重庆理工大学学报(社会科学)
Journal of Chongqing University of Technology(Social Science)
2022 年 第 36 卷 第 3 期
Vol. 36 No. 3 2022
doi: 10.3969/j.issn.1674-8425(s).2022.03.007
经济学

基于多维高频数据和 LSTM 模型的沪深 300 股指期货价格预测

邱冬阳,丁 玲

(重庆理工大学 经济金融学院,重庆 400054)

摘要: 以 2010—2019 年的沪深 300 股指期货为对象,收集日收盘价、5 分钟收盘价,以及影响波动的 5 维度 89 个指标,采用维度删减、间隔采样方法,组合成多个不同维度和不同频率的 LSTM 深度学习模型对沪深 300 股指期货进行预测,并且从空间和时间角度分析维度和频率对股指期货价格波动的影响。研究表明: LSTM 模型可以很好地描绘沪深 300 股指期货多维高频数据的特征; 空间上,变量维度对沪深 300 股指期货价格的预测带来间接影响,预测精度最高的出现在 10 至 20 个交易日区间; 时间上,数据频率的影响更为直接,频率越高预测精度越高。研究结论有助于股指期货参与各方分散和化解金融风险。

关键词: 多维高频数据; 深度学习; LSTM 模型; 沪深 300 股指期货

中图分类号: F83 **文献标识码**: A **文章编号**: 1674-8425(2022)03-0055-15

一、引言

金融时间序列的预测是拥有悠久历史且被学者们持续关注的经典问题,方兴未艾的金融科技和量化投资都在寻求有效的预测方法作为突破口。在人工智能、区块链、云计算、大数据时代背景下,信息的获取、传播与规模达到了前所未有的水平,大数据正潜移默化地改变着金融市场中每位投资者的日常交易方式。早期投资者们获取信息的渠道单一,只能通过证券交易所现场交易,后来可以坐在计算机前观察价格走势,再发展到随时随地通过移动终端借助应用软件实时获取行情信息。重要性日益凸显的金融信息数据正逐渐商品化,催生了大量的金融信息数据公司。以信息数据为底层逻辑的量化投资、智能投顾等新兴的投资手段应运而生并不断扩散。显然,信息获取、交易方式、投资模式的改变意味着股票市场、期货市场乃至整个金融市场的价格波动有新特点、新规律,需要用升级换代的新方法来刻画这种变化。

与此同时,从神经网络、机器学习发展到深度学习的建模、算法伴随着计算机科学和人工智能的发展而广泛应用在工程领域、语音识别、图形处理和金融市场中,其中,深度学习中的众多具体方法在处理非线性海量数据时的优势明显,拓展了经典金融计量学的理论框架。同行研究表明[1]: 深度学习方法能够

收稿日期: 2021-12-02
基金项目: 国家社会科学基金重点项目“基于大数据+深度学习的中国金融市场波动性及预警机制研究”(17AJY028); 重庆市高校哲学社会科学协同创新团队——重庆智能金融研究协同创新团队的支持
作者简介: 邱冬阳,教授,博士,主要从事计量方法、金融市场研究。

本文引用格式: 邱冬阳,丁玲. 基于多维高频数据和 LSTM 模型的沪深 300 股指期货价格预测[J]. 重庆理工大学学报(社会科学),2022(3):55-69.

Citation format: QIU Dongyang, DING Ling. Forecast of CSIF 300 price based on multi-dimensional & high-frequency data and LSTM model[J]. Journal of Chongqing University of Technology(Social Science),2022(3):55-69.

有效刻画、恰当拟合和预测人工智能时代金融市场价格波动的新特征。期货市场属于分散风险的金融衍生品市场,其风险远远高于股票市场,因此对期货市场价格波动的预测、预警更有必要。本文选取沪深300股指期货(简称:CSIF 300)作为样本,采用深度学习的长短期记忆(Long Short-Term Memory,简称LSTM)模型,重点放在变量维度、数据频率处理方面,拟解决如何有效提高期货市场价格及波动性预测精度的问题。后续结构安排如下:第二部分为文献综述,第三部分为研究设计,第四部分为数据处理,第五部分为实证研究,第六部分为结论及启示。

二、文献综述

(一)文献回顾

金融资产价格与波动率的预测是学术界持续探讨的问题之一。多年以来,国内外学者们从不同角度对金融时间序列的预测与分析开展了相关研究。

在研究对象上,现有文献集中于股票价格与波动性的预测,其中不乏以股票市场多只股票、个股为对象进行的研究,如李斌等[2]、史建楠等[3];一些文献选取股指作为样本对象,如刘晓倩等[4]、陈黎明等[5];不少学者从多种角度也对期货市场价格变动及其波动率开展了相关研究,但研究角度主要集中在跨市场行为的信息传导[6]、期现货市场之间的价格相依性[7]。对于期货价格与波动率预测的关注度显然低于股票,研究成果甚至不到股票的一半。

在预测方法上,当前文献多集中于运用统计预测方法如移动平均、指数平滑、线性回归、ARMA模型、ARCH模型、GARCH模型、Monte Carlo方法、随机波动(SV)等对金融时间序列数据进行预测,并伴随着信息获取、交易方式、投资模式的变化仍在不断迭代更新。而股指期货预测的研究文献早期集中于基础统计模型、经典的GARCH及其衍生模型。Baillie等[8]利用GARCH分别计算了期现货市场条件方差的比率。魏宇等[9]通过比较OLS、VAR、VECM和MVGARCH等传统避险模型研究日内高频的避险效率。邱冬阳等[10]运用马尔科夫链进行预测。近年来,学者不断尝试运用非参数模型和其他算法模型,发现算法模型的预测能力往往优于数据模型,计算机软硬件条件的提升和大数据时代的到来,助其在理论和实践中脱颖而出。近段时间,利用机器学习方法预测金融资产价格波动性的研究文献越来越多。王宣承[11]以CSIF 300为样本构建了基于LASSO和神经网络的量化交易系统,而王国长等[12]提出LASSO主要是用来惩罚变量选择,做预测时需要结合ANN模型进行优化;陈标金等[13]构建了随机森林算法机器学习预测模型。但已有文献中所采用的多为常见的方法,只有部分文献开始引入深度学习的方法对金融资产进行预测,预测效果有所提升,这些文献多集中于国外,国内仅有少数学者[1]对这方面开展过相关研究。

在样本数据上,股指期货交易频率极高,交易速度极快,交易量极大,高峰期甚至更大,低频数据并不能准确刻画股指期货日内风险特征。但国内部分文献主要采用低频数据进行研究,基本都是选取日度行情数据,邵振文等[14]研究了日数据和月数据,魏宇[15]是采用分钟级别的分时金融市场价格数据,即高频数据。

在影响因素上,传统的计量模型难以挖掘复杂的输入特征,景楠等[16]仅对收盘价、成交量等单一指标进行研究,却忽略了很多外因,比如宏观经济政策、经济发展状况、投资者情绪等市场相关因素,而尚玉皇等[17]证实了对这些深层因素的考量会大幅增加发现金融市场隐藏的波动规律的能力。

在精度评价上,Cochrane[18]通过实证研究确定资产收益是可预测的。一直以来,不少文献致力于提高资产收益的预测精度,但关于影响预测精度的原因却说法不一。陈标金等[13]将预测精度不高的原因归于因素选取不够充分,Stoll等[19]提出预测频率升高或降低使得精度产生变化,景楠等[16]认为预测方法无法适应金融市场的结构突变。研究者关于模型预测精度比较的评价指标选取也不尽相同,大多数选用的评价指标为均方误差(MSE)、平均绝对误差(MAE)、平均绝对百分比误差(MAPE)等。

(二) 简单述评

受限于当时的技术条件,信息数据获取方式单一,已有文献多是运用统计预测方法描述低维度、低频率的金融数据特征,影响因素的选取也有限。大数据时代下的金融数据特征更为复杂多变,经典的计量方法预测精度不够理想,对诸多非线性和不确定性因素更是无所适从,而且多数文献主要关注股票价格与波动率预测,而以股指期货为研究对象进行分析的文献则很少,仅有的以股指期货为对象的研究结果精度也不够。深度学习的 LSTM 模型有强大的时序数据处理能力,对于处理多维度与高频率数据信息的学习效率与速度、预测精度更优,并且在处理非线性数据时可以提取更为复杂的特征,对于兼备大规模、多维度、高频率等数据特征的股指期货市场及其波动性具有高度适用性。

因此,本文选取 LSTM 模型对 CSIF 300 的价格波动性进行预测,可能的创新之处有:

(1) 将深度学习中处理时序数据表现出色的 LSTM 模型引入到期货市场进行价格预测,可以对新的决策范式下金融数据呈现的新规律和新特征有良好的拟合效果。

(2) 同时选用多维度与高频率两类样本,从空间和时间两个角度全方位挖掘数据特征。一是采用 CSIF 300 期货日收盘价低频数据,全面考虑造成其价格波动的影响因素,囊括五大维度,对应 89 个具体指标,突破了同类研究的最高水平; 二是采用 CSIF 300 日内 5 分钟收盘价高频数据,选取影响更为直接的两个维度、25 个具体指标。并将多维度与高频率相结合,提升了预测精度。

(3) 使用维度逐层删减方法组合成多个预测模型,分析各类型指标对 CSIF 300 的预测能力。对多维度数据集划分成高、中、低不同维度数据,设计多个模型进行预测; 将高频率数据进行不同频率抽样,划分成不同频率的等时间间隔数据再次对模型做预测。

三、研究设计

(一) 理论模型

根据持有成本模型,股指期现货价格之间的关系式可以使用持有成本描述。通常会将股指看成支付股息的投资资产,定义股指提供收益率为 q 的中间收入,无风险利率为 r,则持有成本 c 就可以表示成:

$$c = r - q \tag{1}$$

股指期货的远期价格 F_0 与当前价格 S_0 的关系式为:

$$F_0 = S_0 e^{cT} \tag{2}$$

综合考虑同 CSIF 300 价格有关的影响因素,根据各因素由内而外的相关性依次确定为 CSIF 300 的自身行情、影响 CSIF 300 的内在因素、宏观经济形势、关联金融市场和偶发事件 5 个维度,选择具体的指标进一步细化 5 个维度。

(二) LSTM 模型

统计预测模型往往依赖于历史交易数据,需要满足一定的前提假设,还受到数据维度和频率的多重限制。机器学习对于金融时序数据在维度和频率的要求上有所放宽,但是 CSIF 300 价格的预测属于监督学习的回归,受到不同因素的影响面临巨大不确定性,而深度学习对于该类回归任务具有优势。CSIF 300 收盘价的时序问题,深度学习模型中的循环神经网络(RNN) 和长短期记忆网络(LSTM) 均能够处理,但由于 RNN 同样面临着梯度消失和梯度爆炸的问题,即无法很好地发现其长期依赖关系,因此引入 LSTM 模型预测 CSIF 300 的价格。

在 LSTM 模型构建上,综合金融大数据与互联网交易高并发、多频次、大流量等特征,展开了充分全面的考虑。第一,LSTM 模型解决了循环神经网络会存在的梯度消失和梯度爆炸问题,能够更好地适应 CSIF 300 价格非平稳的数据特征; 第二,CSIF 300 价格具有长期依赖性,即先前的价格和指标均会对之后

产生影响,而具备长短记忆性的 LSTM 模型在处理时间间隔较长或作用效果有延迟的数据上存在明显优势;第三,LSTM 模型可以很好地总结非线性期货价格的内在规律,并准确预测未来期货价格的变动情况。

(三)样本选取

CSIF 300 合约最具代表性,流动性也比较强,因此样本对象确定为中国金融期货交易所 CSIF 300 合约。由于交割日规定比较特殊,相较下月及随后两个季月的合约价格来讲,当月代表性更强,同时保证数据连贯性,确定采用 CSIF 300 主力连续合约价格。

样本数据分为低频和高频数据两类:上市首日涨跌幅无法计算,确定低频数据的样本时间区间为 2010 年 4 月 19 日,截至 2018 年 12 月 28 日,共计 2 118 个交易日;高频数据样本时间区间为 2019 年 1 月 2 日 9 时 35 分,截至 2019 年 12 月 31 日 13 时,共 244 个交易日,以 5 分钟为抽样频率,每日 4 个小时交易时间,有 48 个 5 分钟收益率,最终形成的样本量为 $48 \times 244 = 11\ 712$。

(四)指标选择

根据股指期货理论定价和 CSIF 300 市场运行的实际,选取了 CSIF 300 的自身行情(27 个),影响沪深 300 股指的内在因素(16 个),宏观经济因素(26 个),关联金融市场,偶发事件因素(1 个)5 个维度共 89 个指标。

1. CSIF 300 的自身行情因素(27 个)

反映期货市场自身行情的变动又分为基本交易指标、市场指标与技术指标三类。

(1)基本交易指标,包括开盘价、最高价、最低价、成交量、成交额、均价、价差、结算价、持仓量、未平仓量和剩余交易日共 11 个指标。需要特别说明的是,为使研究更为充分,设计了剩余交易日指标。期货价格不仅会受到标的资产价格、交易量、未平仓合约数量等的影响,也会随着期货合约交割月份的逼近,逐渐收敛到标的资产的即期价格。事实上,处于交割月份中的期货价格波动更为剧烈。

(2)市场指标选取沪深两市的融资余额与融券余额 2 个指标。

(3)技术指标则选取 K、D、J、OBV、CCI、DIF、DEA、MACD、RSI1(6 日)、RSI2(12 日)、RSI3(24 日)、MA1(5 日)、MA2(10 日)、MA3(20 日)14 个关注度较高的技术指标。

2. 影响沪深 300 股指的内在因素(16 个)

(1) 根据理论公式,内在因素首选 CSIF 300 的标的资产为沪深 300 股指收盘价。

(2) 从外部看,沪深 300 股指会受沪深两市的大盘涨跌影响,借助上证综指和深证成指的收盘价 2 个指标反映股票市场的一般走势。

(3) 从内部看,沪深 300 股指各样本股的价格波动会对股指本身产生影响,进而影响股指期货价格,因此考虑沪深两市发行 300 只股票的上市公司的经营情况衡量 CSI 300 内在价值,这是影响 CSIF 300 的间接因素。包括①财务指标,最能反映标的资产估值,主要包括资产负债率、流动比率、净资产收益率(ROE)、开发支出、每股税后现金股利、股利分配率、主营业务收入、总股本数、在外流通股本数、平均市盈率、平均市净率共计 11 个财务指标;②公司治理结构指标,包括前 10 股东占比和董、监、高比例 2 个指标。

3. 宏观经济形势因素(26 个)

(1)主要选取经济增长、物价水平和国际收支 3 个方面。其中,① 衡量经济增长的指标包括 GDP 总量、GDP 增长率、城镇固定资产投资额、外商直接投资额、新增信贷额、制造业采购经理指数、非制造业采购经理指数共 7 个指标;② 度量物价水平的指标可以直接选取价格指数,包括 CPI、PPI、新建房价指数、二手房价指数、企业商品价格指数,同时选取间接运行物价水平的指标,包括 M0、M1、M2、Shibor(隔夜)、存款准备金率、财政收入、税收等 12 个指标;③ 国际收支平衡选用海关出口额和海关进口额 2 个指标。

(2) 度量宏观走势的预期指标,选用消费者信心指数、消费者满意指数、消费者预期指数、企业景气指数、企业家信心指数 5 个指标。

4. 关联金融市场因素(19 个)

(1) 国内金融市场主要考虑债券和期货市场: ① 债券市场类指标选取政府债券发行量和金融债券发行量; ② 期货市场选用资金量相对充足的国债期货和其他股指期货收盘价,包括 5 年期国债期货主连、10 年期国债期货主连、上证 50 股指期货主连、中证 500 股指期货主连共 6 个国内关联金融市场指标。

(2) 海外金融市场中,涵盖股票、期货和外汇 3 个市场类型,具体包括: ① 香港恒生指数、日经 225 指数、道琼斯工业指数、COMEX 黄金库存量和 COMEX 白银库存量 5 个指标; ② 迷你道琼斯指数期货、迷你纳斯达克指数期货、迷你标准普尔指数期货 3 个股指期货市场指标; ③ COMEX 黄金 6 月期货合约、COMEX 黄金期货、NYME 原油期货和 WTI 原油期货 4 个其他期货市场指标; ④ 人民币对美元汇率这一外汇市场指标。各个期货市场指标均选取主连合约的收盘价。

5. 偶发事件因素(1 个)

基于 APT 套利定价模型,期货市场价格会受到突发事件或"黑天鹅"事件带来的不确定性冲击,市场参与者面对其做出的即时反应又会造成期货市场的波动程度加剧,因此引入偶发事件作为一大类别因素输入模型。笔者总结了自 2010 年 4 月至 2018 年 12 月期间发生的可能影响期货市场的网络热点舆论事件 41 个,并根据偶发事件影响程度的大小进行了定量分析。基于次强式有效市场假说,对足以对整个经济运行状况产生影响的重大事件偶发当天,判断其正面或负面效应时分别给予 +3 和 -3 的赋值; 而对于单个公司、某一领域有相对较小影响的事件则在偶发事件当天分别给予 +1 和 -1 的赋值①。

四、数据处理

(一) 数据来源及选择

选取的数据来源于 10 个统计网站或数据库,包括 wind 数据库、雅虎财经、国泰安数据库等②。

输出特征选取了中金所公开的 2010 年 4 月 19 日至 2018 年 12 月 28 日近 9 年 CSIF 300 日收盘价以及 2019 年 1 月 2 日至 2019 年 12 月 31 日一年间的 CSIF 300 日内 5 分钟的收盘价作为原始数据。

输入特征中,低频日数据对应选取 89 个输入指标, 5 分钟高频数据对应选取 25 个输入指标。此外,基于高频数据的可获得性、输入特征的选取考虑 3 个方面原因: 其一,影响 CSI 300 内在因素对于高频数据的波动影响并不大。一天之内,上市公司的经营状况、管理结构等并不会发生频繁巨烈波动,往往是与经理人的长期经营决策相关。其二,宏观因素指标一年内的变动已经不够明显,一天之内更是微乎其微,其影响主要是长期的。其三,偶发事件发生的具体日期可知但确切时间点难以界定,对日内分时价格的影响程度也无法合理判断。基于此,高频数据选取的 25 个输入指标主要为日常交易数据、技术指标以及关联金融市场,包括开盘价、最高价、最低价、成交量、成交额、未平仓量、K、D、J、OBV、CCI、DIF、DEA、MACD、RSI-1、RSI-2、RSI-3、MA5、MA10、MA20、WTI 原油期货主连、COMEX 黄金期货主连、迷你标普指数期货主连、迷你道琼斯指数期货主连、迷你纳斯达克指数期货主连。

(二) 数据预处理

训练过程中,输入数据的质量与预测精度息息相关。根据 LSTM 模型对输入数据的要求,需要对原始数据进行预处理。数据的缺失值、标准化、混频数据等都按常规方法处理③。采集数据出现混频是常见现象,处理需要分多种情况进行,比如: 原始数据采集的时间间隔不一致、国内国外交易时间不同等等,往往需要结合经验综合判断。

① 因篇幅所限,总结的 2010—2018 年 41 个偶发事件此处未详列,如感兴趣可联系作者邮箱。
② 因篇幅所限,收集的原始数据此处省略,如需了解,可与作者联系。
③ 因篇幅所限,删减的数据预处理内容未列举,如有兴趣可与作者联系。

(1) 对于高频数据,由于国内外存在时差,而且国内外期货市场交易时间也不同,因此进行了超前滞后处理。海外市场前一日的交易情况往往会与国内市场当日的价格走势有关联,因此在高频数据中对海外市场数据进行了超前处理,即使用前一日的开盘时间对应当日国内的开盘时间,而且海外市场一天的交易时间与国内不一致,因此对海外市场的数据进行了截取。

(2) 为提高模型的泛化能力和实用性,克服深度学习模型过拟合现象,将数据集划分为训练集、验证集和测试集。在调试超参数阶段,为保证模型精度足够高,适当提高了验证集的占比,将原始数据集划分为8∶1.5∶0.5的训练集、验证集、测试集;在预测阶段,对多维度数据和高频率数据进行了不同处理,由于多维度数据不到10年,样本个数仅有2 118个,因此取95%的数据输入模型进行训练,即将验证集纳入训练集,剩余5%用来预测;而对于5分钟的高频率数据,数据量充足,一年的样本个数已经达到11 712,因此取90%的数据输入模型,剩余10%用来做测试集进行预测。

(三) 横纵向数据的递阶处理

1. 多维度数据

多维度数据分别从五大类别的影响因素出发,共选取了89个具体指标,数据信息量丰富。为比较数据维度的不同是否会对模型预测结果的准确程度造成影响,分析不同类别的影响因素描述CSIF 300价格波动特征的优劣差异,使用逐层剥离的方法将日数据的数据维度不同程度地减少,设计成不同维度的数据分别构建9种CSIF 300预测模型,对LSTM模型的泛化能力进行检验以便做进一步的比较分析。

模型1选取全部89个指标;模型2去除影响CSI 300的内在因素,保留74个指标包括自身行情、宏观经济、关联市场及偶发事件四大类影响因素;模型3去除关联金融市场因素,保留70个指标即其余四大类影响因素;模型4去除宏观经济形势因素指标,保留63个指标;模型5去除影响CSI 300的内在因素和关联金融市场因素,保留55个指标包括自身行情、宏观经济及偶发事件三大类影响因素;模型6去除影响CSI 300的内在因素和宏观经济形势因素,保留48个指标包括自身行情、关联市场及偶发事件三大类影响因素;模型7去除宏观经济形势因素和关联金融市场因素,保留44个指标包括自身行情、股指本身及偶发事件三大类影响因素;模型8去除三大类别的影响因素,仅保留29个指标包括自身行情以及偶发事件因素两大类;模型9仅剩余15个指标包括自身行情中的基本交易指标和偶发事件因素。

2. 高频率数据

高频率数据选用的25个指标属于自身行情和关联市场两个类别影响因素。对于CSIF 300日内5分钟高频数据进行等间隔抽样,设计成5、10、15、20、30、60分钟多种不同频率的分时数据分别输入预测模型,再次比较数据频率的不同是否会对LSTM模型预测精度产生影响。各个不同频率分时数据的样本个数分别为11 712、5 856、3 904、2 928、1 952、976个。

五、实证研究

(一) 实证过程

1. LSTM模型搭建

实证部分运用Tensorflow 2.0开源平台,采用Python 3.7编写程序,使用Keras搭建网络结构,确定的深度学习模型结构由输入层、LSTM隐藏层、输出层组成。

损失函数使用均方误差(MSE),训练过程选Adam优化器进行优化。超参数设置为训练时间步长t、批处理大小batch size,训练次数epochs、隐藏层神经元个数n。通过反复训练,最终确定超参数取值范围分别为:多维度数据的步长t在1~120,高频率数据的步长t在1~60;batch size为64;多维度数据的epochs为2 000次,高频率数据的epochs为1 000次,n均为256个,激活函数均为tanh函数。另外,多次

训练确定隐藏层数,对于多维度的日数据,在 $t<15$ 时,隐藏层选 2 层 LSTM 模型,在 $t\geqslant15$ 时,隐藏层选 1 层 LSTM 模型; 对于高频率数据,均选 2 层 LSTM 模型。

2. LSTM 模型预测精度的评价

为便于和同类研究的模型预测效果比较,选取常用的 3 个预测精度评价指标,分别为均方误差(MSE) 、平均绝对百分比误差(MAPE) 、平均绝对误差(MAE) 。这 3 个定量评价指标的数值越小,则预测值与真实值偏离程度越低,即预测效果越理想。

(二) 多维数据实证结果与分析

1. 预测结果

使用 LSTM 模型分别对 89、74、70、63、55、48、44、29、15 个变量的 2 118 条日数据进行拟合。现有研究大多为追求更高的精度构建各种 LSTM 复合模型,注重各模型之间的比较,预测效果在不断更新。与以往研究不同的是,针对不同维度的影响因素构建 9 种 LSTM 预测模型,并分别选择步长为 1、2、3、5、10、15、20、40、60、120 个交易日构造训练数据,输入模型进行训练,分析训练时间步长对预测效果的影响。为了进一步比较不同维度的预测变量对预测的影响,将训练时间步长为 10 ~ 20 细化至每一日,共形成 $9\times18=162$ 次不同的估计,然后预测未来一个交易日的收盘价,同时计算不同模型的预测精度,对比分析因素维度对模型预测效果产生的影响。模型的预测结果见表 1。

表 1 不同维度对应的 9 种模型预测精度(MAPE) 比较

步长	模型 1	模型 2	模型 3	模型 4	模型 5	模型 6	模型 7	模型 8	模型 9
	维度								
	89	74	70	63	55	48	44	29	15
1	1.336 9	1.331 2	1.567 4	1.318 8	1.450 7	1.376 0	1.647 5	1.625 5	1.504 9
2	1.158 3	1.270 6	1.271 7	1.286 0	1.174 3	1.311 7	1.207 4	1.346 3	1.279 2
3	1.059 1	1.484 9	1.244 5	1.140 8	1.129 0	1.294 9	1.284 9	1.218 9	1.446 8
5	1.178 0	1.281 7	1.209 5	1.156 1	1.158 8	1.202 5	1.174 5	1.308 1	1.359 7
10	1.033 5	1.313 6	1.032 3	1.282 4	1.131 2	1.004 4	1.300 4	1.114 3	1.420 7
11	0.992 4	1.022 4	1.309 3	1.246 8	1.267 5	1.099 5	1.235 7	1.145 5	1.477 9
12	1.108 1	1.191 2	1.073 9	1.114 3	1.048 3	1.028 6	1.137 3	1.086 2	1.413 9
13	1.077 9	0.965 6	1.108 3	1.286 7	1.140 3	0.951 2	1.271 7	1.323 2	1.288 6
14	1.153 5	1.300 1	1.199 4	1.180 8	1.224 2	1.246 2	1.352 1	1.154 2	1.315 1
15	1.065 1	1.135 3	1.126 2	1.108 4	1.248 3	0.983 4	1.082 2	1.060 5	1.207 2
16	1.167 1	1.131 9	1.081 8	1.104 0	1.006 5	1.070 6	1.287 6	1.283 1	1.070 7
17	0.933 8	1.045 3	1.284 8	1.132 4	1.034 6	0.849 7	1.145 8	1.052 8	1.337 0
18	1.170 5	1.046 6	1.085 9	1.095 7	1.570 8	1.147 1	1.222 2	1.207 3	1.239 3
19	1.210 5	1.225 8	1.015 2	1.299 0	1.095 0	0.993 2	1.078 7	1.037 5	1.313 7
20	1.055 9	1.207 1	1.123 0	0.998 5	1.183 3	1.128 5	1.155 2	0.951 7	1.340 6
40	1.342 2	1.125 3	1.150 1	1.140 2	1.301 0	1.058 0	1.362 0	1.267 1	1.486 5
60	1.112 9	1.434 6	1.153 5	1.065 4	1.148 6	1.334 8	1.319 7	1.225 7	1.471 1
120	1.377 9	1.196 3	1.247 5	1.233 0	1.271 9	1.256 9	1.265 7	1.237 9	1.492 5

注: 下划线突出显示各维度的最小 MAPE 值; 阴影部分突出显示全局最小 MAPE 值

从预测精度的结果来看,模型 1 ~ 模型 9 的 MAPE 均在 1.0 左右,表现出良好的预测效果。从整体来看,不同维度的变量之间存在的多重共线性,并不会对模型的预测精度造成较大影响,反而考虑的影响因素越全面,如模型 1 有 89 个特征变量,当步长为 17 个交易日时,MAPE 值达到了最小,低至 0.933 8。模型 2 有 74 个变量,仅仅是对影响沪深 300 股指的内在因素指标直接进行删减,当步长为 13 个交易日时,

预测的 MAPE 达到最小,但最小值反而高于模型 1, MAPE 值为 0.965 6。观察模型 6 的预测结果可以发现,当删除影响 CSI 300 的内在因素指标和宏观经济形势指标时,相对于模型 1 产生更高的预测精度,达到全局最小值,可能是数据集自身属性对预测效果产生的影响,这有待进一步验证。维度删减最多的模型 9 仅有 15 个特征变量,当步长为 16 个交易日时出现 MAPE 最小值 1.070 7,显然模型 1 的预测精度要高于模型 9。各模型的预测精度最小值集中在步长 10 至 20 个交易日之间,说明半个月到 1 个月的历史交易日数据参考价值是最大的。

2. 预测效果对比图

选取各模型 MAPE 值最小的训练时间步长进行了 60 个交易日的预测值与真实值的比较,并进行可视化,预测结果如图 1 ~ 图 9 所示①。

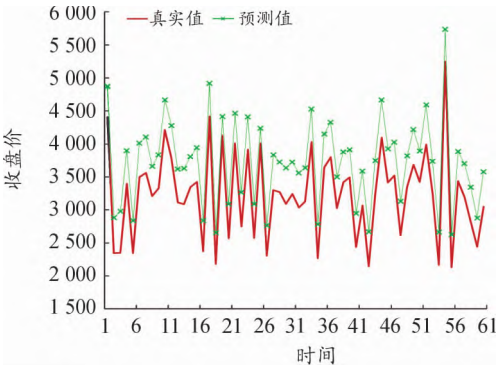

图 1 89 个变量 LSTM 模型预测结果

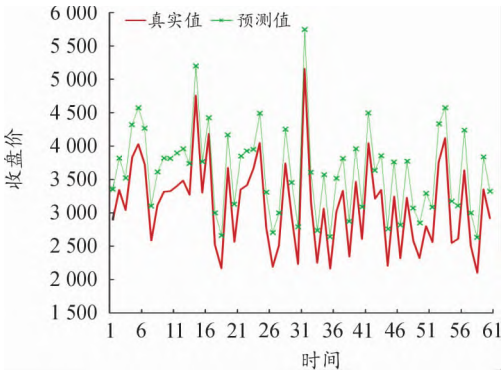

图 2 74 个变量 LSTM 模型预测结果

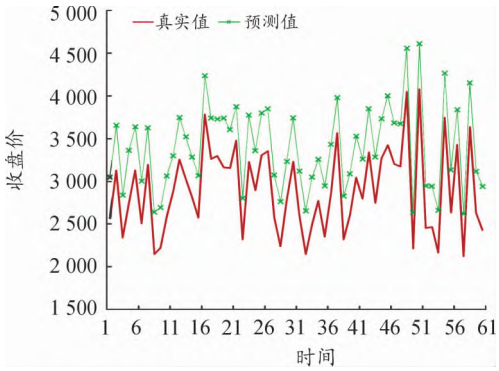

图 3 70 个变量 LSTM 模型预测结果

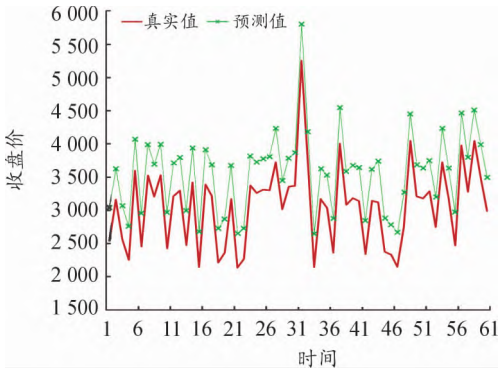

图 4 63 个变量 LSTM 模型预测结果

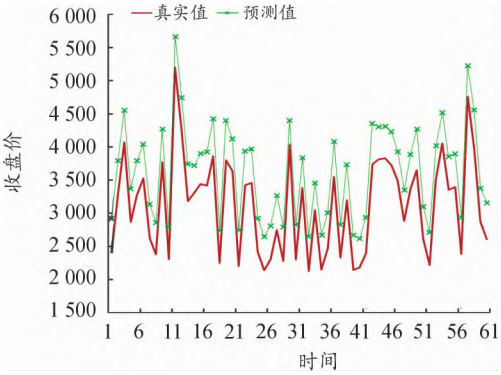

图 5 55 个变量 LSTM 模型预测结果

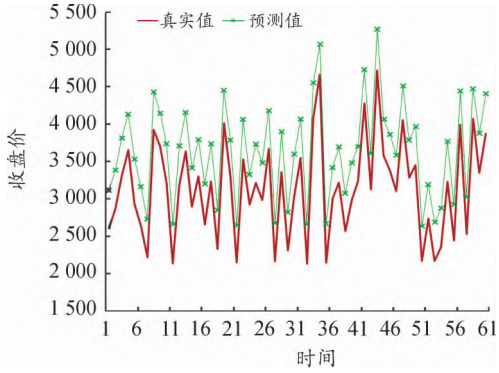

图 6 48 个变量 LSTM 模型预测结果

① 为满足清晰作图要求,在预测值上手动添加 500 展示,特此说明。

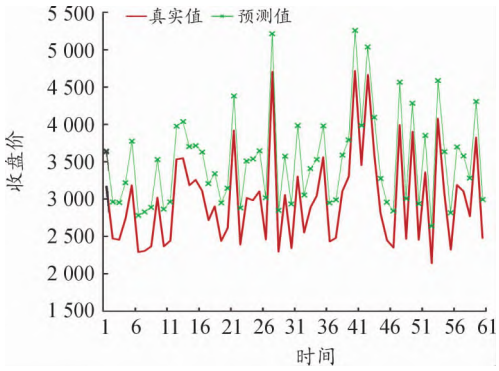

图 7 44 个变量 LSTM 模型预测结果

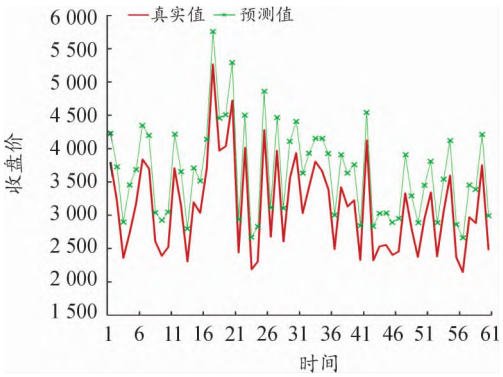

图 8 29 个变量 LSTM 模型预测结果

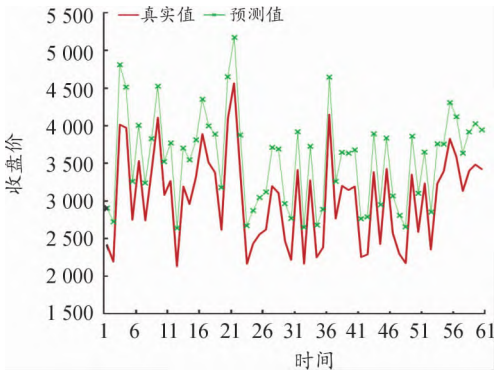

图 9 15 个变量 LSTM 模型预测结果

由图 1 ~ 图 9 可以看出,9 个 LSTM 模型的预测结果均表现良好,预测值与真实值的吻合程度极高,在实际的日收盘价波动较小的地方,预测值与真实值更为接近,而实际的日收盘价波动较大的地方,预测值与真实值相对而言有所偏离。这和宋刚等[20]使用深度学习预测期货和股票价格得出的结论是相符的。整体来看,LSTM 模型对于不同维度的特征变量均获得了优良的预测效果。同时,LSTM 模型的泛化能力也极强,具有较高的适用性。

3. 结果解读

(1) 从因素维度看,因素多少对预测精度有间接影响,但影响并不明显。依赖短期数据预测时,影响因素维度越多,预测精度越高,基本面分析和技术分析相结合对提高 LSTM 模型的预测精度确实有效;依赖中长期数据预测时,因素维度的作用有所减弱。

① 使用短期数据即在步长处于 10 个交易日内时,模型 1 即五大类的 89 维的预测效果显著优于另外 8 个模型。通过计算步长 1 ~ 5 的 MAPE 平均值可以发现,模型 1 的预测效果是最好的,MAPE 均值为 1.183 1,模型 4 次之。结果表明,短期数据支撑预测时纳入的影响因素越全面,预测越精准。这说明基本分析与技术分析相结合在期货价格预测比较有效,基本分析会提高期货价格的预测效果,与陈标金等[12]的研究结论一致。影响沪深 300 股指的内在因素、宏观经济形势因素、关联金融市场因素同其价格在近 5 个交易日之内存在一定的关联性,虽然这些因素对于价格的影响存在滞后期,但大数据时代的到来使投资者获取信息的时效性明显增强,应对更及时,滞后效果的反映等待期不超过 5 个交易日,各个方面的影响因素均会在 5 日内反映完全。

② 使用中期数据即达到 1 个月的历史数据量时,通过计算步长 10 ~ 20 的 MAPE 平均值发现,除模型 9,其他预测效果均明显提升,相对较好的预测效果集中在模型 1 和模型 6,这说明投资者考虑多个方面的因素与只考虑 CSIF 300 自身行情、关联市场以及偶发事件三类因素得到的效果差不多,甚至后者表现更优,因素维度的作用不再像短期数据那么明显。这与实际股指期货交易、理论上的噪声交易是相符

的。历史数据信息充足时,可以只考虑自身的交易行情与关联市场的行情指标,不需要考虑过多庞杂因素,考虑过多反而可能由于信息冗杂对预测效果产生不利影响。模型9仅考虑了基本交易信息和偶发事件因素,虽然基本交易信息一定程度上可以反映价格波动,但想达到精准预测是远远不够的,还需要将CSIF 300自身行情中的技术指标和关联市场行情等方面考虑进去。原因有两点:一是技术指标的计算本身就存在特定的周期性,忽视它就直接损失了其在中期数据预测时的有效性;二是半个月甚至1个月内关联金融市场行情已经完全反映到股指期货价格的变动上,过久的历史数据反而会带来负效应,使同股指期货价格的关联性出现明显下降。

③ 使用长期数据即在训练时间窗口逐渐拉长至120个交易日即半年时,除了模型4出现了轻微提升外,其他模型的预测精度均有所下降,这说明宏观经济形势会对期货价格产生持续性的影响,与Altavilla等[21]的研究结论存在一致性。但太早期的历史数据无论是哪个类别对于分析价格波动规律都不起显著作用,交易信息具有时效性。模型9的预测结果是最差的,说明长期数据预测仅考虑期货市场的基本交易行情远远不够,数据蕴含的信息量过少,数据深度不够,深度学习无法有效挖掘到数据特征。

(2) 从训练时间步长看,近期(举个例子,预测4月1日应该用3月21日至3月31日的数据)10至20个交易日的历史数据达到的预测效果最为理想。分析表1可知,MAPE最小值均出现在10~20个交易日。

① 使用短期数据即步长在10个交易日以内时,MAPE值较大,说明LSTM处理短期数据的表现略逊于中长期数据。

② 使用长期数据即步长超过20个交易日时,各模型的预测效果整体来看均出现了显著下降,只有模型4出现了轻微提升。预测效果虽然受损,但并未出现严重偏离,只是存在信息冗余现象。这和经济计量方法得出的结论是一致的。一般来讲,数据越多,预测的精度会越高,但Hull[22]提出太老的历史数据对于预测未来价格的波动性可能不太相干,一个折中的办法是采用最近90~180天的日收盘价数据。

③ 使用中期数据时,拟合历史数据的步长甚至可以拉近至更短,使用最近10~20交易日的数据预测精度是最高的。这与尚玉皇等[17]的结论基本一致。具体来看,在步长为17时的模型1和6的预测精度提升得更为明显,此时模型6的预测效果达到全局最优。这符合预期和期货市场的实际投资情况,CSIF 300合约的交割日期在每个月第三周的周五,能够对近三周的历史数据最为有效这一结论做出合理解释。

(3) 整体看,模型1~模型9的预测精度均在1.0左右波动,总体表现稳定。比较模型1和模型9在各步长的表现,模型1的预测精度要明显高于模型9,证实了深度学习在处理多维度共线性的海量数据存在优势,维度的增加会使得预测带来一定提升,但提升程度相对有限,指标纳入需要有一个合适的度。LSTM模型确实能够避免长时依赖问题,对于短期和长期数据信息均适用,但效果略微有差别。

(三)高频数据实证结果与分析

1.预测结果

使用2019年1月2日至2019年12月31日的CSIF 300 5分钟高频数据,将其进行等时间间隔抽样成不同频率的数据,分别选择步长为1、2、3、6、12、18、24、30、36、42、60个时间间隔构造训练数据,输入LSTM模型进行训练,分析步长对预测效果的影响,然后预测下一时间间隔的收盘价,预测结果见表2。步长的设计与数据采样频率相关,对于5分钟的数据而言,不同步长分别对应的时间窗口是5分钟、10分钟、15分钟、半小时、1小时、1个半小时、2小时、2个半小时、3小时、3个半小时、5小时。

从表2来看,对于不同频率的数据,MAPE存在数量级差别,但相对日数据,其预测精度明显更高。对于5分钟数据,MAPE均在0.15左右,最小值0.138 2;10分钟MAPE为0.2左右,最小值0.186 6;15分钟MAPE为0.25左右,最小值0.222 1;20分钟MAPE为0.3左右,最小值0.257 1;30分钟MAPE为0.4左右,最小值0.364 1;1小时MAPE为0.6左右,最小值0.493 7。不同频率数据的模型MAPE最小值均出现在步长为24个时间间隔。

表 2 不同频率对应的模型预测精度(MAPE) 比较

步长	频率					
	5	10	15	20	30	60
1	0.171 9	0.232 8	0.306 4	0.347 3	0.521 2	0.726 8
2	0.148 5	0.220 3	0.264 8	0.375 9	0.402 2	0.484 6
3	0.149 9	0.203 8	0.253 4	0.333 7	0.390 9	0.582 4
6	0.148 3	0.224 4	0.280 2	0.321 8	0.426 6	0.563 1
12	0.154 5	0.217 5	0.255 9	0.280 0	0.415 0	0.552 8
18	0.141 7	0.196 8	0.261 8	0.301 1	0.393 5	0.515 8
24	0.138 2	0.186 6	0.222 1	0.257 1	0.364 1	0.493 7
30	0.139 0	0.211 9	0.265 6	0.304 3	0.381 7	0.524 6
36	0.162 0	0.222 2	0.250 1	0.275 3	0.358 0	0.490 2
42	0.146 3	0.189 4	0.250 0	0.354 5	0.355 1	0.481 9
60	0.153 7	0.203 8	0.267 4	0.307 2	0.618 9	3.992 1

注: 下划线突出显示各频率的最小 MAPE 值; 阴影部分突出显示全局最小 MAPE 值

2. 预测效果对比图

同样为了更为直观地展示不同频率数据的预测效果,类比日数据进行预测值与真实值的比较,并进行数据的可视化,预测结果如图 10 ~ 图 15 所示①。

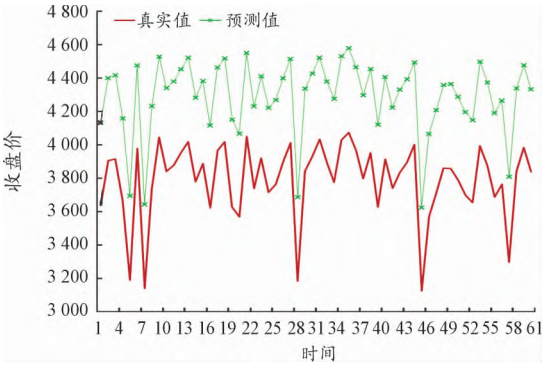

图 10 5 分钟数据的预测结果

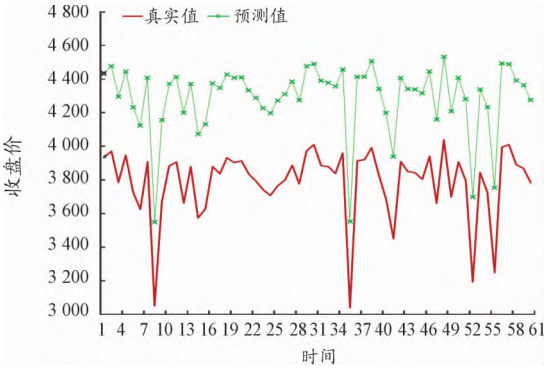

图 11 10 分钟数据的预测结果

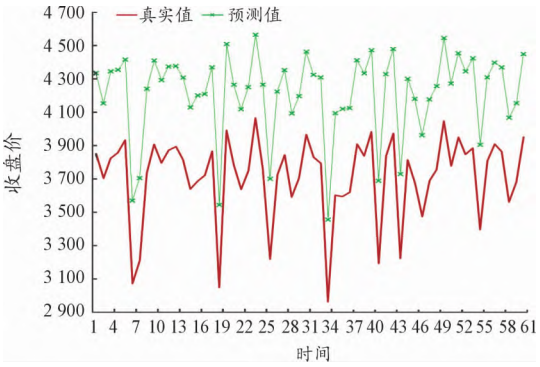

图 12 15 分钟数据的预测结果

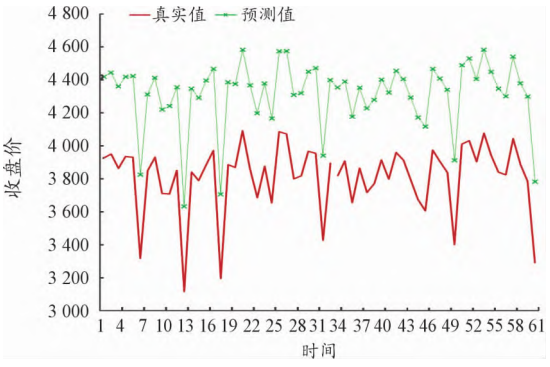

图 13 20 分钟数据的预测结果

① 为满足清晰作图要求,在预测值上手动添加 500 展示,特此说明。

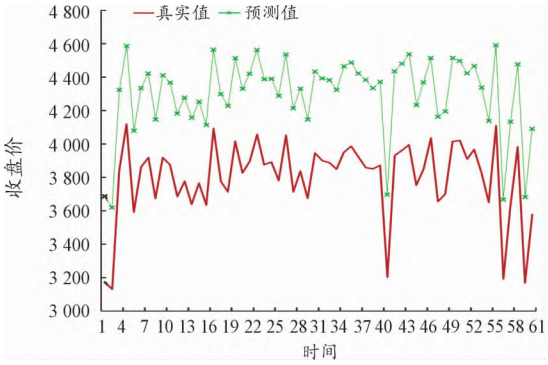

图 14 30 分钟数据的预测结果

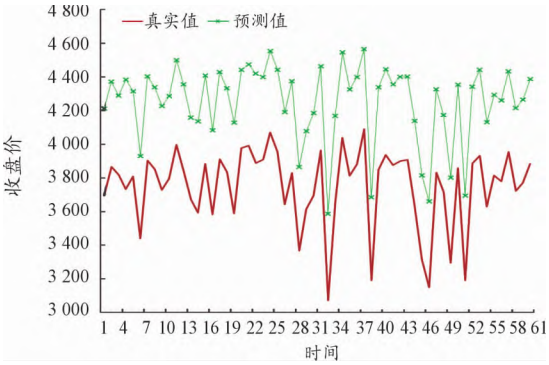

图 15 1 小时数据的预测结果

图 10 ~ 图 15 进行了不同频率数据预测效果比较,通过对比数据量和变量维度个数相近的图 8 和 14 可以看出,高频数据的预测效果是明显优于低频数据的。

3. 结果解读

(1) 从数据频率看,不同数据频率的对模型预测效果产生了直接影响。由表 2 可以看出,数据频率越高,预测精度越高。对于 5 分钟数据的预测精度在 0.15 左右,间隔 10 分钟的在 0.2 左右,间隔 15 分钟的在 0.25 左右,间隔 20 分钟的在 0.3 左右,间隔 30 分钟的在 0.4 左右,间隔 1 小时的在 0.6 左右上下浮动,相较而言波动程度略大,可能是由于数据量过小导致预测不够准确,部分原因是时间间隔过宽损失了交易数据中大量的有效信息。由此可见,数据深度与预测能力存在明显关联,数据频率越高,LSTM 模型对数据的挖掘越深层,就越能够提取更多有用信息。鉴于此,高频数据对于期货市场预测更有效,且作用程度明显。可能的原因是交易信息本身存在时效性,况且股指期货是 T +0 交易机制,频率较低的数据信息无法及时反映市场行情的变化。这与 Silva 等[23] 观点保持一致。早期 Stoll 等[19] 也得出与之相吻合的研究结论,证实利用 5 分钟的股指期货高频数据预测效果要优于 10 分钟。

(2) 从模型的训练时间步长看,考虑近期 24 个步长达到的预测效果最为理想。由表 2 可以看出,无论是高频数据的频率如何,模型均在 24 个步长表现最佳。对于 5 分钟数据,投资者应考虑 2 个小时内或是接近 2 个小时(即半个交易日) 的交易情况,半个交易日以外的数据对于当前的预测效果并不理想; 对于间隔为 10 分钟的数据,投资者应参考近一个交易日的数据,这样信息的有用程度达到最大化,前一日的海外市场也会产生影响; 对于间隔为 15 分钟的数据,投资者考虑近 1 个半交易日的市场信息最为有效; 对于间隔半小时的数据,投资者应考虑近 3 天的交易信息; 对于间隔 1 小时的数据,近 6 天的数据为有效信息,而不同频率的数据如果参考的范围过小或过大,准确程度均会大大降低。尤其是间隔为 1 小时的数据,倘若考虑 3 个月的历史信息反而会使得模型的预测效果呈断崖式下降,这或许因为每个交易日 CSIF 300 的合约有 4 份,合约月份为当月、下月及随后两个季月,因此交易周期为 1 个季度,选取当季的信息预测是更为有效的。

(3) 总体看,对于不同频率的数据,一方面,高频数据相较于日交易数据,维度对模型预测的作用就相对较弱,主要是数据信息量不同导致预测结果有明显差异,频率越高,数据信息量越大,模型的预测精度越高,这也是深度学习对于海量数据处理存在优越性的具体体现; 另一方面,结合表 1 与表 2 分析,可以得出模型预测效果的差异并不仅仅是因为数据量,时间间隔为 30 分钟的收盘价数据与近 10 年的日收盘价数据同为 2 000 左右的样本个数,选用表 2 中 30 分钟的预测结果与表 1 中模型 8 的预测结果进行对

比时,特征变量同为 25 维左右,MAPE 数量级也存在显著差别。30 分钟的数据 MAPE 在 0.4 上下浮动,而日收盘价 MAPE 在 1.0 上下浮动,这足以表明频率对模型预测效果产生了影响,因为高频部分影响因素的选取均为自身市场行情及关联市场行情,排除了不同类别因素的性质对预测结果带来的影响。此外,由于 30 分钟数据和日数据的数据量差距并不大,也可以排除单单是 LSTM 模型自身特征发挥作用的原因。结果表明,30 分钟的数据频率高于日收盘价数据,模型的预测精度得到了大幅度提高。

4. 稳健性检验

为了提升实证结果的可靠程度,增强结果分析的说服力,通过把 MAPE 替换为 MSE、MAE 后,对 LSTM 模型预测结果的稳健性进行检验。检验结果①与前文的研究结论一致[24]。

六、结论及启示

(一)主要结论

结合大数据和深度学习二者的优势,利用 CSIF 300 自正式上市以来不同频率的交易数据,通过构建 LSTM 模型重点研究了人工智能时代 CSIF 300 价格的新波动特征,进行科学有效的预测。结合实际情况筛选出 89 个预测变量,全面涵盖 CSIF 300 的自身行情、影响沪深 300 股指的内在因素、宏观经济形势、关联金融市场行情及偶发事件因素五大类别,层层渗透,逐步深入对 CSIF 300 价格的波动特征进行挖掘,从变量维度和数据频率两个方面探究影响股指期货价格预测精度的深层原因。主要结论如下:

第一,运用多维高频数据与 LSTM 模型的有机融合建立金融预测模型,可以很好地刻画、拟合和预测 CSI 300 价格波动的新特征,变量维度和数据频率均会对 LSTM 模型的预测精度产生影响。

第二,因素的纳入会对 CSIF 300 价格的预测产生间接影响。使用短期数据预测时,变量维度越多,预测精度越高;使用中长期数据预测时,变量维度的影响减弱,此时并非纳入的因素越全面,预测精度越高。

第三,数据频率的差别会对 CSIF 300 价格的预测产生直接影响。数据频率越高,预测精度就越高。高频数据信息包含的信息更为丰富,而频率越低损失的有用信息越多,因此深度学习对于高频数据中隐藏的深层信息可以提取出来,预测结果表现更为优良。

第四,变量维度的增加会使得 LSTM 模型的预测精度带来一定程度的提升,但变量指标纳入量需要有一个合适的度,数据频率的提高对 LSTM 模型的预测精度提升效果十分明显。

第五,LSTM 模型的预测精度也会受到训练时间窗口大小的影响。对于低频的日交易数据,考虑近 10 个交易日至 20 个交易日的数据信息已经可以做出准确预测;对于高频的分时交易数据,需要针对不同时间间隔数据损失的信息程度调整分析的时间范围。

(二)政策启示

基于实证研究结论,对 CSIF 300 市场参与各方的启示如下:

就期货交易所和监管部门而言,精准预测 CSIF 300 的波动特征有助于科学把握市场资金流向,进而精准监管整个期货市场,细化市场交易规则,遏制倒填日期等违规交易行为。同时,可以构建地方金融数据中心,公布更多的期货市场、金融市场的连续数据,充分发挥期货市场价格发现功能。

就金融期货产品的设计而言,股指期货的标的资产在对波动性比较大或者退市的股票做定期样本清理时,可以将频率提高到半个月至 1 个月清理一次,使股指期货的流通性进一步增强。在设计交割月份

① 因篇幅所限,稳健性检验结果此处省略,如需了解,可与作者联系。

时,要综合考虑标的公司经营状况、宏观经济形势等多个类别的影响因素,具体到第一交割通知日和最后交割日之间的时长是否可以考虑延长至1周左右,缩小投机者的套利空间。也可以引进迷你合约吸引小额度投资者,并适当缩小头寸限额防止投机者给期货市场造成不利影响。

就套期保值者及投资者而言,全面考虑大数据时代的各种可获得数据信息来分析股指期货是必要的,但过度的数据、过期的信息不利于其精准预测分析,尤其是套期保值者要关注期货标的资产及宏观经济等综合因素,以便达到优化资产配置规避风险的目的。

需要指出的是,本文虽然充分发挥了深度学习智能算法处理非线性、非平稳、大容量时序数据方面的优势,也引入偶发事件这一具有非结构化特征的大数据源,但异构可变的数据在实际量化处理时仍是带有主观性和经验判断。此外,受到经典计量模型处理共线性海量数据的局限,数据的统计口径和智能算法存在客观差别,CSIF 300价格的预测结果并未和经典计量模型进行比较分析。相应地,这些不足提供了金融期货价格波动问题的后续研究思路:将文本挖掘技术应用到偶发事件因素、宏观经济政策、投资者情绪等没有量化的指标选取与量化上,进一步提高金融市场预测分析能力。

参考文献:

[1] 宋鹏,张森. 国债收益率预测的VAR-LSTM框架[J]. 统计与决策,2021(5):148-152.

[2] 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投资算法[J]. 系统工程理论与实践,2017(5):1089-1100.

[3] 史建楠,邹俊忠,张见,等. 基于DMD-LSTM模型的股票价格时间序列预测研究[J]. 计算机应用研究,2020(3):662-666.

[4] 刘晓倩,王健,吴广. 基于高频数据HAR-CVX模型的沪深300指数的预测研究[J]. 中国管理科学,2017(6):1-10.

[5] 陈黎明,龙灵芝,郑千一. 基于CEEMDAN方法和灰色模糊聚类的汇率预测研究[J]. 重庆理工大学学报(社会科学),2021(6):109-121.

[6] 苑莹,王梦迪,樊晓倩,等. 市场间相依性检验、非对称性及传导方向研究[J]. 系统工程理论与实践,2016(11):2778-2790.

[7] 胡振华,钟代立,王欢芳. 中国铁矿石期货市场的定价影响力研究——基于VEC-SVAR模型的实证分析[J]. 中国管理科学,2018(2):96-106.

[8] BAILLIE R T,MYERS R J. Bivariate garch estimation of the optimal commodity futures hedge[J]. Wiley Subscription Services,Inc. A Wiley Company,1991,6(2):109-124.

[9] 魏宇,赖晓东,余江. 沪深300股指期货日内避险模型及效率研究[J]. 管理科学学报,2013(3):29-40.

[10] 邱冬阳,苏理云. 金融市场随机波动的联动性及预警机制研究:基于马尔科夫链蒙特卡洛抽样方法[M]. 北京:经济科学出版社,2017:62-67.

[11] 王宣承. 基于LASSO和神经网络的量化交易智能系统构建——以沪深300股指期货为例[J]. 投资研究,2014(9):23-39.

[12] 王国长,梁焙婷,王金枝. 改进的自适应Lasso方法在股票市场中的应用[J]. 数理统计与管理,2019(4):750-760.

[13] 陈标金,王锋. 宏观经济指标、技术指标与国债期货价格预测——基于随机森林机器学习的实证检验[J]. 统计与信息论坛,2019(6):29-35.

[14] 邵振文,侯丹. 我国股指期货市场非对称性波动与下行风险研究[J]. 经济纵横,2018(3):108-113.

[15] 魏宇. 沪深300股指期货的波动率预测模型研究[J]. 管理科学学报,2010(2):66-76.

[16] 景楠,吕闪闪,江涛. 基于HMM和GARCH模型的中国期货市场波动性研究[J]. 管理科学,2019(5):152-162.

[17] 尚玉皇,郑挺国. 基准收益率曲线与宏观经济:基于混频DSGE模型的研究[J]. 经济研究,2018(6):36-51.

[18] COCHRANE J H. Presidential address: Discount rates [J]. Journal of Finance,2011,6(3):165 - 219.

[19] STOLL H R,WHALEY R E. The dynamics of stock index and stock index futures returns [J]. The Journal of Financial and Quantitative Analysis,1990,25(4):441 - 468.

[20] 宋刚,张云峰,包芳勋,等. 基于粒子群优化 LSTM 的股票预测模型 [J]. 北京航空航天大学学报,2019(12):2533 - 2542.

[21] ALTAVILLA C,GIANNONE D,MODUGNO M. Low frequency effects of macroeconomic news on government bond yields [J]. Journal of Monetary Economics,2017,92:31 - 46.

[22] HULL J C. Options,futures,and other derivatives [M]. Fifth Edition. Upper Saddle River: Prentice Hall,2003:523 - 549.

[23] SILVA E,CASTILHO D,PEREIRA A,et al. A neural network based approach to support the market making strategies in high-frequency trading [C]//2014 International Joint Conference on Neural Networks. IEEE,2014:845 - 852.

[24] 陈卫华. 基于深度学习的上证综指波动率预测效果比较研究 [J]. 统计与信息论坛,2018(5):99 - 106.

Forecast of CSIF 300 price based on multi-dimensional & high-frequency data and LSTM model

QIU Dongyang, DING Ling

(School of Economics and Finance, Chongqing University of Technology, Chongqing 400054, China)

Abstract: Taking the CSIF 300 from 2010 to 2019 as the object, this paper collects daily closing price, 5-minute closing price, and 89 5-dimensional indicators affecting fluctuation, and uses the methods of dimension deletion and interval sampling to combine them into LSTM deep learning models with different dimensions and different frequencies to predict the closing price of CSIF 300. It also analyzes the impact of dimension and frequency on the price fluctuation of stock index futures from the perspective of space and time. The research shows that the LSTM model can well describe the characteristics of multidimensional high-frequency data of CSIF 300. Spatially, the variable dimension has an indirect impact on the prediction of the price of CSIF 300. The highest prediction accuracy occurs in the range of 10 to 20 trading days. In terms of time, the influence of data frequency is more direct. The higher the frequency, the higher the prediction accuracy. The research conclusion is helpful for all parties involved in stock index futures to disperse and resolve financial risks.

Key words: multi-dimensional and high-frequency data; deep learning; LSTM model; CSIF 300

(责任编辑 张佑法)