

采用机器学习与二维伽马函数的 股票指数量化交易策略

柴昱白, 陈伟, 赵舒欣, 毛新越

(西安交通大学电子与信息学部, 710049, 西安)

摘要: 为了通过短线交易获取对标股票指数的超额收益,提出一种基于机器学习的股票指数增强型量化交易策略,并实现程序化自动交易。首先生成股票指数的初始特征集,通过最大互信息系数法筛选得21维特征;然后设计双阈值涨跌不平衡标签与3种预测分类方式(T、B-B、B-T),组合构建不同机器学习模型,并对比择优得到多空交易方向的最优机器学习模型分别为LSTM-B-T模型与RF-B-T模型;接着设计基于日内涨跌幅的二维伽马函数;最后使用经二维伽马函数计算得到的类概率判别阈值与最优机器学习模型预测的类概率进行涨跌类别判定,得到交易信号。将该策略应用于中证500指数的股指ETF进行回测与模拟盘交易验证,实验结果表明:相较于随机建仓策略,采用该策略使交易评价指标得到整体性提升;在回测与为期3个月的模拟盘交易验证中使用该策略均能获得对标指数的理想超额收益,分别为11.24%、11.08%。

关键词: 股票指数;指数增强;量化交易;程序化自动交易;二维伽马函数;机器学习

中图分类号: F832.51;TP181 **文献标志码:** A

DOI: 10.7652/xjtuxb202305020 **文章编号:** 0253-987X(2023)05-0204-09

Quantitative Trading Strategy for Stock Index Based on Machining Learning and 2D Gamma Function

CHAI Yubai, CHEN Wei, ZHAO Shuxin, MAO Xinyue

(Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: This paper proposes a stock index enhanced quantitative trading strategy based on machine learning and realizes programmed automatic trading to obtain the excess return of the benchmark stock index through short-term trading. First, the initial feature set of the stock index is generated, and the 21-dimensional features are screened by the maximum information coefficient method. Then a double-threshold unbalanced label and three prediction classification methods (T, B-B, B-T) are designed. In addition, different machine learning models are built and compared to obtain the optimal machine learning models for the long-short trading direction: LSTM-B-T model and RF-B-T model respectively. After that, a two-dimensional gamma function based on intraday rise and fall is designed. Finally, ups and downs are determined and trading signals obtained using the class probability discrimination threshold calculated by the two-dimensional gamma function and the class probability predicted by the optimal machine learning models. This strategy is applied to the backtesting and simulation trading verification of the CSI 500

ETF. The experimental results show that, compared with the random position opening strategy, the trading strategy proposed in this paper makes the transaction evaluation index improve on the whole. This strategy can obtain the ideal excess return of the benchmark index in both backtesting and three-month simulation trading verification, which are 11.24% and 11.08% respectively.

Keywords: stock index; exponential augmentation; quantitative trading; programmatic trading; 2D gamma function; machine learning

随着我国证券市场的不断迅猛发展,投资者面临着更加动态多变的投资可行集和市场环境^[1]。因此,制定什么样的投资交易策略始终是一个研究的热点问题。近年来,使用计算机技术进行量化投资的方式为投资者交易提供了新思路。量化交易是投资者使用金融数学模型对证券市场交易数据进行量化分析而完成交易^[2],其中投资择时依靠投资者对资产价格趋势的有效预判,选择适当时机买卖资产,以实现规避风险、获取超额收益^[3]。计算机快速处理数据并计算得出结果的能力正适用于此。越来越多的量化交易者通过程序将交易过程集成化,以程序化自动交易完成整个交易过程。

机器学习方法在金融量化交易中得到了广泛的应用。Lohrmann 等使用随机森林(random forest, RF)对标普 500 指数构建了一种四分类涨跌预测模型,证明此模型比只分涨跌模型更有助于改善交易策略^[4]。王淑燕验证了随机森林量化选股模型在中国股票市场上有较好的性能^[5]。李斌等设计了一套基于机器学习和技术指标的量化投资算法,证明由此构建的投资组合收益率远超大盘指数^[6]。张茂军等设计并研究了一种基于决策树的量化交易择时策略,对比证明机器学习模型使用其构造的区间突破分类标签进行训练具有更好的预测效果^[7]。高妮等使用集成经验模态分解和不敏感支持向量回归构建低频分量预测模型,构建了良好的指数择时策略^[8]。

除了传统机器学习模型,神经网络模型在量化交易方面也得到了广泛研究。Nemes 等将神经网络模型应用于罗马尼亚股票的短期趋势预测,提出了可为投资者交易决策提供预测模型^[9]。Thomas 等发现利用长短期记忆神经网络模型(LSTM)预测非线性金融时序的精度显著优于随机森林等模型^[10]。杨青等构造深层 LSTM 神经网络并对全球 30 个股票指数进行不同期限的预测研究,发现 LSTM 能够有效控制误差波动,泛化能力很强^[11]。

现有的量化交易研究有一定的不足:一是标签构造较为简单,仅基于价格涨跌标记为(1:涨,2:

跌),没有考虑具体涨跌幅数值^[4,7];二是交易策略较为基础,通常仅使用简单建仓策略,并没有使用机器学习信号结合交易实时情况进行更细化的设计^[7-8,12];三是交易验证不充分,现有文献通常仅使用历史数据进行回测实证,并没有使用实时获取的行情数据验证,对于其策略可在真实交易市场中获取理想收益难具说服力。

本文旨在通过探讨不同分类方式下机器学习对于股票指数涨跌的预测效果选择多空两方向表现最好的预测模型,在此基础上设计与机器学习预测信号紧密结合的实际交易策略,然后使用历史数据回测以及最接近真实交易的模拟盘实时数据进行验证分析,为投资者形成一种实际可用的交易策略提供参考。从而将机器学习的方法真正有效运用在金融交易市场中,并为形成稳定理性的股票交易市场提供行之有效的技术支持与智力支持。

1 随机森林和长短期记忆神经网络

1.1 随机森林

随机森林算法是 Bagging 类算法思想的一种改进算法,其采用并行计算的学习方式,各个学习分类器之间相互独立。随机森林通过组合多个决策树进行分类或者回归^[13],得到广泛使用。决策树算法无需数据的归一化处理,简单易行,但是面对繁多冗余的数据存在容易过拟合的缺陷,预测分类能力较差。随机森林通过自举采样和随机子空间分割赋予的双重随机性解决了决策树泛化能力弱的缺点^[14]。

1.2 长短期记忆神经网络

LSTM 是循环神经网络(RNN)的变体^[15],它通过门机制使网络具有长期记忆的功能^[16]并且缓解了 RNN 的梯度问题^[17]。LSTM 由记忆块递归连接而成,每个块的基础结构又包括一个存储单元和 3 个乘法单元。具体来说,乘法单元又根据实现的具体功能分别被称为输入门、输出门以及遗忘门。输入门决定输入信号能否改变存储器单元的状态,其可以短暂存储相关信息,是 LSTM 不同于 RNN 的最大特点^[15]。输出门的功能是决定存储单元能

否改变其他存储单元的状态。遗忘门的功能是决定记住先前的状态还是进行遗忘丢弃,如果上一条网络传递过来的输入数据具体值没有达到对应阈值,则丢弃此数据。LSTM 神经网络基本结构如图 1 所示。

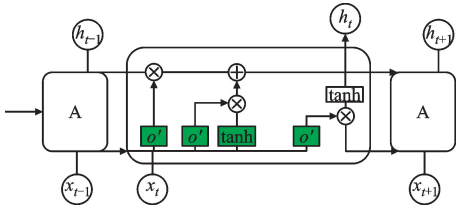


图 1 LSTM 神经网络基本结构

Fig. 1 Basic structure of LSTM neural network

2 量化择时系统设计

2.1 机器学习模型

2.1.1 特征选取

本文所使用的目标数据包括 15 min 线级别股票指数的逐笔资金流数据和价格数据,具体包括 7 个维度:主动卖出额、主动买入额、成交额、开盘价、收盘价、最高价和最低价。每日的交易时间为 240 min,即一天共计 16 根 15 min 线数据。在最终 7 维度数据的基础上,本文通过不同处理方式计算得到了总体的特征指标共 44 维,所用方式基于不同指标的时间周期,不同指标相互间的不同运算而变化。

若输入特征过多会使模型复杂化^[18],所以需要 对计算的特征进行预处理。本文采用最大互信息系数法(MIC)^[19]剔除特征相关性较大的指标,筛选后特征的数量由 44 维降至 21 维。

2.1.2 标签定义

本文提出了一种双阈值涨跌不平衡型标签,其计算方式如下

$$L_i^{(t)} = \begin{cases} 1, & r_{i,\max}^{(t)} > T_b; r_{i,\min}^{(t)} > T_s \\ 2, & r_{i,\min}^{(t)} < -T_b; r_{i,\max}^{(t)} < -T_s \\ 0, & \text{其他} \end{cases} \quad (1)$$

式中: T_b 、 T_s 分别为两个给定阈值,二者区分于数值大小, $r_{i,\max}^{(t)}$ 、 $r_{i,\min}^{(t)}$ 分别为 T 周期内未来收益的最大值与最小值。当 T 周期内未来收益的最大值大于较大的给定阈值 T_b 且其最小值同时大于给定阈值 T_s 时,标记为 1(涨),当 T 周期内未来收益的最小值小于 $-T_b$ 且其最大值仍小于 $-T_s$ 时,标记为 2(跌),而两方向的阈值均未达到时,标记为 0(平)。若出现

两种判定条件都满足的情况时,标记为 0(平),此种情况说明行情波动剧烈,不具有明显的趋势性,为了不给模型增加噪声,将其标记为 0(平)。阈值选择的方式为对训练集进行试探性标注,选取标注后平涨跌三类标签数量比例较为合适的标签阈值 T_b 、 T_s 。经验证,以年为最小单位的不同时期数据集在给定阈值下的标签分布大致相同,本文两阈值的绝对值分别为 0.004 5、0.003 5,打标签后平:涨:跌数量的比例大致为 3:1:1。

使用两个阈值的原因是:让 1(涨)与 2(跌)的判定更加严格,从阈值来看三类别的区分度更高,能够更便于模型进行类别特点的学习;这样的标签在实际交易中具有更高的容错率:因为其严格的判定条件,本在阈值相同型标签中可以标记为 1(涨)或 2(跌)标签的样本,若其收益的小值没有超过小阈值,仍被标记为 0(平),这样就存在 0(平)标签仍可能获利 的情况,更有利于实际的交易。

2.1.3 分类方式

在传统机器学习随机森林算法的尝试中发现,直接进行涨、跌、平(1、2、0)三分类,会将大量样本分类为 0,即平标签,实际上损失了很多可供交易的机会,造成有效交易信号的减少,所以在三分类的基础上本文还进行了 B-B 分类、B-T 分类两次分类,其流程图分别如图 2、图 3 所示。本文的机器学习模型分类方式共有 3 种。

(1)T 方式。直接使用各个模型进行基础的三分类训练。

(2)B-B 方式^[20]。训练集原始标签为 0(平)、1(涨)、2(跌),第一次训练过程为:将其中 1(涨)、2(跌)两类视为一类,标记为新的 1,然后与原始标签 0(平)组合成新的训练集进行训练。第二次训练过程为:只取出原始数据中标签为 1(涨)、2(跌)的样本作为训练集进行训练。

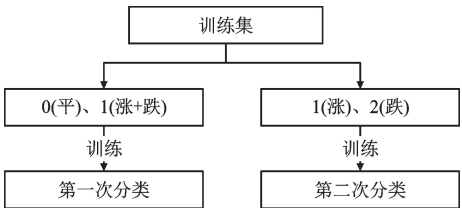


图 2 B-B 分类方式

Fig. 2 B-B classification

(3)B-T 方式。训练集原始标签为 0(平)、1(涨)、2(跌),第一次训练过程为:将原始数据集中标签为 1(涨)、2(跌)的样本取出进行训练,训练方式等同于

B-B方式中的第二次训练方式。第二次训练过程为:将原始数据集样本直接进行涨跌平(1、2、0)三分类。

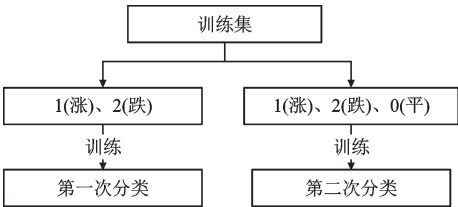


图 3 B-T 分类方式
Fig. 3 B-T classification

2.2 交易策略

简单建仓策略中交易信号只用于开仓,即预测信号为买入时进行买入操作,达到止盈止损阈值时进行卖出。为最大化发挥机器学习模型预测信号在交易过程中的指导作用,将其与交易策略紧密结合,同时考虑到 ETF 交易中手续费较大的问题,本文在简单建仓策略的基础上进行了进一步改进。

2.2.1 持仓中的更新

本交易策略增添了根据正向交易信号进行持仓更新的操作,更新对象包括参考价格、止盈止损价格与此方向的持仓时间。具体执行过程为:若此时持有仓位且交易线程获得开仓正向预测信号,同时,此时获取的最新收盘价已优于上一次更新的参考价格,则更新参考价格为此时的最新收盘价,以参考价格与所定止盈止损阈值比例计算最新的止盈止损价格,更新持仓时间为 0。

2.2.2 开仓阈值自适应调整

涨跌幅限制是为了抑制过度投机行为和从众行为而采取的一种价格稳定机制^[21],此制度是指在一个交易日内证券的交易价格只能在指定的范围内波动^[22]。对于股票指数来说,单日内的涨跌幅不会过于夸张,因此当行情在多(空)方向已有较大涨(跌)幅时,应对其同方向的开仓进行限制以避免行情反转带来的亏损。本文统计所有数据多空方向的日内相对于开盘价的涨跌幅分布统计结果,具体见表 1,统计结果所依据的数据集为表 2 中的神经网络训练集数据。

表 1 中: N_L 为数据集中涨与跌方向的样本数; C_{max} 为此方向绝对值最大的涨跌幅,将数据集中涨跌幅为 25%~75% 的数量分位数所对应的涨跌幅表示为 $C_{25\%} \sim C_{75\%}$ 。

表 1 中证 500 指数单日内涨跌幅分布统计结果

Table 1 Distribution statistics of intraday price of CSI 500 index					
方向	N_L	C_{max}	$C_{25\%}$	$C_{50\%}$	$C_{75\%}$
涨	5 263	0.035 0	0.001 9	0.004 1	0.007 3
跌	4 463	-0.057 2	-0.001 6	-0.003 7	-0.007 4

可以直观得出:统计范围内,中证 500 指数单日内涨跌幅的整体分布集中在百分之一以内,日内最大跌幅大于日内最大涨幅,分别为(涨:0.035 0,跌:-0.057 2)。据此,本文借鉴图像处理领域中一种使用光照分量对光照不均匀图像进行自适应矫正的算法^[23],提出一种根据日内涨跌幅对开仓方向的类概率判别阈值进行自适应调整的二维伽马函数,其公式如下式

$$\begin{cases} P_t = 0.5, & |p_c| \leq |m| \\ P_t = 0.36(1.35)^{\alpha} + 0.01, & |p_c| > |m| \\ \alpha = (1/2)^{|m| - |p_c| / |m|}, & |p_c| > |m| \end{cases} \quad (2)$$

式中: P_t 为此时要求的类概率判别阈值; m 为表 1 中所统计的样本 75% 分割点对应的涨跌幅,此处两方向判别阈值的绝对值均取近似值 0.007 5; p_c 为此时刻获取的最新收盘价相对于今日开盘价的涨跌幅。此函数为分段函数,阈值一般情况下为 0.5,只有在 $|p_c| > |m|$ 时才应用式(2),其对阈值的自适应调整效果如图 4 所示。

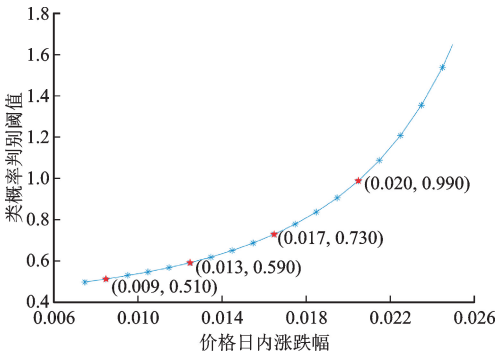


图 4 二维伽马函数校正后的类概率判别阈值
Fig. 4 Output threshold corrected by the 2D Gamma function

图 4 中类概率判别阈值初始为 0.5,以日内涨跌幅正值为例,在 $p_c > 0.007 5$ 时类概率判别阈值与 p_c 呈正相关,其增长趋势为先缓慢增加而后迅速上涨,代表类别判据越来越严格。在机器学习预测过程中,只有在预测概率大于给定判别阈值时才能预测为对应类别,而在横坐标日内涨跌幅 p_c 为 0.02

时,纵坐标类概率判别阈值 P_i 为 0.99,即只有预测的概率大于 0.99 才可判为此类,已接近判别阈值的极限 1。这意味着在 $p_c > 0.02$ 后基本已不存在判为此类的可能,在交易过程中则表示为无此方向的交易信号,不进行持仓更新或开仓。据表 1 所用数据集统计, $p_c > 0.02$ 的样本仅占有所有样本的 1.67%,所以,经此函数调整阈值,只丧失了少于 1.67% 的交易机会,说明此种算法对在交易方向上已出现较大日内涨跌幅值行情时的开仓进行限制的同时,并没有影响实际交易过程中所需要的美好召回率。

3 实验分析

本文先通过机器学习模型的对比分析选定交易策略中多空方向分别使用的机器学习预测模型,然后,为了验证本策略的收益能力,还设计了简易量化交易系统,实现了交易策略的程序化自动交易,不仅可以使使用历史数据进行回测验证,还可以在交易日获取实时行情数据进行模拟盘的实时验证。量化交易系统总体框架如图 5 所示。

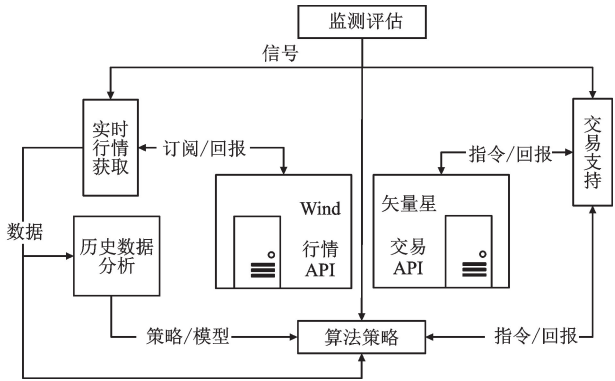


图 5 量化交易系统总体框架
Fig. 5 Overall system framework

3.1 实验数据

本文样本集基于中证 500 指数日内 15 min 线级别数据。随机森林模型数据集共 7 647 条数据。研究表明使用近一年的交易日即约 250 天交易数据进行训练具有良好的效果^[7,10,24],据此,本文选取训练集为:2020 年 1 月 29 日至 2021 年 1 月 29 日共 4 147 条数据。验证集为:2021 年 1 月 29 日至 2021 年 3 月 29 日共 562 条数据。测试集为 2021 年 3 月 29 日至 2021 年 12 月 29 日共 2 938 条数据。

参考文献[10]对比实验,本文设置神经网络模型中深度神经网络(DNN)为 LSTM 网络的参照模

型。因神经网络模型的学习能力依赖于样本数,而本文的 15 min 线级别数据每日只有 16 条数据,一年(250 交易日)约 4 000 条训练数据,数据较少,而神经网络模型的学习性能依赖于更充分的训练样本,所以本文在神经网络模型训练预测中使用时间更久远的历史数据共 9 565 条。其中,训练集:2019 年 7 月 3 日至 2021 年 1 月 29 日共 6 081 条数据,验证集与测试集与随机森林模型相同。

表 2 模型数据集标签分布
Table 2 Label distribution of dataset

模型	样本集	标签频次		
		0	1	2
随机森林	训练集	1 933	1 120	1 094
	验证集	239	145	178
	测试集	1 787	539	612
神经网络	训练集	2 988	1 589	1 504
	验证集	239	145	178
	测试集	1 787	539	612

3.2 模型预测结果

随机森林模型几乎无需调整参数便具有很好的预测性能^[10],因此本文选用模型默认参数进行实验。DNN、LSTM 模型则参考文献[12,25]中的设置,设置结构如下:一个输入层(神经元个数等于输入特征数),两个隐含层(神经元个数分别为 32、16),隐含层间设置 Dropout 层(失活概率 0.1)以防止过拟合,一个输出层(神经元个数等于各分类方式中的目标分类个数)。

表 3 与表 4 是本节所列预测模型分别在验证集与测试集数据中预测表现的评价指标。常用的机器学习评价指标包括精确率、错误率、召回率,考虑到实际交易中除了追求预测的准确性,还需要一定的召回率保证稳定的交易机会,否则,盲目追求预测正确而使召回率过低会导致开仓数量急剧减少,浪费交易机会,因此本文还重点关注召回率。

表 3 验证集结果显示多方向 LSTM-B-T 分类具有最高的精确率 0.307 7,虽然其错误率 0.283 5 比 RF-B-T 分类 0.269 2 略高,但其召回率 0.274 0 是 RF-B-T 分类 0.137 6 的二倍左右,因此认为 LSTM-B-T 分类模型在验证集多方向预测表现最佳;做空方向 DNN 模型预测表现不佳,出现预测精确率为 0 的结果。其原因是由于 DNN-T 分类模型将待测样本全部预测为 0,导致预测精确率为 0,而 DNN-B-T 分类模型的第二步分类是 T 分类,所以

其最终精确率也为 0。其他做空方向模型的预测结果中:RF-B-T 分类的预测精确率 0.485 3,大幅领先其他预测模型且其召回率理想,因此认为 RF-B-T 分类在验证集空方向预测表现最佳。

表 4 测试集结果显示虽然多方向预测 RF-B-T 分类具有最高的精确率 0.324 3,但其召回率仅有 0.047 6,表现不佳,而召回率表现优异的 LSTM-B-T 分类精确率 0.246 7,错误率 0.186 4,召回率 0.244 0,综合表现相比于 LSTM-T 精确率 0.247 2,错误率 0.216 1,召回率 0.325 9,其精确率相近且错误率更低,因此认为 LSTM-B-T 分类在测试集多方向预测表现最佳;同理做空方向排除多数召回率低

于 0.1 的模型后,RF-B-T 分类在召回率 0.453 0 表现优异的情况下其预测精确率 0.225 6,错误率 0.170 6,精确率高于错误率 5 个百分点,预测结果是理想的,因此认为 RF-B-T 分类模型在测试集空方向预测表现最佳。

综上所述,本文通过验证集与测试集在多空方向的评价指标进行对比归纳得出,在做多方向上 LSTM-B-T 模型最为理想,做空方向上 RF-B-T 模型的预测表现最为理想。其相应方向评价指标与原待测样本中数据分布比例的对比结果显示两模型都对精确率与错误率有较好的提升,同时保证了稳定的召回率,适用于实际交易策略的形成。

表 3 模型在验证集的预测效果评价

Table 3 Evaluation of the prediction effect of the model in the validation set

模型类别	分类方式	精确率		错误率		召回率		F_1	
		涨	跌	涨	跌	涨	跌	涨	跌
RF	T	0.241 9	0.402 9	0.403 2	0.238 8	0.103 5	0.221 7	0.142 8	0.222 1
	B-B	0.280 7	0.418 9	0.403 5	0.283 7	0.220 1	0.173 7	0.253 2	0.257 8
	B-T	0.288 4	0.485 3	0.269 2	0.321 6	0.137 6	0.192 7	0.151 0	0.270 7
DNN	T	0.261 3	0	0.396 0	0	0.365 3	0	0.301 2	0
	B-T	0.259 5	0	0.419 8	0	0.234 4	0	0.256 9	0
LSTM	T	0.301 3	0.454 6	0.294 9	0.272 7	0.321 9	0.056 8	0.311 4	0.106 1
	B-T	0.307 7	0.450 0	0.283 5	0.250 0	0.274 0	0.051 1	0.295 6	0.092 4

表 4 模型在测试集的预测效果评价

Table 4 Evaluation of the prediction effect of the model in the test set

模型类别	分类方式	精确率		错误率		召回率		F_1	
		涨	跌	涨	跌	涨	跌	涨	跌
RF	T	0.280 0	0.206 1	0.260 0	0.180 7	0.030 0	0.571 7	0.050 0	0.300 0
	B-B	0.211 7	0.215 9	0.224 1	0.186 1	0.220 1	0.243 7	0.223 2	0.227 8
	B-T	0.324 3	0.225 6	0.243 2	0.170 6	0.047 6	0.453 0	0.081 5	0.285 6
DNN	T	0.197 2	0.437 5	0.217 3	0.187 5	0.642 7	0.011 4	0.301 4	0.025 6
	B-T	0.210 1	0.437 5	0.216 7	0.187 5	0.293 1	0.010 7	0.241 2	0.027 6
LSTM	T	0.247 2	0.373 9	0.216 1	0.138 2	0.325 9	0.075 5	0.324 9	0.130 9
	B-T	0.246 7	0.328 1	0.186 4	0.156 2	0.244 0	0.034 5	0.253 7	0.065 3

3.3 交易回测结果

本文使用中证 500 指数 ETF 完成多空方向的回测与模拟盘交易验证。需要说明的是为使用指数 ETF 完成空方向的交易,需要在前一交易日提前持有底仓,因此本文的交易策略收益因底仓的存在而跟随中证 500 指数波动,此策略追求获得对标指数基准收益的超额收益,是一种指数增强型策略。

本文对比实验中的交易策略如下:①买入并持有策略,后文以 M1 表示,此策略的收益率等于指数此段周期内股票指数的价格增长率,是实验的基准策略;②随机建仓策略,即随机给定此刻预测信号,结合简单建仓策略进行买入卖出的策略,以 M2 表示;③持仓更新策略,即使用机器学习预测模型与持仓更新策略进行交易的策略,以 M3 表示;④完整

策略,即结合持仓更新策略与二维伽马函数对开仓信号的类概率判别阈值进行自适应调整的策略形成的最终完整策略,以 M4 表示。

各交易策略在 2021 年 1 月 29 日至 2021 年 12 月 29 日的回测收益曲线如图 6 所示。

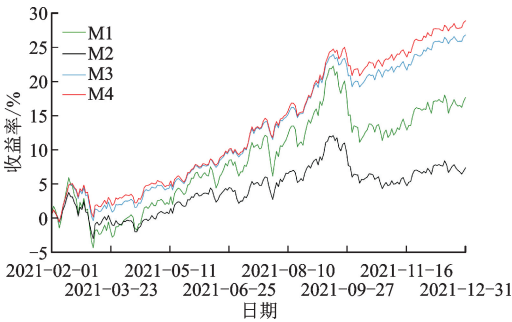


图 6 不同策略整体回测收益曲线

Fig. 6 Overall return curve of different strategies

图 6 中 M2 与 M3 对比可得,相比于随机判断,机器学习模型的预测信号提供了正确的指导作用,获取了正向收益。M3 与 M4 对比可得,加入开仓信号的类概率判别阈值自适应调整策略实现了更小的回撤以及更大的收益。不同策略整体回测收益评价指标见表 5,本文的完整策略交易指标整体最优,策略收益率为 28.84%,相比于持仓更新策略提升了约 2 个百分点,相对于基准策略的超额收益率为 11.24%,实现了超越参考标的中证 500 指数的增强收益。

图 7 为本文所提完整交易策略在验证集与测试集时间区间内于多空两向整体交易回测的开平仓交易点示意图。

蓝色曲线为交易标的每时刻收盘价连起来的价格走势曲线。可以看出,在验证集与测试集时间区间内的交易情况中多空两方向都表现为开仓点较为密集,证明开仓次数理想,没有损失较多交易机会。且在 2021 年 3 月 10 日与 2021 年 9 月 22 日附近的流畅行情中,蓝色曲线上空方向开平仓点较少,证明持仓更新的设定在流畅行情中有效增加了持仓时间,

避免频繁开平仓而造成手续费的磨损,从而产生了相对较多的收益。

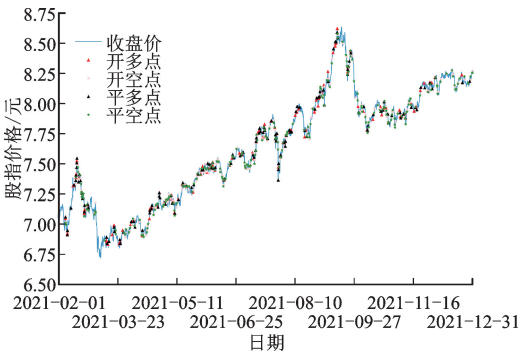


图 7 完整策略多空方向整体回测交易点

Fig. 7 The overall backtesting trading point of the complete strategy

3.4 模拟盘结果

图 8 为 2022 年 3 月 15 日至 6 月 15 日的模拟盘测试图,点画线从零开始呈逐渐上升趋势,代表超额收益始终为正值,收益表现良好。具体的模拟盘测试评价指标见表 6,胜率大于 50%且盈亏比接近 2.1,证明开仓创造的收益远大于开仓导致的亏损,这 3 个月的超额收益为 2.599%,折算年化收益为 13.38%,指数增强年化收益为 11.08%,夏普比率为 1.476。

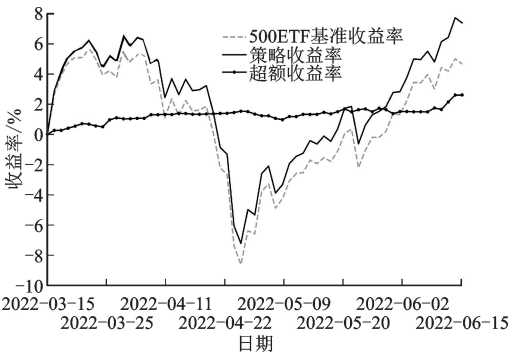


图 8 模拟盘测试收益

Fig. 8 Overall return curve of simulation trading verification

表 5 不同策略整体回测收益评价指标

Table 5 Overall return evaluation indicators of different strategies

策略	胜率/%	盈亏比	最大回撤率/%	基准收益率/%	策略收益率/%	超额收益率/%	夏普比率
随机建仓	37.20	0.878	6.520	17.60	7.287	-10.31	0.385
持仓更新	42.25	1.916	5.143	17.60	26.76	9.161	2.558
完整策略	42.59	2.010	4.771	17.60	28.84	11.24	2.740

表 6 模拟盘测试评价指标

Table 6 Evaluation indicators of simulation trading verification

胜率/%	盈亏比	最大回撤率/%	策略收益率/%	超额收益率/%	年化收益率/%	夏普比率
62.50	2.126	12.92	7.367	2.599	13.38	1.476

4 结 论

本文设计了多空方向采用不同机器学习模型的日内股指 ETF 程序化自动交易策略并完成回测和模拟盘验证。研究发现:本文给定的分类方式中,B-T 方式具有较好的预测表现,多空方向最终确定的 LSTM-B-T 模型、RF-B-T 模型都以此分类方式为基础;使用历史数据的回测对比实验结果显示,本文提出的结合持仓更新及二维伽马函数的完整交易策略使交易评价指标得到整体性提升;最终,使用实时交易数据进行程序化自动交易的模拟盘测试获取了 11.08% 的指数增强年化收益,证明此策略是一个对标指数可获得良好超额收益的量化交易策略,具有重要的工程应用价值。

参考文献:

[1] 陈伟忠,金以萍,汪应洛. 证券资产动态定价模型与机制研究 [J]. 西安交通大学学报, 1997, 31(10): 107-113.
CHEN Weizhong, JIN Yiping, WANG Yingluo. A capital asset pricing model for emerging market with changeable opportunity set [J]. Journal of Xi'an Jiaotong University, 1997, 31(10): 107-113.

[2] NEELY C J, RAPACH D E, TU Jun, et al. Forecasting the equity risk premium: the role of technical indicators [J]. Management Science, 2014, 60 (7): 1772-1791.

[3] CHEN Yiwen, CHOU R K, LIN Chubin. Investor sentiment, SEO market timing, and stock price performance [J]. Journal of Empirical Finance, 2019, 51: 28-43.

[4] LOHRMANN C, LUUKKA P. Classification of intra-day S&P500 returns with a random Forest [J]. International Journal of Forecasting, 2019, 35 (1): 390-407.

[5] 王淑燕,曹正凤,陈铭芷. 随机森林在量化选股中的应用研究 [J]. 运筹与管理, 2016, 25(3): 163-168, 177.
WANG Shuyan, CAO Zhengfeng, CHEN Mingzhi. Research on application of random forests in the quantitative stock selection model [J]. Operations Research

and Management Science, 2016, 25 (3): 163-168, 177.

[6] 李斌,林彦,唐闻轩. ML-TEA: 一套基于机器学习和技术分析的量化的投资算法 [J]. 系统工程理论与实践, 2017, 37(5): 1089-1100.
LI Bin, LIN Yan, TANG Wenxuan. ML-TEA: a set of quantitative investment algorithms based on machine learning and technical analysis [J]. Systems Engineering-Theory & Practice, 2017, 37(5): 1089-1100.

[7] 张茂军,饶华城,南江霞,等. 基于决策树的量化交易择时策略 [J]. 系统工程, 2022, 40(2): 118-130.
ZHANG Maojun, RAO Huacheng, NAN Jiangxia, et al. Timing selection of quantitative trading based on decision tree [J]. Systems Engineering, 2022, 40(2): 118-130.

[8] 高妮,贺毅岳,马新成. 基于低频分量 EEMD-SVR 预测建模的指数择时策略 [J]. 统计与决策, 2022, 38 (2): 140-145.
GAO Ni, HE Yiyue, MA Xincheng. Exponential timing strategy based on EEMD-SVR predictive modeling of low frequency component [J]. Statistics & Decision, 2022, 38(2): 140-145.

[9] NEMES M D, BUTOI A. Data mining on Romanian stock market using neural networks for price prediction [J]. Informatica Economica Journal, 2013, 17 (3): 125-136.

[10] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predictions [J]. European Journal of Operational Research, 2018, 270(2): 654-669.

[11] 杨青,王晨蔚. 基于深度学习 LSTM 神经网络的全球股票指数预测研究 [J]. 统计研究, 2019, 36(3): 65-77.
YANG Qing, WANG Chenwei. A study on forecast of global stock indices based on deep LSTM neural network [J]. Statistical Research, 2019, 36(3): 65-77.

[12] 贺毅岳,高妮,韩进博,等. 基于长短记忆网络的指数量化择时研究 [J]. 统计与决策, 2020, 36(23): 128-133.
HE Yiyue, GAO Ni, HAN Jinbo, et al. Research on the exponential quantitative timing based on long short term memory [J]. Statistics & Decision, 2020, 36 (23): 128-133.

- [13] 陈玲, 琚亚平, 张楚华. 多级轴流压气机末级静子叶片串列改型方法及流动机理研究 [J]. 西安交通大学学报, 2022, 56(10): 170-179.
- CHEN Ling, JU Yaping, ZHANG Chuhua. Tandem blading and flow mechanism of last stator of multi-stage axial flow compressor [J]. Journal of Xi'an Jiaotong University, 2022, 56(10): 170-179.
- [14] 张西宁, 张雯雯, 周融通, 等. 采用单类随机森林的异常检测方法及应用 [J]. 西安交通大学学报, 2020, 54(2): 1-8, 157.
- ZHANG Xining, ZHANG Wenwen, ZHOU Rong-tong, et al. Anomaly detection method based on one-class random forest with applications [J]. Journal of Xi'an Jiaotong University, 2020, 54(2): 1-8, 157.
- [15] 陈保家, 陈学力, 沈保明, 等. CNN-LSTM 深度神经网络在滚动轴承故障诊断中的应用 [J]. 西安交通大学学报, 2021, 55(6): 28-36.
- CHEN Baojia, CHEN Xueli, SHEN Baoming, et al. An application of convolution neural network and long short-term memory in rolling bearing fault diagnosis [J]. Journal of Xi'an Jiaotong University, 2021, 55(6): 28-36.
- [16] 任君, 王建华, 王传美, 等. 基于正则化 LSTM 模型的股票指数预测 [J]. 计算机应用与软件, 2018, 35(4): 44-48, 108.
- REN Jun, WANG Jianhua, WANG Chuanmei, et al. Stock index forecast based on regularized LSTM model [J]. Computer Applications and Software, 2018, 35(4): 44-48, 108.
- [17] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training recurrent neural networks [C]// Proceedings of the 30th International Conference on Machine Learning. Pittsburgh, PA, USA: IMLS, 2013: 2347-2355.
- [18] 陈佳, 刘冬雪, 武大硕. 基于特征选取与 LSTM 模型的股指预测方法研究 [J]. 计算机工程与应用, 2019, 55(6): 108-112.
- CHEN Jia, LIU Dongxue, WU Dashuo. Stock index forecasting method based on feature selection and LSTM model [J]. Computer Engineering and Applications, 2019, 55(6): 108-112.
- [19] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting novel associations in large data sets [J]. Science, 2011, 334(6062): 1518-1524.
- [20] 张园园. 基于资金流及股指短期指数变化的程序化交易软件设计与实现 [D]. 西安: 西安交通大学, 2021.
- [21] 陆珩琪, 姚恒宇. 涨跌幅限制变动对市场波动性及流动性影响的仿真研究 [J]. 系统工程理论与实践, 2022, 42(4): 897-905.
- LU Hengzhen, YAO Hengyu. A simulation study on the effect of price limits change on market volatility and liquidity [J]. Systems Engineering-Theory & Practice, 2022, 42(4): 897-905.
- [22] 李梦雨. 证券价格涨跌幅限制制度的存在原因: 基于反市场操纵的视角 [J]. 西安交通大学学报(社会科学版), 2015, 35(1): 34-39.
- LI Mengyu. Reasons for existence of system of securities price limits: from a visual angle against market manipulation [J]. Journal of Xi'an Jiaotong University (Social Sciences), 2015, 35(1): 34-39.
- [23] 刘志成, 王殿伟, 刘颖, 等. 基于二维伽马函数的光照不均匀图像自适应校正算法 [J]. 北京理工大学学报, 2016, 36(2): 191-196, 214.
- LIU Zhicheng, WANG Dianwei, LIU Ying, et al. Adaptive adjustment algorithm for non-uniform illumination images based on 2D gamma function [J]. Transactions of Beijing Institute of Technology, 2016, 36(2): 191-196, 214.
- [24] MANURUNG A H, BUDIHARTO W, PRABOWO H. Algorithm and modeling of stock prices forecasting based on long short-term memory (LSTM) [J]. ICIC Express Letters, 2018, 12(12): 1277-1283.
- [25] KRAUSS C, DO X A, HUCK N. Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500 [J]. European Journal of Operational Research, 2017, 259(2): 689-702.

(编辑 武红江 刘杨)