

基于机器学习的牛熊市分市场股票预测

刘勇福, 潘淼鑫, 熊金波

(福建师范大学 计算机与网络空间安全学院, 福建 福州 350117)

摘要: 股票收益预测是量化投资领域的关键课题。传统的股票收益预测方法大多基于全市场数据, 未充分考虑牛市与熊市的市场差异。基于此, 提出一种基于机器学习的牛熊市分类预测方法。首先, 将股票市场划分为牛市和熊市两个阶段; 其次, 在不同市场状态下筛选出对股票收益率有显著影响的关键因子, 并分别构建对应的预测模型, 使预测模型能够更精准地反映各阶段的市场动态; 最终, 通过比较牛市和熊市分市场模型与统一市场环境下预测模型的效果, 验证该方法的有效性。实验结果表明, 分市场模型在预测表现上显著优于统一模型, 基于牛市和熊市的划分策略能够显著提升机器学习模型在股票收益预测中的应用效果。

关键词: 股票收益预测; 机器学习; 牛市; 熊市; 因子筛选

DOI: 10.11907/rjdk.251370

中图分类号: TP181

文献标识码: A

文章编号: 1672-7800(2026)009-0058-09

扫描二维码阅读全文:



Machine Learning Based Stock Prediction for Bull and Bear Markets

LIU Yongfu, PAN Miaoxin, XIONG Jinbo

(College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China)

Abstract: Stock return prediction is a key topic in the field of quantitative investment. Traditional stock return prediction methods are mostly based on overall market data and often overlook the differences between bull and bear markets. Based on this, we propose a machine-learning-based approach for bull - bear market classification and forecasting. First, the stock market is partitioned into two regimes: bull and bear markets. Second, under each regime, key factors that significantly affect stock returns are identified, and corresponding predictive models are constructed separately so that the models can more accurately capture the market dynamics in each phase. Finally, by comparing the performance of regime-specific models (bull and bear) with that of a single model trained under a unified market setting, the effectiveness of the proposed approach is validated. Experimental results show that the regime-specific models significantly outperform the unified model, indicating that the bull-bear market segmentation strategy can substantially enhance the predictive performance of machine-learning models in stock return forecasting.

Key Words: stock return prediction; machine learning; bull market; bear market; factor screening

0 引言

股票收益预测一直是金融市场的重要课题, 尤其在量化投资领域, 精准的预测对投资者至关重要。Fama^[1]指出, 金融市场的复杂性和动态性主要源于环境高度的不确定性与波动性。在瞬息万变的市场中, 投资者需要依赖有效策略来制定决策和优化资产配置。由于股票市场具有

难以预判的波动性, 准确预测市场趋势成为一大挑战。

近年来, 随着机器学习与数据分析技术的迅速发展, 相关方法已被广泛应用于提升股票预测的准确性, 以应对市场的复杂性与不确定性。目前, 许多机器学习模型通常基于全市场数据进行预测。例如, Hoseinzade 等^[2]提出基于卷积神经网络(CNN)的框架, 用于预测多个股指的次日走势; Lu 等^[3]构建结合 CNN 与 LSTM 的模型, 先提取时间序列特征, 再预测次日收盘价; Wang 等^[4]则基于 Transform-

收稿日期: 2025-05-29

基金项目: 国家自然科学基金面上项目(62272102); 福建省自然科学基金重点项目(2023J02014)

作者简介: 刘勇福(1999-), 男, 福建师范大学计算机与网络空间安全学院硕士研究生, 研究方向为深度学习、量化投资; 潘淼鑫(1987-), 女, 博士, 福建师范大学计算机与网络空间安全学院讲师, 研究方向为机器学习、无人机路径规划、物联网; 熊金波(1981-), 男, 博士, 福建师范大学计算机与网络空间安全学院教授, 研究方向为数据安全与隐私保护、人工智能安全。本文通讯作者: 潘淼鑫。

er架构,依据近期收盘价数据预测次日价格;邓飞燕等^[5]使用LSTM预测价格趋势,输入包括全市场的开盘价、收盘价、最高价、最低价和成交量等特征;Abdul等^[6]则提出通过多层LSTM网络优化预测收盘价的方法;王德广等^[7]构建融合技术指标与基本面指标的混合模型,提升整体预测能力。然而,这些模型普遍忽略了市场状态(如牛市和熊市)对因子表现和市场行为的影响。在不同市场环境下,统一建模可能带来较大的预测偏差。例如,在牛市中因子作用机制与熊市显著不同,若不加区分,可能导致模型在某些市场状态下效果较差。

事实上,不同市场状态下的投资者行为往往存在显著差异:牛市中情绪高涨、资金持续流入,推动价格上涨;熊市中则易出现恐慌抛售,导致市场剧烈下行。若预测模型未能有效区分这些状态,尤其在极端行情下,预测误差可能显著放大。因此,构建能够适应不同市场状态的预测模型十分必要。针对不同市场环境下的因子表现进行深入研究,通过分别在牛市和熊市中筛选对股票收益率有显著影响的因子,可以构建出更加精准的预测模型,从而更好地捕捉市场波动特征。因子选择不应仅基于整体市场表现,还需考虑市场分化状态下的动态变化。基于上述问题,本文拟从市场状态出发,构建面向牛市与熊市的机器学习预测模型,以提升收益预测的精度和鲁棒性。本文主要贡献如下:

(1)提出一种基于高低点标记法的市场状态划分方法,将股票市场划分为牛市与熊市,更真实地反映市场波动与趋势。

(2)为提高划分精度,引入滑动窗口机制,设计不同时间尺度的实验(短期6个月,中期12个月,长期24个月),分析窗口长度对市场划分与预测性能的影响,从中确定最优窗口。

(3)在牛市和熊市环境下,分别筛选对收益率显著有效的因子,构建具备市场适应性的预测模型,从而提升预测精度。

(4)系统比较分市场模型(牛市与熊市分别建模)与统一市场模型在不同市场条件下的表现,检验其实际预测效果差异,并验证分市场建模的优势。

1 相关工作

股票收益预测方法主要可分为两类^[8-10]:①传统模型。如自回归移动平均(ARMA)和广义自回归条件异方差(GARCH),基于历史价格和波动率进行预测,适用于波动较小的市场环境,但对复杂非线性关系的建模能力有限;②机器学习算法。如随机森林和支持向量机,擅长捕捉收益与多种影响因素之间的非线性关系,能够有效应对金融数据中的噪声和复杂结构。然而,这类算法对训练样本依赖性较强,在数据分布发生变化时适应性较差,尤其

在高波动的金融市场中表现不稳定。因此,传统机器学习方法在高频交易和实时决策场景中存在一定局限^[11]。相比之下,深度学习模型具备更强的非线性特征提取能力和自动学习能力,在处理复杂金融数据时表现更优。但已有研究指出,该类模型在应对突发市场事件时适应性不足,容易在历史数据上产生过拟合,影响实际应用效果^[12]。因此,在不同市场状态下采用分市场建模策略,有助于提升深度学习模型的灵活性和预测能力。

股票市场具有明显的阶段性波动特征,通常表现为牛市与熊市的交替轮动。牛市指股市整体呈上涨趋势,投资者情绪积极、信心强;熊市则表现为整体下跌,市场情绪普遍低迷。由于不同市场阶段下价格行为差异显著,准确识别牛熊市的转换点对于股票收益及其预测具有重要意义。目前,学术界尚未就牛市与熊市的划分达成统一标准,常见方法包括:基于技术分析指标、基于价格峰谷的划分法以及高低点标记法。Menkhoff^[13]的研究表明,尽管许多基金经理使用技术分析工具识别市场转折点,且该方法在短期交易中更具优势,但在高波动性或快速变化的市场环境下容易导致误判。Candelon等^[14]研究了多市场间的牛熊市同步性,指出峰谷划分在高波动市场中存在极值判定困难的问题,可能影响识别准确性。相比之下,Harding等^[15]提出的高低点标记法能更有效地划分市场或经济的周期阶段。其优点在于:可根据市场数据的动态变化自动调整划分,更适用于研究长期趋势。在实际投资中,市场阶段对投资策略表现影响显著。Paramita等^[16]发现,在牛市中,善于市场时机选择的基金通常表现出色;而在熊市中,尽管择时效果受限,但仍有助于减少损失,从而实现更稳定的回报。Li等^[17]指出,在牛市中波动率预测的准确性较高,而熊市中因波动性加剧,模型预测效果显著下降。此外,不同因子在牛熊市中的预测能力存在差异:如动量因子在牛市中效果更佳,而估值因子在熊市中的表现更优。Tran等^[18]研究了中国大陆和台湾省市场中宏观经济变量对熊市的预测能力。通过使用马尔可夫转换模型,发现通胀率、实际汇率变化和外汇储备增长是预测中国大陆熊市的关键因子,而期限利差和失业率等是台湾省熊市的主要预测因子。

在金融领域,因子筛选是研究股票收益率与风险关系的核心问题。传统方法多基于资产定价理论,如Fama等^[19]和French等^[20]提出的三因子模型(市场因子、规模因子、账面市值比因子)及其扩展的五因子、六因子模型。这些模型在金融因子的筛选与构建中占据重要地位,强调了价值因子与规模因子对股票收益率的显著影响。近年来,随着机器学习和市场状态分割方法的发展,研究者开始更加关注在不同市场状态下,筛选对收益率具有显著解释力的因子。Ang等^[21]探讨了波动率因子在牛市和熊市中的表现,指出若不对市场状态进行区分,模型可能忽视牛熊市的结构性差异,进而降低预测准确性。在整体市场上训

训练的模型,往往对特定市场条件产生过拟合,从而在其他状态下预测效果不佳。相比之下,分市场建模策略能够更有效地识别市场环境的变化,如牛市、熊市或震荡行情等,并据此调整模型结构与参数,从而提升预测精度。这种方法具备更强的灵活性与适应性,有助于解决模型在不同市场阶段的稳定性问题。然而,目前多数研究仍集中于统一市场环境下的预测,对市场的阶段性特征关注不足。尽管部分研究考虑了波动率的影响,但明确针对牛熊市进行分市场建模的文献仍相对有限,有待进一步系统探索与深入研究。

2 牛熊市分市场预测方法

2.1 数据集及预处理

(1)数据来源。本文使用的数据来源于 Tushare Pro 平台,下载内容为前复权价格数据。该类数据已充分考虑拆股、分红和配股等公司行为的影响,能够有效反映股票在不同时间点的真实价格。例如,未经复权处理的 2:1 拆股可能导致次日价格出现-50%的虚假回报,前复权处理能有效避免此类误差。数据集涵盖 2005—2024 年的每日股票交易信息,包括交易日期、开盘价、收盘价、最高价、最低价、成交量及成交额等。

(2)数据清洗。为保证数据质量,对缺失值进行了处理。由于缺失数据占比极小,采用直接删除法,不会对样本总体构成造成实质性影响。

(3)特征工程。研究中所使用的因子基于前复权数据计算获得。所有因子均进行了 MinMaxScaler 归一化处理,以消除量纲差异并提升模型训练效果。归一化公式如公式(1)所示:

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中, x 是原始数据; $\hat{x}$ 是归一化后的数据; $x_{\min}$ 是原始数据的最小值; $x_{\max}$ 是原始数据的最大值。

2.2 牛市与熊市的划分

高低点标记法用于将市场划分为牛市和熊市,其核心在于通过识别价格序列中的局部高点与低点,确定市场转折点及阶段起始。具体做法是基于滑动窗口技术,在时间序列中滑动固定长度的窗口,动态识别局部极值(最高价和最低价)。通过实验对比确定窗口长度,以保证划分的合理性与准确性。

(1)定义时间窗口。时间窗口 $W(t, j)$ 为给定时间 t 的前后 j 个时段,即 $w(t, j) = \{x(t-j), \dots, x(t), \dots, x(t+j)\}$ 。

(2)定义高点和低点。如果 $x(t) = \max(W(t, j))$,则时间 t 为高点;如果 $x(t) = \min(W(t, j))$,则时间 t 为低点。

(3)边界处理。如果在时间序列的边界(如开头或结尾),窗口范围超出了时间序列的可用数据,则只考虑针对可用的部分进行比较。

(4)连续高点或低点处理。如果连续出现多个高点,且之前没有出现低点,或者它们之间没有间隔点,则只保留第一个高点,其他的高点不作标记;同理,如果连续出现多个低点,且之前没有出现高点,或者它们之间没有间隔点,则只保留第一个低点,其他低点不作标记。

(5)趋势生成。从低点 t_1 到下一个高点 t_2 的区间,满足 $x(t_2) > x(t_1)$ 被识别为上升趋势。这意味着从某一低点开始,价格逐渐攀升至高点,市场呈现牛市状态;从高点 t_1 到下一个低点 t_2 的区间,满足 $x(t_2) < x(t_1)$ 被识别为上下下降趋势。从某一高点开始,价格持续下降到低点,市场处于熊市状态。

为实现基于高低点标记的牛熊市划分,该流程被拆分为“极值点识别”与“趋势区间生成”两个阶段:首先,算法 1 基于滑动窗口,在价格序列中逐时刻搜索局部极大值与极小值,并输出高点列表 H 与低点列表 L (必要时结合边界截断与重复极值的去冗规则,以保证拐点集合的稀疏性与一致性);其次,算法 2 以 H 和 L 为输入,按照时间顺序对高低点进行交替配对,生成由若干连续区间构成的市场趋势列表 T 。其中,从低点到其后相邻高点构成的上升区间被标记为牛市,从高点到其后相邻低点构成的下降区间被标记为熊市,从而得到完整的牛熊市阶段划分结果。

算法 1 高低点检测

输入:价格序列 $p[1 \cdots n]$,窗口大小 j
输出:高点列表 H ,低点列表 L
1. $H, L \leftarrow [], []$ //初始化高点和低点列表
2. for $t \in \{j+1, n-j\}$ do //循环遍历
3. window = $p[t-j : t+j+1]$ //设定窗口范围
4. if $p[t] == \max(\text{window})$ and $p[t] \neq p[t-1]$ then
 $H.append(t)$ //添加高点
5. if $p[t] == \min(\text{window})$ and $p[t] \neq p[t-1]$ then
 $L.append(t)$ //添加低点
6. return H, L

算法 2 市场趋势分类

输入:高点列表 H ,低点列表 L
输出:市场趋势列表 T
1. $T \leftarrow [], i \leftarrow 0, k \leftarrow 0$ //初始市场趋势列表及索引
2. while $i < |H|$ 且 $k < |L|$ do //循环遍历
 if $L[k] < H[i]$ then
 if 存在 $L[k+1]$ 且 $L[k+1] < H[i]$ then $k \leftarrow k+1$ //移动低点索引到下一个位置
 else then T .添加("牛市", $L[k], H[i]$) //添加牛市趋势到列表 T
 $k \leftarrow k+1, i \leftarrow i+1$ //移动低点和高点索引到下一个位置
 3. else then
 if 存在 $H[i+1]$ 且 $H[i+1] < L[k]$ then $i \leftarrow i+1$ //移动高点索引到下一个位置
 else then T .添加("熊市", $H[i], L[k]$) //添加熊市趋势到列表 T
 $i \leftarrow i+1, k \leftarrow k+1$ //移动高点和低点索引到下一个位置

4.return T

(6)起始点和终点的趋势判断。起始点与终点分别是时间窗口内第一个与最后一个局部极值点(高点或低点)。如果起始点恰好是高点或低点,则根据高低点判断起始趋势。如果起始点是高点,则起始点到下一个低点之间被识别为下降趋势;如果起始点是低点,则起始点到下一个高点之间被识别为上升趋势。如果起始点不是高点或低点,则通过向前寻找到的第一个局部极值点(高点或低点)来判断趋势。如果起始点前的第一个局部极值是高点,则从该高点到起始点的区间被判断为下降趋势;如果是低点,则为上升趋势。同理,终点后的处理也应该类似,根据终点后的第一个局部极值点来判断趋势。

(7)牛市与熊市的划分。如果从一个低点到一个高点之间的价格呈上升趋势,则这一阶段标记为牛市,意味着价格持续上涨;如果从一个高点到一个低点之间的价格呈下降趋势,则这一阶段标记为熊市,意味着价格持续下跌。

2.3 因子筛选

本文分别对牛市和熊市阶段的数据采用普通最小二乘法(Ordinary Least Squares, OLS)对因子与股票收益率进行回归分析。通过回归系数的显著性,筛选出在不同市场状态下的有效因子。所选因子包括动量(Momentum)、波动率(Volatility)和市值(Market Capitalization)等,这些因子在股票回报分析中应用广泛。

(1)数据准备。首先从数据集中提取相关因子(如 bias1, bias2),并进行归一化处理,以消除量纲差异带来的影响。为确保回归模型包含截距项,采用 add_constant 方法添加常数项。假设有 N 个因子,则回归模型形式如公式(2)所示:

$$y_{t+1} = \alpha + \beta_1 \cdot x_{1,t} + \beta_2 \cdot x_{2,t} + \cdots + \beta_N \cdot x_{N,t} + \epsilon_t \quad (2)$$

其中, $x_{i,t}$ 表示第 t 期第 i 个因子的值; ϵ_t 是误差项。

(2)构建因变量(y)。将月度收益率 y_{t+1} 作为因变量,这表示用第 $t+1$ 期的收益率来回归当前时间点(即第 t 期)的因子值。

(3)OLS 回归模型构建。在每一个时间点(例如每月),通过最小化残差平方和的方法进行 OLS 回归,拟合数据。最小化残差平方和如公式(3)所示:

$$\min_{\alpha, \beta_1, \dots, \beta_N} \sum_{t=1}^T \left(y_{t+1} - \alpha - \sum_{i=1}^N \beta_i \cdot x_{i,t} \right)^2 \quad (3)$$

其中, T 是时间长度(样本的期数)。

(4)回归系数和统计量。OLS 回归的结果会给出每个因子的回归系数 β_i 和对应的 t 统计量 t_i 。这些值用于评估因子对收益率的解释能力。回归系数通过公式(4)求得:

$$\beta = (X^T X)^{-1} X^T y \quad (4)$$

其中, X 是自变量矩阵(因子矩阵,包括常数项); y 是因变量(收益率)。 t 统计量用于判断每个回归因子是否显著,计算公式如公式(5)所示:

$$t_i = \frac{\beta_i}{\text{标准误差}(\beta_i)} \quad (5)$$

标准误差由残差的方差估计得到,用于衡量估计的准确性。

(5)因子的表现分析。回归系数反映因子对收益率的影响方向与强度:系数越大,影响越显著。

统计量(t -values)用于判断因子的显著性。一般来说, $|t| > 1.96$ 时,因子在 95% 置信水平下被认为是显著的。

(6)筛选因子的过程。通过对回归系数与 t 值的比较,分别筛选出在牛市与熊市中显著影响收益率的因子。若某因子在某一市场状态下的 $|t|$ 值显著(>1.96),则说明该因子在该市场环境下具有较强的预测能力。通过对比牛熊市下同一因子的回归结果,可进一步识别在不同市场状态中更具解释力的因子。

(7)逐步双向 F 检验。对初步筛选出的显著因子,需进一步筛选其中最具解释力的因子。该方法结合前向选择(逐步引入显著性强的因子)和后向剔除(移除不显著的因子),分别在牛市和熊市阶段独立进行,从而构建适应不同市场状态的最优因子模型,以提高模型在特定市场环境下的预测精度与稳健性。

2.4 模型构建与评估

(1)模型构建。本文采用长短期记忆网络(Long Short-Term Memory, LSTM)进行时间序列预测。LSTM 是递归神经网络(RNN)的改进形式,具备良好的长期依赖建模能力,适用于捕捉金融数据中复杂的时序特征。图1展示了该模型的结构,包括输入层、三层 LSTM 单元、全连接层和输出层。此外,模型中还包括损失函数和优化器的配置。其中,10代表时间步长、4代表输入特征(输入特征随着因子的数量改变)、50代表神经元的数量、1代表最终输出层的神经元数量(即预测的股票收益率)。

(2)模型训练。在数据预处理阶段,采用 MinMaxScaler

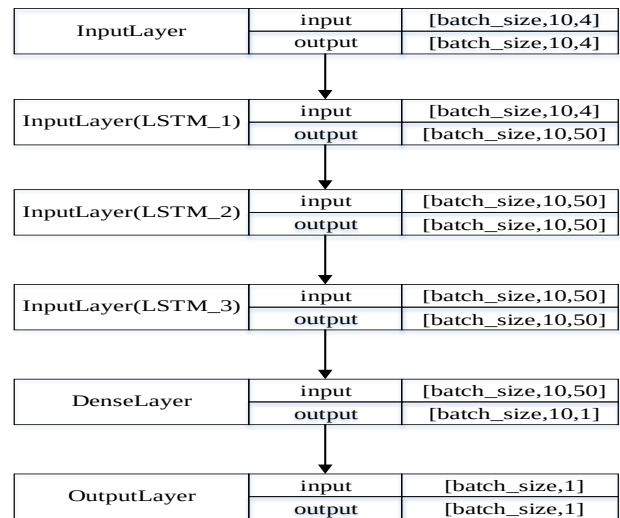


Fig. 1 Model structure

图1 模型结构

er 对输入特征和目标变量进行归一化处理,使其数值范围压缩至 $[0, 1]$ 区间,旨在消除不同尺度间的影响,以提升模型的训练效率与稳定性。数据集的划分严格遵循时间顺序,区分为训练集与测试集,以确保测试阶段不泄露未来信息。训练过程中,模型使用 RMSProp 优化器对参数进行更新,并以均方误差(MSE)作为损失函数,最小化预测值与真实值之间的差异。为抑制过拟合,引入早停机制:若测试集损失在连续若干轮次内无显著改善,则提前终止训练。每轮训练结束后,模型在测试集上的表现均被评估与记录。若当前轮次的测试损失优于此前最佳记录,则保存该轮模型参数,作为最终模型的备选方案,以确保模型在泛化能力上的最优表现。

(3)模型评估。模型性能采用均方根误差(RMSE)、平均绝对百分比误差(MAPE)和 R^2 (决定系数)进行评估,具体如下:

均方根误差(RMSE)表示预测值与真实值之间的平方误差的平均值,用来衡量预测值的波动性,如公式(6)所示:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2} \quad (6)$$

平均绝对百分比误差(MAPE)表示预测值相对真实值的平均绝对误差,反映模型预测的相对误差,如公式(7)所示:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}}{y_i} \right| \quad (7)$$

R^2 (决定系数)用来衡量模型预测的拟合优度,数值越接近1,表明模型的预测效果越好,如公式(8)所示:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

其中, y_i 是真实值; $\hat{y}$ 是预测值; $\bar{y}$ 是平均值。

3 实验设计与结果

3.1 实验平台和工具

本文实验在集成开发环境 PyCharm 中完成,采用 Python3.6 编程语言,基于 TensorFlow 库构建并实现深度学习模型。数据可视化部分使用 matplotlib3.4 进行图表绘制与结果分析。硬件环境包括 64 GB 内存、Intel i7-12650H@2.30GHz 处理器及 NVIDIA GeForce RTX 4060 Laptop GPU,为模型训练与计算提供了强有力的加速支持。

3.2 实验设置

(1)实验数据。本文所用数据来源于 Tushare Pro,涵盖沪深 300 指数及 4 只具有代表性的前复权股票:贵州茅台、紫光股份、招商银行和中国石化。选择这 4 只股票的具体原因包括:①行业覆盖广,分别代表消费、科技、金融和能源等关键领域,可全面反映不同产业在牛市与熊市中的表现差异;②各股票在所属行业中具备较高市场份额和

影响力,能够有效反映行业及整体市场波动;③上述公司在不同市场阶段下的收益特征差异显著,为研究市场的阶段性变化提供了有代表性的样本。其中,招商银行代表市值高、波动小的金融股;贵州茅台代表稳健增长的蓝筹股及高端消费品;中国石化为传统能源龙头;紫光股份则体现科技成长股的特征。训练集时段为 2005 年 4 月—2019 年 12 月,测试集时段为 2020 年 1 月—2024 年 9 月。

(2)股票因子。根据资产定价研究的常见方法,本文选用了趋势、流动性、波动性、动量及基本面等 5 类因子,具体如下:①趋势因子包含 DI1(正向指标)、DI2(负向指标)、ADX(平均趋向指数)及 ADXR(平均趋向变动速率);②流动性类因子包含换手率和单日成交量;③波动性因子包含 6 个月波动率、9 个月波动率、平滑异同移动平均线(MACD)、移动平均线(MA5、MA10 和 MA20)、过去 30 天日收益率标准差、30 天内最大日度回报、随机指标 KDJ、前 1 个月涨跌幅、前 3 个月涨跌幅以及前 6 个月涨跌幅;④动量类因子包含相对强弱指数(RSI;RSI1、RSI2、RSI3 分别表示 6 日、12 日、24 日周期的 RSI)和威廉指数(W%R);⑤基本面因子包含市盈率、市净率、市销率、流通股、流通市值、三重指数平滑均线(TRIX)、TRIX 的移动均线(TRMA)以及乖离率(BIAS 包括了 3 种,其中,BIAS1 代表 6 日移动平均线计算乖离率,BIAS2 代表 12 日移动平均线计算乖离率,BIAS3 代表 24 日移动平均线计算乖离率)。

(3)数据集。基于沪深 300 指数,按照时间窗口划分牛熊市,分别提取 4 只股票的牛市和熊市数据集,同时保留其全市场数据。针对每只股票,筛选牛市显著因子作为牛市数据集的输入特征,用于训练和预测;熊市和全市场数据采用相同方法处理。

3.3 实验结果

3.3.1 不同市场的实验结果

(1)实验方法。基于沪深 300 指数,以 6 个月为时间窗口划分训练集与测试集的牛熊市状态。针对每只股票,分别筛选牛市、熊市及全市场的显著因子,并利用 LSTM 模型(参数见表 1)对不同市场状态的月收益率进行训练。随后,在测试集上,根据各自市场的显著因子,使用训练好的 LSTM 模型分别预测牛市、熊市及全市场表现。

Table 1 LSTM model parameters settings

表 1 LSTM 模型参数设置

模型参数	数值大小	说明
时间步长	10	输入序列长度,表示使用连续 10 个交易日的数据作为输入特征
隐藏层大小	50	每层 LSTM 单元包含 50 个神经元,用于特征提取和时序模式学习
层数	3	堆叠 3 层 LSTM 结构以增强模型复杂模式的捕捉能力
学习率	0.000 1	RMSProp 优化器的初始学习率
训练轮数	200	每个训练批次包含 16 个样本,兼顾内存效率和梯度更新稳定性
提前停止	40	监控指标:验证集 MSE;停止条件:连续 40 轮无改善;保存策略:最佳模型

(2)实验结果与分析。图2—图5分别呈现了4只股票在3种市场状态下的月收益率预测结果。由图可知,区分牛熊市后,预测值与真实值的拟合效果较好。

表2列出了4只股票的详细评估指标。在RMSE指标

上,全市场条件下的RMSE较高,而牛熊市条件下普遍较低,其中,熊市的预测误差最小;在MAPE指标上,全市场条件下的MAPE较高,牛熊市条件下较低,说明未区分市场状态时预测误差较大;在R²指标上,牛熊市的R²普遍高

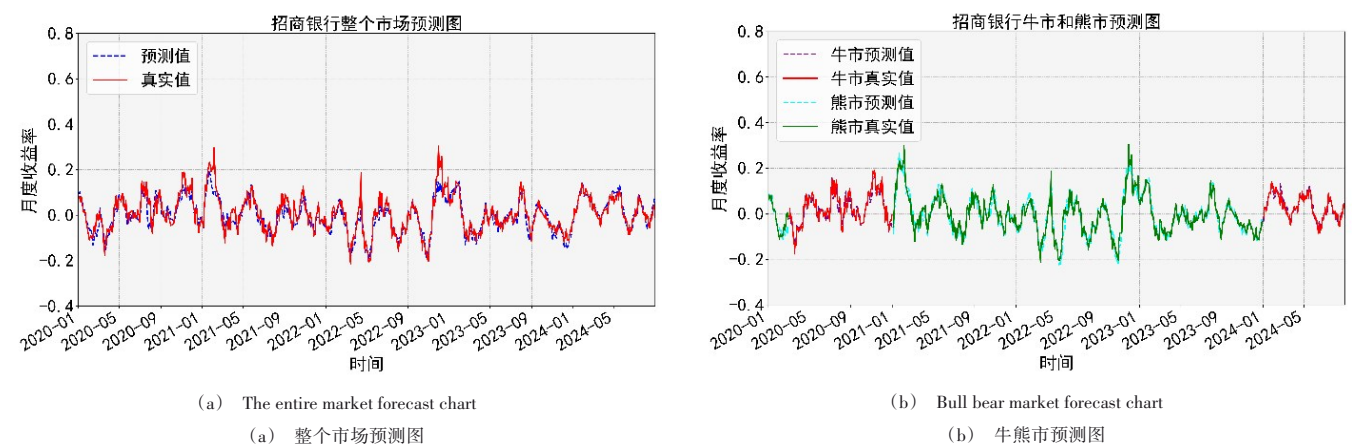


Fig. 2 China merchants bank forecast chart

图2 招商银行预测图

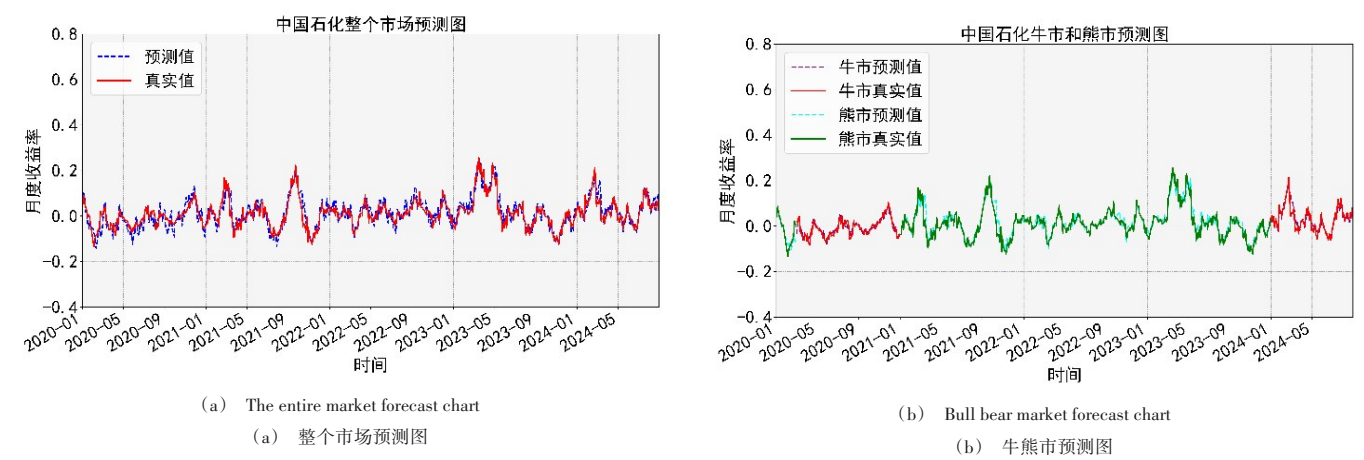


Fig. 3 Sinopec forecast chart

图3 中国石化预测图

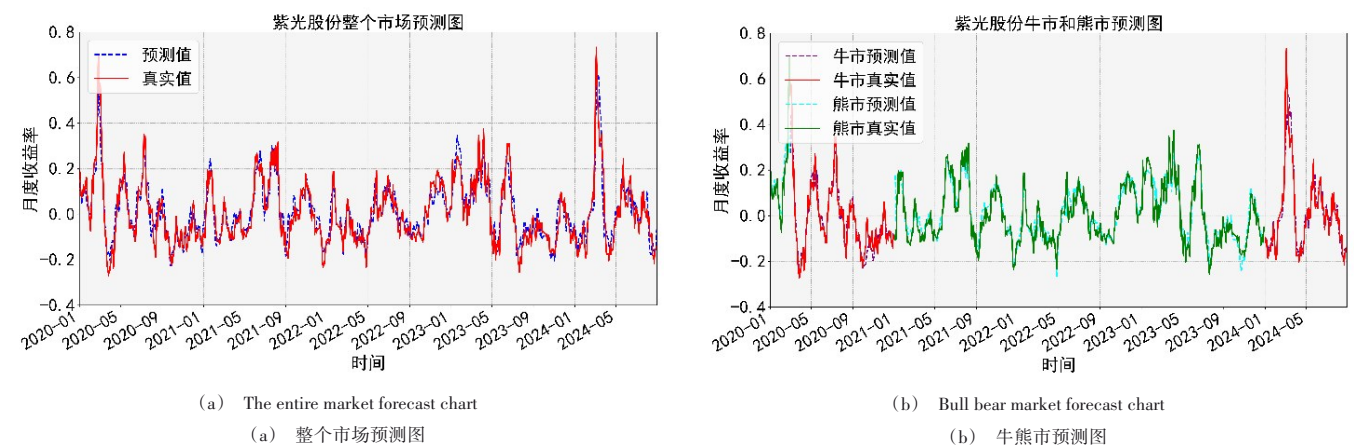


Fig. 4 Ziguang stock prediction chart

图4 紫光股份预测图

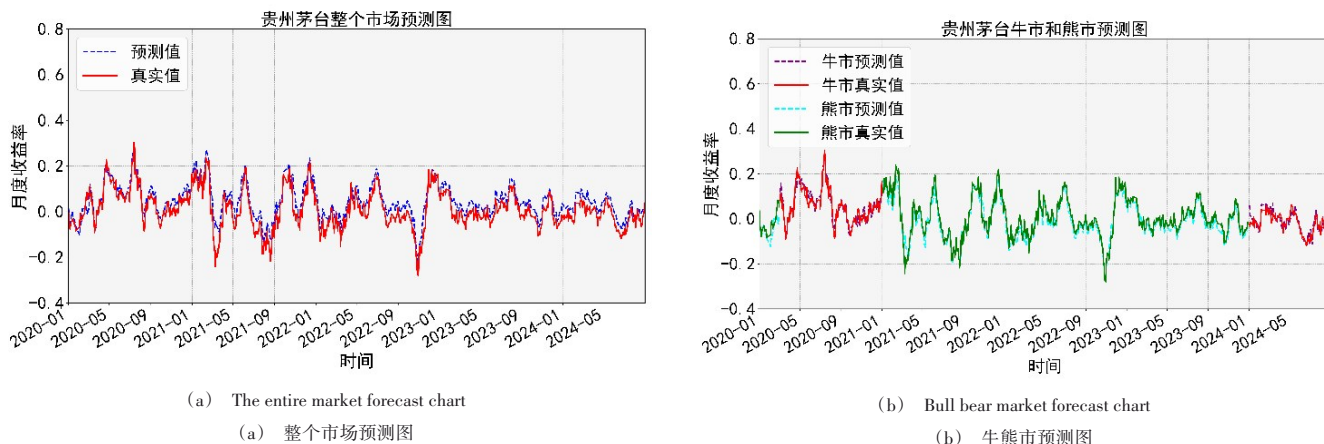


Fig. 5 Kweichow Moutai forecast chart

图5 贵州茅台预测图

于全市场,其中,熊市的拟合效果最佳。

综上,市场状态划分显著提升了模型的预测性能。尤其在牛熊市中,模型能够更精准地捕捉收益率变化趋势;相对而言,全市场条件下的预测效果则较弱。

Table 2 Comparison of prediction performance metrics for each stock model under different market regimes

表2 不同市场状态下各股票模型预测性能指标对比

股票	市场状态	RMSE	MAPE	R ²
招商银行	牛市	0.002 708 5	0.004 474 1	0.914 64
	熊市	0.003 749 9	0.006 235 5	0.903 81
	整个市场	0.004 722 1	0.007 507 6	0.790 65
中国石化	牛市	0.002 074 1	0.034 293 6	0.894 86
	熊市	0.002 250 4	0.038 338 5	0.902 49
	整个市场	0.002 907 4	0.057 910 4	0.804 38
紫光股份	牛市	0.001 448 8	0.076 690 4	0.891 17
	熊市	0.001 814 4	0.054 695 1	0.917 77
	整个市场	0.001 659 2	0.080 353 7	0.863 25
贵州茅台	牛市	0.003 719 1	0.058 690 2	0.878 65
	熊市	0.003 542 5	0.056 328 5	0.887 69
	整个市场	0.005 867 1	0.131 354 0	0.736 31

3.3.2 不同时间窗口的实验结果

本文旨在探讨不同滑动窗口长度对牛熊市划分效果及LSTM模型预测性能的影响,并进一步阐明如何通过实验确定最优的窗口参数。

(1)采用高低点标记法划分市场趋势。在沪深300指数价格序列上,以滑动窗口长度为 j 的窗口滑动,通过识别局部极值点来判断趋势拐点,从而实现对牛市和熊市的区分。为全面评估窗口长度的影响,选取了6个月(短期)、12个月(中期)和24个月(长期)3个典型周期进行对比实验。选择依据具体包括:①短期窗口。对市场阶段性变化反应迅速,能够识别更细粒度的波动,适用于短期策略;②中期窗口。金融实践中常用周期,平衡了短期扰动与长期趋势,具有一定通用性;③长期窗口。强调趋势

稳定性,适合观察宏观走势,减少过度波动的干扰。

在每个窗口下完成牛熊市划分后,基于划分结果构建了LSTM模型。通过对招商银行等4只代表性股票的月度收益率进行预测,分别考察牛市、熊市及整体市场3种状态下的预测效果。模型结构、参数及数据划分方法与第3.3.1节保持一致。

(2)实验结果与分析。图6—图8展示了不同窗口长度下的牛熊市划分效果。由图可知,短期窗口在趋势拐点识别上更为敏感,而长期窗口则更为平滑,但可能产生延迟反应。

不同窗口下的模型评估指标汇总如表3所示。从结果分析可知:在RMSE与MAPE指标上,6个月窗口在多数股票的牛市与熊市阶段均表现最佳,说明该窗口能够更准确捕捉市场波动;在R²指标上显示,6个月窗口下模型拟合度普遍更高,例如,紫光股份在熊市下的R²高达0.921 84,优于12个月和24个月窗口;与未区分市场状态的整体预测相比,将市场划分为牛市与熊市进行建模,显著提升了模型的预测效果。

综上,6个月的滑动窗口在牛熊市划分与预测效果之间达到了最优平衡,其较短的周期有助于捕捉市场趋势变化,从而提升模型预测能力。实验结果表明,高低点标记



Fig. 6 Short term market segmentation chart of bulls and bears

图6 短期窗口牛熊市划分图

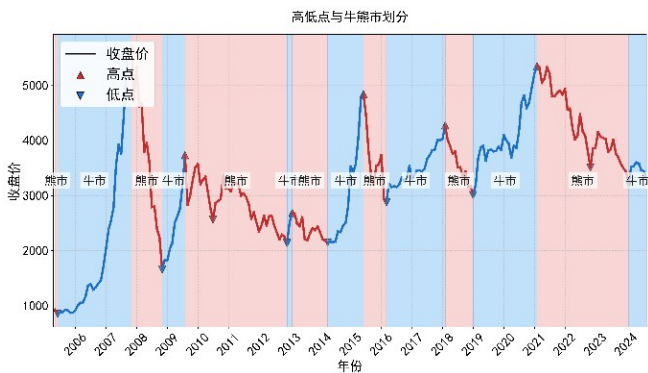


Fig. 7 Mid term market segmentation chart of bulls and bear
图 7 中期窗口牛熊市划分图

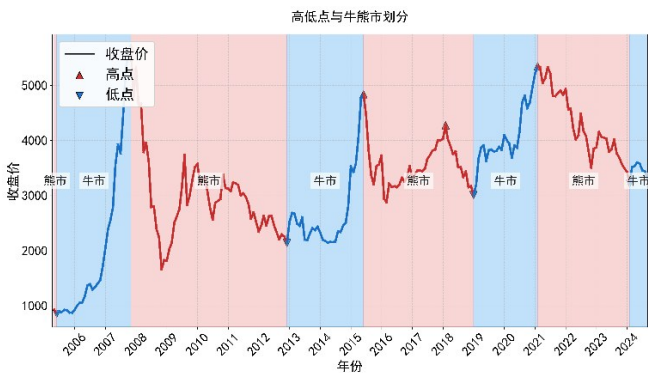


Fig. 8 Long term market segmentation chart of bulls and bears
图 8 长期窗口牛熊市划分图

法中的滑动窗口参数应基于预测性能并通过实证对比来确定,而非依据经验设定。这一结论为后续相关研究提供了方法论上的参考与指导。

4 结语

本文提出了一种基于牛市和熊市划分的股票收益率预测方法,旨在通过结合市场阶段性波动特征,从而提升机器学习模型的预测性能。通过将市场划分为牛市和熊市两个阶段,并分别针对不同市场状态筛选出对股票收益率具有显著影响的因子,构建分市场预测模型。实验结果表明,与整个市场模型相比,分市场模型在预测精度、拟合效果等方面有显著提升。这一研究证实了市场划分策略对于机器学习模型在股票收益预测中的有效性,表明在不同市场环境下使用分市场的预测模型,能够更好地捕捉市场变化,从而提升预测准确性。

未来研究可以进一步深入探讨更多复杂的市场分类方法,如细化市场的阶段划分或引入更多的市场环境因素。同时,随着时间的推移和更多数据的积累,也可以应用这一方法来进行跨市场的比较研究,探索不同国家或地区市场中牛熊市划分的适用性及其对股票收益预测的影响。上述探索方向将为量化投资中的股票收益预测提供更精准、更具动态适应性的解决方案。

Table 3 Evaluation indicators for each window
表 3 各窗口的评估指标

股票	时间窗口	市场状态	RMSE	MAPE	R ²
招商银行	6个月	牛市	0.002 137 5	0.034 152 1	0.909 36
		牛市	0.002 111 4	0.033 302 2	0.909 64
		牛市	0.002 656 9	0.057 632 9	0.899 94
	12个月	熊市	0.003 575 8	0.058 043 9	0.916 91
		熊市	0.003 571 4	0.060 425 1	0.902 62
		熊市	0.003 606 1	0.059 528 8	0.909 94
	24个月	整个市场	0.003 945 7	0.060 616 1	0.843 72
		牛市	0.001 220 2	0.065 049 3	0.858 19
		牛市	0.005 344 4	0.085 292 8	0.841 02
紫光股份	6个月	牛市	0.003 208 2	0.079 652 1	0.834 52
		熊市	0.000 925 1	0.052 054 3	0.921 84
		熊市	0.001 474 1	0.091 982 9	0.805 87
	12个月	熊市	0.001 025 4	0.058 476 8	0.902 42
		整个市场	0.001 592 9	0.088 923 1	0.858 61
		牛市	0.002 498 4	0.045 829 4	0.866 61
	24个月	牛市	0.002 641 9	0.049 090 5	0.848 38
		牛市	0.002 710 2	0.068 699 7	0.842 41
		熊市	0.002 630 2	0.047 211 4	0.919 94
贵州茅台	6个月	熊市	0.006 045 1	0.118 925 1	0.773 28
		熊市	0.004 441 2	0.097 016 7	0.881 93
		整个市场	0.004 541 9	0.100 504 3	0.846 09
	12个月	牛市	0.002 573 3	0.043 769 8	0.852 51
		牛市	0.002 666 6	0.045 631 5	0.831 44
		牛市	0.002 895 7	0.057 461 0	0.848 87
	24个月	熊市	0.002 294 5	0.037 974 9	0.901 61
		熊市	0.002 290 1	0.039 124 7	0.905 64
		熊市	0.002 183 8	0.038 041 0	0.909 95
中国石化	6个月	整个市场	0.002 883 2	0.056 532 5	0.820 67
		牛市	0.002 573 3	0.043 769 8	0.852 51
		牛市	0.002 666 6	0.045 631 5	0.831 44
	12个月	牛市	0.002 895 7	0.057 461 0	0.848 87
		熊市	0.002 294 5	0.037 974 9	0.901 61
		熊市	0.002 290 1	0.039 124 7	0.905 64
	24个月	熊市	0.002 183 8	0.038 041 0	0.909 95
		整个市场	0.002 883 2	0.056 532 5	0.820 67
		牛市	0.002 573 3	0.043 769 8	0.852 51

参考文献:

[1] FAMA E F. Efficient capital markets: a review of theory and empirical work[J]. The Journal of Finance, 1970, 25(2): 383-417.

[2] HOSEINZADE E, HARATIZADEH S. CNNpred: CNN-based stock market prediction using a diverse set of variables[J]. Expert Systems with Applications, 2019, 129: 273-285.

[3] LU W, LI J, LI Y, et al. A CNN-LSTM-based model to forecast stock prices[C]//Proceedings of the 2020 International Conference on Artificial Intelligence and Big Data (ICAIBD), 2020: 58-63.

[4] WANG C, CHEN Y, ZHANG S, et al. Stock market index prediction using deep Transformer model[J]. Expert Systems with Applications, 2022, 200: 116935.

[5] DENG F Y, CEN S Q, ZHONG F Q, et al. Short term price trend prediction based on LSTM neural network[J]. Computer System Applications, 2021, 30(4): 187-192.

邓飞燕, 岑少琪, 钟凤琪, 等. 基于 LSTM 神经网络的短期价格趋势预测[J]. 计算机系统应用, 2021, 30(4): 187-192.

[6] ABDUL Q M, SANJIT K, CHRIS J A V, et al. Novel optimization approach for stock price forecasting using multi-layered sequential LSTM

- [J]. *Applied Soft Computing*, 2023, 134: 109830.
- [7] WANG D G, MA H R, LIANG Y, et al. Stock price prediction based on ATLG hybrid model [J]. *Computer System Applications*, 2023, 32 (3): 123-130.
- 王德广, 马恒锐, 梁叶, 等. 基于 ATLG 混合模型的股票价格预测 [J]. *计算机系统应用*, 2023, 32(3): 123-130.
- [8] DWIVEDI S A, ATTRY A, PAREKH D, et al. Analysis and forecasting of time-series data using S-ARIMA, CNN and LSTM [C]//*Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021: 131-136.
- [9] PARMAR K, SINGH S, KUMAR J. Soft computing models coupled with statistical models estimate the future of the stock market [J]. *Neural Computing and Applications*, 2021, 33: 7629-7647.
- [10] XIANG Y. Using ARIMA-GARCH model to analyze fluctuation law of international oil price [J]. *Mathematical Problems in Engineering*, 2022, 11: 3936414.
- [11] YIN L, LI B, LI P, et al. Research on stock trend prediction method based on optimized random forest [J]. *CAAI Transactions on Intelligent Technology*, 2021, 8: 274-284.
- [12] AWAD A L, ELKAFFAS S M, FAKHR M W. Stock market prediction using deep reinforcement learning [J]. *Applied System Innovation*, 2023, 6(6): 106.
- [13] MENKHOFF L. The use of technical analysis by fund managers: international evidence [J]. *Journal of Banking & Finance*, 2010, 34 (3): 563-577.
- [14] CANDELON B, PIPLACK J, STRAETMANS S. On measuring synchronization of bulls and bears: the case of East Asia [J]. *Journal of Banking & Finance*, 2008, 32(5): 959-970.
- [15] HARDING D, PAGAN A. Dissecting the cycle: a methodological investigation [J]. *Journal of Monetary Economics*, 2002, 49(2): 365-381.
- [16] PARAMITA S, JAFAR B, SIREGAR W I. Market timing and stock selection performance of mutual fund in bull and bear market condition [J]. *International Journal of Monetary Economics and Finance*, 2017, 10 (4): 309-321.
- [17] LI X, ZAKAMULIN V. Stock volatility predictability in bull and bear markets [J]. *Journal of Forecasting*, 2020, 39(8): 1149-1167.
- [18] TRAN V P D, LIN S H, LAI H H, et al. Macroeconomic variables for predicting bear stock markets of Taiwan and China [J]. *Journal of Forecasting*, 2020, 39(5): 675-688.
- [19] FAMA E F, FRENCH K R. Common risk factors in the returns on stocks and bonds [J]. *Journal of Financial Economics*, 1993, 33(1): 3-56.
- [20] FAMA E F, FRENCH K R. A five-factor asset pricing model [J]. *Journal of Financial Economics*, 2015, 116(1): 1-22.
- [21] ANG A, HODRICK R J, XING Y, et al. The cross-section of volatility and expected returns [J]. *The Journal of Finance*, 2006, 61 (1): 259-299.

(责任编辑:陈维捷)