第5卷 第1期 计 量 经 济 学 报 Vol. 5, No. 1
2025年1月 China Journal of Econometrics Jan., 2025

doi: 10.12012/CJoE2024-0425

基于阈值距离加权损失的股票收益预测

孔艳蕾 [1,2], 秦祎辰 [3], 李 扬 [1,2]

(1. 中国人民大学应用统计科学研究中心, 北京 100872; 2. 中国人民大学统计学院, 北京 100872;
3. 辛辛那提大学运筹商业分析与信息系统系, 辛辛那提 45221)

摘 要 股票收益预测的准确性对投资决策具有关键影响. 随着深度学习模型的引入, 收益预测的性能得到了显著提升. 但是, 现实世界中的股票序列常常包含多种模式的异常值, 这些异常值在一定程度上扭曲了数据的关键统计指标, 掩盖了序列的真实趋势, 削弱了深度学习模型的预测能力, 严重时可能导致后续投资决策的失误. 鉴于此, 本文在考虑异常值存在的场景和梯度下降算法的学习特性的基础上, 提出了一种新颖的损失函数: 阈值距离加权损失 (threshold distance weighted loss, TDW). 该损失通过为样本施加差异化的权重, 有效减少了模型对异常值的敏感度. 模拟研究和实证分析验证了该方法的预测准确性和稳健性, 证明其能够为投资组合带来稳定的正向收益, 并为实际金融投资决策提供有力支持.

关键词 股票收益预测; 稳健损失函数; 深度序列模型; 量化投资策略

Stock Return Prediction Based on Threshold Distance Weighted Loss

KONG Yanlei[1,2], QIN Yichen[3], LI Yang[1,2]

(1. Center for Applied Statistics, Renmin University of China, Beijing 100872, China; 2. School of Statistics, Renmin University of China, Beijing 100872, China; 3. Department of Operations, Business Analytics, and Information Systems, University of Cincinnati, Cincinnati 45221, USA)

Abstract The accuracy of stock return prediction has a critical impact on investment decisions. The advent of deep learning models has markedly improved the accuracy of return forecasts. However, stock market sequences are often observed with anomalies that can distort key statistical measures, obscure the true trends of the data, and diminish the predictive capabilities of deep learning models. In extreme

收稿日期: 2024-12-10
基金项目: 国家自然科学基金 (72271237)
Supported by National Natural Science Foundation of China (72271237)
作者简介: 孔艳蕾, 博士研究生, 研究方向: 量化投资, 鲁棒深度学习, E-mail: kongyanlei@ruc.edu.cn; 秦祎辰, 美国辛辛那提大学运筹商业分析与信息系统系教授, 博士, 研究方向: 复杂数据分析, 模型不确定性评价与可视化, 临床试验设计等, E-mail: qinyn@ucmail.uc.edu; 通信作者: 李扬, 中国人民大学统计学院教授, 应用统计科学研究中心研究员, 博士, 研究方向: 相关型数据分析, 模型选择与不确定性评价, 潜变量建模, 临床试验设计等, E-mail: yang.li@ruc.edu.cn.

cases, these anomalies can result in erroneous investment decisions. Based on the presence of anomalies and the learning dynamics of gradient descent algorithms, this paper introduces a novel loss function, the threshold distance weighted loss (TDW), which is designed to mitigate the susceptibility of the model to outliers by assigning variable weights to data samples. The TDW loss function has been tested through simulation studies and empirical analysis. These evaluations have confirmed the improved predictive accuracy and robustness of the method, highlighting its potential to deliver consistent positive returns to investment portfolios and to bolster informed financial investment decisions.

Keywords stock return prediction; robust loss; deep sequence model; portfolio construction

1 引言

股票收益预测在金融领域中扮演着举足轻重的角色, 其准确性对于投资决策具有决定性的影响, 同时也是金融风险管理和资产配置的关键因素 (Lin et al., 2011; Barak et al., 2017). 传统预测方法多依赖于统计模型和经济指标, 对因子、收益率及残差等做出一定的分布假设. 然而, 实际数据常常违背这些假设. 例如, 尽管研究中常假设股票收益率服从正态分布, 但实际的股票收益率分布往往具有 “尖峰厚尾” 的特征, 异常值的发生概率高于正态分布所描述的概率, 从而破坏了数据服从正态分布的假设, 导致传统模型在预测真实市场波动时表现不佳. 此外, 随着存储技术的进步, 可供分析的数据规模和数据复杂性显著增加, 传统模型在捕捉复杂序列关系方面显得力不从心. 因此, 机器学习模型逐渐成为建模的主流选择 (袁靖等, 2021). 特别是长短期记忆网络 (long-short term memory network, LSTM) 等深度序列网络, 因其卓越的序列处理能力, 在股票收益率预测中得到了广泛应用 (李斌等, 2019; 韩莹等, 2023; 任佳屹和王爱银, 2023).

股票序列中常常包含大量异常值, 这些异常值可能会削弱深度学习模型的预测性能. 异常值指数据中显著偏离常规模式的点, 主要可以分为点异常, 上下文异常和集体异常三种类型 (Choi et al., 2021). 具体来说, 点异常指少数数据点与其他数据点相比显著偏离的现象 (图 1a). 这种情况可能是由股票价格的飙升或暴跌, 或者交易量的异常增加引起的. 上下文异常指短时间内序列波动与预期的正常模式有所偏离, 导致在时间维度上序列出现漂移现象 (图 1b). 集体异常描述的是一组股票或整个市场在某一时间段内表现出的异常行为 (图 1c). 这种情况下, 同一行业或同一类型的股票可能会同步异常波动. 这些异常值的存在在一定程度上扭曲了数据的关键统计指标, 掩盖了序列的真实趋势. 而深度学习模型由于参数众多, 在训练这类数据时, 容易过拟合到异常值点, 产生预测偏差, 导致模型的泛化能力降低 (Fitters et al., 2021; Bas et al., 2024).

如何有效地处理这些异常值, 以实现稳健而精确的预测, 是一个亟待解决的问题. 异常值检测和处理是一个常用的解决思路. 通常, 先综合利用小波变换 (Wu et al., 2022; Safa et al., 2024))、聚类算法 (朱映秋和张波, 2021; 徐業釗等, 2021), 以及序列预测 (Naidoo and Du, 2022; Iqbal and Amin, 2024) 等技术识别异常值点. 进而针对识别出的异常值点, 特别是点异常模式的数据点, 进行去除或者纠正. 一方面, 这个识别-处理-再预测的过程高度依赖于异

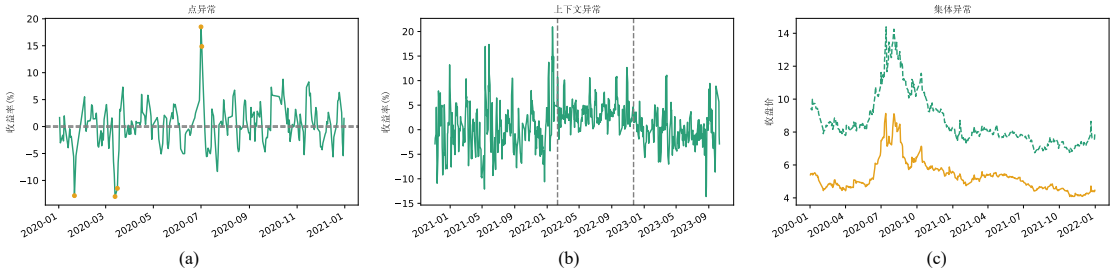

图 1　股票序列数据中的异常类型

常值识别的准确性. 当数据噪声较大时, 检测性能会受到影响. 另一方面, 异常值可以在时间维度或者横截面维度上反映排序信息, 直接去掉会降低样本量, 改变数据分布, 造成信息损失, 这对于股票预测以及基于预测结果进行的量化选股操作是非常不利的. 此外, 真实世界中时间序列数据的多样性导致了多种异常同时存在, 现有检测算法很难有效的同时处理这些不同的异常情况 (Cheng et al., 2024).

稳健损失函数是同时处理多种异常的一个解决思路. 回归分析和深度学习建模中常规使用的损失函数是均方误差 (mean squared error, MSE), 其在数据平稳、遵循高斯分布的时候表现优异, 收敛速度快 (Kumar et al., 2022; Ji et al., 2024). 皮尔逊信息系数 (pearson information coefficient, PIC) 旨在最大化预测值和真实值之间的线性相关性, 也具有类似的性质 (Saidi et al., 2019; Thakkar et al., 2021). 当异常值存在时, 数据的分布假定被破坏, 常用平均绝对误差损失 (mean absolute error, MAE) 进行优化. MAE 对异常值敏感度较低, 结果比较稳健. 但 MAE 存在收敛到局部最优的情况, 无法取得更精确的预测 (Romero, 2018). Fang et al. (2020) 提出了可导斯皮尔曼信息系数 (derivable spearman information coefficient, SIC), 采用平滑函数解决了秩相关系数不可导的问题, 最大化秩相关性, 在非偏分布数据上取得了良好的预测表现. Cheng et al. (2024) 沿着深度学习污染标签的损失函数设计思路, 提出 RobustTSF 方法, 序列存在单点异常时该方法取得了不错的表现, 但对于序列异常的其他情况未做深入讨论. Li et al. (2021); Chen et al. (2022) 等研究者提出了最大 L_q 估计函数, 交替运用极大似然估计及其一阶泰勒展开形式进行样本优化. 这一分段策略显著提高了模型在面对多种数据污染时变量选择的稳健性. 该方法的提出, 为在复杂多变的股票市场中实现稳健预测提供了新的视角和工具.

针对包含多种异常模式的股票数据, 为了提高收益预测任务的表现, 本文提出了一种具有鲁棒性的损失函数. 具体贡献主要包括以下几个方面. 首先, 考虑到异常值在数据中所占比例较小, 结合梯度下降算法沿合力方向优化的特性, 本文提出了阈值距离加权损失 (threshold distance weighted loss, TDW). 该损失通过设定预测值和真实值之间的距离阈值, 对处于阈值范围内的样本赋予固定权重, 以此保持模型对规律样本的学习能力, 减少对异常值的敏感性. 其次, 针对三天累积收益率的预测问题, 本文采用混合高斯分布对因变量进行模拟, 以点异常为主要模拟设置, 兼顾考虑了上下文异常和集体异常, 丰富了模拟场景. 通过模拟试验, 证明了各种场景下阈值距离加权损失的预测精确性和稳健性. 最后, 本文对沪深股市 A 股的实际收益进行了预测, 并基于此构建了投资组合, 验证了 TDW 能够带来稳定的正向

收益, 在实际应用场景中具有一定的优势.

本文的结构安排如下. 第二部分详细介绍了阈值距离加权损失函数, 并阐述了所使用的序列预测模型. 第三部分描述了数据处理和模型设定的过程. 第四部分展示了在不同异常模拟条件下, TDW 的预测表现. 第五部分呈现了 TDW 在实际收益率预测中的有效性. 第六部分是结论与启示.

2 方法

2.1 符号和模型

本研究采用股票的收盘价格计算收益. 假设股票池内共有 n 支股票, 第 i 支股票在第 t 天的收盘价为 P_{it}, 可以得到当日收益: $r_{it} = (P_{it} - P_{i(t-1)})/P_{i(t-1)}$, $i \in \{1, \cdots, n\}$. 记第 t 天开始, 累积 d 天的收益率为 $y_{it}^{(d)}$, 则:

$$y_{it}^{(d)} = r_{it} + r_{i(t+1)} + \cdots + r_{i(t+d)},$$

其中 $d \geq 0$. 假设每只股票都有 p 个指标, 且可采集到各个指标过去连续 k 期的信息, 则因子张量为 $\boldsymbol{X}_t = (\boldsymbol{X}_{t-k}, \cdots, \boldsymbol{X}_{t-2}, \boldsymbol{X}_{t-1})$. 对于 $j \in \{t-k, \cdots, t-1\}$, 有 $\boldsymbol{X}_j = (\boldsymbol{x}_{1j}, \cdots, \boldsymbol{x}_{nj})' \in \boldsymbol{R}^{n \times p}$, 表示 n 支股票在第 j 天的解释变量矩阵. 而 $\boldsymbol{x}_{ij} \in \boldsymbol{R}^p$ 代表第 i 支股票在第 j 天时收集到的 p 维预测变量. 则:

$$\boldsymbol{y}_t^{(d)} = g(\boldsymbol{X}_t, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_t,$$

其中 $\boldsymbol{y}_t^{(d)} \in \boldsymbol{R}^n$, 模型 $g(\cdot)$ 是可以捕捉时间序列动态规律的连续函数, 对应参数为 $\boldsymbol{\theta}$, $\boldsymbol{\epsilon}_t$ 为误差项.

本文采用长短期记忆网络 (long-short term memory network, LSTM) 作为估计模型 $g(\cdot)$. 其主要包含三个门: 输入门, 遗忘门和输出门. LSTM 模型首先通过遗忘门决定是否丢弃过去的数据, 得到 $\boldsymbol{f}_t$; 再结合输入门添加新的信息 $\boldsymbol{i}_t$, 更新记忆单元状态 $\boldsymbol{c}_t$. 同时, 输出门负责计算当前时间步 t 下 LSTM 的输出 $\boldsymbol{o}_t$, 结合 $\boldsymbol{c}_t$ 更新隐状态 $\boldsymbol{h}_t$, 传递给下一个时间步. 用 $(\boldsymbol{W_f}, \boldsymbol{W}_i, \boldsymbol{W}_o, \boldsymbol{W}_c, \boldsymbol{U}_f, \boldsymbol{U}_i, \boldsymbol{U}_o, \boldsymbol{U}_c)$ 表示不同模块的权重矩阵, $(\boldsymbol{b}_f, \boldsymbol{b}_i, \boldsymbol{b}_o, \boldsymbol{b}_c)$ 表示对应的偏置向量参数, σ 表示 sigmoid 激活函数, 则一个时间步下的主要计算流程如下:

$$\begin{aligned}
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}_f \times \boldsymbol{X}_t + \boldsymbol{U}_f \times \boldsymbol{h}_{t-1} + \boldsymbol{b}_f), \\
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}_i \times \boldsymbol{X}_t + \boldsymbol{U}_i \times \boldsymbol{h}_{t-1} + \boldsymbol{b}_i), \\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}_o \times \boldsymbol{X}_t + \boldsymbol{U}_o \times \boldsymbol{h}_{t-1} + \boldsymbol{b}_o), \\
\boldsymbol{c}_t' &= \tanh(\boldsymbol{W}_c \times \boldsymbol{X}_t + \boldsymbol{U}_c \times \boldsymbol{h}_{t-1} + \boldsymbol{b}_c), \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \cdot \boldsymbol{c}_t', \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \cdot \tanh(\boldsymbol{c}_t).
\end{aligned}$$

2.2 损失函数

为了减轻异常值对深度学习模型训练的负面影响, 本文提出了阈值距离加权损失 (threshold distance weighted loss, TDW). 该方法针对数据不满足平稳高斯分布假设的情况, 对传统

的均方误差损失函数进行了创新性优化. 对单个样本, 真实收益表示为 y_i, 预测收益为 $\hat{y}_i$. 给定阈值 $\tau \in \boldsymbol{R}$, 定义 $|y_i - \tau|$ 为阈值距离, 则 TDW 具有如下表达形式:

$$\mathcal{L}_i = \begin{cases} (y_i - \hat{y}_i)^2, & |y_i - \hat{y}_i| \geq |y_i - \tau|, \\ |y_i - \tau||y_i - \hat{y}_i|, & |y_i - \hat{y}_i| < |y_i - \tau|. \end{cases}$$

该损失通过设置距离阈值 $|y_i - \tau|$, 对不同预测距离的样本施加不同的优化力度, 从而调节整体样本的收敛速度. 下面考虑从梯度的角度进行分析. 深度学习模型通过梯度下降算法更新网络参数, 以最小化损失函数. 其优化方向由梯度平均得到: $\bar{g} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathcal{L}_i}{\partial \theta}$. 其中, $\mathcal{L}_i$ 表示每个样本对应的损失函数值, $\frac{\partial \mathcal{L}_i}{\partial \theta}$ 表示损失函数关于参数 θ 求导. 梯度越大的样本, 对 $\bar{g}$ 贡献越多, 越容易被模型学习. 分析阈值距离加权损失的梯度. 当 $|y_i - \hat{y}_i| \geq |y_i - \tau|$, 有:

$$\frac{\partial \mathcal{L}_i}{\partial \theta} = -2(y_i - \hat{y}_i)\nabla G_i.$$

而当 $|y_i - \hat{y}_i| < |y_i - \tau|$,

$$\frac{\partial \mathcal{L}_i}{\partial \theta} = (-1)^{\boldsymbol{I}[|y_i - \hat{y}_i| \geq 0]} \times |y_i - \tau| \nabla G_i.$$

其中, ∇G_i 表示 $\hat{y}_i$ 关于参数 θ 求导得到的导数.

在训练过程中, 阈值距离加权损失通过参数 τ 的选取, 对样本施加差异化的权重, 有效抑制异常值对模型收敛速度的影响. 具体而言, 对于接近 τ 的观测点, 其预测误差通常较大, 因此更倾向于采用均方误差损失函数进行优化. 相反, 对于距离 τ 较远的样本 y_i, 则更多地采用加权平均绝对误差, 并赋予固定的权重参与训练. 对于那些与 τ 距离适中的样本, 即阈值距离与预测距离的相对大小容易波动的样本, 可以根据预测表现自动选择相应的权重参与训练. 参数 τ 的取值位置在一定程度上反映了模型对序列的关注区域. 观测值越接近 τ, 模型越可能采用均方误差的形式进行优化, 实现对应样本的快速学习和收敛. 因此, τ 的取值应当适中, 以减少模型对异常值的关注, 尤其是对正数异常值的关注, 从而在保持模型对关键信息的敏感性的同时, 增强对异常值的鲁棒性.

针对阈值距离与预测距离相对大小容易波动的样本, 阈值距离加权损失能够有效维持单个样本的梯度参与度. 考虑样本 y_i 的训练动态. 样本训练的早期阶段, 通常满足 $|y_i - \hat{y}_i| \geq |y_i - \tau|$. 此时样本学习尚不充分, 预测值距离真实值差距较远, 需要较大的权重参与梯度平均, 以促进对应样本的收敛. 而当 $|y_i - \hat{y}_i| < |y_i - \tau|$, 说明模型对样本 i 已具备稳定的预测能力, 继续采用均方误差损失训练, 将导致 $(y_i - \hat{y}_i) \to 0$, $\frac{\partial \mathcal{L}_i}{\partial \theta} \to 0$, 使得该样本的梯度贡献越来越小. 因而, 改用加权的平均绝对误差损失, 样本 i 的梯度权重为阈值距离 $|y_i - \tau|$, 始终保持大于 0, 因此在参与反向传播的过程中维持样本的参与度, 防止模型遗忘样本规律, 降低预测偏倚.

3 数据和模型

3.1 研究数据

本研究的数据来源于开源数据平台 tushare. 采集对象为沪深股市 A 股日频交易的量价数据, 涵盖时间段从 2015 年 1 月 1 日至 2023 年 10 月 11 日. 通过对每支股票年平均成交

额的计算, 本研究筛选出平均成交额超过一千万人民币的股票纳入分析, 以此确保所分析的股票具备相当的规模和经营稳定性.

本文聚焦于三天累积收益的预测问题, 即考虑 $d = 3$, 因变量为 $\boldsymbol{y}_t^{(3)} = \boldsymbol{r}_t + \boldsymbol{r}_{t+1} + \boldsymbol{r}_{t+2}$. 采集到的原始数据共包含 8 个量价指标, 包括开盘价 (open)、最高价 (high)、最低价 (low)、收盘价 (close)、成交量 (volume)、成交额 (amount)、成交量加权平均价格 (vwap) 和换手率 (turnover rate). 为了丰富预测变量, 对上述基本指标进行大规模数据变换, 最终得到了 147 个潜在预测因子. 所有因子依据新版申万行业分类进行了行业中性化处理. 在模型训练阶段, 如果直接使用所有因子, 由于变量之间的冗余, 可能会降低模型的预测性能. 因此, 需要先从候选因子中筛选出重要因子, 再将这些重要因子用于模型构建. 计算上述候选因子的 SHAP 值 (Scott and Lee, 2017) 并排序, 选出排名靠前的指标. 最终 26 个解释变量被纳入预测模型. 表 1 详细报告了 26 个因子的具体变换公式及对应的数据分布情况. 特别地, $\mathrm{Corr}(\boldsymbol{x}_t, \boldsymbol{y}_t^{(3)})$ 展示了各个预测变量与因变量 $\boldsymbol{y}_t^{(3)}$ 在研究区间内的平均线性相关性. 分析显示, 在未进行行

表 1 预测变量描述性统计

	均值	标准差	偏度	峰度	$\mathrm{Corr}(\boldsymbol{x}_t, \boldsymbol{y}_t^{(3)})$
sin(vwap)	0.6046	0.3648	−1.2068	1.0495	0.0032
arctan(close)	1.4519	0.0886	−1.9201	9.3749	−0.0012
slog1p(turnover)	1.0116	0.6313	0.9252	0.4224	−0.0620
logdiff(high)	−0.0003	0.0261	−1.0512	63.0692	0.0009
logdiff(close)	−0.0001	0.0280	−1.2229	67.5722	0.0082
logdiff(open)	−0.0001	0.0287	−1.0704	56.1662	0.0045
logdiff(amount)	−0.0038	0.4338	0.5111	2.5192	−0.0171
logdiff(vwap)	−0.0001	0.0248	−1.7786	109.2321	0.0105
logdiff(vol)	−0.0033	0.4259	0.5056	2.6749	−0.0181
logdiff(low)	−0.0001	0.0253	−2.0097	121.5221	0.0094
$\mathrm{ts\_rank(high)}_{10}$	0.5281	0.3324	0.0778	−1.4921	−0.0213
$\mathrm{ts\_rank(open)}_{10}$	0.5437	0.3279	0.0294	−1.4783	−0.0187
$\mathrm{ts\_rank(vwap)}_{10}$	0.5430	0.3354	0.0291	−1.5032	−0.0178
$\mathrm{ts\_rank(low)}_{10}$	0.5569	0.3304	−0.0180	−1.4842	−0.0185
$\mathrm{ts\_skewness(low)}_{10}$	−0.0003	0.6596	0.0120	0.3653	−0.0128
$\mathrm{ts\_skewness(vwap)}_{10}$	0.0502	0.6505	0.0801	0.2784	−0.0154
$\mathrm{ts\_ir(low)}_{10}$	46.1200	29.7890	1.2872	3.0218	0.0123
$\mathrm{ts\_ir(vwap)}_{10}$	46.4307	29.9154	1.2445	2.6314	0.0191
$\mathrm{ts\_std(close)}_{10}$	0.4370	0.4592	6.6817	280.6276	−0.0174
close − vwap	0.2807	2.4492	18.6044	622.4051	−0.0107
high − close	0.0078	0.1132	56.9756	8693.3831	−0.0185
low − vwap	0.2731	2.3684	18.1605	585.2265	−0.0094
high − vwap	0.2885	2.5244	18.9334	649.3893	−0.0111
low − close	−0.0076	0.1216	−51.7076	5847.5272	0.0271
open − close	−0.0005	0.1200	−7.8820	4251.1013	0.0111
open − low	0.0072	0.1078	50.6435	6408.1076	−0.0176

业中性化之前, 因子的数量分布存在差异, 并且不符合正态分布特征. $\mathrm{Corr}(\boldsymbol{x}_t, \boldsymbol{y}_t^{(3)})$ 显示单个变量与因变量的相关性不强, 这意味着单个变量对因变量的预测能力有限, 需要综合多个因子的信息来共同提升预测准确性.

针对三天累积收益的预测任务, 数据将按照滚动窗口的方式进行模型训练. 从 2015 年开始, 每六年构成一个完整的训练预测周期, 逐年向后递进, 共有 4 个完整周期. 在每个周期中, 前四年的数据用于模型训练; 紧随其后的一年数据用于模型验证, 辅助参数调整; 最后一年数据进行样本外预测. 测试年份为 2020 年初到 2023 年 10 月中旬. 表 2 详细列出了每个周期的数据划分情况.

为了评估数据集中因变量异常值的比例, 本文参考华泰证券处理异常值的方法对收益率序列的异常值水平进行了评估. 具体操作为: 在时刻 t, 假设一个预测变量在所有股票上的观察序列为 $\boldsymbol{y}_t$. 其中, y_m 为该横截面序列的中位数. $\boldsymbol{y}_t - y_m$ 的中位数为 y_{m1}. 设定 f 表示指定离散倍数, 定义 $\tau_u = y_m + f y_{m1}$, $\tau_l = y_m - f y_{m1}$. 将满足 $\{y_{it} < \tau_l\} \cup \{y_{it} > \tau_u\}$ 的观测集合定义为异常值集. 通过计算该异常值集中包含样本占全部样本的比例, 可以了解收益率序列的异常分布情况.

表 2 展示了不同训练周期下因变量的异常值比例. 随着 f 值的增大, 异常值定义变得更加严格, 异常值的比例也相应减少. 通常取 $f = 5$ 进行分析. 在这种情况下, 异常值数量占比超过 1%, 可能对模型的训练造成干扰. 此外, 可以观察到, 在同一个训练周期内, 训练集的异常值比例高于验证集和测试集. 而在不同的训练周期中, 训练集的异常值呈现逐渐下降的趋势, 说明可能出现数据分布随时间进程发生了改变, 存在上下文异常数据, 在模型构建过程中会潜在的降低样本外预测表现. 因此, 这样的数据集训练需要一个稳健的损失函数, 以降低模型对异常值的敏感性.

表 2　不同训练周期下因变量异常值比例

数据类型	开始日期	结束日期	总样本量	异常比例		
				$f=3$	$f=5$	$f=7$
训练集	20150101	20181231	2659269	0.1256	0.0376	0.0139
验证集	20190101	20191231	783740	0.0939	0.0150	0.0018
测试集	20200101	20201231	776687	0.1038	0.0157	0.0019
训练集	20160101	20191231	2883238	0.1067	0.0203	0.0041
验证集	20200101	20201231	776687	0.1038	0.0157	0.0019
测试集	20210101	20211231	771340	0.1183	0.0122	0.001
训练集	20170101	20201231	3028843	0.1036	0.0175	0.0026
验证集	20210101	20211231	771340	0.1183	0.0122	0.0010
测试集	20220101	20221231	770842	0.1009	0.0121	0.0015
训练集	20180101	20211231	3084957	0.1036	0.015	0.0019
验证集	20220101	20221231	770842	0.1009	0.0121	0.0015
测试集	20230101	20231011	575112	0.1053	0.0137	0.0013

3.2 对比方法和评价指标

将本文提出的损失函数与以下方法进行对比: 均方误差损失 (mean squared error, MSE), 平均绝对误差损失 (mean absolute error, MAE), 皮尔逊信息系数 (pearson information coefficient, PIC), 可导斯皮尔曼信息系数 (derivable spearman information coefficient, SIC). 其中, MAE 和 SIC 具有稳健性. 而 MSE 和 PIC 对异常值敏感.

模型得到的预测值应当对未来收益具有一定的预测能力. 因此, 本文采用日均秩相关系数 rankIC 和日均线性相关系数 IC 作为评价准则. 两者的定义如下:

$$\text{rankIC} = \frac{1}{T}\sum_{t=1}^{T}\left(1 - \frac{6\sum_{i=1}^{n_t}(y_{it} - \hat{y}_{it})^2}{n_t(n_t^2 - 1)}\right),$$

$$\text{IC} = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{i=1}^{n_t}(\hat{y}_{it} - \bar{\hat{y}}_{it})(y_{it} - \bar{y}_{it})}{(n_t - 1)\sqrt{D(\hat{y}_{it})}\sqrt{D(y_{it})}}.$$

二者的绝对值越大, 表明预测值对股票收益的预测能力越强. 值得注意的是, IC 衡量的是预测序列和真实序列之间的线性相关性, 在数据满足正态分布的情况下度量更为准确. rankIC 比较的是两序列的排名顺序相关性. 在量化研究领域, 尤其关注横截面上股票预测排名和实际排名的一致性, 以便后续通过选股构建稳定盈利的投资组合. 因此, 从辅助选股的角度出发, 本文利用验证集上的 rankIC 表现作为模型设置和参数调整的评价指标.

同时考虑对预测误差进行评价, 计算预测值和真实值之间的 MSE、MAE 和 R^2. 由于各个损失函数的优化目标不同, 故每个指标挑选的最优方法并不一致. 且不同的指标规模不同, 不可直接比较. 基于此, 考虑对各个指标采用秩次进行综合评价. 首先对每个方法计算各个指标. 进而针对单个指标, 对五个损失函数进行排名, 该排名即为损失函数对应的秩次. 排名越小表示在该指标下方法表现越优秀. 最后将 RankIC、IC、MSE、MAE 和 R^2 下的秩次加和, 得到秩和得分, 即多个指标的总排名分数. 分数越小, 代表综合表现越好. 在下文的工作中, 本文将通过总排名分数来评估各个方法的预测误差.

3.3 模型设定

本文采用同样的模型架构和参数配置, 评估不同损失函数的性能. 经过参数调整, 本文采用了单层 LSTM 模型, 隐藏层的神经元数量为 64. 设定滞后期 $k = 20$, 则 LSTM 模型的输入窗口宽度为 20. 输出维度设定为 1. 同时, 采用 Adam 优化算法, 权重衰减系数设定为 0.0001, 学习率设定为 0.001, 每个训练批次的样本量设定为 128.

4 模拟研究

4.1 因变量模拟

本文针对因变量, 即三天累积收益 $\boldsymbol{y}_t^{(3)}$ 进行模拟. 首先利用真实因变量得到预测因变量. 针对数据集 $\mathcal{D} = \{\boldsymbol{y}_t^{(3)}, \boldsymbol{X}\}$, 考虑利用三层结构的多层感知机 (multilayer perceptron, MLP) 进行训练. 当滞后期参数 $k = 20$, 有 $\hat{\boldsymbol{y}}_t^{(3)} = \text{MLP}(\tilde{\boldsymbol{X}})$. 其中, $\tilde{\boldsymbol{X}}$ 为 $\boldsymbol{X}$ 二维展开得到的因子矩阵.

模拟主要针对点异常进行. 采用混合高斯分布对 $\hat{\boldsymbol{y}}_t^{(3)}$ 添加异常值. 假设数据量为 n, 异常比例为 η, 则随机抽取 $n\eta$ 个样本, 添加非对称高斯噪声. 剩余样本则作为正常样本, 添加白噪声. 在生成时考虑上下文异常的情形, 设定异常值在训练集和验证集的离散程度高于测试集的离散程度. 故训练验证集施加 $N_{o1}(\mu_1,\sigma_1^2)$ 噪声, 测试集施加 $N_{o2}(\mu_2,\sigma_2^2)$ 噪声. 其中, $\mu_1>\mu_2,\sigma_1^2>\sigma_2^2$. 该类生成异常值的方式称为单一噪声法, 具体地施加情况如表 3 所示.

此外, 考虑集体异常的情形. 异常值序列在同一个行业更有可能呈现相似的规律, 而不同行业之间差异显著. 针对这种情况, 考虑设定一组异常分布, 按照行业从中随机抽取一个分布, 行业内样本噪声均服从该分布. 本文参照申万一级行业分类对股票进行划分, 按上述混合噪声策略进行数据生成 (表 3).

结合异常值统计表 2, 异常比例 η 分别取 0.005、0.01、0.04、0.07、0.1、0.2. 设定 2020 年 1 月 1 日作为 $N_{o1}(\mu_1,\sigma_1^2)$ 和 $N_{o2}(\mu_2,\sigma_2^2)$ 的生成分界线, 按上述步骤对 $\hat{\boldsymbol{y}}_t^{(3)}$ 加噪得到模拟因变量, 采用滚窗训练策略进行训练和测试.

在模拟实验中, 首先考虑参数的选取策略. 阈值参数 τ 决定了模型关注序列数据的关键区域. 本研究考虑采用横截面序列的分位数作为 τ 的候选值, 进而针对横截面序列中真实收益排名靠前的样本, 评估不同阈值的预测效果, 以此来探究损失函数在捕捉高排名样本方面的效能. 进一步地, 本研究在多种异常情况下, 对不同的损失函数进行了全面的比较分析, 以此探究阈值距离加权损失函数的适用场景.

表 3　异常序列模拟设定

	组别	$N_{o1}(\mu_1,\sigma_1^2)$	$N_{o2}(\mu_2,\sigma_2^2)$
单一噪声	S1	$N(10,10^2)$	$N(5,5^2)$
	S2	$N(10,20^2)$	$N(5,10^2)$
	S3	$N(5,5^2)$	$N(3,3^2)$
	S4	$N(5,10^2)$	$N(3,5^2)$
混合噪声	S5	$[N(10,10^2),N(10,20^2),N(20,20^2),$ $N(20,50^2),N(50,50^2),N(50,100^2)]$	$[N(2,2^2),N(2,5^2),N(5,5^2),$ $N(5,10^2),N(10,10^2),N(10,20^2)]$

4.2 模拟结果分析

表 4 展示了在 S1 生成的各类异常水平下, 采用不同分位数阈值训练的模型在测试集上的表现. 比较模型在不同异常水平下的表现. 可以看到, 随着异常值比例上升, 模型整体性能出现下降趋势. 而在特定的异常水平下, rankIC 和 IC 指标随着分位数的提高呈现出先增后减的趋势. 当阈值设定在 70% 至 90% 分位数之间时, 模型预测达到最佳表现. 然而, 随着分位数阈值进一步上升, 超过 90% 分位数时, rankIC 和 IC 指标下降. 尽管如此, 其表现在大多数情况下依然优于采用 30% 分位数阈值的模型. 这一观察验证了 τ 在反映模型对序列关注区域方面的重要性. 当训练目标是在横截面序列前半部分实现更好的预测时, 应当考虑 50% 以上的分位数作为阈值. 同时, 分位数选取不可过高, 否则模型会更多关注正向异常值, 反而削弱模型的整体性能.

表 5 展示了五种损失函数在各种异常值场景下的预测表现. 其中, 阈值距离加权损失 $\mathcal{L}_{70}$ 采用了横截面 70% 分位数进行模型训练和预测. 可以看到, $\mathcal{L}_{70}$ 在各类设置下相较于对

表 4 不同阈值类型下模型的预测表现

阈值	rankIC				IC			
	0.01	0.04	0.07	0.1	0.01	0.04	0.07	0.1
99	0.0289	0.0332	0.0279	0.0274	0.0319	0.0318	0.0268	0.0247
90	0.0337	0.0338	0.0274	0.0262	0.0364	0.0343	0.0265	0.0240
80	0.0347	0.0341	0.0280	0.0281	0.0372	0.0349	0.0267	0.0252
70	0.0352	0.0347	0.0309	0.0280	0.0385	0.0355	0.0293	0.0251
50	0.0337	0.0345	0.0296	0.0277	0.0371	0.0352	0.0273	0.0248
30	0.0301	0.0305	0.0271	0.0253	0.0347	0.0307	0.0268	0.0220

表 5 不同异常水平下的各个损失函数的对比结果

	损失函数	rankIC						IC					
		0.005	0.01	0.04	0.07	0.1	0.2	0.005	0.01	0.04	0.07	0.1	0.2
S1	$\mathcal{L}_{70}$	0.0377	0.0352	0.0347	0.0309	0.0280	0.0230	0.0420	0.0385	0.0355	0.0293	0.0251	0.0178
	MSE	0.0369	0.0348	0.0339	0.0304	0.0272	0.0221	0.0413	0.0385	0.0350	0.0286	0.0247	0.0174
	MAE	0.0365	0.0338	0.0333	0.0304	0.0274	0.0210	0.0401	0.0373	0.0345	0.0284	0.0247	0.0168
	SIC	0.0362	0.0324	0.0332	0.0304	0.0278	0.0225	0.0396	0.0337	0.0337	0.0285	0.0241	0.0169
	PIC	0.0362	0.0332	0.0341	0.0281	0.0271	0.0228	0.0397	0.0350	0.0340	0.0248	0.0235	0.0178
S2	$\mathcal{L}_{70}$	0.0367	0.0365	0.0333	0.0308	0.0288	0.0220	0.0405	0.0408	0.0335	0.0288	0.0261	0.0159
	MSE	0.0363	0.0362	0.0332	0.0304	0.0284	0.0185	0.0400	0.0400	0.0332	0.0281	0.0257	0.0142
	MAE	0.0355	0.0350	0.0318	0.0305	0.0287	0.0216	0.0392	0.0390	0.0318	0.0284	0.0260	0.0158
	SIC	0.0353	0.0336	0.0313	0.0291	0.0283	0.0200	0.0385	0.0375	0.0312	0.0270	0.0245	0.0142
	PIC	0.0354	0.0352	0.0314	0.0285	0.0288	0.0205	0.0389	0.0385	0.0311	0.0259	0.0253	0.0140
S3	$\mathcal{L}_{70}$	0.0369	0.0369	0.0356	0.0330	0.0302	0.0235	0.0405	0.0409	0.0367	0.0328	0.0286	0.0206
	MSE	0.0357	0.0362	0.0350	0.0320	0.0295	0.0218	0.0403	0.0407	0.0364	0.0327	0.0282	0.0202
	MAE	0.0343	0.0330	0.0328	0.0302	0.0261	0.0185	0.0389	0.0381	0.0348	0.0312	0.0261	0.0164
	SIC	0.0343	0.0356	0.0341	0.0317	0.0296	0.0221	0.0382	0.0393	0.0346	0.0318	0.0279	0.0202
	PIC	0.0357	0.0359	0.0342	0.0309	0.0296	0.0227	0.0389	0.0394	0.0350	0.0310	0.0277	0.0204
S4	$\mathcal{L}_{70}$	0.0375	0.0364	0.0338	0.0317	0.0293	0.0254	0.0412	0.0404	0.0344	0.0311	0.0270	0.0208
	MSE	0.0373	0.0362	0.0320	0.0307	0.0290	0.0229	0.0412	0.0402	0.0336	0.0305	0.0264	0.0184
	MAE	0.0351	0.0351	0.0329	0.0309	0.0288	0.0249	0.0391	0.0389	0.0344	0.0308	0.0263	0.0203
	SIC	0.0364	0.0358	0.0321	0.0287	0.0280	0.0238	0.0389	0.0396	0.0328	0.0286	0.0245	0.0191
	PIC	0.0372	0.0346	0.0332	0.0307	0.0280	0.0226	0.0407	0.0374	0.0343	0.0301	0.0250	0.0171
S5	$\mathcal{L}_{70}$	0.0366	0.0360	0.0337	0.0292	0.0296	0.0227	0.0407	0.0401	0.0344	0.0281	0.0274	0.0175
	MSE	0.0356	0.0357	0.0326	0.0280	0.0284	0.0211	0.0397	0.0391	0.0337	0.0273	0.0261	0.0168
	MAE	0.0347	0.0347	0.0326	0.0273	0.0279	0.0189	0.0388	0.0380	0.0335	0.0269	0.0260	0.0157
	SIC	0.0351	0.0351	0.0330	0.0282	0.0289	0.0221	0.0372	0.0381	0.0338	0.0269	0.0264	0.0171
	PIC	0.0355	0.0354	0.0323	0.0280	0.0286	0.0218	0.0388	0.0385	0.0331	0.0270	0.0257	0.0167

比方均有优势.

S1 异常类型下, 异常比例越小, 则方法的预测表现越好. 在每一个异常比例下, $\mathcal{L}_{70}$ 的 rankIC 和 IC 指标均优于其他方法. 值得注意的是, 尽管 PIC 和 SIC 旨在最优化信息系数,

它们并未在测试集中实现最高的 IC 和 rankIC 值. MSE 在异常值比例较低 ($\eta < 0.1$) 时都是次优表现. 而 MAE 作为稳健损失函数, 在各个比例下未显出明显优势. S2、S3 和 S4 的模拟设定下, 预测表现也遵循类似的规律. S5 是由多个高斯分布组合而成的, 它模拟了行业间异常值分布的多样性. 即使在这类复杂的异常情况下, $\mathcal{L}_{70}$ 依然展现了稳定的预测性能. 图像2进一步强调了 TDW 在预测任务中的优越性. 该图展示了在 S5 类型的不同异常值水平下, 五种方法的总得分排名. 在所有情况下, TDW 得分最低. 而 PIC 和 SIC 在多种异常条件下的得分较高. 同时, MSE 和 MAE 方法排名居中, 且随着异常比例升高, 它们的次序也有所升高. 综合图表分析, MAE 缺乏预测优势, MSE 在低异常水平下表现较好. PIC 和 SIC 对异常值较为敏感, 稳定性不足. 相比之下, TDW 在 rankIC, IC 和指标得分方面均显出了优势, 这充分说明了当序列存在异常时, 阈值距离加权损失训练模型的有效性和稳定性.

图 3 展示了在不同的异常情况下, 相比于其他损失函数, 阈值距离加权损失 (TDW) 在预测表现中的绝对优势大小. 该图计算了 TDW 与其他次优损失函数的 rankIC 差值. 差值越大, 表明在该特定情况下, TDW 相较于其他方法的优势越显著. 相比于 S2, TDW 在 S1 下展现出更明显的优势. 而 S3 比 S1 呈现出更高的净优势. 这表明在单一噪声模式下, 异常值的过度分散会对模型性能产生负面影响, 各类损失函数的优化能力可能失效, 预测表现趋于一致. 而在 S5 情形下, TDW 始终保持着较为明显的预测优势. 说明在复杂的混合异常机制下, 尤其是数据中存在集体异常时, 阈值距离加权损失函数相较于其他方法, 能够实现更为稳健和精确的预测.

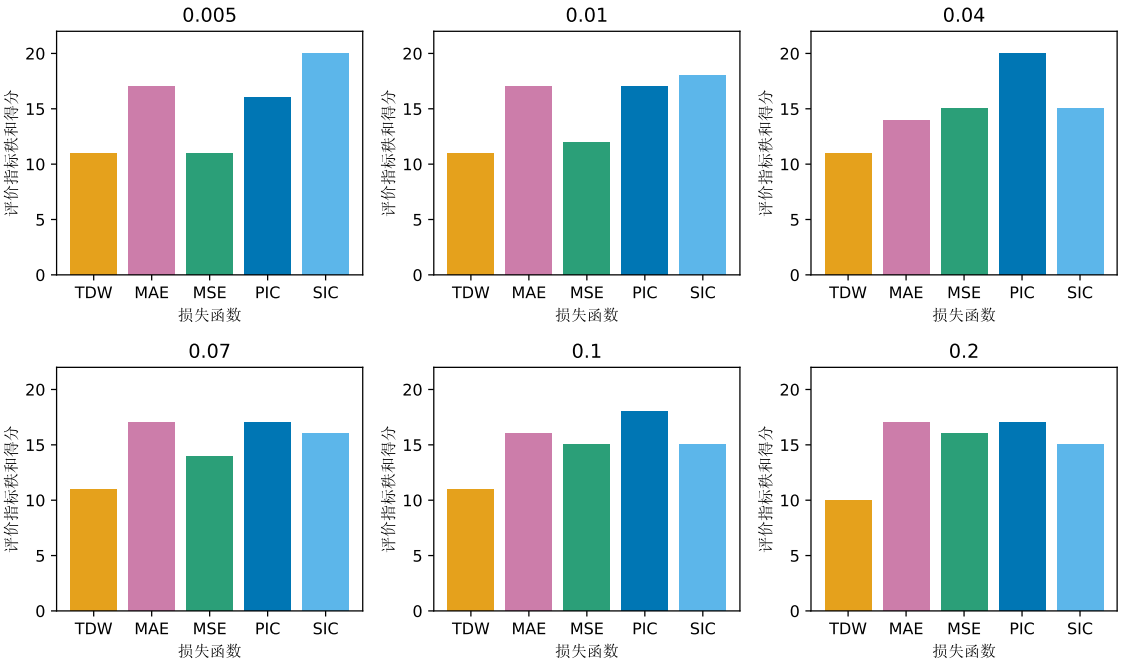

图 2　不同异常水平下的模型评价指标总排名得分

综合以上分析, 阈值距离加权损失在面对多样化的异常情况以及不同比例的异常值时, 均显示出了优秀的预测效果和稳定性, 表明该方法在缓解模型对异常值的敏感度, 提升预测

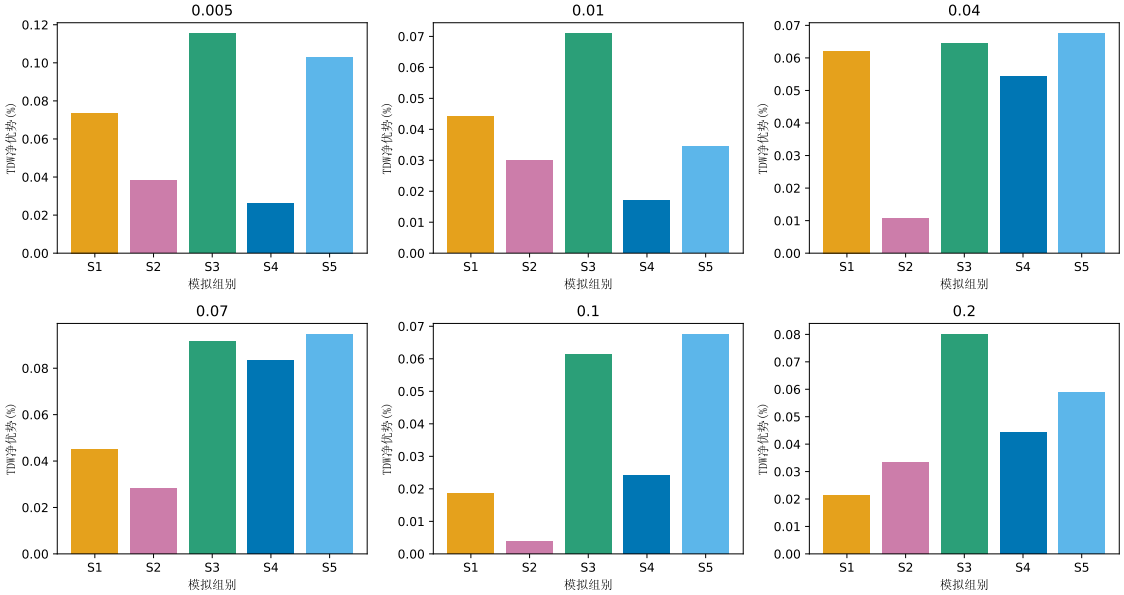

图 3　不同异常类型下阈值距离损失和次优损失的 rankIC 差值

稳健性方面具有良好的效果, 充分说明了其在不同数据特性和异常分布条件的使用灵活性和有效性.

5　实证研究

本文利用实际市场数据对真实的三日累积收益率进行预测. 研究首先聚焦于分位数阈值的选取效果. 通过细致地分析不同时间周期内的数据表现, 本文提出了一种自适应阈值策略. 该策略能够根据市场动态调整阈值, 以期在不断变化的市场条件下优化预测性能. 随后, 本文对这一策略的预测效果进行了全面的评估, 以确保其在实际应用中的有效性. 进一步地, 运用预测所得的收益率来构建投资组合. 具体而言, 在使用长短期记忆网络模型得出收益率估计值后, 依据这些预测值对股票进行排序, 选择排名最高的 $r\%$ 的股票作为投资组合的构成部分. 在选股的基础上, 计算该组合的关键绩效指标, 包括年化收益率、年化波动率、夏普比率以及最大回撤. 这些指标为评估投资组合表现提供了全面的视角.

5.1　实际数据中阈值 τ 的影响

在模拟研究中, 同一支股票在整个时间序列中遵循一致的生成模式. 这使得在整个分析期间应用一个统一的分位数阈值能够实现令人满意的预测效果. 然而, 在真实股票收益率预测中, 数据的异常模式更为复杂, 且概念漂移的现象时有发生, 可能导致不同时间周期内最优阈值的选择出现差异.

为了验证该问题, 本文采取了一种基于滚动预测周期的年度评估机制, 发现不同年份适用的分位数阈值的确存在不一致. 如图 4 所示, 在测试集上分别计算 IC、rankIC 以及依据前 10% 选股策略构建投资组合所获得的年化收益率和夏普比率. 每个面板中绘有 4 条曲线, 分别代表了相应测试周期的表现. 曲线中的每个数据点分别对应横轴上的一个分位数阈值. 可

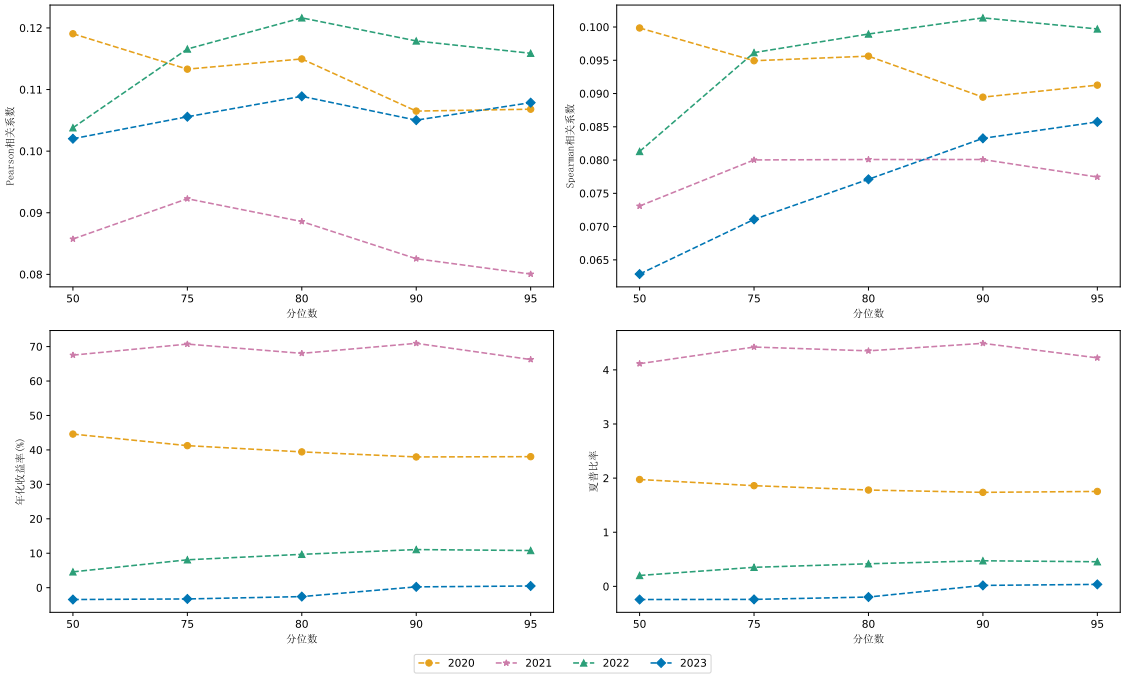

图 4　不同分位数阈值在 2020 年至 2023 年测试集上的表现

以明显观察到, 每年度的曲线走势并非单调模式, 且波动的规律也不尽相同. 这表明不同年份的最优分位数阈值各异, 暗示了不同测试周期内数据分布的变异性. 因此, 每个滚动周期应采用各自的最佳阈值. 此外, 尽管 rankIC 的整体水平略低于 IC, 但其最佳分位数点与年化收益率, 夏普比率的最佳分位数阈值吻合, 而 IC 的趋势则不总是如此. 这表明 rankIC 能够较好地反映投资组合的表现, 可作为在验证集上确定最优阈值的评估指标.

基于验证集 rankIC 的表现, 本研究为每个测试周期选定了最佳阈值. 2020 至 2023 年分别采用 50%、75%、80%、90% 分位数计算阈值距离, 并将该策略与仅使用单一分位数阈值的策略进行比较, 结果如表 6 所示. 其中, $\mathcal{L}_{\text{adj}}$ 代表上述自适应阈值策略, $\mathcal{L}_{50}$、$\mathcal{L}_{75}$、$\mathcal{L}_{80}$ 和 $\mathcal{L}_{90}$ 分别对应固定使用 50%、75%、80% 以及 90% 分位数作为阈值的损失函数. 研究结果清晰地表明, $\mathcal{L}_{\text{adj}}$ 在 rankIC 和 IC 水平上均超越了其他基于单一分位数阈值的策略. 投资组合表现上, 尽管各种策略在年化波动率和最大回撤方面的表现相差无几, 但 $\mathcal{L}_{\text{adj}}$ 在年化收益率和夏普比率上展现出优势, 说明该方案能够带来更丰富的正向收益. 同时可以看出, 由于数据分布在时间维度上的分布差异性, 根据验证集为每个测试周期选取最佳阈值的策略是有效的. 通过这种细致入微的动态策略调整, 能够确保模型在不断变化的市场环境中保持最佳的预测性能和投资决策能力.

5.2 不同损失函数的比较分析

如表 6 所示, 阈值距离加权损失 (TDW) 取得的 IC 水平高于其他损失函数. 特别是 PIC 损失函数, 旨在最大化皮尔逊相关性, 在 IC 值预测上拥有较高的起点. 即便如此, TDW 在 IC 水平上依然展现出了竞争力. 而 MSE、MAE 和 SIC 在 IC 表现上明显不如 TDW. 而在

表 6 不同损失函数下的预测和投资组合表现

损失函数	IC	rankIC	年化收益率 (%)	年化波动率 (%)	夏普比率	最大回撤 (%)
$\mathcal{L}_{\text{adj}}$	0.1098	0.0911	30.6630	19.4909	1.5732	20.4063
$\mathcal{L}_{90}$	0.1029	0.0889	28.5602	19.4794	1.4662	20.3643
$\mathcal{L}_{80}$	0.1085	0.0887	27.9097	19.4231	1.4369	20.4633
$\mathcal{L}_{75}$	0.1071	0.0866	28.2851	19.5149	1.4494	20.6919
$\mathcal{L}_{50}$	0.1027	0.0804	27.4362	19.7554	1.3888	20.6690
MSE	0.0986	0.0794	24.9308	20.6594	1.2068	21.5091
MAE	0.0804	0.0759	21.3275	19.8724	1.0732	21.0061
SIC	0.0900	0.0792	26.6673	19.6777	1.3552	20.6775
PIC	0.1014	0.0805	23.3096	20.9281	1.1138	21.4855

rankIC 的表现上, TDW 的优势更为显著. 其他方法的 rankIC 在 0.08 左右, TDW 大部分稳定在 0.086 以上, 显示出其在预测排名上的能力. 特别是 $\mathcal{L}_{\text{adj}}$, 与对比损失函数相比, 提升幅度超过 1%.

基于不同模型得到的收益率估计进行排序, 选择前 10% 的股票构建投资组合, 回测结果如表 6 所示. 无论 TDW 方法采取何种分位数阈值, 其投资组合得到的年化收益率均超过 27%, 显著高于其他损失函数. 夏普比率呈现出类似的优势, 保持在 1.38 以上, 而其他方法均在 1.36 以下. 尽管 TDW 的年化波动率和最大回撤受阈值的影响较小, 变化范围不大, 但与其他损失函数函数相比, 仍保持了轻微的优势.

总体而言, $\mathcal{L}_{\text{adj}}$ 在各个关键指标上都展现出了不错的投资潜力. 无论是在预测精度、投资组合的年化收益率, 还是在风险调整后的回报上, 都证明了其作为一种高效预测工具的价值. 这一全面的分析不仅验证了 TDW 方法的有效性, 也为投资者提供了一个强有力的工具, 以实现更优的投资决策.

5.3 $\mathcal{L}_{\text{adj}}$ 策略下的投资组合表现

在采纳了 $\mathcal{L}_{\text{adj}}$ 策略并获取了收益率预估之后, 本文进一步分析了选股比例对投资组合表现的影响. 此处覆盖了从 1% 到 50% 的不同选股比例, 用以构建多样化的投资组合, 具体效果如表 7 所示. 结果显示, 年化收益率和夏普比率随着选股比例的减少而呈现出近乎单调递增的趋势. 这表明精选较少数量的股票能够带来更高的年化收益率和更佳的风险调整后回报. 然而, 这种高收益的背后也隐藏着更高的风险. 随着选股比例的降低, 年化波动率也随之增大. 例如, 选择排名在前 1% 的股票时, 波动率可达 21%. 最大回撤走势较为稳定. 当选股比例超过 40%, 最大回撤超过 22%. 综合考虑, 当选股比例维持在 5% 至 20% 的区间时, 各项指标能够保持在较优的水平, 实现收益与风险的平衡. 基于这个发现, 本文采用前 10% 的选股比例进行其他分析, 旨在在确保较高年化收益的同时, 保持相对稳定的年化波动率. 图 5 展示了采用 $\mathcal{L}_{\text{adj}}$ 策略选股的回测收益. 本文通过将回测累积收益与相应年份的中证 1000 (ZZ1000)、中证 500 (ZZ500) 和沪深 300 (HS300) 指数收益进行对比, 揭示了 $\mathcal{L}_{\text{adj}}$ 策略的选股优势. 从对比曲线中可以明显看出, 采用 TDW 训练的模型在长期投资中不仅能够实现稳定收益, 而且在与主要市场指数的比较中显示出了明显的选股优势.

表 7　不同选股比例下的投资组合表现

选股比例 (%)	年化收益率 (%)	年化波动率 (%)	夏普比率	最大回撤 (%)
1	36.3517	21.3965	1.6990	19.5657
5	30.6197	19.8430	1.5431	20.2858
10	30.6630	19.4909	1.5732	20.4063
20	28.1399	19.5681	1.4381	21.6301
30	25.2548	19.4995	1.2952	21.8875
40	22.9186	19.4770	1.1767	22.4125
50	20.7698	19.4782	1.0663	22.5216

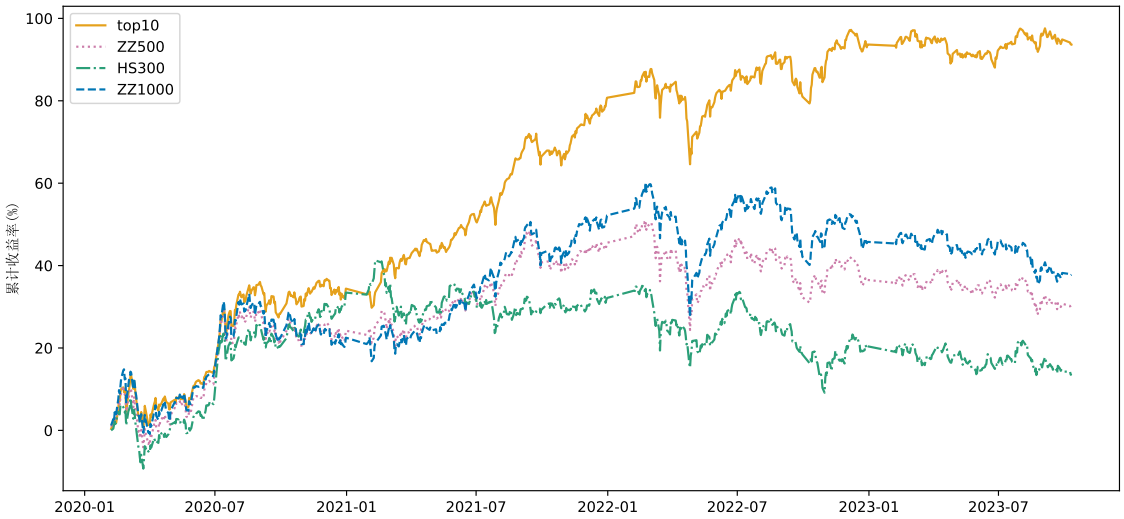

图 5　$\mathcal{L}_{\text{adj}}$ 策略在 2020 年至 2023 年的累积收益

通过这些细致的分析, 我们不仅进一步验证了 TDW 的有效性, 而且展示了 $\mathcal{L}_{\text{adj}}$ 策略在不同选股比例下的特点, 为投资者提供了在追求最大化收益的同时如何有效控制风险的宝贵见解. 这能够为投资者在构建投资组合时提供科学的决策依据, 有助于实现更为稳健和高效的资产管理.

6　结论与启示

本研究聚焦于股票收益率预测领域, 特别关注金融时间序列中异常值对神经网络训练的干扰, 以及这些干扰如何削弱模型的预测效能并增加构建投资组合的难度. 为解决这一问题, 本文提出了阈值距离加权损失函数 (TDW), 并将其应用于长短期记忆网络 (LSTM) 中, 以减少模型对异常值的敏感性, 并提高预测性能. 研究基于沪深股市 A 股日频交易的量价数据, 聚焦于三天累积收益的预测任务. 在研究的初始阶段, 首先对 8 个基础量价指标进行数据预处理和数据变换, 创建了一个庞大的候选因子库. 随后, 通过计算候选因子的 SHAP 值, 筛选出 26 个关键变量, 这些变量在后续的 LSTM 模型训练中发挥了重要作用. 通过最小化 TDW 损失函数, 完成模型训练和预测. 最后根据预测结果, 选取排名靠前的股票构建投资组

合. 分析结果清晰地显示, 与其他的损失函数相比, 本文提出的 TDW 损失在面对存在异常值的数据时, 能够显著提升模型的预测和选股能力, 同时保持了一定的预测稳定性. 在滚动预测方案中, 通过为不同测试周期定制不同的阈值距离, 可以进一步提高模型的表现. 实际数据分析显示, 阈值距离加权损失能够带来更丰厚的年化收益和夏普比率, 同时维持相对较低的波动率, 展现出在长期投资中的卓越市场表现.

尽管本研究在存在异常值的股票收益率预测方面取得了一定的进展, 但仍有许多领域值得未来进一步探索. 例如, 本文采用 SHAP 值评估变量重要性并选择关键的预测变量, 这一过程不仅耗时, 而且需要人工经验的辅助. 探索如何在模型训练过程中自动进行变量选择, 并对因子进行有效解释, 提高因子利用效率, 是一个关键的研究方向. 其次, 由于数据类型的限制, 本文只考虑了日频数据建模的效果. 而高频因子数据量庞大, 通常蕴含更丰富的信息. 如何利用神经网络模型拟合高频因子, 并从更复杂的异常机制中捕捉出序列规律, 是深度学习领域面临的另一大挑战. 此外, 如何更加精确的定义和量化异常值, 以及如何对不同类型的异常值进行针对性处理, 也是未来研究的重要课题.

参 考 文 献

韩莹, 张栋, 孙凯强, 谈昊然, 陆超, (2023). 结合长短时记忆网络和宽度学习的股票预测新模型研究 [J]. 运筹与管理, 32(8): 187–192.

Han Y, Zhang D, Sun K Q, Tan H R, Lu C, (2023). Research on a New Stock Price Prediction Model Combining LSTM and BLS[J]. Operations Research and Management, 32(8): 187–192.

李斌, 邵新月, 李玥阳, (2019). 机器学习驱动的基本面量化投资研究 [J]. 中国工业经济, (8): 61–79.

Li B, Shao X Y, Li Y Y, (2019). Research on Machine Learning Driven Quantamental Investing[J]. China Industrial Economics, (8): 61–79.

任佳屹, 王爱银, (2023). 融合因果注意力 Transformer 模型的股价预测研究 [J]. 计算机工程与应用, 59(13): 325–334.

Ren J Y, Wang A Y, (2023). Causal Attention Transformer Model for Stock Price Prediction[J]. Computer Engineering and Applications, 59(13): 325–334.

徐業釗, 张晗, 冯贯昂, 韩立言, 庞华, (2021). 基于系统聚类法的基金资产配置策略研究 [J]. 中国商论, (17): 102–104.

Xu Y Z, Zhang H, Feng G A, Han L Y, Pang H, (2021). Research on Fund Asset Allocation Strategy Based on System Clustering Method[J]. China Journal of Commerce, (17): 102–104.

袁靖, 刘响, 董雅菁, 马腾, (2021). 机器学习方法拟合宏观经济变量优于传统统计方法吗——基于经济政策不确定指数的拟合 [J]. 统计理论与实践, (12): 9–14.

Yuan J, Liu X, Dong Y J, Ma T, (2021). Does Machine Learning Outperform Traditional Statistical Methods in Fitting Macroeconomic Variables: An Examination Based on the Economic Policy Uncertainty Index[J]. Statistical Theory and Practice, (12): 9–14.

朱映秋, 张波, (2021). 基于已实现波动率的上证综指异常时序检测 [J]. 系统工程理论与实践, (3): 625–635.

Zhu Y Q, Zhang B, (2021). Time Series Anomaly Detection of Shanghai Composite Index Based on Realized Volatility[J]. Systems Engineering — Theory & Practice, (3): 625–635.

Barak S, Arjmand A, Ortobelli S, (2017). Fusion of Multiple Diverse Predictors in Stock Market[J]. Information Fusion, 36: 90–102.

Bas E, Egrioglu E, Cansu T, (2024). Robust Training of Median Dendritic Artificial Neural Networks for Time Series Forecasting[J]. Expert Systems with Applications, 238: 122080.

Chen J, Bie R, Qin Y, Li Y, Ma S, (2022). L_q-based Robust Analytics on Ultrahigh and High Dimensional Data[J]. Statistics in Medicine, 41(26): 5220–5241.

Cheng H, Wen Q, Liu Y, Sun L, (2024). RobustTSF: Towards Theory and Design of Robust Time Series Forecasting with Anomalies[J]. arXiv Preprint: 2402.02032.

Choi K, Yi J, Park C, Yoon S, (2021). Deep Learning for Anomaly Detection in Time-series Data: Review, Analysis, and Guidelines[J]. IEEE Access, 9: 120043–120065.

Fang J, Lin J, Xia S, Xia Z, Hu S, et al. (2020). Neural Network-based Automatic Factor Construction[J]. Quantitative Finance, 20(12): 2101–2114.

Fitters W, Cuzzocrea A, Hassani M, (2021). Enhancing LSTM Prediction of Vehicle Traffic Flow Data via Outlier Correlations[C]// 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2021: 210–217.

Iqbal A, Amin R, (2024). Time Series Forecasting and Anomaly Detection Using Deep Learning[J]. Computers & Chemical Engineering, 182: 108560.

Ji Y, Luo Y, Lu A, Xia D, Yang L, et al. (2024). Galformer: A Transformer with Generative Decoding and a Hybrid Loss Function for Multi-step Stock Market Index Prediction[J]. Scientific Reports, 14(1): 23762.

Kumar D, Sarangi P K, Verma R, (2022). A Systematic Review of Stock Market Prediction Using Machine Learning and Statistical Techniques[J]. Materials Today: Proceedings, 49: 3187–3191.

Li Y, Li R, Qin Y, Lin C, Yang Y, (2021). Robust Group Variable Screening Based on Maximum L_q-likelihood Estimation[J]. Statistics in Medicine, 40(30): 6818–6834.

Lin W Y, Hu Y H, Tsai C F, (2011). Machine Learning in Financial Crisis Prediction: A Survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4): 421–436.

Naidoo V, Du S, (2022). A Deep Learning Method for the Detection and Compensation of Outlier Events in Stock Data[J]. Electronics, 11(21): 3465.

Romero R A C, (2018). Generative Adversarial Network for Stock Market Price Prediction[EB/OL]. [2024-01-07]. https://cs230.stanford.edu/projects_fall_2019/reports/26259829.pdf.

Safa K, Belatreche A, Ouadfel S, Jiang R, (2024). WALDATA: Wavelet Transform Based Adversarial Learning for the Detection of Anomalous Trading Activities[J]. Expert Systems with Applications, 255: 124729.

Saidi R, Bouaguel W, Essoussi N, (2019). Hybrid Feature Selection Method Based on the Genetic Algorithm and Pearson Correlation Coefficient[J]. Machine Learning Paradigms: Theory and Application, 2019: 3–24.

Scott M, Lee S, (2017). A Unified Approach to Interpreting Model Predictions[J]. Advances in Neural Information Processing Systems, 30(2017): 4765–4774.

Thakkar A, Patel D, Shah P, (2021). Pearson Correlation Coefficient-based Performance Enhancement of Vanilla Neural Network for Stock Trend Prediction[J]. Neural Computing and Applications, 33: 16985–17000.

Wu D, Wang X, Wu S, (2022). A Hybrid Framework Based on Extreme Learning Machine, Discrete Wavelet Transform, and Autoencoder with Feature Penalty for Stock Prediction[J]. Expert Systems with Applications, 207: 118006.