一个神经元包括三个门：输入门、遗忘门和输出门。其中遗忘门确定模型丢弃了哪些单元状态信息，它接受前一个时间步 ht−1 的输出和当前时间步新的输入。它的主要功能是记录从前一个单元状态到当前单元状态所保留的信息量。它将输出一个介于 0 到 1 之间的值，其中 0 表示完全保留，1 表示完全放弃。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

输入门决定了当前时间步有多少输入保留到新的细胞状态 Ct，它避免了将不重要的信息输入到当前存储细胞。它有三个不同的组件:（1）获取必须更新的单元格状态;（2）创建一个新的细胞状态;（3）将细胞状态更新为当前细胞状态。

$$i_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_i)$$

$$\hat{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t$$

输出门控制新创建的细胞状态将被丢弃多少，输出信息首先由一个 sigmoid 层确定，然后由 tanh 处理新创建的细胞状态，再加上 sigmoid 输出来确定最终输出。

$$O_t = \sigma(W_\sigma \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = O_t * tanh(C_t)$$

三 量化交易因子构建

3.1 结构化数据处理

3.1.1 数据描述性统计分析

通过三个行业的股市 K 线数据，能够初步得知行业在近三年内的股价波动情况以及行业整体变化，能够对后续工作建立初步探索与认识。对于基础 K 线数据，三个行业的基本数据描述性统计分析结果如下：

Table 5 钢铁行业基本数据描述性统计结果

	open	high	low	close	preclose	volume	amount	turn	pctChg
mean	6.240355	6.344876	6.146907	6.246490	6.245810	3.463070e+07	1.776242e+08	1.132182	-0.000138
std	4.514298	4.615145	4.431133	4.527743	4.516018	6.019549e+07	3.083422e+08	1.261160	2.276565
min	1.040000	1.070000	1.040000	1.050000	1.050000	0.000000e+00	0.000000e+00	0.001956	-10.112360
25%	3.310000	3.350000	3.270000	3.310000	3.310000	6.401910e+06	3.285440e+07	0.370005	-1.094681
50%	4.840000	4.910000	4.770000	4.840000	4.840000	1.839977e+07	7.869681e+07	0.706868	0.000000
75%	7.700000	7.870000	7.560000	7.720000	7.720000	4.095853e+07	1.974350e+08	1.408768	1.023113
max	50.500000	54.220000	49.530000	54.220000	50.550000	1.811060e+09	7.482788e+09	15.098500	10.169500

Table 6 化学原料基本数据描述性统计结果

	open	high	low	close	preclose	volume	amount	turn	pctChg
mean	9.042592	9.227777	8.875259	9.050525	9.051959	1.711293e+07	1.149292e+08	2.008151	0.002091
std	7.580783	7.786596	7.402381	7.592042	7.589264	4.001627e+07	2.915607e+08	2.748831	2.706357
min	1.310000	1.320000	1.300000	1.310000	1.310000	0.000000e+00	0.000000e+00	0.007800	-13.403500
25%	4.530000	4.600000	4.450000	4.540000	4.540000	3.187226e+06	2.196533e+07	0.610623	-1.321596
50%	6.900000	7.000000	6.790000	6.900000	6.900000	7.016596e+06	4.995892e+07	1.113090	0.000000
75%	10.690000	10.930000	10.510000	10.730000	10.740000	1.682434e+07	1.156659e+08	2.238024	1.245403
max	58.790000	60.680000	57.320000	58.440000	58.440000	1.126629e+09	1.202904e+10	52.687932	19.988800

Table 7 种植业基本数据描述性统计结果

	open	high	low	close	preclose	volume	amount	turn	pctChg
mean	8.973875	9.193097	8.806625	9.006014	9.001554	2.271825e+07	2.056679e+08	3.744096	0.082051
std	4.952246	5.107602	4.842262	4.979712	4.973248	2.932984e+07	3.071057e+08	4.905663	3.148524
min	2.030000	2.060000	1.930000	2.060000	2.060000	0.000000e+00	0.000000e+00	0.037871	-11.692300
25%	5.010000	5.110000	4.910000	5.020000	5.012500	5.410970e+06	3.339385e+07	0.893700	-1.472951
50%	8.120000	8.290000	7.960000	8.140000	8.135000	1.262963e+07	9.549523e+07	1.951132	0.000000
75%	11.490000	11.770000	11.247500	11.510000	11.510000	2.820829e+07	2.474215e+08	4.508800	1.377892
max	36.510000	36.510000	36.510000	36.510000	36.510000	3.485447e+08	4.234999e+09	47.751965	15.197400

此外，从其它数据部分可以得知，A 股共计钢铁行业个股 34 支，化学原

料个股 33 支，种植业个股 19 支。

根据结果可以初步看出，化学原料的平均股价与最大股价均最高，但标准差较大，相对而言钢铁行业平均股价较低，但总体差异相对较小。三个行业均有成交量为 0 的个股，存在影响模型的可能，在后续处理中应当删除或对其进行特殊处理。在成交量与成交总额规模上来看，三个行业比较接近，钢铁行业的成交规模相对较大，种植业的成交规模差距相对较大。换手率来看，种植业换手率相对较高，钢铁行业则相对较低。

总的来说，三个行业整体市场表现相近，交易相对活跃，具有较好的研究价值。

3.1.2 结构化数据可视化

对于三个行业的相对占比，绘制饼状图如下所示：

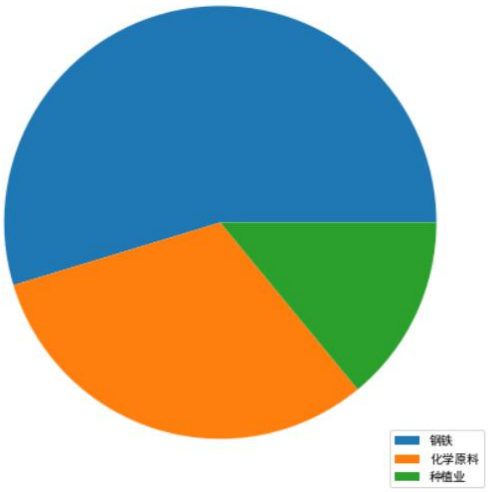


Figure 4 三行业成交总额相对占比饼状图

其中最上面为钢铁，左下角为化学原料，右下角为种植业。

通过饼状图可以看出，钢铁行业的成交总额占比最大，化学原料其次，种植业最少，这与各个行业的相对股票个数有关。去除股票个数影响后，相对占比如下所示：

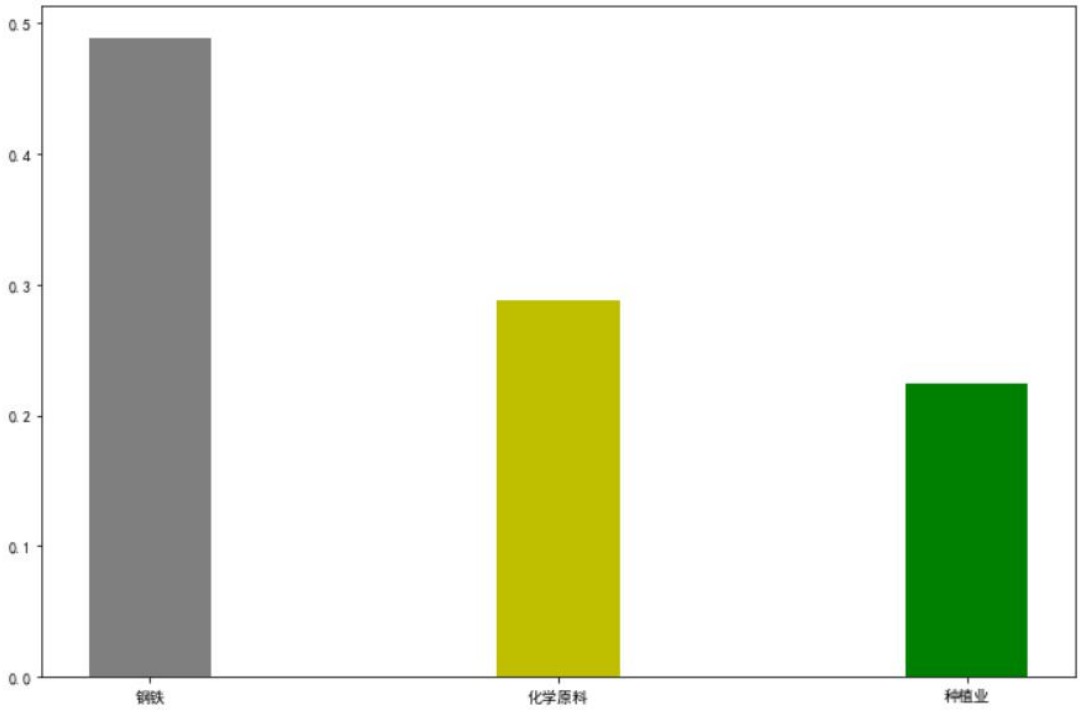


Figure 5 三行业成交总额相对占比饼状图

可以看出，钢铁行业整体平均占比也最大，化学原料其次，种植业最少。在研究过程中，可以分别认为钢铁行业代表规模相对较大的行业股市模型，化学原料行业代表规模相对中等的行业股市模型，种植业代表规模相对较小的行业股市模型。

研究三个行业的日 K 线数据，首先对整体进行研究，绘制 K 线图如下所示：

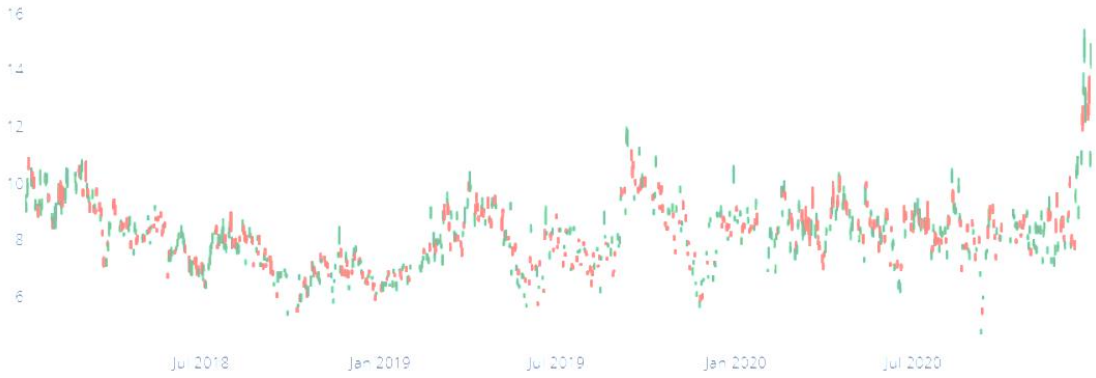


Figure 6 钢铁行业 K 线图

Figure 7 化学原料 K 线图

Figure 8 种植业 K 线图

为了对比整体行业情况，特别给出 A 股整体趋势图，如下所示

Figure 9 A 股整体趋势图

初步可以看出，行业股票走势与总体走势有一定部分的相似性，这初步验证了我们先验认为总体指标能够影响行业股票走势的假设。

在股市研究中，股市技术指标为研究员提供了良好的依据，研究员可以通过多种类型的技术指标，判断股票未来可能的趋势，做出买入或卖出的判断。

对于所采用的股市技术指标，绘制其在三年内的趋势如下所示：

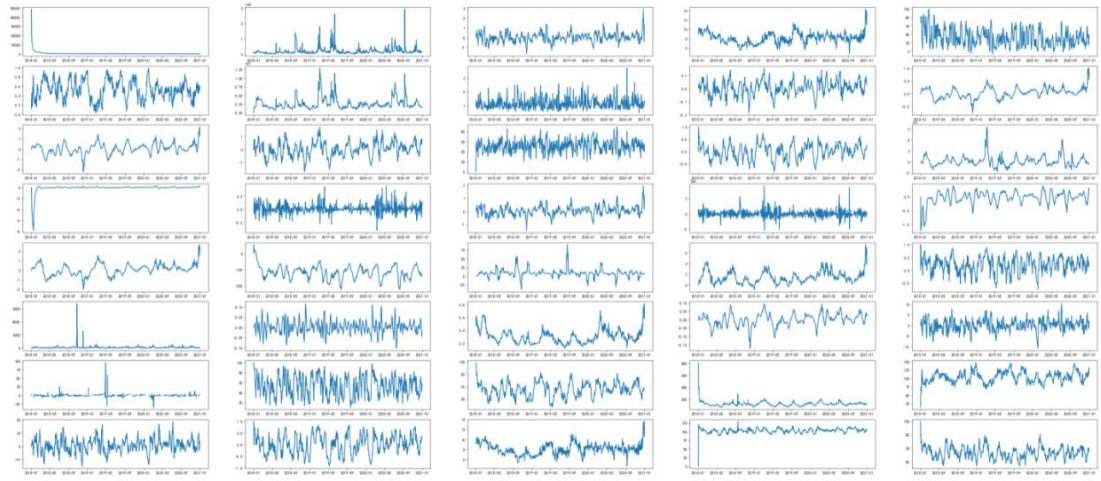


Figure 10 钢铁行业股市技术指标

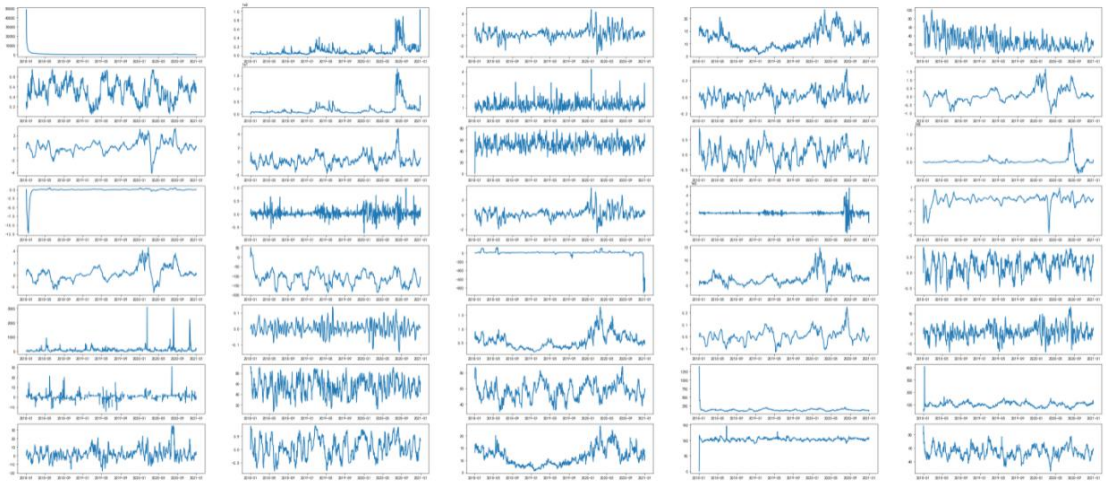


Figure 11 化学行业股市技术指标

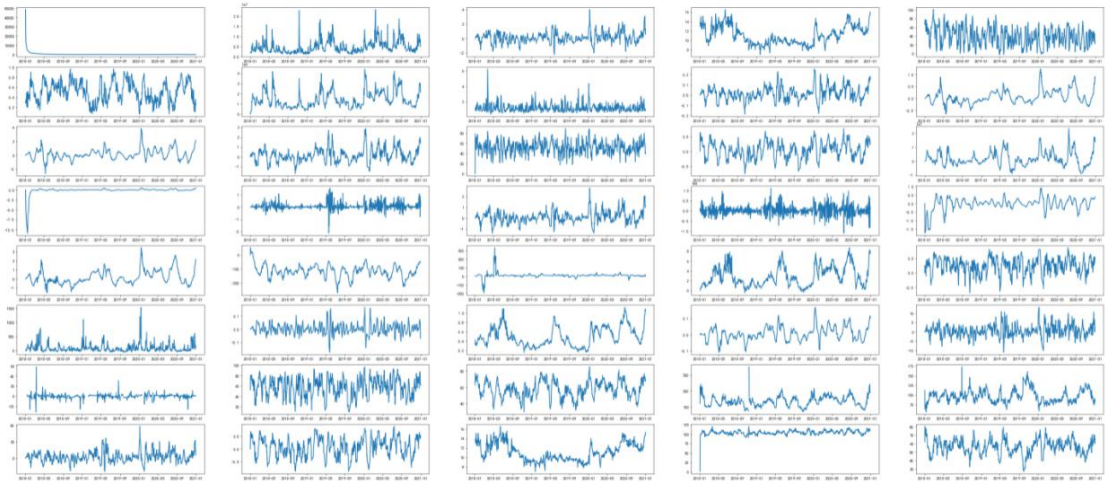


Figure 12 种植业股市技术指标

对于完整的股市技术指标，在因子构造部分会进行详细描述。大体可以看出，各个指标趋势均呈现波动态势，对比而言能够捕捉股市中的波动信息。对

于这些因子，需要进行进一步的因子有效性检测，对检测结果进行分析，筛选出较好结果的因子进行后续分析。

3.2 非结构化数据处理

3.2.1 非结构化数据描述性统计分析

按价值分析的思路，对非结构化数据进行因子构造如下图所示

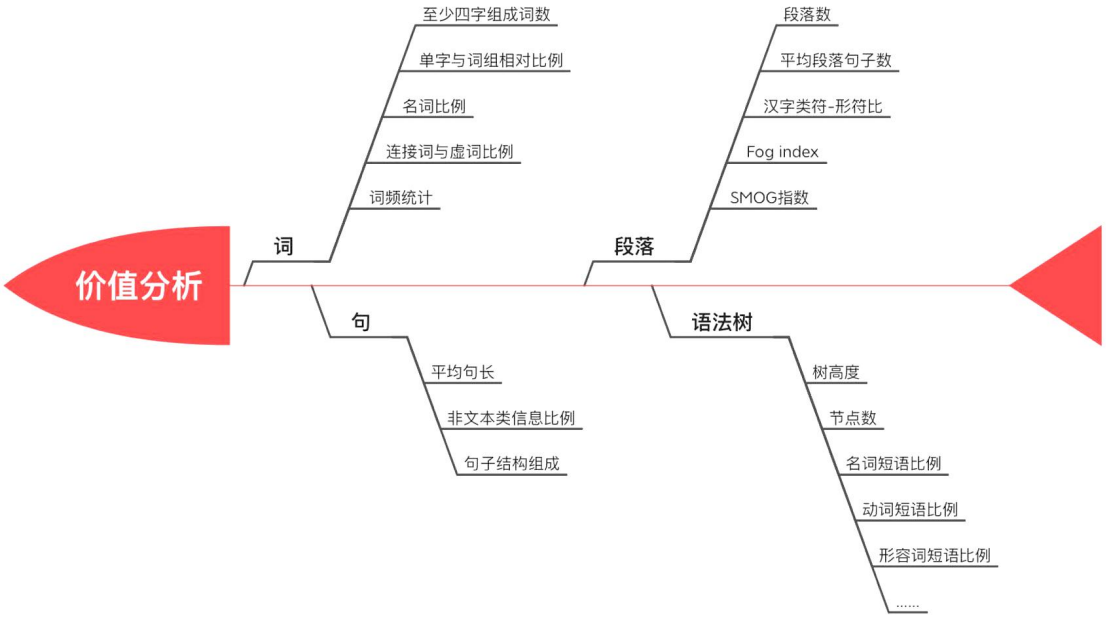


Figure 13 非结构化数据构造因子鱼骨图

最终得到的因子，通过因子有效性检测后，保留 16 个有效因子，其中对于因子名称进行给定并解释如下：

Table 8 非结构化数据构造因子名称释义表

非结构化数据因子名称	因子释义
avg_sentence	平均句长
avg_stroke	平均笔画数
four_word	至少四字组成词数
function_ratio	连接词数与虚词数
n_ratio	名词比例
nine_stroke	至少九笔字数
no_text	非文本信息比例

sen_per_para	平均段落句子数
sum_para	段落数
word_max	每句词最大频率
word_mean	每句词平均频率
word_phrase	单字与词组比例
word_var	每句词频率方差
quality_val	信息熵
pos_num	积极情感词数
neg_num	消极情感词数

对于上述因子，进行描述性统计分析结果如下表所示

Table 9 非结构化数据描述性统计分析

	mean	std	min	25%	50%	75%	max
avg_sentence	86.514857	151.633947	0.0	0.0	0.0	109.598684	1173.116909
avg_stroke	13.322164	23.643381	0.0	0.0	0.0	15.220779	164.857515
four_word	0.048907	0.086600	0.0	0.0	0.0	0.060025	0.566168
function_ratio	0.121677	0.209245	0.0	0.0	0.0	0.161241	1.499095
n_ratio	0.420929	0.735430	0.0	0.0	0.0	0.517241	5.194628
nine_stroke	0.422189	0.747420	0.0	0.0	0.0	0.506550	5.059347
no_text	0.463204	0.862659	0.0	0.0	0.0	0.535610	6.425982
sen_per_para	6.167862	11.397189	0.0	0.0	0.0	7.812500	81.866392
sum_para	17.269497	30.277005	0.0	0.0	0.0	23.000000	237.000000
word_max	23.516884	40.971631	0.0	0.0	0.0	32.000000	303.000000
word_mean	2.994463	5.301544	0.0	0.0	0.0	3.743323	36.900841
word_phrase	0.456381	0.844753	0.0	0.0	0.0	0.517165	6.547497
word_var	5.144481	9.332853	0.0	0.0	0.0	6.670065	71.370224
quality_val	0.917431	2.372050	0.0	0.0	0.0	0.000000	14.804044
pos_num	3.664431	10.229447	0.0	0.0	0.0	0.000000	97.000000
neg_num	1.418429	4.002717	0.0	0.0	0.0	0.000000	34.000000

非结构化数据的最大特点之一就是其稀疏性。从表格数据可以看出，几乎所有的非结构化数据都存在超过 50%以上的缺失情况，这是行业研究报告与个

股研究报告的稀疏性所导致的。在这种情况下，数据处理提供三种解决方案：

（1）不进行处理，通过模型自我识别特征，保留数据真实性；

（2）对数据进行高斯平滑处理，降低其稀疏性，保证数据有效性；

（3）对数据进行基于信息熵权重的记忆累加，即对于某个时刻的研报因子数据，在第二天不会归零，而是继续保留，但会衰减。衰减速率基于信息熵大小确定——信息熵越大，代表非结构化数据越有价值，衰减速率也越小，保证数据记忆性。

经过仔细对比讨论，结合实际数据集，我们选择第一种方案。对于本次数据，我们考虑通过给予模型多源异构数据进行量化交易分析，对于非结构化数据部分，起到一个信息增强作用。对于某个时间点，非结构化数据对个股的股价变化趋势有增强作用。另外，所使用的深度学习 LSTM 模型与 CNN 模型能够识别稀疏信息的特征，已经保证了数据的有效性；同时模型评估过程需要基于真实的数据来推导结论，故在该数据处理中，考虑方案一。另外两种方案也存在一定合理性，保留进一步探讨空间。

3.2.2 非结构化数据可视化

对于非结构化数据所得到的因子，进行可视化处理，通过图像数据进行进一步探索性分析。

字平均笔画数而言，通过论文研究表明，文本内平均笔画数越大，一般代表着该文本越难懂，并且笔画数一定程度上服从 Γ 分布。其直方图与核密度图如下所示

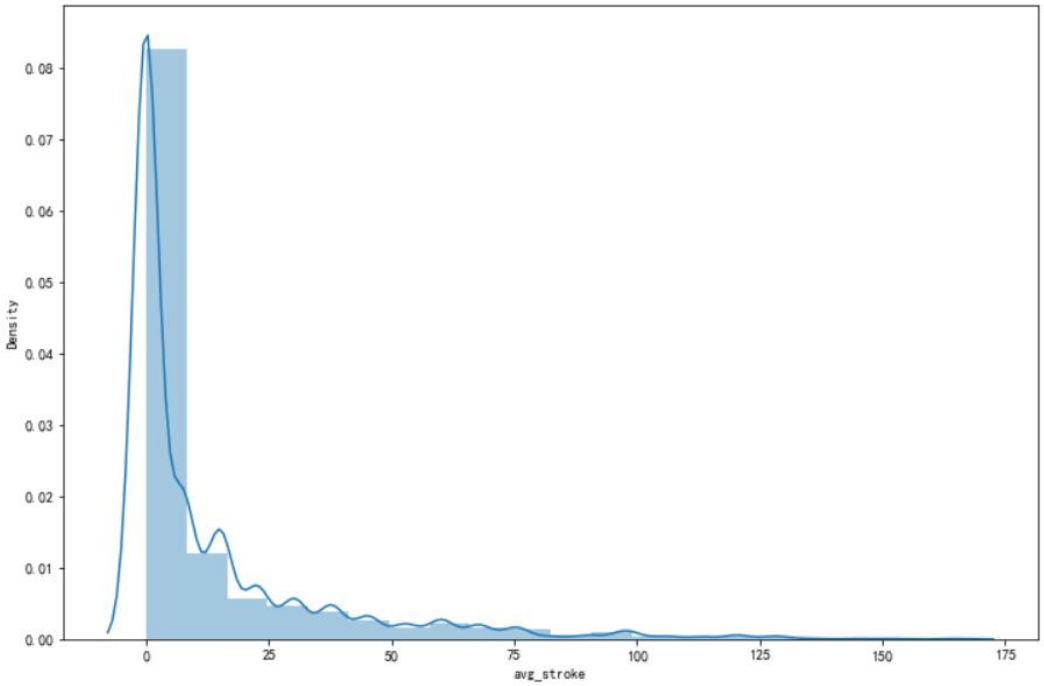


Figure 14 字平均笔画数直方图与核密度图

从图中我们可以看出，除去缺失部分数据，字平均笔画数呈现明显右拖尾现象，对比一般情形的分布，我们可以发现，相较于常规文章而言，股市行业研报相对更加复杂，阅读难度较高，解析难度也相对较高，需要进行更加细致、准确的处理分析。

对于其词频统计，考虑其去除停用词后的均值。在均值分布中，如果平均词频越大，说明词重复次数越多，对于一篇文章，过多的重复与过少的重复都代表该行业研报的质量不高。对于三个行业的研报内容，分析其词频并计算词频均值，绘制直方图与核密度图如下图所示：

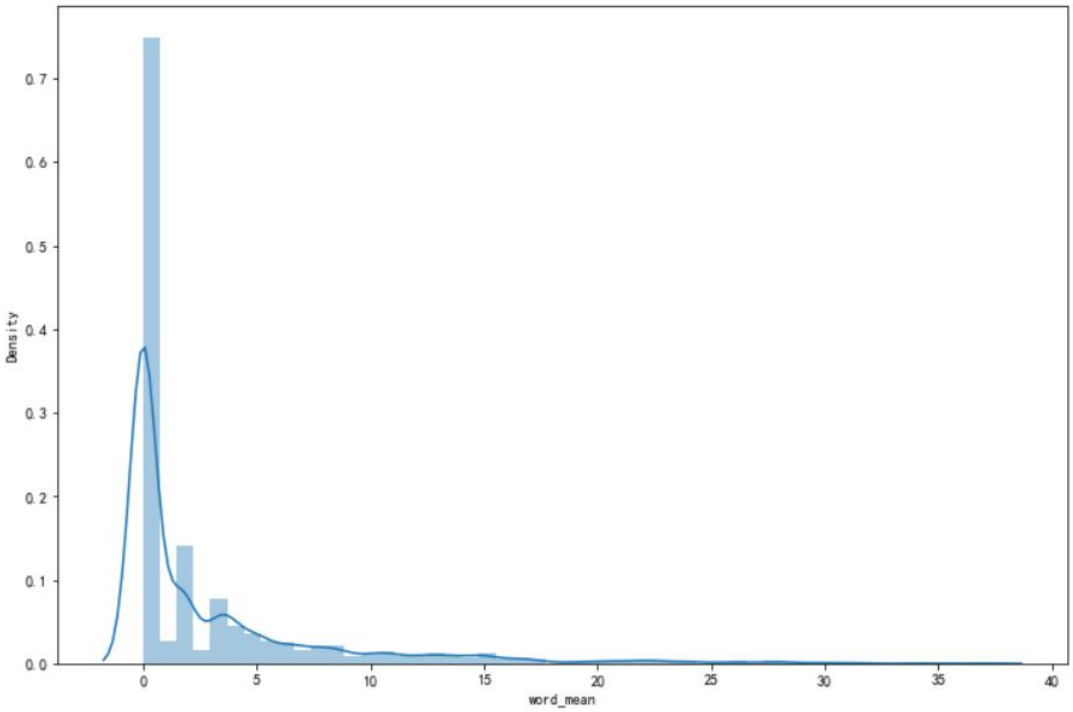


Figure 15 词频均值直方图与核密度图

此外，情感与信息熵是我们所构造的最为重要的两个非结构化数据因子。信息熵能够反映一篇文章所包含内容信息程度，侧面反映一篇研报的重要程度；另一方面，信息熵也能够反映一篇研报内容的冗余程度，过高的信息熵可能会带来相反的效果，即代表研报内容冗余、存在干货不足的可能，在生活中。经过计算，信息熵分布的直方图与核密度图如下所示：

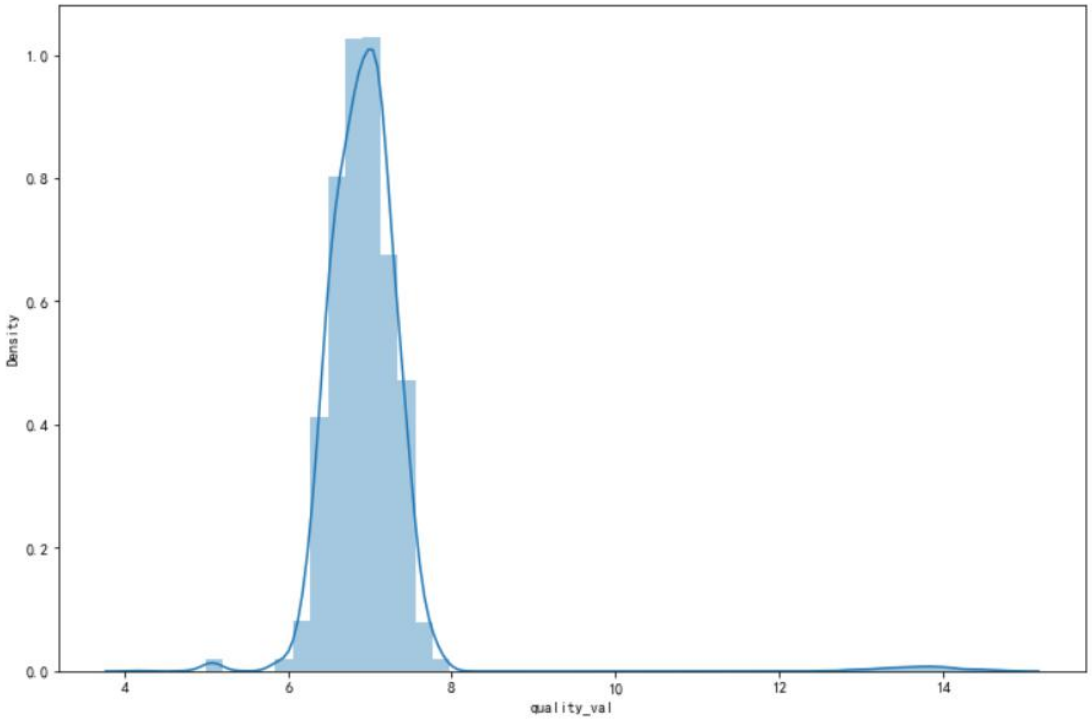


Figure 16 信息熵直方图与核密度图

而情感分析部分，通常分析文本情感是基于词典对照或者词袋模型-神经网络模型进行学习，此处我们选择金融行业研报的情感词词典，对研报文本的积极情感词与消极情感词进行统计，最终统计结果呈现整体平均占比如下所示：

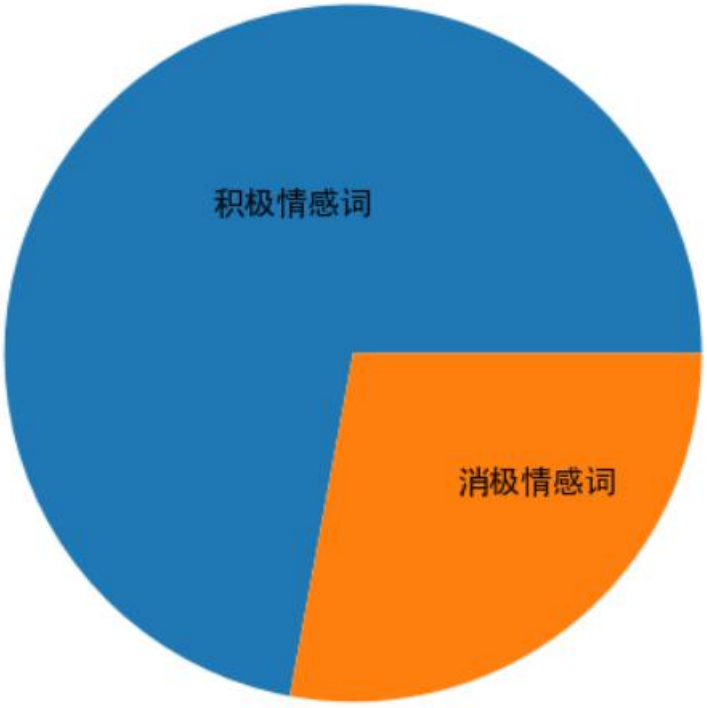


Figure 17 积极情感词与消极情感词相对占比

从图中我们可以看出，股市行业研报以积极情感词为主，消极情感词相对较少。相关论文研究表明，股市相关文章内容存在严重的春秋笔法，即报喜不报忧，通常情况下认为积极情感词占文章主导为正常现象。

3.3 数据预处理

数据预处理的整体流程如下所示：

Figure 18 数据预处理整体流程图

具体流程如下所示：

（1）导入数据

分别导入"钢铁","化学原料","种植业"三个行业 2018-2020 的股市数据、股市技术指标数据，对数据进行合并处理，以"code"为区分个股指标，以"date"与"code"唯一识别样本。对于整体数据，除"date"与"code"外 •，其余均为数值数据。

（2）合并五日数据

按照时间顺序，对连续五天的因子数据进行合并，将数据整理为一行。最终所得到的数据因子相对于原先多了五倍，确保样本包含历史股市信息。

（3）数据缺失处理

对于三个行业的多源异构数据，主要存在两种缺失情况。

第一种是因子计算过程中，由于 K 线数据特殊性导致因子超出阈值或者异常取 0，这种情况考虑进行多项式插值填补。

第二种情况是在非结构化数据中，存在某个时间段的研报缺失，这种情况在数据描述性统计中进行了详细的探讨，最终采用不处理的方式。

（4）数据转换

对于多源异构数据，需要对不同来源的数据进行数据转换，使得数据的框架保持一致。数据主要由日 K 线数据、结构化数据因子构造、非结构化数据因子构造三部分构成，对于数据部分，按照”date”进行拼接，对数据的类型进行对应转换，保证数据的类型与结构正确。

（5）特征构建

本部分仅针对机器学习数据而言，对于数据进行特征构建。特征构建考虑进行特征组合，特征组合此处使用暴力交叉的构造方法，包括一阶的加减乘除，以及二阶的正态变化、平方和、平方差和绝对平方根处理。特征组合内容如下所示：

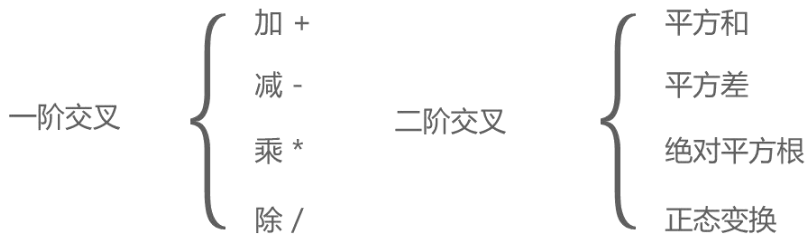


Figure 19 特征组合示意图

其中对于除法，为了防止除零错误，每个特征量作为除数时都需要加一个极小量，此处基于数据考虑取

$$\delta = 10^{-5}$$

最终对于数据样本得到 6730 个特征，后续会进行特征筛选。

（6）标准化

对于数据特征，为了使模型拟合过程中损失函数在各个方向上下降速度一致，首先进行归一化处理。此处采用的计算公式如下所示

$$\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

经过归一化处理后，数据整体落在[0,1]范围内。

（7）特征筛选

特征并不是越多越好。特征多会对模型的求解带来困难。并且在实际中，不是所有的特征都对于目标是有帮助的，或者有些特征包含的信息被其他特征覆盖，即产生了冗余。特征选择可以增强模型的泛化能力，减少过拟合。如果使用回归、SVM 这种需要遍历的算法，本身的运算量很大，计算时间也较长，特征太多会会使其计算过于复杂。于是进行特征选择是十分必要的。

此次模型数据的结构化数据因子构造部分，我们通过因子回测的方法，进行因子有效性检测，选择出其中有效的 40 个因子，而对于非结构化数据因子与机器学习特征组合出的新特征，我们主要采用了过滤式和嵌入法进行特征筛选。

Filter（过滤式）是先设计一个过滤方法进行特征选择，再去训练学习器。而这个过滤方式是设计一个“相关统计量”，对特征进行计算，最后设定一个阈值去选择，这个“相关统计量”可以选择相关系数、方差、MIC 值等。过滤法的主要目的是：在维持算法表现的前提下，帮助算法们降低计算成本。对于此数据，考虑采用方差过滤法进行简单的特征筛选，优点是计算简便、速度快。

Embedded（嵌入式）方法可以一边进行模型训练，一边完成特征选择。我们先使用某些机器学习的算法和模型进行训练，得到各个特征的权重系数，根据权值系数从大到小选择特征，这些权值系数往往代表了特征对于模型的某种贡献或某种重要性，比如决策树和树的集成模型中的 feature_importances_属性，可以列出各个特征对树建立的贡献。我们可以基于这种贡献的评估，找出对模型建立最有用的特征，因此相对于过滤法，嵌入法的结果会更加精确到模型的效用本身，对提高模型效力有更好的效果，并且，由于考虑特征对模型的贡献，因此无关的特征和无区分度的特征都会因为缺乏对模型的贡献而被删除掉。一般常用的嵌入式模型有正则化模型与树模型，这里考虑采用 lightGBM 的 GBDT 树集成模型进行特征重要程度比较，其结果如下：

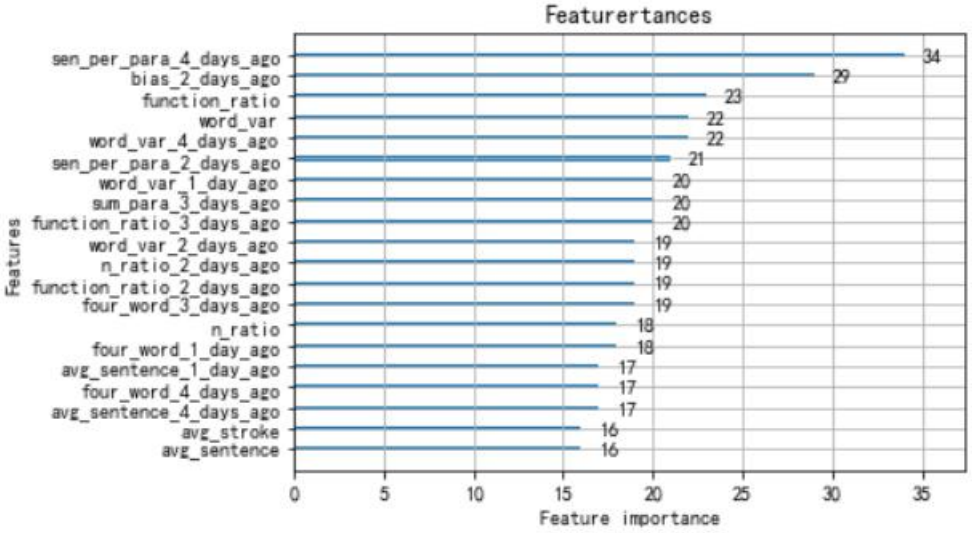


Figure 20 特征重要程度

可以看出，在机器学习分类模型中，非结构化数据因子部分占有非常重要的重要程度。随着训练次数增加与模型调参调优，非结构化数据因子表现也非常优异，对于模型拟合提供了重要信息，可以认为非结构化数据因子在机器学习模型中是相对有效的。

最终通过过滤式与嵌入式筛选，在机器学习部分保留了 2566 个特征。

3.4 量化交易因子构建

符号设定如下：

符号	意义
open	开盘价
high	最高价
low	最低价
close	收盘价
preclose	前收盘价
volume	成交量
amount	成交金额
turn	换手率=[指定交易日的成交量(股)/指定交易日的股票的流通股总股数(股)]*
pctChg	日涨跌幅=[(指定交易日的收盘价–指定交易日前收盘价)/指定交易日前收盘
TP	(最高价+最低价+收盘价)/3
MA	近*N*日收盘价的累计之和/N
MD	近N日（MA-收盘价）的绝对值的累计之和/N
C	当日收盘价
CN	N日前的收盘价
D	当日加权指数
HF	移动平均线HF=(open+close+high+low)/4

根据在同花顺平台对指标的有效性检测的结果，按照指标的有效性从高到低排列，选择有效性高的指标介绍如下：

Table 10 **指标介绍**

指标	名称	公式	含义
CCI	顺	$\mathrm{CCI}=\frac{1}{0.015}\frac{TP-MA}{MD}$	测量股价是否已超出常

	势指标		态分布范围。
TAPI	指数点成交值	$TAPI = AMOUNT + D$	探讨每日成交量值与指数间的关系。
MTM	动量指标	$\mathrm{MTM}=C - CN$	研究股价波动的中短期技术分析工具。
VMA	变异平均线	$VMA = HF$的M日简单移动平均线	每日的开盘价、收盘价、最高价和最低价得到的数据计算平均线。
KDJ	随机指标	K值、D值和J值点位连接形成	股票分析的统计体系。
OBV	能量潮	$OBV = \frac{(close - low) - (high - close)}{(high - low)}$	通过统计成交量变动的趋势来推测股价趋势。

1. CCI（顺势指标）

①指标含义

CCI 指标专门测量股价、外汇或者贵金属交易是否已超出常态分布范围。属于超买超卖类指标中较特殊的一种。CCI 指标是根据统计学原理，引进价格与固定期间的股价平均区间的偏离程度的概念，强调股价平均绝对偏差在股市技术分析中的重要性，是一种比较独特的技术指标。

$$\mathrm{CCI} = \frac{1}{0.015}\frac{TP - MA}{MD}$$

②运用原理：

CCI 指标的运行区间也分为三类：+100 以上为超买区，—100 以下为超卖区，+100 到—100 之间为震荡区：

a. 当 CCI 指标曲线从上向下突破+100 线而重新进入常态区间时，表明市场价格的上涨阶段可能结束，将进入一个比较长时间的震荡整理阶段，应及时平多做空。

b. CCI 指标曲线从下向上突破-100 线而重新进入常态区间时，表明市场价格

的探底阶段可能结束，有可能进入一个盘整阶段，可以逢低少量做多。

c. 当 CCI 指标曲线从上向下突破-100 线而进入另一个非常态区间（超卖区）时，表明市场价格的弱势状态已经形成，将进入一个比较长的寻底过程，可以持有空单等待更高利润。

d. CCI 指标曲线从下向上突破+100 线而进入非常态区间(超买区)时，表明市场价格已经脱离常态而进入强势状态，如果伴随较大的市场交投，应及时介入成功率将很大。

2. TAPI（指数点成交值）

① 指标含义

TAPI 指标是根据股票的每日成交值与指数间的关系，来反映股市买气的强弱程度及未来股价展望的技术指标。需要注意的是，TAPI 指标必须与其它指标结合研判，不能单独作用。

$$TAPI = AMOUNT + D$$

② 运用原理

a. 上涨过程，在股价的明显转折处，若 TAPI 值异常缩小，是向下反转讯号，应逢高卖出；连续下跌中，在股价明显转折处，若 TAPI 值异常放大，是向上反转讯号，可逢低买进。
b. 发生背离现象。即指数上涨，TAPI 值下降，此为卖出讯号，可逢高卖出；反之，为买进信号。

3. MTM（动量指标）

① 指标含义

动量指标从股票的恒速原理出发，考察股价的涨跌速度，以股价涨跌速度的变化分析股价趋势的指标。动量指数以分析股价波动的速度为目的，研究股价在波动过程中各种加速，减速，惯性作用以及股价由静到动或由动转静的现象。动量指数的理论基础是价格和供需量的关系，股价的涨幅随着时间，必须日渐缩小，变化的速度力量慢慢减缓，行情则可反转。反之，下跌亦然。

$$\mathrm{MTM}=C-CN$$

② 运用原理

a. MTM 由上向下跌破中心线时为卖出时机，相反，MTM 由下向上突破中心线时为买进时机。
b. 若股价与 MTM 在低位同步上升，显示短期将有反弹行情；若股价与 MTM 在高位同步下降，则显示短期可能出现股价回落。

4. VMA(变异平均线)

① 指标含义

变异平均线则是用每日的开盘价、收盘价、最高价和最低价相加后除以 4 得

出的数据计算平均线。

② 运用原理

a. 股价高于 VMA，视为强势；股价低于 VMA，视为弱势。

b. VMA 向上涨升，具有助涨力道；VMA 向下跌降，具有助跌力道。

5. KDJ（随机指标）

① 指标含义

随机指标 KDJ 是以最高价、最低价及收盘价为基本数据进行计算，得出的 K 值、D 值和 J 值分别在指标的坐标上形成的一个点，连接无数个这样的点位，就形成一个完整的、能反映价格波动趋势的 KDJ 指标。它主要是利用价格波动的真实波幅来反映价格走势的强弱和超买超卖现象，在价格尚未上升或下降之前发出买卖信号的一种技术工具。

要选择周期（n 日、n 周等），再计算当天的未成熟随机值（即 RSV 值），然后再计算 K 值、D 值、J 值等。

RSV 的计算公式：

$$RSV = \frac{C - L_n}{H_n - L_n} \times 100$$

其中，L_n 为之前 n 日内的最低价，H_n 为之前 n 日之内的最高价。

计算 K_i：

$$K_i = \frac{2}{3} K_{i-1} + \frac{1}{3} RSV_i$$

K_i, RSV_i 分别表示某一天当天的 K 值和 RSV 值。

计算 D_i：

$$D_i = \frac{2}{3} D_{i-1} + \frac{1}{3} K_i$$

D_i, K_i 分别表示当天的 D 值和 K 值。

计算 J 值：

$$J_i = 3K_i - 2D_i$$

② 运用原理

a. K 与 D 值永远介于 0 到 100 之间。D 大于 80 时，行情呈现超买现象。D

小于 20 时，行情呈现超卖现象。

b. 上涨趋势中，K 值大于 D 值，K 线向上突破 D 线时，为买进信号。下跌趋势中，K 值小于 D 值，K 线向下跌破 D 线时，为卖出信号。

5. OBV(能量潮)

①指标含义

能量潮是将成交量数量化，制成趋势线，配合股价趋势线，从价格的变动及成交量的增减关系，推测市场气氛。其主要理论基础是市场价格的变化必须有成交量的配合，股价的波动与成交量的扩大或萎缩有密切的关连。通常股价上升所需的成交量总是较大；下跌时，则成交量可能放大，也可能较小。价格升降而成交量不相应升降，则市场价格的变动难以为继。

$$OBV = [(close - low) - (high - close)] \div (high - low) \times volume$$

②运用原理

a. 当股价上升而 OBV 线下降，表示买盘无力，股价可能会回跌。

b. 股价下降时而 OBV 线上升，表示买盘旺盛，逢低接手强股，股价可能会止跌回升。

c. OBV 线缓慢上升，表示买气逐渐加强，为买进信号。

d. OBV 线急速上升时，表示力量将用尽为卖出信号。

四 深度学习模型量化交易

4.1 深度学习模型原理

股票价格收到众多因素的共同影响，是一个极为复杂的动力学系统，具有非线性、非平稳性、低信噪比以及长记忆性等特点。

深度学习是一种基于多层神经网络的机器学习方法，其由神经网络输出层和输入层以及两者之间的一系列堆叠的隐含层所构成，通过逐层的传递提炼出学习对象高度抽象、复杂的特征，并以此特征矩阵作为数据的表现形式，获得输入数据与输出数据之间高度复杂的非线性函数，从而最终提升分类或预测的准确性，具有处理高纬度、非线性、非平稳的结构数据上的优势。

许多研究结果已经表明深度学习对于金融时间序列数据能够取得更好的预测效果，并将预测结果与传统方法进行了对比。Fischer 和 Krauss（2018）通过标准普尔 500 指数对比发现，循环神经网络中的长短时记忆模型对于金融时间序列数据具备更好的适应性，比随机森林、前馈神经网络等模型的预测准确率更高，涨跌准确率可以达到 54.3%。杨青和王晨蔚（2019）将长短时记忆模型应用于全球 30 种股票指数，发现其在不同的预测期限上均较支持向量机、多层感知机以及 ARIMA 模型有更高的平均预测精度，且预测结果更为稳定。欧阳红兵等（2020）将小波分析与长短时记忆模型相结合，对道琼斯工业指数日收盘价进行预测，发现其预测精度比多层感知机、支持向量机、K 近邻以及 GARCH 四种模型更高。

深度学习在金融时间序列数据中的应用优势可以归纳为以下三点：

- 不受维度限制，可以将相关数据都作为输入数据纳入模型之中；
- 具备更好的非线性拟合能力，更适应于金融时间序列数据的自身特点；
- 能够有效减少过拟合以及陷入局部最优解的问题（Heaton et al.，2016）。

4.1.1 卷积神经网络预测模型

卷积神经网络（CNN）是一种具有局部连接、权重共享等特性的深层前馈神经网络。卷积神经网络是受生物学上感受野的机制而提出的，克服了全连接前馈网络参数太多和无法提取局部不变性特征的问题。目前的卷积神经网络一般由卷积层（Convolutional layer）、池化层(Pooling layer)和全连接层(Fully connected layer)

交叉堆叠而成的前馈神经网络，使用反向传播算法进行训练。

卷积神经网络具有三个结构上的特性：局部连接（局部感受野）、权重共享和汇聚（池化）。局部感受野可以提取局部、初级的特性；权值共享可以使网络拥有更少的自由参数，在降低网络模型复杂度的同时具有减少过拟合、提高泛化能力的优点；池化层可以减少特征的维度。实现对位移、缩放和扭曲的不变性。

在卷积神经网络中，卷积层和池化层的交替作用能够从大量的数据中挖掘出具有区分度的深层次特征。全连接层前接最后一个池化层，用于整合交替卷积、池化所提取的特征，输出层用于输出最终的结果。

卷积神经网络具有强大的特征提取和识别能力，在图像数据和时间序列数据的分类任务中得到了成功的应用。本文中我们首先使用 CNN 模型来构建股票价格的预测模型。传统卷积神经网络的输入数据的行数和列数通常相等。而我们处理得到的股票时序数据显然难以满足此条件，为了适应股票时序数据的形式，参照 Lee 等(2017)[1]的研究，对传统 CNN 的结构进行了调整。

1. 卷积层

卷积层输出的结果由多个特征面（特征映射 Feature Map）构成，特征面中的每一个取值都代表一个神经元，且每个神经元都与上一层特征面中 M*5 的区域相连接。特征面与卷积核一一对应，且特征面中每个神经元的取值都通过对应卷积核计算得到。

第一个卷积层：

$$y_{k,j}^{(1)} = f(\sum_{s=1}^{N}\sum_{i=1}^{M} w_{k,s,t}^{(1)} y_{j+t-1}^{(0)} + b_k^{(1)})$$

$y_{k,j}^{(1)}$ 为第一个卷积层中第 k 个特征面的第 j 个神经元的输出值；f 为 ReLU 激活函数；$w_{k,s,t}^{(1)}$ 是第 k 个卷积核中第 s 行第 t 列的权值；$y^{(0)}$ 为输入数据，$b_k^{(1)}$ 为第 k 个卷积核对应的偏置值。

2. 池化层

每个卷积层后都连接池化层，在不增加训练参数的前提下对卷积层输出的特征面进行进一步的降维，进行特征选择，降低特征数量，以减少网络的参数，提高模型的鲁棒性。

卷积层虽然可以显著减少网络中连接的数量，但特征映射组中的神经元数并

没有显著减少。如果后面接一个分类器，分类器的输入维数依然很高，很容易出现过拟合。在卷积层之后加上一个汇聚层，从而降低特征维数，避免过拟合。

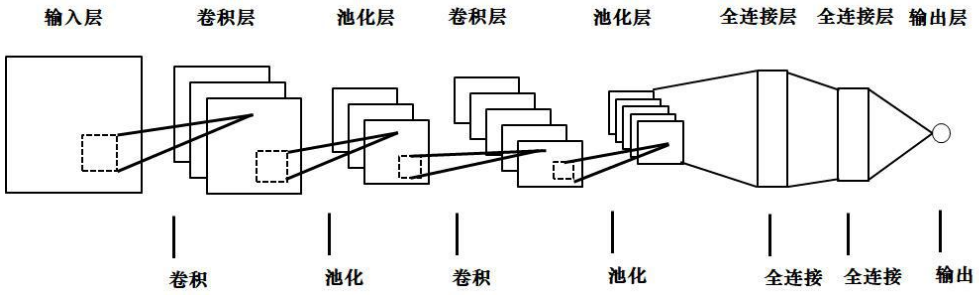
在池化层中将输入的特征面划分成多个区域，可以重叠，也可以不重叠。池化就是对每个区域进行下采样（Down Sampling）得到一个值作为这个区域的概括。进行池化运算，池化层输出的每一个压缩特征面都对应于上一层的卷积层所输出的一个特征面。此处我们选用常用的最大值池化（Maximum Pooling）进行运算，即提取池化阈中元素的最大值组成压缩特征面：

$$Y_{m,n}^{d} = \max_{i \in R_{m,n}^{d}} x_i$$

3. 全连接层

经过反复的卷积、池化后，全连接层将最后一个池化层输出的所有压缩特征面串联成特征向量，并输入到全连接层中。全连接层中的每一个神经元于前一层的神经元进行全连接，能够对卷积层和池化层提取的特征进行整合，从而获得更有区分度的特征。

由于我们做的是预测模型，所以对处理分类问题的 CNN 模型进行了改进，将最后一个全连接层前的激活函数取消，变为直接进行线性连接，得到股票价格预测值。



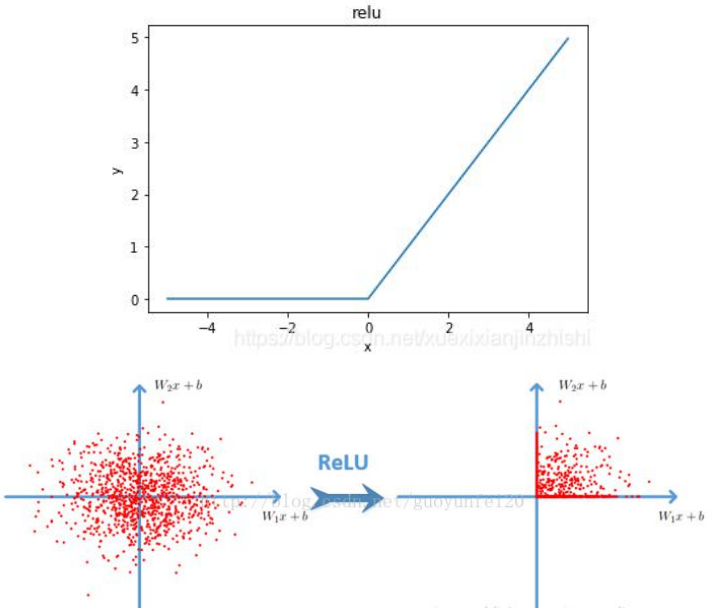
4. 激活函数

激活函数在神经元中非常重要。在 CNN 中，卷积层、全连接层后都有激活函数的身影。如果不使用激活函数，每一个网络层的输出都是一种线性输出，而我们所处的真实场景，更多的是各种非线性的分布。常用的激活函数有 Sigmoid 型函数、ReLU 函数、ELU、SoftPlus 函数等。

这里我们选用 ReLU 函数，是一个斜坡函数：

$$\mathrm{Re}LU = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

ReLU 函数具有以下优点：采用 ReLUctant 的神经元只需要进行加、乘和比较的操作，计算上更加高效；ReLU 函数被认为有生物上的解释性，比如单侧抑制、宽兴奋边界；具有很好的稀疏性，Sigmoid 函数会导致一个非稀疏的神经网络，而 ReLUc 函数大约只有 50%的神经元会处于激活状态；在优化方面，Sigmoid 函数两端饱和，而 ReLU 函数为左饱和，在 x>0 时导数为 1，在一定程度上缓解了神经网络的梯度消失问题，加速梯度下降的收敛速度。

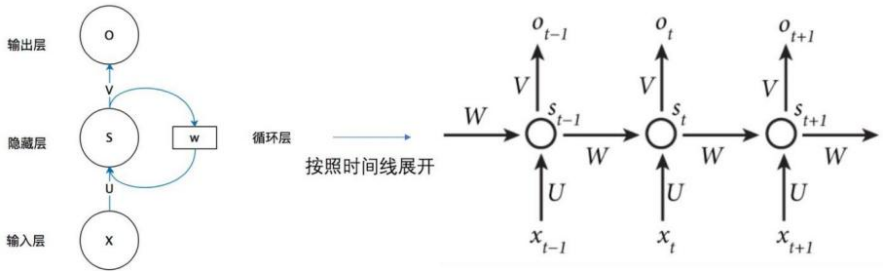


5. 深度学习框架

在深度学习中，一般通过误差反向传播算法来进行参数学习，此外，深度学习模型需要的计算机资源比较多，一般需要在 CPU 和 GPU 之间不断进行切换，开发难度较大。所以一些支持自动梯度计算、无缝 CPU 和 GPU 切换等功能的深度学习框架应运而生，这里我们选用 Pytorch 框架进行深度学习模型的建立。

4.1.2 LSTM 模型

股票市场上的数据多为时间序列数据，数据之间的时序性无法被 BP 神经网络捕捉学习到。递归神经网络不同于传统的 BP 神经网络的结构只在层与层之间建立了连接，它在同一层的不同神经元之间也建立了连接，这样的结构使得 RNN 可以处理序列变化的数据。



在优化 RNN 的过程中，由于时间序列的长期依赖情况，一些参数在优化的过程中是一系列小于 1 的数字相乘，使得出现梯度消失的现象。除此之外，若网络参数的值太大，也可能是一系列大于 1 的数字相乘从而出现梯度爆炸。解决这个问题的办法就是 LSTM。

LSTM（Long Short-Term Memory，长短期记忆网络）在 1997 年首先被 Sepp Hochreiter 和 Jurgen Schmidhuber 提出，它在 RNN 的基础上在每个神经元内加入了三个门来控制信息的流入、存储和流出。

LSTM 是为了解决循环神经网络模型由于输入序列过长，而产生的梯度爆炸和消失问题而发展出来的一种机器学习神经网络，主要由记忆细胞、输入门、输出门、遗忘门组成。三个门的激活函数均为 Sigmoid 。输入门用来控制当前时刻神经单元的输入信息，遗忘门用来控制上一个时刻神经单元中存储的历史，输出门用来控制当前时刻神经单元的输出信息。

1.LSTM 算法

LSTM 引入一个新的内部状态， c_t 专门进行线性的循环信息传递，在每个时刻 t，LSTM 网络的内部状态 c_t 记录了到当前时刻为止的历史信息。 c_{t-1} 为上一时刻的记忆单元，非线性输出信息给隐藏层的外部状态 h_t ：

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

$$h_t = o_t \odot \tanh(c_t)$$

$\tilde{c}_t$ 是通过非线性函数得到的候选状态，

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

f_t, i_t, o_t 为三个门来控制信息传递的路径，三个门分别为输入门 i_t ，遗忘门 f_t 和输出门 o_t 。

输入门 i_t 决定了当前时间步有多少输入保留到新的细胞状态 c_t，它避免了将不重要的信息输入到当前存储细胞。它有三个不同的组件:（1）获取必须更新的单元格状态;（2）创建一个新的细胞状态;（3）将细胞状态更新为当前细胞状态。

$$i_t = \delta(W_i x_t + U_i h_{t-1} + b_i)$$

其中，W_o 代表 sigmoid 激活函数层的权重参数；b_o 代表 sigmoid 激活函数层的偏置参数。

遗忘门 f_t 确定模型丢弃了哪些单元状态信息，它接受前一个时间步 ht−1 的输出和当前时间步新的输入。它的主要功能是记录从前一个单元状态到当前单元状态所保留的信息量。它将输出一个介于 0 到 1 之间的值，其中 0 表示完全保留，1 表示完全放弃。计算方式分别为：

$$f_t = \delta(W_f x_t + U_f h_{t-1} + b_f)$$

输出门 o_t 控制新创建的细胞状态将被丢弃多少，输出信息首先由一个 sigmoid 层确定，然后由 tanh 处理新创建的细胞状态，再加上 sigmoid 输出来确定最终输出。

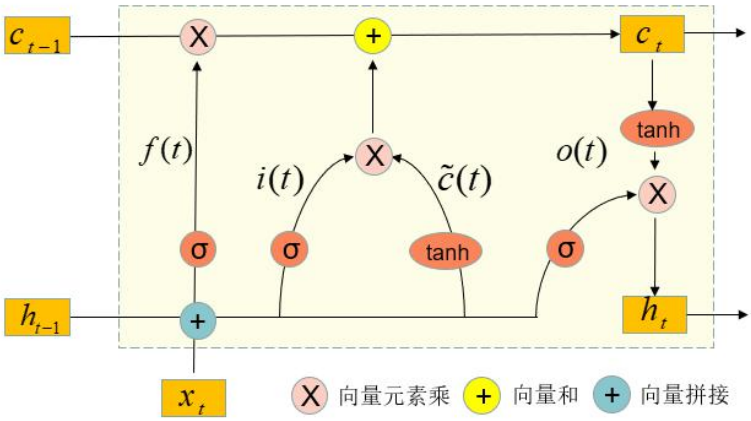
$$o_t = \delta(W_o x_t + U_o h_{t-1} + b_o)$$

2. LSTM 循环神经网络的计算过程

Step1: 首先利用上一时刻的外部状态 h_{t-1} 和当前时刻的输入 x_t，计算出三个门，以及候选状态 $\tilde{c}_t$；

Step2： 结合遗忘门 f_t 和输入门 i_t 来更新记忆单元 c_t；

Step3： 结合输出门 o_t，将内部状态的信息传递给外部状态 h_t；



图表 1 LSTM 循环单元结构

4.1.3 CNN-LSTM 模型

为了更好地应用深度学习方法对股票价格进行预测，本文利用同时提取卷积神经网络与长短期记忆神经网络的特征优势，将这两种深度学习模型融合进行预测，以更好地处理和分析金融时间序列数据。

CNN-LSTM 模型进行股票指数预测的过程可划分为两个阶段。第一阶段为 CNN 模型的特征提取阶段，第二阶段为 LSTM 模型根据时序数据进行预测阶段。

由于 CNN 中全连接层之前的作用是提取特征，全连接层的作用是分类。全连接过程相当于用全局卷积来实现这一过程，卷积的重要作用是把分布式特征映射到样本标记空间，即把特征整合到一起，输出为一个值。为了将两种深度学习模型融合到一起，我们对 CNN 模型进行修改，保留卷积层、非线性层进行特征提取，删去最后的将特征整合到一起的全连接层。从而能够利用卷积神经网络的优势提取特征。将提取出的保留时序关系的特征作为输入数据，使用 LSTM 专门来处理这些金融时间序列数据。

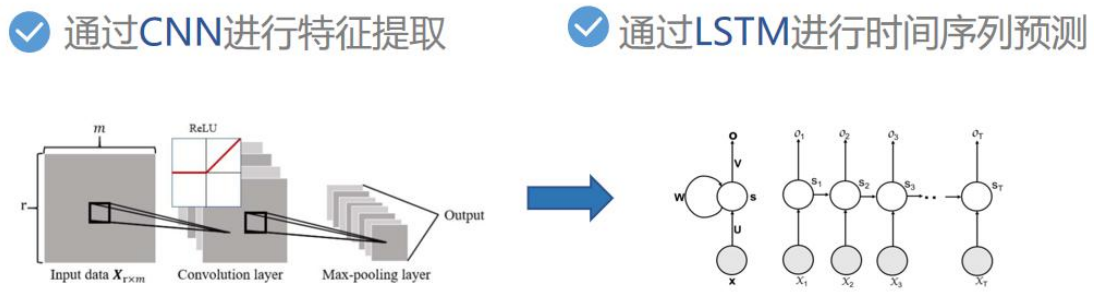


Figure 21 CNN-LSTM 模型原理示意图

4.1.4 评价指标

对模型的拟合优度做一个数量化的判断，从而表明我们所考虑的自变量能够在多大程度上结束应变量。定义判决系数：

$$R^2 = 1 - \frac{SSE}{SST}$$

其中 $SST = \sum (y_i - y)^2$ 为总平方和； $SSE = \sum (y_i - \hat{y}_i)^2$ 为残差平方和； R^2 表示误差能被模型解释的部分占总误差的百分比。一般认为 R^2 越大，模型拟合效果越好。

4.1.5 模型建立

本文选取了 2018 年 1 月 1 日至 2020 年 12 月 31 日的股票相关信息，对结构化数据和非结构化数据进行处理，进行量化因子的构建后，得到各支股票每日 49 个特征的数据。在时间尺度上使用的时间步长为 5，即利用前 5 个交易日的价格信息预测后一天股票的价格。因此，本实验中共有 62780 个可用样本。将其中的 90%作为训练集，20%作为测试集，则训练集样本数为 56502 个，测试集样本数为 6278 个。对于获取的股票数据，本文采用（0，1）标准化方法对数据进行预处理。（0，1）标准化方法是基于原始数据的最大值、最小值对数据进行标准化处理：

$$x'_i = \frac{x_i - \min}{\max - \min}$$

目前常用的卷积神经网络主要是二维卷积神经，其通过二维卷积核可以有效提取数据中的空间关系和特征，但是一般的金融时间序列数据的存储结构中并不存在有意义的二维空间关系。

一维卷积神经网络与二维卷积神经网络的主要区别在于卷积核的尺寸和窗口滑动方式，在二维卷积神经网络中，卷积窗口会按照横向以及纵向两个方向移动提取输入数据特征。但是使用二维卷积网络处理金融时间序列数据时，存在纵向移动会导致时间上信息的损失的问题。所以本文中我们最终选择使用一维卷积神经网络模型进行特征提取与预测。

在一维卷积神经网络中，通过将卷积核的尺寸设为与输入数据的尺寸同宽，从而卷积窗口仅做纵向的滑动。这种网络结构主要用于提取输入数据在单个空间方向上的平移特征，所以对于金融时间序列数据具有更好的适用性。

本文采用了 CNN 模型、LSTM 模型、CNN-LSTM 混合神经网络分别进行预测对比结果。具体数据处理和特征学习过程如下：

（1）输入层。经过结构化和非结构化数据的处理，我们对构建的量化因子进行筛选后，最终得到 49 个特征。输入的数据为经过（0，1）标准化处理后的样本数据。每一个样本包含 5 个交易日的特征数据，本文使用 5 个交易日预测第 6 个交易日。输入数据的单个样本为 49*5 的矩阵。

（2）卷积层。卷积层的作用是提取一个局部区域的特征，不同的卷积核相当于不同的特征提取器。每批选取 20 条数据，为了使用一维卷积，将数据格式进行转变，将 49 个特征作为宽，天数 5 作为长。滤波器的步长即滑动时的时间间隔设置为 1，使用零填充（zero padding）参数设置为 1。以尺寸为 3 的一维卷积核对输入数据进行卷积处理，卷积核个数为 32，即每一个卷积核对输入空间按照 49*3 的窗口大小横向滑动提取特征，训练 32 个卷积核，共提取出 32 种不同的特征。其中经过一维卷积过后，得到 32 个 1*5 的分布式特征。

对这 32 个张量进行拼接后使用全连接层得到预测结果，即是 CNN 模型的方法。这 32 个张量作为输入数据放入 LSTM 中，使用循环层进行预测，即是 CNN-LSTM 混合神经网络模型的方法。

（3）循环层

将卷积层的输出作为该层的输入，即输入数据为特征维度为 32 的数据结构，使用隐藏层状态大小为 16 的一层 LSTM 模型对输入数据进行处理，进行股价的预测。

4.2 深度学习模型实现与求解

4.2.1 实验统一设置

利用 5 个连续交易日的数据预测下一个交易日的数据，使用滑动窗口的方法对数据进行切分，每支股票三年的交易日数据可划为 725 条。按时间排列后，取前百分之九十的数据作为训练集，后百分之十的数据作为测试集，对模型进行训练和测试。

随机数种子设置为 1122，最大训练轮数设置为 10，一批数据有 20 条。使用早停法防止模型过拟合，若连续 5 轮测试集损失都为下降，就停止训练。损失函数使用 MSELoss，优化器使用 Adam，学习率为 0.0001。

4.2.2 模型结构

LSTM：由于原数据特征不多，使用一层 LSTM 的模型结构进行预测。隐藏层大小为 64。

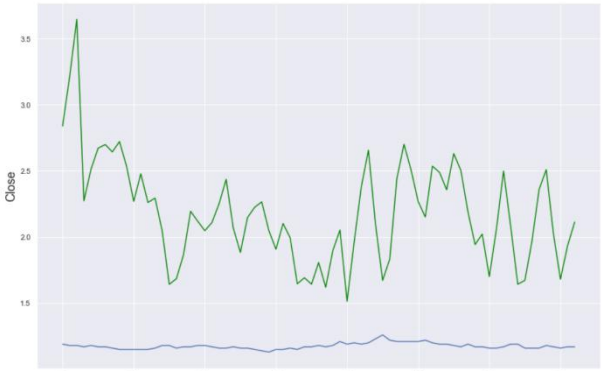
CNN：可把时间序列数据视为二维数据，长为天数，宽为特征数，输入通道为 1，输出通道为 32，卷积核大小为 3。由于使用一维卷积，因此卷积核实际大小为特征数×3。在填充后的数据上进行卷积后，输出 32 个 1×5 的张量。对这 32 个张量进行首尾拼接后再进行全连接，得到预测结果。

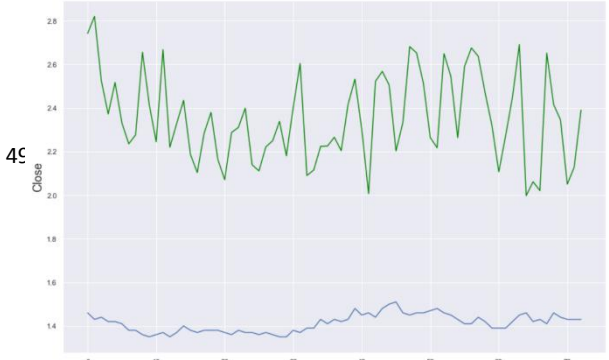
CNN-LSTM：首先使用 CNN 提取特征，卷积核设置同 CNN 模型，得到 32 个 1×5 的张量后拼成 5×32 的二维张量。送入隐藏层状态大小为 16 的一层 LSTM 中得到预测结果。

4.2.3 实验结果分析

（1）关于 R^2 结果的讨论

最开始采用拟合优度（R^2）作为模型的评价指标，这样可以与其他学者的结果进行比较。训练完成后分别计算测试集中每支股票的拟合优度，取平均作为最终评价结果。由于拟合优度的取值范围为(-∞,1]，且有效值仅能取(0,1]，因此少量异常值就能使得最终结果异常。且在一次实验结果中有大量股票的取值都小于 0，极端最值为-2321.7734085731786。将几支效果极差的股票预测结果与真实结果可视化，结果如下。





我们猜测原因如下：原始数据中，特征由三部分组成，分别是交易因子数据、历史交易数据与文本特征。文本特征由两类研报数据处理而得，分别是个股研报与行业研报。个股研报的影响施加在这个研报研究的个股上，行业研报的影响施加在这个行业的所有个股上。文本特征仅在当日出现相关研报的情况下有值，其他时候都为 0。总的来看在文本特征部分数据非常稀疏。坚持这样处理的原因是我们把没有文本信息时的特征当作正常情况，而出现文本特征时就能额外考虑这些信息，让这些信息对预测过程产生较大影响。但对于一些股价波动不大、关注者不多的小企业的股票（我们在实验中认为平均收盘价小于 2.5 元的股票为此类股票）来说，行业研报的信息并不会对其股价产生较大影响。因此，现实情况中这类股票的股价是稳定维持在一个很小的水平上，但因为额外的行业研报信息的加入，预测值隔一段时间就突变一次。波动幅度不大的历史交易数据无法提供很多信息，更放大了行业研报的影响。

选取拟合优度小于-500 的同一行业的三支股票，将预测结果反映在一张图上。

可以发现虽然是三支不同的股票，且真实值都在 1~2 之间波动，但预测值却有相同的剧烈变化的趋势。这正是每隔一段时间就加入的行业研报信息的影响。

（2）关于 R^2 评价指标选取的讨论

由于拟合优度是预测回归问题中唯一一个可以在不同数据集上互相比较的指标，因此我们最开始使用拟合优度来评价我们的结果。

4.2.4 调整后的实验

由于不受市场影响的小企业的股票在训练集中会扰乱模型的学习效果，因此我们把小企业的股票剔除后进行实验。实验结果如下：

Table 11 实验结果

cnn1D（一层卷积无池化层）					
MSE	train:0.9 所有 test：0.1 所有	train：0.9 种植业 test：0.1 种植业	train:0.9 化学原料 test：0.1 化学原料	train：0.9 钢铁 test：0.1 钢铁	平均
完整特征 66	3.038416269	0.969350094	6.059466803	0.833293457	2.725131656
去掉情感 63	2.790978395	0.970583831	3.901409451	0.803169398	2.116535269
去掉信息熵 65	2.828525726	0.951207711	5.307856417	0.983280578	2.517717608
去掉情感信息熵 62	2.629605325	1.153336684	2.307915864	1.203217341	1.823518803
去掉文本特征 51	2.765849803	0.995193224	4.615836532	1.176041647	2.388230302
平均	2.810675104	1.007934309	4.438497014	0.999800484	2.314226728

lstm（一层）					
MSE	train:0.9 所有 test：0.1 所有	train:0.9 种植业； test：0.1 种植业	train:0.9 化学原料 test：0.1 化学原料	train：0.9 钢铁 test：0.1 钢铁	平均
完整特征 66	2.474699205	1.189232067	3.053862851	0.935942163	1.913434072
去掉情感 63	2.4450411	1.117112853	2.41027684	0.812603747	1.696258635
去掉信息熵 65	2.513052044	1.079240741	2.95255368	1.05214881	1.899248819

去情感信息熵 62	2.47882115	0.968446921	2.279391937	0.663015863	1.597418968
去掉文本特征 51	2.583805058	1.061332279	2.838422515	0.964809417	1.862092317
平均	2.499083712	1.083072972	2.706901565	0.885704	1.793690562

cnn（1 层卷积无池化层）+lstm(1 层)					
MSE	train:0.9 所有 test：0.1 所有	train:0.9 种植业； test：0.1 种植业	train:0.9 化学原料 test：0.1 化学原料	train：0.9 钢铁 test：0.1 钢铁	平均
完整特征 66	2.411676088	0.83489926	1.43779116	1.232044652	1.47910279
去掉情感 63	2.536727302	0.883865728	1.674591996	1.439165476	1.633587625
去掉信息熵 65	2.121168359	0.929081971	2.316662377	1.478167093	1.71126995
去情感信息熵 62	2.404599077	0.969142756	1.458767817	1.263899493	1.524102286
去掉文本特征 51	2.37969336	1.013250026	1.448511962	1.244130283	1.521396408
平均	2.370772837	0.926047948	1.667265062	1.3314814	1.573891812

由以上结果我们可以看出：

（1）信息熵与情感指标对股价预测的准确度有负面影响。

在观察原始数据、特征值分布后我们猜测原因如下：1、情感值方面。投资研报与财经新闻、股吧评论等不同。后者发布频率高，时效性短，根据事实的好坏有不同的情感倾向，且十分明显。而前者经常倾向于研究利好趋势，或是使用一些“春秋笔法”。因此情感值大多为正向，无法客观地反映现实中的情况，也就无法为真正的股价变化提供有效的信息。2、信息熵方面。从信息熵的值来看，由于投资研报都由专业人员编写，有类似的撰写方式，因此信息熵的值集中在一个很小的区域内，非常相近，可以视为一个无效特征。加入模型中还会为股价的变化带来错误的信息。从信息熵的定义来看，信息熵能够反映一篇文章所包含内容信息程度，侧面反映一篇研报的重要程度；另一方面，信息熵也能够反映一篇文章的冗余程度，过高的信息熵可能会带来相反的效果，即代表研报内容冗余、存在干货不足的可能。因此，仅仅通过一个值，无法判断研报是有效信息多，还是冗余程度大。若想利用这方面的特征，统计学的方法以无法解决，需要结合研

报的语义特征来处理。

（2）按行业分类后的预测更为准确。

在大多数情况下，用单个行业训练的模型预测此行业内的股票，结果比三个行业混合使用更好。这是因为在同一行业内，各股会受到相同行业波动的影响，因此会产生相似的变化趋势。并且同一行业内的各企业关系密切，比如产品互为替代品的竞争关系，因此各股之间的变化可能是相互牵连的。如果用一个行业的股票训练专门预测此行业股价的模型，模型会学到内在关系，从而产生更好的预测结果。

（3）文本特征对结果的提升效果不显著。

加与不加文本特征，对效果没用特别显著的影响。我们推测原因有：1、新浪财经网站的研报数据具有一定范围的局限性。此网站无法涵盖所有股票研报报告，并且研究员在撰写研报时根据自己的主观想法，无法保证绝对的客观和全面。并且只用一种文本数据（投资研报）作为文本信息的来源稍显单薄。2、数据处理的问题。文本特征存在大量缺失值，用 0 填充的方法未考虑研报信息的时效性、相互作用的关系等因素，因此无法真正地发挥研报数据对股价预测的促进作用。文本量化只有 11 个浅层特征与 3 个情感特征、1 个信息熵特征，未把研报中所有有价值的信息挖掘出来。3、模型问题。模型的结构不适合处理大量缺失的数据，可以从模型内部结构做出改进。

（4）模型的特点与有效性分析。

CNN 的效果优于 LSTM，而我们将二者结合的模型具有最好的效果。CNN 的卷积操作可以提取数据中的有效特征，为其加权，并减小特征数。LSTM 适合处理时序数据，通过 Cell 状态保留数据之间的时序关系。因此通过 CNN 提取特征并简化，同时保留其中的时序关系，送入 LSTM 中进行最后的预测，达到了最好的效果。

五 总结与展望

5.1 本文研究主要内容与贡献

本文主要基于多源异构数据的分析方法，对国内股指数据、研报数据与新闻数据进行研究，最终以 A 股钢铁行业、化学原料行业以及种植业为例，实现了对未来股价的初步预测工作，预测结果准确度相对较高。工作总结具体如下：

（1）通过多渠道收集多源异构数据。通过 Wind、国泰安、Tushare、Resset 以及 Nasdaq 等方式进行股指数据的收集，借助新浪财经、网易财经、东方财富等网站收集个股与行业研报数据、新闻数据等。

（2）对数据进行预处理。结构化数据部分，进行指标因子的构造，例如 OBV、CCI 等共计 40 个特征，分别对这些特征进行因子有效性检验，均通过测试。非结构化数据部分，对研报数据与新闻数据提取情感值与信息熵，以及构造的文本语句特征等，用来评估行情舆论对股市的影响。

（3）模型求解。分别使用 LSTM、CNN 与 CNN-LSTM 模型进行求解，用五天的历史数据来预测第六天的股价或者涨跌情况，最终最佳效果为 CNN-LSTM 模型，MSE 评价为 0.83489926，股价预测方面准确度较高

（4）结果分析。对于模型求解的后验分析，我们可以发现，情感与信息熵对模型评估有负面作用，考虑是因为缺失值过多，以及特征构造不够充分；文本特征对模型提升效果不明显，因为所收集的新闻与研报数据过少，特征稀疏，不能够给模型带来较明显的作用。其它方面，整体建模会对微型上市企业预测带来过大影响，因为行业特征会影响其预测趋势，但微型企业的股价几乎不波动，后续考虑剔除微型企业，整体建模，效果有较大提升。

最终我们实现了基于数据驱动的量化交易，对收集数据、处理数据、模型求解、模型评估以及修正等理论方法有进一步的学习理解，创新的实现了 CNN 与 LSTM 模型的融合，模型能够一定程度上进行选股与模拟交易策略，在量化交易方面做出了进一步贡献。

5.2 未来的工作

（1）考虑更深层次的语义特征。若要使用文本数据促进股价预测，不能仅

仅通过统计学的方法计算关于词语的浅层特征，还需要进一步考虑其中的语义信息。例如，从情感分析上来看，仅仅计算积极词语与消极词语的个数不能完全反应文本的情感倾向，并且目前的方法仅是文档级别的情感分析。后续可以加入句子级别的情感分析或方面级别的情感分析特征。

（2）加入新闻等文本数据，考虑其时效性与相互作用的关系。在进行特征的预处理时，当天的文本数据特征仅拼接在当天的交易数据特征之后，且未出现研报的日期中都用 0 填充。若一天中同一股票有多份研报，仅仅是将其特征值相加。后续可以调研更多文献，处理出研报信息的衰退效应，使得特征更加平滑。或是将向量改成高维张量，存储不同层面数据之间的相互作用关系。

（3）加入网络关系。股票市场各企业都是相互关联的，存在各种各样的网络关系。比如供应链网络、持股网络，若能将网络中蕴含的信息加入模型中，能更好地学到市场中的信息，对股价进行更准确的预测。

（4）设计更适合金融市场的模型。目前金融科技仍在起步阶段，所用的 CNN、LSTM 模型，最早的提出是针对图像、时序数据的。而金融市场的数据还有很多自己独特的特征，如股票市场中的动量溢出效应。目前的模型不能够有效捕捉金融市场的信息也在意料之中，未来可以针对金融市场数据的本质特征来设计针对金融领域的模型。

参考文献

[1] Chen, Y. W., Chen, K., Yuan, S. Y., & Kuo, S. Y. (2016). Moving object counting using a tripwire in H. 265/HEVC bitstreams for video surveillance. Ieee Access, 4, 2529-2541.

[2] Liu, Y., Liu, Q., Zhao, H., Pan, Z., & Liu, C. (2020, April). Adaptive quantitative trading: an imitative deep reinforcement learning approach. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 02, pp. 2128-2135).

[3] Kim, J. Y. (2015). Analysis of Loss Expectancy on Personal Information Leakage using Quantitative Invest Decision Model. Journal of Society for e-Business Studies, 20(2).

[4] Chen, C., Zhang, P., Liu, Y., & Liu, J. (2020). Financial quantitative investment using convolutional neural network and deep learning technology. Neurocomputing, 390, 384-390.

[5] Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015, April). Long short term memory networks for anomaly detection in time series. In Proceedings (Vol. 89, pp. 89-94). Presses universitaires de Louvain.

[6] Gers, F. (2001). Long short-term memory in recurrent neural networks (Doctoral dissertation, Verlag nicht ermittelbar).

[7] Chen, Y. W., Chen, K., Yuan, S. Y., & Kuo, S. Y. (2016). Moving object counting using a tripwire in H. 265/HEVC bitstreams for video surveillance. Ieee Access, 4, 2529-2541.

[8] Ta, V. D., Liu, C. M., & Tadesse, D. A. (2020). Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. Applied Sciences, 10(2), 437.

[9] Fang, Y., Chen, J., & Xue, Z. (2019). Research on quantitative investment strategies based on deep learning. Algorithms, 12(2), 35.

[10] Cheng, R., & Li, Q. (2021). Modeling the Momentum Spillover Effect for Stock Prediction via Attribute-Driven Graph Attention Networks.

[11] 潘莉,徐建国.A 股市场的风险与特征因子[J].金融研究,2011(10):140-154.

[12] 孔翔宇,毕秀春,张曙光.财经新闻与股市预测——基于数据挖掘技术的实证分析[J].数理统计与管理,2016,35(02):215-224.

[13] 耿立校,刘丽莎,李恒昱.多源异构数据融合驱动的股票指数预测研究[J/OL].计算机工程与应用: 1-10[2021-05-16]. http://kns.cnki.net/kcms/detail/11.2127.TP.20210419.1324.017.html.

[14] 黄创霞,温石刚,杨鑫,文凤华,杨晓光.个体投资者情绪与股票价格行为的互动关系研究[J].中国管理科学,2020,28(03):191-200.

[15] 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].首都经济贸易大学,2016.

[16] 刘泽羲,王文俊,潘林.基于多层复杂网络理论的海洋货运网络的抗毁性研究[J].海洋通报,2018,37(06):652-658.

[17] 赵一鸣,吴林容,任笑笑.基于多知识图谱的中文文本语义图构建研究[J].情报科学,2021,39(04):23-29.

[18] 查韵洁. 基于混合因子选股的交易策略研究[D].电子科技大学,2020.

[19] 王成龙,王曦.基于投资者情绪的量化投资策略研究[J].中国物价,2021(03):82-85.

[20] 马长峰,陈志娟,张顺明.基于文本大数据分析的会计和金融研究综述[J].管理科学学报,2020,23(09):19-30.

[21] 杨妥,李万龙,郑山红.基于新闻信息的股票指数预测[J].长春工业大学学报,2020,41(01):47-52.

[22] 高雅,冯爽.结合注意力机制的新闻文本情感分析算法[J].新型工业化,2020,10(07):15-18.

[23] 唐振鹏,吴俊传,冉梦,张婷婷.考虑投资者情绪的中国股市自激发效应研究[J].中国管理科学,2020,28(07):1-12.

[24] 贺康,宋冰洁,刘巍.年报文本信息复杂性与资产误定价——基于文本分析的实证研究[J].财经论丛,2020(09):64-73.

[25] 谢赤,边慧东,王纲金.牛熊市视角下股票关联网络动态拓扑结构研究——以上证 50 指数为例[J].复杂系统与复杂性科学,2017,14(01):66-74.方建武,安宁.中美股市的联动性分析及预测[J].经济问题探索,2010(04):80-86.