Advances in Applied Mathematics 应用数学进展, 2022, 11(11), 7850-7855
Published Online November 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1111830

Hans汉斯

基于LSTM的棉花期货价格预测研究

苏郅宏，陈　翔*，陈茹君，陈嘉颖

闽江学院数学与数据科学学院(软件学院)，福建　福州

收稿日期：2022年10月14日；录用日期：2022年11月8日；发布日期：2022年11月17日

摘　要

本文主要基于LSTM神经网络对棉花期货价格进行预测研究。结果表明，LSTM模型预测的均方根误差(RMSE)、平均绝对误差(MAE)以及平均绝对百分比误差(MAPE)均小于Xgboost模型，故LSTM的预测精度比XGboost更好。LSTM模型在棉花期货价格上的预测表现出较好的性能。

关键词

棉花期货，LSTM神经网络，Xgboost模型，预测

Research on Prediction of Cotton Futures Price Based on LSTM

Zhihong Su, Xiang Chen*, Rujun Chen, Jiaying Chen

College of Mathematics and Data Science (Software College), Minjiang University, Fuzhou Fujian

Received: Oct. 14th, 2022; accepted: Nov. 8th, 2022; published: Nov. 17th, 2022

Abstract

In this paper, LSTM neural network is used to predict cotton futures prices. The experiment results show that the root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) of the LSTM model were smaller than the Xgboost model, so the forecast accuracy of LSTM model is better than Xgboost model, and LSTM model has good performance in the prediction of cotton futures prices.

Keywords

Cotton Futures, LSTM, Xgboost Model, Prediction

*通讯作者。

文章引用: 苏郅宏, 陈翔, 陈茹君, 陈嘉颖. 基于 LSTM 的棉花期货价格预测研究[J]. 应用数学进展, 2022, 11(11): 7850-7855. DOI: 10.12677/aam.2022.1111830

1. 引言

近几个月以来，全球棉花产量、消费量环比调减，国内棉花销售持续缓慢，库存走高，因而国内棉花的价格波动大，总体上进一步萎缩，其主要原因在于国内厂商因为成本较高而减少订单量，其次是欧美国家疯狂加息，棉织品消费市场极度不稳定。本文基于长短期记忆神经网络模型(LSTM)对棉花期货的价格进行预测研究，为投资者提供价格指导，为企业提供价格预警，合理应对市场风险。

统计学中，针对期货价格预测中效果最好、应用最广的方法主要是回归分析法和神经网络分析法。郭婷婷[1]运用 PCA-ELM 方法对我国粮食价格进行预测研究，其模型对传统的神经网络和时间序列进行改良，客观上为我国粮食价格的预测提供了一种新方法。王进[2]通过 BP 神经网络模型对肉鸡价格进行研究，其研究成果不但能预测未来某一时刻的价格是上涨还是下跌，还可以通过预测模型来预测未来某一时刻的价格，大幅降低预测误差，预测精度更高。目前，也有很多基于 LSTM 模型对农产品价格等作预测的研究[3] [4] [5] [6]。

大多数金融时间序列可能存在异方差，波动集聚性，我们所常用来预测的时间序列模型虽然能比较快速的预测出短期内期货价格，但其模型的精确程度远不如神经网络预测模型。传统的神经网络由于对样本具有较强的依赖性，预测能力和训练能力的可能存在矛盾等问题。LSTM 模型作为循环神经网络模型(RNN)的一种变体，其最大的特点就是具有时间循环结构，可以很好地刻画具有时间关联的金融序列数据。

本文主要是基于 LSTM 神经网络对棉花期货价格的预测研究。而近年来，由陈天奇团队开发的 XGboost 提升算法在许多研究中有着不错的成效，因此本文将分别通过两种算法来进行预测研究，比较出哪一种算法的预测精度更好。

2. 研究理论

2.1. 长短期记忆网络模型(LSTM)

传统 RNN 循环网络可以实现对数据的短期记忆并进行连续预测，但在处理长序列数据中，会使时间展开步更长，在反向传播更新参数时，梯度要按时间步连续相乘，会导致梯度消失，所以出现了长短期记忆网络。长短期记忆网路引入了如下概念：

$$
\begin{aligned}
&\text{输入门}: i_t = \sigma\left(W_i \bullet \left[h_{t-1}, x_t\right] + b_i\right)\\
&\text{遗忘门}: f_t = \sigma\left(W_j \bullet \left[h_{t-1}, x_t\right] + b_f\right)\\
&\text{候选态}: \tilde{C}_t = \tanh\left(W_c \bullet \left[h_{t-1}, x_t\right] + b_c\right)\\
&\text{细胞态}: C_t = f_t * C_{t-1} + i_t * \tilde{C}_t\\
&\text{输出门}: o_t = \sigma\left(W_o \bullet \left[h_{t-1}, x_t\right] + b_o\right)\\
&\text{记忆体}: h_t = o_t * \tanh\left(C_t\right)
\end{aligned}
$$

其中下标 t 表示时刻，x 为 LSTM 单元的输入信号，h 为 LSTM 单元的输出信号 C 为 LSTM 单元状态，W 为权重矩阵，b 为偏置向量，σ,tanh 分别为 sigmoid 和 tanh 激活函数。LSTM 单元中有三个门，分别用来控制增加或者删除信息，从而实现记忆或遗忘的功能，是一种具有信息选择性的结构。细胞态也称作长期记忆，是上个时刻的长期记忆乘遗忘门，加上当前时刻归纳出的新知识，乘以输入门的结果；记忆

体也称作短期记忆；候选态也称作归纳出的新知识。

传统 RNN 神经网络模型因为存储量不足，存在着很难长时间学习并保存信息的问题，而 LSTM 模型是 RNN 模型的一个优秀的变体，它继承了大部分 RNN 模型的特性，其采用了特殊隐式单元的存储方法能够做到长期的学习和保存输入信息。但同时也存在缺点是：若时间序列的跨度太大，网络深度太高，这时候 LSTM 模型的计算量会非常大，耗时成了一个比较棘手的问题。

2.2. XGboost (eXtreme Gradient Boosting)

XGboost 模型是由陈天奇开发的一个开源机器学习项目。本质上，XGboost 模型是对 GBDT 模型进行提升与改进，目的是将速度和效率提升到极致。使用上，改进效果十分明显，使用起来简单，学习速度快，被广泛运用在学术界，工业界中。

XGboost 基本思路是就是通过学习不断降低模型的偏差，也就是不断生成新的树，每棵树的生成是基于前面树的预测结果和实际值的残差进行学习，算法如下：

$$\begin{aligned}
&\hat{y}_k^{(0)} = 0\\
&\hat{y}_k^{(1)} = f_1(x_k) = \hat{y}_k^{(0)} + f_1(x_k)\\
&\hat{y}_k^{(2)} = f_1(x_k) + f_2(x_k) = \hat{y}_k^{(1)} + f_2(x_k)\\
&\quad\vdots\\
&\hat{y}_k^{(t)} = \sum_{j=1}^{t} f_j(x_k) = \hat{y}_k^{(t-1)} + f_t(x_k)
\end{aligned}$$

其中 $\hat{y}_k^{(t)}$ 是前 t 棵树时模型总的预测值，$\hat{y}_k^{(t-1)}$ 是前 $t-1$ 棵树模型总的预测值，$f_t(x_k)$ 是第 t 棵树的预测值。

XGboost 模型的一个优点在于允许样本缺失。针对某一样本的某一维度中，若存在样本缺失，则 XGboost 模型会将缺失的样本放入左子树计算一次信息熵，再放入右子树计算一次信息熵，根据两次信息熵的结果再来决定放在哪个子树中。

3. 模型构建与结果分析

3.1. 数据选取及指标构建

我国棉花期货在 2004 年在郑州上市。棉花产业产量高、需求大，价格容易受多种因素所影响，从上市到现在 18 年的时间里只有三次较大幅度的期货上涨。第一波是 2010 年的超级牛市，第二波是 2016 年至 2018 年初，股市大跌之后的反弹，第三波则是新冠疫情以后的大规模上涨。其他绝大多数时间均以波动为主。而在上市初期的棉花期货市场波动为主，由于受元旦节假日的影响，本文选取数据的时间为 2006 年 1 月 4 日至 2022 年 9 月 30 日，数据来自 Tushare (https://www.tushare.pro/)，该网站提供了许多免费、开源的财经数据接口，研究者能够较为方便的对所需数据进行采集、清洗以及存储，最后进行学术研究。本文选取 80%的总体数据作为训练集，选取 20%的总体数据作为测试集。

现实中，影响棉花期货价格的宏观因素例如农产品指数、采购经理指数等更新时间较长，用于预测效果低，意义不大，因此本文选取的预测指标为棉花期货交易数据的开盘价、最高价、最低价、成交量和持仓量五个指标。

3.2. 数据归一化及预测方法

本文所构建的棉花期货价格预测指标中不同指标间数值差别较大，在将总体数据放到 LSTM 进行训练前我们对其进行 Min-Max Normalization 归一化，将总体数据收缩至[0, 1]之间，计算公式如下：

$$x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

其中，x_i'表示第 i 时刻的期货价格经过标准化处理后的数据，x_i表示第 i 时刻的期货价格，$\max(x_i)$和$\min(x_i)$表示第 i 个时刻的期货价格的最大值和最小值。

为了避免由于数据滞后期的延长导致模型对棉花期货的预测效果产生影响，本文选取的是前 10 天交易日的预测数据对第 11 天棉花期货的收盘价进行预测，通过不断训练到合适的次数后得到最优模型，将这个最优模型对预测集进行预测后，对比预测出的数据和原始数据，客观上能够评价预测的精度。

3.3. 基于 LSTM 的棉花期货价格预测

棉花期货交易中开盘价、成交量、最高价、最低价和持仓量五个指标作为模型输入，模型输出为棉花期货的预测价格。通过对 LSTM 模型的调校后，能够较为理想的去预测棉花期货的价格。综合考虑后，模型参数设置如下：研究中搭建了两层 LSTM 模型，第一层采用 200 个神经元，第二层采用 100 个神经元，batch_size = 20，单次训练的迭代次数 epochs = 150，损失函数 loss =“mse”，优化器 Optimizer =“Adam”。使用 LSTM 模型对棉花期货价格的预测结果如图 1 所示。

Figure 1. Fitting result on cotton futures testing set based on LSTM
图 1. LSTM 在棉花期货测试集上的拟合情况

3.4. 对比分析

同样的，通过构建 XGboost 方法将训练集放入学习后，去预测棉花期货的价格，使用 XGboost 对棉花期货价格的预测结果如图 2 所示。

为了考量 LSTM 模型和 XGboost 模型的效果，我们采用平均绝对误差(MAE)，均方误差(RMSE)和平均绝对百分比误差(MAPE)作为预测模型的评价指标，误差越小说明拟合效果越好，模型准确率越高。评价指标如表 1 所示。通过表 1 我们可以看出，基于 LSTM 神经网络的模型 RMSE、MAE、MAPE 均比 XGboost

方法还要小，因此可以推断出，LSTM 的预测效果要优于 XGboost 方法。

Figure 2. Fitting result on cotton futures testing set based on XGboost
图 2. XGboost 在棉花期货测试集上的拟合情况

Table 1. Prediction errors of the two models
表 1. 两种模型的预测误差

评价指标	LSTM	XGboost
R^2	0.994	0.993
RMSE	258.695	267.528
MAE	178.296	184.672
MAPE	1.132	1.205

4. 结语

本文是基于棉花期货价格的特点，利用 LSTM 长短期记忆网络进行价格的预测，又运用了近年来应用比较广泛的 XGboos 梯度提升算法对结果进行预测。实验中得到，LSTM 长短期记忆网络在预测棉花期货价格方面总体误差更小，预测精度更高，性能优于 XGboost 方法，从整体考虑 LSTM 比 XGboost 更具优势。本文的 LSTM 模型客观上可以为企业与投资者提供价格参考，合理应对市场风险。

棉花的价格容易受到众多因素而产生波动，而本文仅选取了棉花期货价格作为模型训练的数据，通过单步预测实现了对大豆价格的短期预测。未来研究中，会尝试增加时间跨度去预测棉花价格，例如运用前 20 天的价格数据去预测后 5 天的价格数据，使得整个预测的空间进一步扩大，模型更具实用性和现实意义。

基金项目

闽江学院“课程思政”教育教学改革精品项目(MJU2021KC529)。

参考文献

[1] 郭婷婷. 基于 PCA-ELM 的我国粮食价格预测研究[D]: [硕士学位论文]. 太原: 太原理工大学, 2016.

[2] 王进. 基于 BP 神经网络的肉鸡价格预测的研究[D]: [硕士学位论文]. 广州: 华南农业大学, 2018.

[3] 刘锦源. 面向农产品期货价格预测的改进 LSTM 方法[J]. 江苏科技信息, 2019(27): 48-52.

[4] 袁铭涓, 孙若莹. 基于 LSTM 神经网络的大宗农产品价格预测研究[J]. 海峡科技与产业, 2021(11): 43-47+60.

[5] 张宁, 方靖雯, 赵雨宣. 基于 LSTM 混合模型的比特币价格预测[J]. 计算机科学, 2021, 48(S2): 39-45.

[6] 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, 36(3): 55-69.