DOI: 10.19699/j.cnki.issn2096-0298.2020.17.090

基于标的股指及机器学习的股指期货价格预测

浙江财经大学 蔡泽栋

摘 要: 使用机器学习算法对复杂的金融市场数据进行预测，是近年来一个热门的研究方向。本文以沪深300股指期货为价格预测对象，首先构建VAR模型发现标的股指价格对股指期货价格具有显著影响，其次辅助脉冲响应分析结果确定预测模型中的具体特征，最后基于XGBoost算法，使用历史数据训练模型并进行测试。结果表明：模型预测效果较好，且与不含标的股指历史交易信息的预测结果相比更加准确，从而得出结论：标的股指历史交易数据对股指期货价格预测有重要作用。

关键词: 价格预测 股指期货 VAR模型 XGBoost算法

中图分类号: F832.5 **文献标识码**: A **文章编号**: 2096-0298(2020)09(a)-090-03

1 引言

随着人工智能技术与大数据技术的兴起火热，金融市场预测与机器学习算法的结合成为近年来热门的研究方向。股价预测是金融时间序列预测问题中最常见的一类，决策树[1]、神经网络[2]、逻辑回归、支持向量机(SVM)[3]等经典的机器学习算法都有被应用于研究中，且均有较完善的研究体系。

然而股指期货价格的预测却有所不同。股指期货是以股票指数为标的物的标准化期货合约，是一种较为成熟、级别较高的金融投资工具，与股指现货市场密切相关[4]。因此，在对股指期货价格进行预测建模时，除了其自身的历史交易信息之外，标的股指价格作为输入特征来训练模型也是很有必要的。

此外，选取过少的历史交易信息会导致拟合精度的下降，选取过多又会导致数据过拟合的问题，因此模型输入特征中包含目标预测日前多少天的历史交易信息也是值得探究的。

本文以沪深300股指期货为研究对象，首先使用VAR模型对股指期货现货价格变动的相关关系进行研究与验证，并使用脉冲响应函数分析价格变动的冲击影响与时效，从而找到最优的历史交易信息天数；其次基于在决策树基础上发展而来的XGBoost算法，将相应的变量作为特征训练模型，并调整出最优参数，再用部分历史数据测试最优参数模型的拟合效果，从而作出相应的评价。

2 股指期现货价格变动关系的实证分析

本部分使用的数据为沪深300股票指数日结算价以及沪深300股指期货连续合约(IF0)日结算价，数据来源均为同花顺iFinD金融终端；数据区间为2017年第一个交易日(1月3日)到2020年春节前最后一个交易日(1月23日)，经缺失值和异常值剔除后，共得724组数据。

模型构建过程如下：第一，将两序列数据进行对数化处理；第二，通过ADF方法检验平稳性，确定将两者的一阶差分作为变量构建VAR模型；第三，按照SBIC准则确定模型的最优滞后阶数为三阶。

模型应用结果如下：首先，稳定性判别图(图1左)显示构建的模型具有很好的稳定性，说明股指期现货价格变动之间存在稳定的相互关系；其次，格兰杰因果检验显示，股指现货价格变动是股指期货价格变动的因，而股指期货价格变动不是现货价格变动的因，这说明将股指现货价格作为特征来训练模型的思想是正确的；最后，脉冲响应图(图1右)显示现货对期货有三期的显著影响，因此可确定模型输入特征中应包含股指期货价格目标预测日前3天的标的股指价格结算价信息。

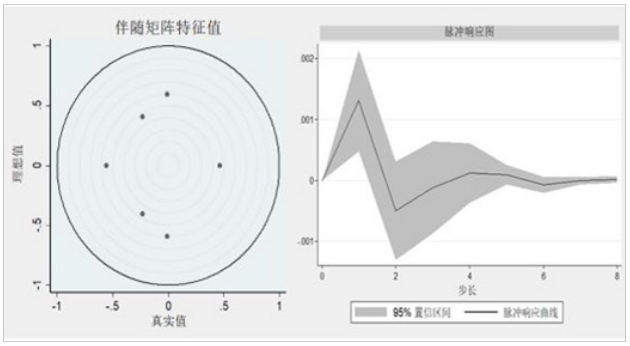

图1 模型应用结果

3 基于XGBoost的股指期货价格预测模型

3.1 XGBoost原理及优势

首先，XGBoost算法是在梯度下降树(GBDT)算法的基础之上经过改进得到的，目的是为了解决GBDT算法中的缺陷。XGBoost是Boosting中的代表性算法，首先以原始数据集为基础训练出第一个弱学习器，计算得到预测值和真实值之间的残差，并将其作为下一个弱学习器的学习对象，每一个弱学习器都以减小真实值与预测值之间的差距为目的[5]。这样看来，XGBoost就是一种迭代的决策树算法[6]，模型输出可以表示为N个弱学习器的叠加(式(1))：

$$\widehat{y_i} = \sum_{k=1}^{N} f_k(x_i) \tag{1}$$

其次，XGBoost目标函数$Obj^{(t)}$中(式(2))包含了正则项$\Omega(f_t)$(式(3))，其中T表示树中的叶子节点数量，w_j^2表示树中

作者简介：蔡泽栋 (1999-)，男，汉族，浙江绍兴人，本科在读，金融工程专业。

某一叶子节点j得分w的L2模平方。对w进行L2正则化，相当于针对每个叶节点的得分增加L2平滑，具有防止过拟合的作用。这种特点增强了拟合和泛化能力与稳定性，因此更适合股指期货价格的预测。

$$Obj^{(t)}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_i)+C \quad (2)$$

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^{\ 2} \quad (3)$$

最后，许多较新的算法及机器学习平台，如卷积神经网络(CNN)[7]、长短期记忆网络(LSTM)[8]、谷歌研发的TensorFlow[9]等都在金融时序预测问题中有过出色的表现。这些算法在预测准确率上优于传统的机器学习算法，但存在着迭代次数多、运行较慢、电脑配置要求较高等局限性。相较而言，XGBoost算法将传统机器学习算法进行了改良与增强，且运行速度较快，是权衡精确度与效率之后更好的选择。

3.2 模型训练与测试

3.2.1 数据下载与预处理

利用Python自带的Tushare包，下载2017年第一个交易日(1月3日)至2020年春节前最后一个交易日(1月23日)沪深300指数日结算价，同时通过request库爬取新浪财经网页中同一数据区间的沪深300股指期货合约开盘价、收盘价、最高价、最低价、成交量等数据。经数据预处理后，最终获得的有效交易数据为724条。

3.2.2 数据降噪

由于市场动态较为复杂，数据含有不确定性的噪声，因此需要将每列数据均通过小波变换去除噪声。这里使用pywt库来进行有关处理。

3.2.3 模型特征选取

基于上一部分的分析，本文使用每个目标预测日过去前3天的股指期货合约结算价、最高价和最低价的差值、开盘价和收盘价差值、成交量以及过去前3天的标的股指结算价共15个特征构建特征工程，来对沪深300股指期货合约结算价进行预测。

3.2.4 特征缩放

将特征序列进行缩放是非常重要的，若没有进行特征缩放，模型对于超出范围的验证集将预测的非常不准确。因此，本文先使用sklearn库中的StandardScaler函数将序列集合的均值缩放为0，方差为1，然后使用这些缩放后的特征来作预测，最后再将获得的缩放后的预测值通过对应的均值和方差实现逆变换获得原值。这种扩展方式便可提供较好的性能。

3.2.5 模型训练

使用xgboost库中的XGBRegressor函数生成模型，并设置初始参数进行模型的训练与交叉验证。这里设置训练集占比为70%，共505条数据；交叉验证集占比为15%，共108条数据；剩下108条数据为测试集。最终得到部分训练集预测结果如图2所示，可见预测训练效果还是比较好的。

3.2.6 模型参数优化

按照平均绝对百分比误差(MAPE)值最小法则优化如下几个参数：

参数优化1：最大迭代次数(n_estimators)与树的最大深度(max_depth)。树的最大深度值用于控制过拟合，范围一般为3~10。最大深度越大，模型学习得更加具体。

参数优化2：学习率(learning_rate)与子节点最小样本权重和(min_child_weight)。如果一个叶子节点的样本权重和小于min_child_weight，则拆分过程结束。min_child_weight越大，算法越保守。

参数优化3：gamma指定了节点分裂所需的最小损失函数下降值，参数值越大，算法就越保守，因为gamma值越大时，损失函数下降更多才可以分裂节点；subsample参数用于控制对于每棵树中随机采样的比例，减小该参数的值，算法会更加保守，避免过拟合，但如果设置得过小可能会导致欠拟合。

参数优化4：colsample_bytree用来控制每颗树随机采样的特征数量的比例，范围在0~1；colsample_bylevel用来确定每棵树每次节点分裂时特征采样的比例，默认值为1。

最终结果：图3显示了各个参数优化的过程，参数初始值与优化后值对比如表1所示，可见所有参数的值都发生了变化，对交叉验证集使用新参数进行预测之后，MAPE值也有所下降，进一步提高了预测的精度。

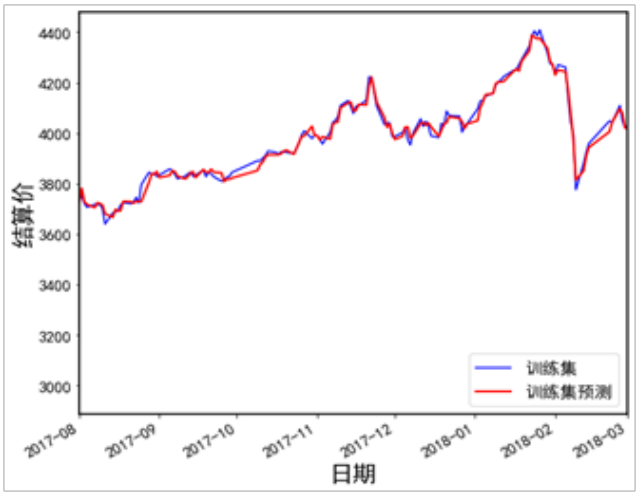

图2 训练集预测结果

表1 参数优化前后数值对比

参数	初始值	最优值
n_estimators	100	260
max_depth	3	2
learning_rate	0.1	0.3
min_child_weight	1	5
subsample	1	0.8
colsample_bytree	1	0.9
colsample_bylevel	1	0.9
gamma	0	0.9
MAPE	0.997%	0.989%

3.2.7 模型测试与比较

将上文所得拥有最优参数的XGBoost模型应用于测试集数据的预测中，得到MAPE的值为0.644%，相较于交叉验证集更

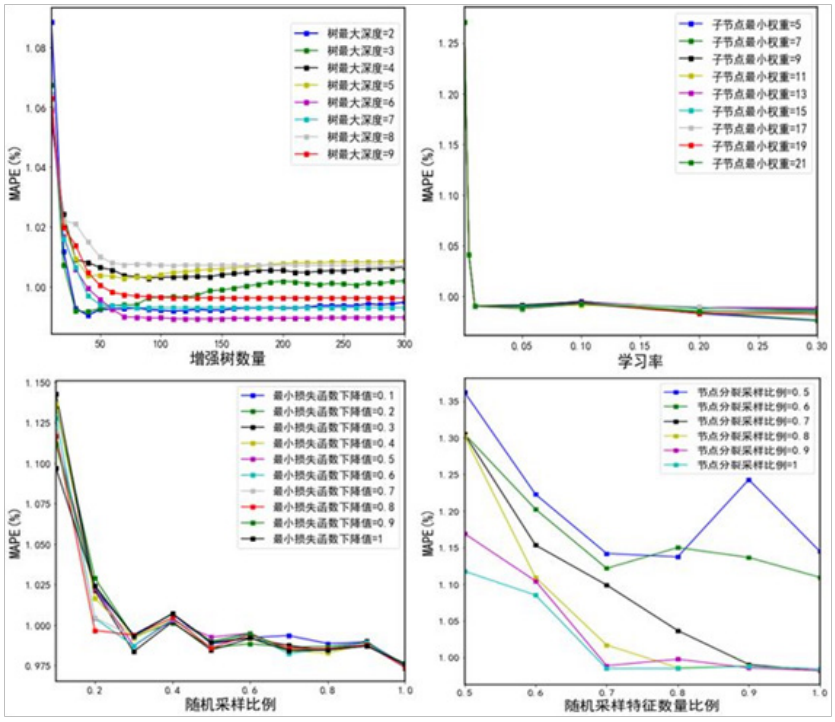

图3 各参数优化过程

低，说明在测试集中模型的预测效果更好。最终预测结果及比较如图4所示，可以发现，与不包含标的股指价格历史交易信息的预测结果相比(图右侧)，左侧精度更高，对价格变动中一些细小趋势的把握更准确，这再次验证了标的股指历史交易信息对股指期货价格预测影响的重要性。

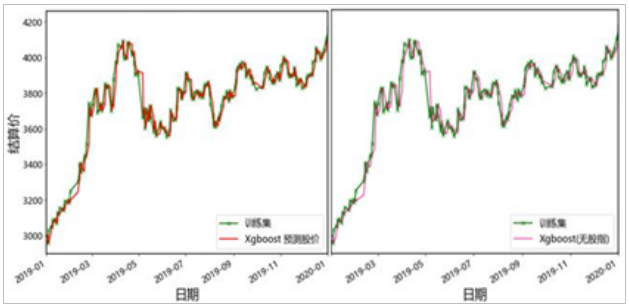

图4 XGBoost测试集预测结果及比较

4 结语

本文以沪深300股指期货为研究对象，通过构建VAR模型发现了标的股指现货价格对股指期货价格预测的重要作用，并依此构建模型特征，使用XGBoost算法对沪深300股指期货结算价进行了预测，最终得到了较好的预测结果，验证了XGBoost算法在股指期货价格预测中的有效性以及标的股指历史交易信息对股指期货价格预测的重要性。

对于股指期货市场来说，标的股指的近期历史交易信息与期货自身的市场交易数据同等重要。将两者的交易数据进行结合，同时借助XGBoost机器学习算法，便可高效准确地基于历史信息对未来的期货价格进行预测。然而，金融市场是复杂多变的，一些政策面的消息、极端事件的发生，都会导致预测误差甚至失灵。因此，投资者若想更准确地预测股指期货价格，后期在考虑加入对期价影响大的市场冲击以提升模型稳定性及预测准确度的同时，还应该训练自身对金融市场的信息敏感度，以便在市场发生非预期波动时，能结合计算机分析作出更正确、更及时的决策。

参考文献

[1] 沈金榕.基于决策树的逐步回归算法及在股票预测上的应用[D].广州:广东工业大学,2017.
[2] 李腾.基于小波分析和BP神经网络相结合的股票波动预测方法研究[D].天津:天津大学,2018.
[3] 黄秋萍,周霞,甘宇健,等.SVM与神经网络模型在股票预测中的应用研究[J].微型机与应用,2015,34(05).
[4] 王康.我国股指期货影响因素的研究——基于VAR模型的实证分析[J].时代金融,2018(09).
[5] 周徐,方东旭,文冰松.一种基于GBDT机器学习的算法及应用研究[J].电信工程技术与标准化,2019,32(11).
[6] 王子通.基于XGBoost的沪深300股指期货交易策略研究[D].西安:西北大学,2019.
[7] 陈祥一.基于卷积神经网络的沪深300指数预测[D].北京:北京邮电大学,2018.
[8] 彭燕,刘宇红,张荣芬.基于LSTM的股票价格预测建模与分析[J].计算机工程与应用,2019,55(11).
[9] 韩山杰,谈世哲.基于TensorFlow进行股票预测的深度学习模型的设计与实现[J].计算机应用与软件,2018,35(06).