基于 XGBoost 和 LSTM 模型的金融时间序列预测

黄 颖，杨会杰

（上海理工大学 管理学院，上海 200093）

摘要：随着人工智能快速发展，深度学习模型预测金融时间序列成为热点问题。数据及特征选取是决定模型效果的重要环节，用 XGBoost 模型进行特征优化并预测黄金价格涨跌趋势，再与 LSTM 模型比较预测效果。用 XGBoost 分析动量因子特征重要性并选取有效指标；形态因子做历史回测并选取胜率较高的 K 线指标，预测准确率提升 1.5%。以相同因子为 LSTM 模型特征值预测准确率提升 6.5%，达到 80%。以欧元和浦发银行股价数据为样本均证实 K 线指标有效且 LSTM 模型预测效果优于 XGBoost。

关键词：XGBoost；技术因子；K 线；LSTM；特征工程

中图分类号：TP389.1 **文献标志码**：A **文章编号**：1671－1807(2021)08－0158－05

股票市场是企业筹集资金的重要渠道之一，也是资本市场的重要组成部分。它的波动同经济市场的盛衰息息相关，有效的趋势预测可为企业和个人带来可观的利润。该研究成为学术界一项重要课题。在股市预测中，常见的方法包括前馈神经网络、反向传播神经网络和循环神经网络等模型，许多学者致力于用此类方法预测金融市场。

在传统机器学习领域，Prasaddas 和 Padhy 证明支持向量机 SVM 对未来价格预测的效果优于反向传播算法[1]；Tay 和 Cao 比较了 SVM 和多层反向传播算法[2]，依然证实 SVM 效果更佳；Jain 等比较 XGBoost 模型和传统回归模型[3]，发现 XGBoost 预测准确率更高；Sawon 和 Hosen 研究 XGBoost 和线性回归、随机森林回归等算法在大规模数据中的预测效果[4]，同样证实 XGBoost 算法的优越性。

在神经网络模型方面，Wilson 和 Sharda 比较了神经网络和传统的多元判别分析在公司破产问题上的表现[5]；Hsieh 等将小波转换和循环神经网络 RNN 结合应用于股市预测[6]；Kamijo 和 Tanigawa 用 RNN 识别股票模式[7]；Akita 等证明了 LSTM 比其他模型更能捕捉时间序列信息[8]；Chen 等用 GRU 预测金融时间序列并在误差较小的情况下效果良好[9]。

本文运用预测效果较好的传统机器学习 XGBoost模型对黄金价格进行预测，根据对 K 线形态因子在历史数据上回测结果的统计分析、动量因子在模型特征重要性中的表现，优化特征选取，降低特征选取的主观性；将优化后的特征用于能够有效记忆长短期信息并克服 RNN 模型梯度爆炸或梯度消失问题的 LSTM 模型，探索能否进一步提升预测效果。

1 基本理论

1.1 XGBoost 极端梯度提升模型

XGBoost 简称为 Extreme Gradient Boosting，由华盛顿大学的陈天奇博士提出。它是由若干棵树集成的模型，每棵树对样本预测值之和为 XGBoost系统对该样本的预测值，函数定义为

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

式中：K 为树的总数；$f_k(x_i)$ 为第 k 棵树对第 i 个 x_i 的预测结果；$\hat{y}_i$ 为 XGBoost 系统对第 i 个样本的预测结果。

模型目标函数为

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

式中：$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$；$\hat{y}_i$ 为样本预测值；y_i 为样本标签真实值；T 为每棵树的叶子节点数；w_j 为第 j 个叶子节点的权重；γ 和 λ 为系数，实际应用中需调参。

目标函数右式的第 1 项为损失函数，第 2 项为

收稿日期：2021-04-02

作者简介：黄颖(1996—)，女，江苏淮安人，上海理工大学管理学院，硕士研究生，研究方向为数量经济学；通信作者杨会杰(1968—)，男，河北保定人，上海理工大学管理学院，教授，博士研究生导师，研究方向为系统生物学、定量经济与定量金融学、统计物理。

正则化项。损失函数即训练误差，类似于用于回归的均方误差 MSE，是可微的凸函数；正则化项是为了防止模型过拟合而存在的，是一个惩罚项，控制模型复杂度。因此，模型所要达到的效果是使目标函数 $L(\phi)$ 最小化。

XGBoost 优化了 GBDT 算法。在求解最小化目标函数的模型 f_x 过程中，GBDT 只采用一阶导数，而 XGBoost 算法对损失函数做二阶泰勒展开；由于传统的枚举所有可能分割点以寻找最佳分割点的贪心算法效率太低，XGBoost 还提出了根据百分位法列举几个候选分割点，再根据公式求解最佳分割点的优化方法。

1.2 LSTM 长短期记忆模型

LSTM 模型[10]是对 RNN 的改进，既能同时储存长期和短期的信息，又能有效避免 RNN 模型存在的梯度爆炸或梯度消失问题。主要应用于 Nature Language Processing，speech recognition 等时间序列数据。

LSTM 网络包括输入层、隐藏层和输出层，每一层包含许多单元；隐藏层中每个单元都有记忆细胞，输出门、遗忘门和输出门共同决定输出值的生成。隐藏层单元内部结构[11]如图 1 所示。

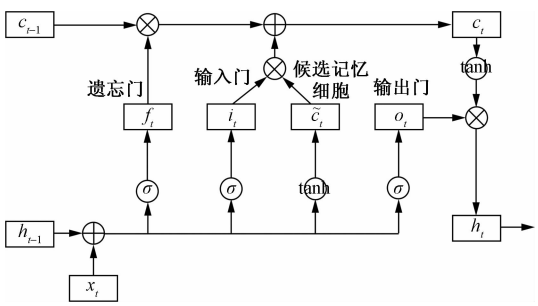

图 1 隐藏层单元结构

图 1 中，x_t 为 t 时刻输入；c_t 为 t 时刻记忆细胞；h_t 为 t 时刻隐藏单元；f_t 为 t 时刻遗忘门；i_t 为 t 时刻输入门；$\tilde{c}_t$ 为 t 时刻候选记忆细胞；o_t 为 t 时刻输出门；tanh 和 σ 均为激活函数。

图 1 可知，f_t、i_t 和 o_t 的输入均为当前时刻的 x_t 与上一时刻的 h_{t-1}，加入 σ 激活函数后输出，取值范围为 $[0,1]$；$\tilde{c}_t$ 输入不变，但激活函数取 tanh 函数，取值范围为 $[-1,1]$。

c_t 和 h_t 均包含上一时刻的 c_{t-1} 和当前时刻的 $\tilde{c}_t$ 两部分信息。f_t 限制 c_{t-1} 保留至 c_t 的信息，i_t 限制 $\tilde{c}_t$ 保留至 c_t 的信息，o_t 限制当前时刻 c_t 到 h_t 的信息流动。具体计算公式为

$$\begin{cases} i_t = \sigma(x_t w_{xi} + h_{t-1} w_{hi} + b_i) \\ f_t = \sigma(x_t w_{xf} + h_{t-1} w_{hf} + b_f) \\ o_t = \sigma(x_t w_{xo} + h_{t-1} w_{ho} + b_o) \\ \tilde{c}_t = \tanh(x_t w_{xc} + h_{t-1} w_{hc} + b_c) \end{cases} \tag{3}$$

$$\begin{cases} c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{4}$$

式中：w_{xf}、w_{xi}、w_{xo}、w_{xc} 为 x_t 到 f_t、i_t、o_t、$\tilde{c}_t$ 的权重；w_{hf}、w_{hi}、w_{ho}、w_{hc} 为 h_{t-1} 到 f_t、i_t、o_t、$\tilde{c}_t$ 的权重；b_f、b_i、b_o、b_c 为 f_t、i_t、o_t、$\tilde{c}_t$ 的偏置；$\odot$ 为按元素相乘。

此算法可有效解决 RNN 模型的梯度衰减问题。例如当 f_t 接近 1 且 i_t 接近 0，则过去信息将会一直保留，便于更好地捕捉序列中时间步较大的样本关系。o_t 越接近 1，c_t 将有越多的信息传递至 h_t，并体现于输出层结果；反之，越接近 0，c_t 传输至 h_t 的信息越少。

2 模型建立与求解

2.1 数据处理

由于获取黄金价格数据 2010—2020 年为每分钟线的开盘价、收盘价、最高价、最低价数据，但步长太小时数据波动较少，近似于求解微分方程，训练效果可信度不高，因此将数据处理为一小时线数据。以第 1 min 的开盘价为小时线开盘价，第 60 min的收盘价为小时线收盘价，60 min 内最高价为小时线最高价，60 min 内最低价为小时线最低价。删除有缺失值的数据，并进行归一化处理，得到样本量为 62 440 条数据。

2.2 XGBoost 模型和技术指标选取

技术因子是以价格、成交量等作为计算因子，通过一定数学计算得到的判断价格走势的数据、图形等，可用于判断买卖时间和价格。以 40 个技术指标及黄金的高开收低价作为特征构建 XGBoost 模型。首先根据每小时的开盘价和收盘价计算收益率 q_i，计算公式为

$$q_i = \frac{\text{close\_price}_i - \text{open\_price}_i}{\text{open\_price}_i} \tag{5}$$

式中：close_price_i 为 i 时刻的收盘价；open_price_i 为 i 时刻的开盘价。根据收益取值分布为样本涨跌趋势打标签，取值分布直方图如图 2 所示。

可见大多数样本集中在 $[-2.5,2.5]$ 区间内，总体符合正态分布特征。取值为总样本前 20% 判为涨，后 20% 判为跌，其余为震荡。建立 XGBoost 模型并用网格搜索法调参，树的最大深度为 10，学习率为 0.1，叶子节点最小权重和为 3，每棵树随机采样的特征比例为 0.7，目标函数使用 softmax 的

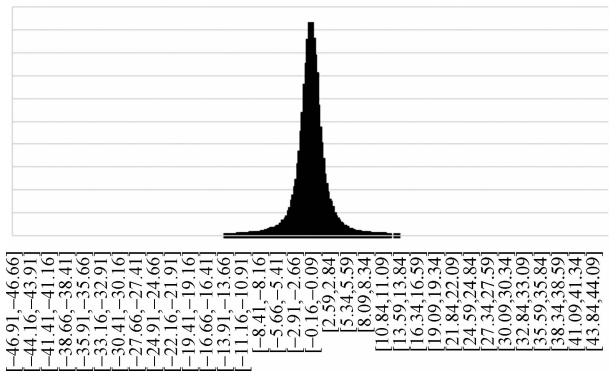

图 2　股市收益取值分布直方图

多分类器，类别数目为 2。输出模型的特征重要性并剔除掉对结果影响较小的特征，最终保留 21 个特征，重要性分布均匀，无冗余特征，见表 1。

为使预测结果有统计意义，在前 60 000 组数据中，以步长为 5 000、训练样本量为 15 000 滑动选取样本，共 9 次训练。例如第 1 次训练样本为第 1～15 000 组，第 2 次训练样本为第 5 000～20 000 组，以此类推；测试样本均为后 15 组数据，对 9 次结果取均值，平均预测准确率为 72%。

2.3　K 线形态因子统计分析

K 线形态因子又称蜡烛线，是股票分析的重要工具，有 61 个指标，包含信息丰富，可直观表示股价趋势强弱、买卖双方力量平衡的情况。从 Talib 官网获取 K 线形态的 C 语言源代码，分析各指标对于涨跌趋势的判断逻辑。用黄金历史数据做回测，从不同持仓时间和预测周期的维度，统计各指标出现次数、胜率及平均收益，结果如图 3 所示。

表 1　技术因子选取

技术指标	RSI	MOM1	PPO	SAR	MACDDEA	MACDHIST	RSIMA
特征重要性	0.033	0.030 7	0.035	0.035	0.040 659 42	0.035 344 1	0.032 46
技术指标	CCI	ROC	CMO	ADX	ULTOSCMA	MACDDIF	CCIMA
特征重要性	0.034	0.034 2	0.034	0.035	0.036 321 02	0.035 532 34	0.038 83
技术指标	TRIX	MOM	BOP1	BOP	ULTOSC	AROONOSC	ADXMA
特征重要性	0.035	0.036	0.037	0.031	0.032 728 53	0.039 952 7	0.031 86

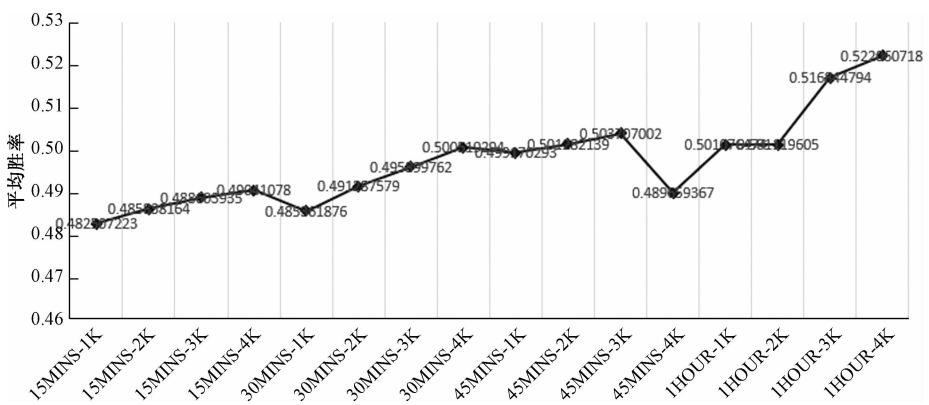

图 3　不同维度的指标平均胜率

可见持仓时间越长、预测周期越长，整体胜率越高，在时间维度上，持仓时间为 1 h 的平均准确率最高。依据一小时线的平均胜率、平均收益率以及总出现次数对各指标进行排序并打分。发现大部分指标的预测准确率在 48%～52%，认为概率随机；选取预测准确率在 52%以上的 8 个指标作为特征，具体指标见表 2。

将上述 8 个特征加入 XGBoost 模型中进一步训练，9 次训练的平均准确率为 73.5%；观察特征重要性，发现 21 个技术因子和 8 个形态因子的重要度均在 0.03 左右，无冗余特征。证明胜率较高的形态因子的加入使模型效果提升，但幅度不是很大。

表 2　K 线形态因子选取

指标	胜率/%	出现次数
CLOSINGMARU	56	753
MARUBOZU	55	268
ENGULFING	54	652
HIKKAKE	54	854
LONGLINE	54	1 185
RICKSHAWMAN	53	812
SHORTLINE	53	1 226
SPINNINGTOP	53	1 658

2.4 LSTM 模型

为使 XGBoost 模型和 LSTM 模型预测结果具有可比性，在建立 LSTM 模型时，保持选取的技术因子和形态因子不变，样本选取和训练集、测试集划分方式均不变，构建两层 LSTM 和两层全连接层的模型。通过网格搜索法选取最佳参数。

模型结构设置：第 1 层 LSTM 层的输出神经元数为 16，第 2 层 LSTM 层的输出神经元数为 4，第 1 层全连接层输出神经元数为 4，最后为 1 个神经元的输出层。参数设置：滑动窗口为 22，即每次迭代用前 22 组特征值预测后一时刻的涨跌趋势；损失函数为最小均方误差 MSE；优化器为 adam；batch_size 为 200，epoch 取 10，表示每训练完 200 组样本为一次迭代并更新一次权重，训练完全部 15 000 组样本为一个 epoch，训练完 10 个 epoch 则权重更新 (15 000/200)×10＝750 次。此为一次训练过程完整结束。

以步长 5 000 滑动选取样本，9 次训练全部完成后取均值，平均预测准确率达到 80%，再次提升了 6.5%，且结果具有统计意义。

3 结果分析

为了证明特征工程及模型结果的有效性，避免数据选取偶然性带来的不确定性，选取欧元和浦发银行股价数据，按照同样的方法进一步做实证研究。结果见表 3。

表 3 实证研究结果分析

数据类别	预测准确率/%			个数
	技术	K 线	LSTM	
黄金	72.0	73.5	80.0	8
欧元	59.5	60.7	67.5	5
SPD	69.3	70.0	78.2	6

表 3 中第 1 列为选取的 3 组数据，第 2 列为根据特征重要性筛选技术因子后的 XGBoost 模型预测准确率，第 3 列为根据历史数据回测筛选出效果较好的形态因子加入 XGBoost 模型后的预测准确率，第 4 列为将优化后的特征用于 LSTM 模型的预测准确率，第 5 列为 3 组数据分别加入的形态因子个数。从整体来看，由于欧元的初始分钟线数据只有 104 万条，经处理后的小时线数据量大约为 16 000条，因此滑动选取的训练样本量为 5 000 条，预测准确率绝对值低于另外两组数据。

分别观察 3 组数据预测结果变化：黄金数据加入 8 个形态因子后，预测准确率提升 1.5%；欧元数据加入 5 个形态因子后，预测准确率提升 1.2%；浦发银行数据加入 6 个形态因子后，预测准确率提升 0.7%。比较优化特征后的 XGBoost 模型和 LSTM 模型：黄金数据预测准确率提升了 6.5%；欧元数据提升了 6.8%；浦发银行数据提升了 8.2%。可见，在历史数据回测中表现优越的少数 K 线形态因子加入模型对预测效果有一定提升作用，LSTM 模型也比 XGBoost 模型预测准确率更高。

4 结论

通过对 K 线形态因子和技术因子的分析及特征优化工作，进一步比较 XGBoost 和 LSTM 模型在预测金融时间序列方面的能力，得出以下结论：

1)将 K 线形态 61 个因子在历史数据中做回测，筛选出不同持仓时间和预测周期的平均胜率和收益率较高且出现次数较多的因子仅有不到 10 个，可见大部分因子的判断效果近似随机；但将少部分表现较好的因子作为特征应用到模型中，仍使预测效果提升 1%左右，提升力度不大。

2) LSTM 模型效果明显优于 XGBoost 模型，由于 LSTM 模型可同时记忆长期和短期信息，并能有效解决梯度消失和梯度爆炸问题，更适用于跨度较长的时间序列数据预测问题。

参考文献

[1] PRASADDAS S，PADHY S. Support vector machines for prediction of futures prices in Indian stock market[J]. International Journal of Computer Applications，2012，41(3)：226.

[2] TAY F E H，CAO L. Application of support vector machines in financial time series forecas ting[J]. Omega，2001，29(4)：309317.

[3] JAIN A，MENON M N，CHANDRA S. Sales forecasting for retail chains[R]. India，2015.

[4] SAWON M，HOSEN M. Prediction on large scale data using extreme gradient boosting[R]. Dhaka：BRAC University，2016.

[5] WILSON R L，SHARDA R. Bankruptcy prediction using neural networks[J]. Decision Support Systems，1994，11(5)：545－557.

[6] HSIEH T，HSIAO H，YEH W. Forecasting stock markets using wavelet transforms and recurrent neural networks：An integrated system based on artificial bee colony algorithm[J]. Applied Soft Computing，2011，11(2)：2510－2525.

[7] KAMIJO K，TANIGAWA T. Stock price pattern recognition：a recurrent neural network approach[C]//In Proceedings of the International Joint Conference on Neural Networks (IJCNN). Washington：1990：215221.

[8] AKITA R,YOSHIHARA A,MATSUBARA T,et al. Deep learning for stock prediction using numerical and textual information[C]// Computer and Information Science. IEEE/ACIS 15th International Conference on,IEEE, 2016: 1－6.

[9] CHEN W,ZHANG Y,YEO C K,et al. Stock market prediction using neural network through news on online social networks[C]//Smart Cities Conference (ISC2),2017 International,2017:11－2.

[10] OMER B S,MEHMET U G,AHMET M O. Financial time series forecasting with deep learning: A systematic literature review:2005－2019[J]. Applied Soft Computing Journal, 2020,90:106181.

[11] 甘文娟,陈永红,韩静. 基于正交参数优化的 LSTM 结构变形预测模型[J]. 计算机系统应用,2020,29(9):212－218.

Financial Time Series Forecasting Based on XGBoost and LSTM Models

HUANG Ying, YANG Hui-jie

(Business School,University of Shanghai for Science and Technology,Shanghai 200093,China)

Abstract: With the rapid development of artificial intelligence, deep learning model prediction of financial time series has become a hot issue. The selection of data and features is important for the effect of the model. XGBoost is used to optimize the features and predict the trend of gold price trend, and then the prediction effect is compared with LSTM. XGBoost is used to analyze the importance of momentum factors and select effective features. The morphological factors are tested on historical data,and the candlestick chart with higher accuracy rate are selected, the prediction accuracy is increased by 1.5 percent. The same factors are used in LSTM, and the prediction accuracy is increased by 6.5 percent to 80 percent. Taking Euro and SPD Bank stock price data as samples, it is proved that the prediction effect of LSTM is better than that of XGBoost.

Key words: XGBoost; technical factors; K-line; LSTM; feature engineering