计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(10):10−21 [doi: 10.15888/j.cnki.csa.009262]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 XGBoost 和 TCN-Attention 的棉花价格多影响因素选择及预测①

王世杰[1,2], 王兴芬[1,3], 岳　婷[1,3]

[1](北京信息科技大学 商务智能研究所, 北京 100192)
[2](北京信息科技大学 计算机学院, 北京 100192)
[3](北京信息科技大学 信息管理学院, 北京 100192)
通信作者: 王兴芬, E-mail: xfwang@bistu.edu.cn

摘　要: 棉花价格受多种因素影响而复杂多变, 通过选择合适的数据特征和预测模型可提高棉花价格预测精度. 本文以棉花日现货价格数据为研究目标, 采集了供需关系、国际市场、宏观经济、产业链这 4 个方面的 9 项影响因素作为特征, 使用极限梯度提升 (XGBoost) 算法对棉花价格影响因素进行特征评估筛选, 选取其中 5 项特征后, 采用引入注意力机制 (Attention) 的时间卷积网络 (TCN) TCN-Attention、TCN、LSTM、GRU 等模型对棉花价格进行预测. 通过消融实验和对比实验, 结果表明: (1) 经过 XGBoost 特征筛选后, TCN-Attention 价格预测的平均绝对误差 (*MAE*) 和均方根误差 (*RMSE*) 为 41.47 和 58.76, 与未筛选相比分别降低了 77.57% 和 76.49%. (2) 与 TCN、LSTM、GRU 相比, 本文提出的 TCN-Attention 模型预测结果更准确, *MAE* 和 *RMSE* 均降低 50% 以上, 运行时间较 LSTM、GRU 缩短 60%.

关键词: 价格预测; XGBoost; TCN; Attention; 消融实验

引用格式: 王世杰,王兴芬,岳婷.基于 XGBoost 和 TCN-Attention 的棉花价格多影响因素选择及预测.计算机系统应用,2023,32(10):10–21. http://www.c-s-a.org.cn/1003-3254/9262.html

Selection and Prediction of Multiple Influencing Factors of Cotton Price Based on XGBoost and TCN-Attention

WANG Shi-Jie[1,2], WANG Xing-Fen[1,3], YUE Ting[1,3]

[1](Institute of Business Intelligence, Beijing Information Science & Technology University, Beijing 100192, China)
[2](School of Computer, Beijing Information Science & Technology University, Beijing 100192, China)
[3](School of Information Management, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract: Cotton price is complex and changeable due to many factors, and the prediction accuracy of cotton price can be improved by selecting appropriate data features and prediction models. In this study, the daily spot price data of cotton are taken as the research target, and nine influencing factors in four aspects of supply and demand, international market, macroeconomy, and industrial chain are collected as features. The extreme gradient boosting (XGBoost) algorithm is used to evaluate and screen the influencing factors of cotton price, and five of them are selected. This study adopts the time convolution network (TCN) with an attention mechanism (Attention), namely TCN-Attention, TCN, long short-term memory (LSTM), gate recurrent unit (GRU), and other models to predict cotton price. Through ablation experiments and comparative experiments, the results show that: (1) After XGBoost feature screening, the mean absolute error (*MAE*) and root mean square error (*RMSE*) of TCN-Attention price prediction are 41.47 and 58.76, which are 77.57% and 76.49% lower than those before screening; (2) compared with TCN, LSTM, and GRU, the TCN-Attention model proposed in this

① 基金项目: 国家重点研发计划 (2019YFB1405003)
收稿时间: 2023-03-01; 修改时间: 2023-04-07; 采用时间: 2023-05-11; csa 在线出版时间: 2023-08-09
CNKI 网络首发时间: 2023-08-10

study has more accurate prediction results. *MAE* and *RMSE* are reduced by more than 50%, and the running time is shortened by 60% compared with LSTM and GRU.

Key words: price prediction; extreme gradient boosting (XGBoost); time convolution network (TCN); Attention; ablation study

棉花从生产者到消费者, 需要经过生产、收购、加工、储存、运输、零售等产业链上的一个或多个环节[1], 任何一个环节的价格变化都会导致全产业链中产品的价格波动, 而这也使得棉花价格在形成过程中受到多种因素的影响. 然而近几年, 由于全球疫情持续蔓延, 世界经济复苏动力不足, 导致棉花价格持续上升[2]. 再加上地缘政治冲突、国际贸易摩擦以及极端气候等众多事件的共同作用, 国内棉花供需再次发生变动, 各种因素对棉花的影响变得更加复杂, 价格也一直保持高位波动. 如何通过科学的方法, 基于多种因素对棉花价格进行精准预测, 对于避免价格波动损害棉花产业链中棉农、纺织企业的切身利益, 影响市场稳定运行方面有着重要意义.

1 相关研究

目前, 针对棉花价格有许多研究人员开展了广泛的研究. 在棉花价格影响因素方面. 国家发改委学术委员会办公室课题组[3] 通过 1999–2012 年棉花价格数据, 详细分析流通体制改革前后棉花价格的波动特点, 指出在流通体制改革后棉花价格持续上涨且波动剧烈, 主要因为供需关系、极端天气变化、国际棉花价格及政府政策调控等因素. 褚志磐等[4] 利用 VAR 模型分析棉花产量、棉花生产成本、美元汇率及棉花进口数量对棉花价格的影响. 结果表明, 对棉花价格影响最强的因素为汇率, 棉花产量、棉花进口、棉花生产成本次之. 彭新宇等[5] 通过脉冲响应函数、方差分解等方法分析国际原油价格变动对大豆、棉花、花生仁和油菜籽的影响, 结果表明, 国际原油对花生仁、棉花价格波动的贡献率较大. Bodjongo[6] 利用描述性统计技术和 VAR 模型验证降雨量、温度等气候变化和国际市场棉花价格波动对棉花生产和棉花价格的影响, 最终发现世界棉花价格的上涨和气温的显著变化会对棉花价格造成影响. 此外指数投资、目标政策、货币发行量、居民消费指数、棉花储备量、市场情绪、产业链等[7–11] 因素都对棉花价格具有一定的影响.

在棉花价格预测研究方法方面, 有许多研究人员相继提出多种预测理论与方法. 包括自回归移动平均模型 (ARIMA)、马尔科夫链模型、支持向量机 (SVM)、基于模糊信息粒化和粒子群优化支持向量回归机 (PSO-SVR)[2,12–14] 等. 其中, ARIMA 模型应用最为广泛, 部分研究人员将该模型与经验模式分解 (EMD)、小波分析、H-P 滤波分析、指数自回归条件异方差模型[15–18] 相结合. 这些方法虽然可以很好地对时间序列数据进行预测, 但是这些方法不能很好地处理大规模复杂数据. 因此, 以人工神经网络 (ANN) 为主的方法逐渐应用于价格预测. 包括反向传播神经网络 (BPNN) 、径向基函数网络 (RBF)[9,19] 等. 但是人工神经网络对输入数据点进行独立处理, 没有考虑输入数据之间的相关性, 因此在对序列数据建模方面存在一定的局限性.

近些年, 使用深度学习建模方法进行棉花及其他农产品的价格预测研究也得到学者们的广泛关注. 江知航等[20] 引入种植面积、进出口、汇率、气候等因素, 采用双向长短期记忆网络 BiLSTM 和 SWA 优化算法对棉花价格进行预测, 使用长短期记忆网络 LSTM 和 LightGbm 模型进行对比试验, 得到了不错的预测效果. 但是其他深度学习方法在棉花价格领域的应用较少. 事实上, 深度学习模型中的循环神经网络 (RNN) 及变体模型在农产品价格预测的研究中已经获得了一定进展. Shin 等[21] 利用 LSTM 对大葱、洋葱、西葫芦、大米和菠菜的价格进行了预测, 选择气候数据、油价、产品综合指数、各年度农产品产量等因素作为影响变量, 最后通过 RMSE 指标验证模型有效性. Gu 等[22] 构建双输入注意长短时记忆模型 (DIA-LSTM), 利用气象数据、交易量数据等影响农产品价格的变量进行训练, 并通过白菜和萝卜的数据进行价格预测. Li 等[23] 根据注意机制开发了异构 GRU 神经网络 (AH-GRU), 引入静态信息对价格波动的影响, 提高了对异构数据的处理能力. 实验结果表明, 该方法对牛羊肉猪肉价格的预测精度、趋势预测和方法收敛性等方面均优于主流的畜产品价格预测方法. 虽然上述方法都取得了很好的

预测精度, 但是它们仍然需要大量的时间和计算机内存来训练模型.

2 分析方法及模型构建

2.1 研究思路及分析方法

本文提出了一种基于极限梯度提升 (XGBoost) 和引入注意力机制的时间卷积网络 (TCN) 的多因素预测模型.

在特征选择方面, 根据现有的研究, 极端天气变化、国际棉花价格、美元汇率等都对棉花价格有较为显著的影响, 所以本研究将棉花价格的影响因素分为供需关系、宏观经济、国际市场这 3 类, 其中: 供需关系是导致棉花价格上下波动的主要原因, 并且价格变化也会引起供给和需求的变化, 影响因素包括棉花的种植面积、产量、气候变化等. 国内经济稳定程度、国际收支平衡等整个宏观经济形势对棉花市场运行和价格波动也有着不可忽略的影响, 影响因素包括指数投资、货币发行量、汇率等. 近几年我国积极发展与世界多个国家的贸易合作伙伴关系, 增加贸易往来, 国内市场与国际市场的联动性不断增强, 国际市场变化对国内棉花市场有着很大的影响, 影响因素包括国际棉花价格、原油等.

结合近期时事, 综合从供需关系、国际市场、宏观经济中分别引入如下变量对棉花价格预测. 因为棉花产业链较长, 从生产到产出的过程中产业链里的各参与者会根据自身成本和市场情况来定价, 造成价格的传递[11], 对棉花的价格波动具有直接影响, 但目前价格预测研究中少有引入该类因素, 因此特选择棉花产业链上游的棉籽和棉粕作为棉花产业链代表数据进行预测.

事实上, 影响因素对价格的影响强度各有不同, 如果利用一些不相关的因素建立预测模型, 模型的性能会因此受到影响. 为此本研究引入了 XGBoost 算法进行关键特征筛选, 有效地从复杂的原始数据中提取重要的特征信息提高预测性能.

在预测方法方面, Bai 等[24] 提出了时间卷积网络 (TCN), 它结合了 CNN 和 RNN 的优点, 在能源预测、功率预测[25,26] 等领域问题中, 都证明了 TCN 模型的预测精度和预测效率优于 LSTM 和 GRU 模型, 能够有效地利用卷积提取复杂时间序列数据的特征进行预测.因此本文选择使用 TCN 模型进行棉花价格预测. 同时, 为了增强 TCN 模型对长时间序列数据学习的能力, 在模型中引入注意力机制, 通过对输入特征赋予不同的权重来缓解重要局部特征随着步长的增加而消失的问题[27]. 预测流程如图 1 所示.

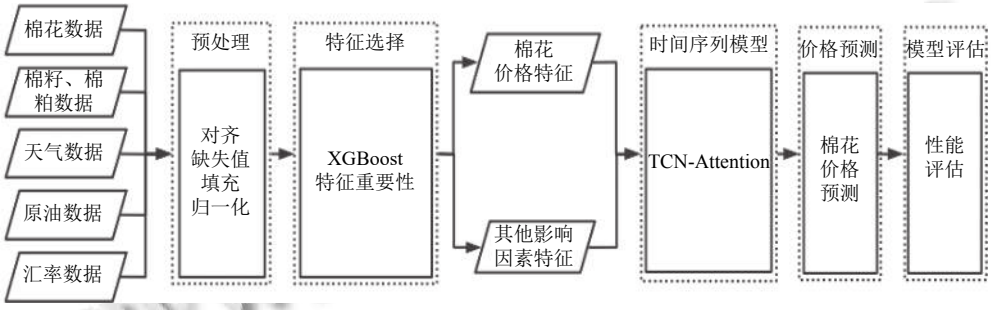


图 1 预测流程

2.2 模型构建

本文提出的 XGBoost+TCN+Attention (X-TCN-ATT) 神经网络模型的结构如图 2 所示, 分为输入层、特征选择层、TCN 网络层、注意力机制层和输出层. 各网络层具体任务如下.

第 1 层为输入层, 将预处理好的棉花价格数据和其他影响因素数据输入到模型中.

第 2 层为特征选择层. 使用 XGBoost 算法计算出每个特征的特征重要性分数, 根据特征重要性排序选择输入特征以降低模型的复杂度.

第 3 层为 TCN 网络层. 接受筛选出来的多变量数据, 通过使用卷积网络的变形结构, 在不加深网络深度的情况下捕获长期依赖关系并通过残差模块实现跨层信息传递.

第 4 层为注意力机制层. 通过计算特征的注意力分数, 为每个特征赋予不同的权重, 忽略对价格影响较低的特征, 提高预测精度.

最后一层为输出 6 层, 通过全连接层得到价格预测结果.

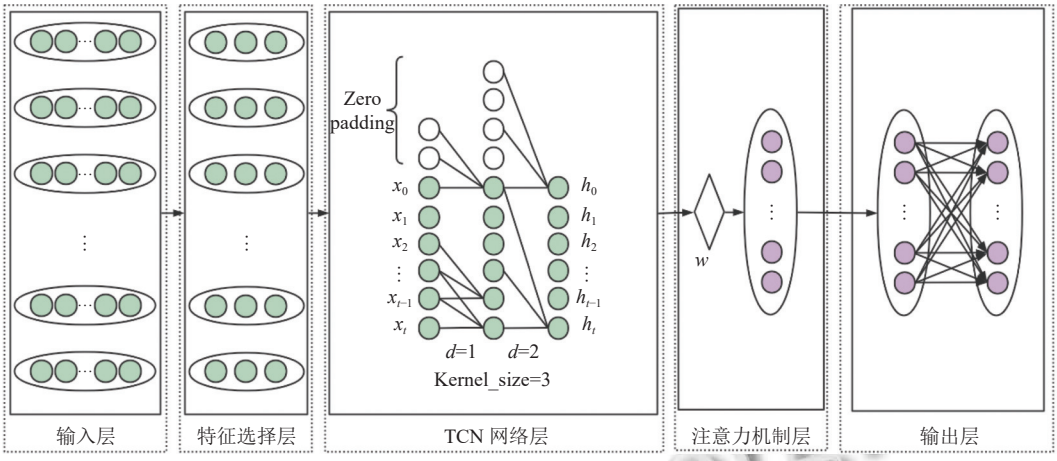


图 2 网络结构

2.2.1 基于 XGBoost 的棉花价格特征评估

XGBoost 评估特征在特征集中的重要性通常有增益、覆盖度量和频率这 3 种方式[28], 其中增益是解释每个特征的相对重要性的最相关属性, 所以本文主要通过枚举所有不同树结构的贪心法求得增益进行特征评估.

对于经过数据预处理之后的棉花数据集 $M=\{(x_i,y_i)\}(i=0,1,\cdots,t)$, 其中 $x_i=\left(x_0^i,x_1^i,\cdots,x_{n-1}^i,x_n^i\right)$ 表示第 i 个样本的 n 维特征向量, 即输入的棉花数据的 n 个特征值, y_i 表示第 i 个样本的标签值. 将棉花数据集 M 输入到 XGBoost 中进行训练得到 K 棵树, 可以表示为:

$$\widehat{y_i}=\sum_{k=1}^{K}f_k(x_i),\ f_k\in F \tag{1}$$

$$F=\left\{f(x)=w_{q(x)}\right\} \tag{2}$$

其中, $\widehat{y_i}$ 表示第 i 个样本的预测结果, F 是基学习器即 K 棵树的集合, f_k 表示第 k 棵回归树, $w_{q(x)}$ 表示叶子节点 q 的分数. 训练过程中的目标函数为:

$$O=\sum_{i=0}^{t}l(y_i,\widehat{y_i})+\sum_{k=1}^{K}\Omega(f_k) \tag{3}$$

其中, l 是用于计算预测值和真实值之间的误差的损失函数, Ω 是表示树的复杂度的函数, 值越小复杂度越低, 泛化能力越强, 其表达式为:

$$\Omega(f_k)=\gamma N+\frac{1}{2}\lambda\|w\|^2 \tag{4}$$

其中, N 表示叶子节点的个数, w 表示节点的数值. 直观上看, 目标要求预测误差尽量小, 且叶子节点 N 尽量少, 节点数值 w 尽量不极端.

通过在梯度方向上不断优化, 目标函数越来越低, 因为预测值 $\widehat{y_i}$ 在第 T 次迭代之后可以由前 $T-1$ 次迭代输出值总和以及第 T 次迭代计算得到的树结构值 $f_T(x_i)$ 相加得到, 所以将目标函数 O 转换成:

$$O^T=\sum_{i=0}^{t}l(y_i,\widehat{y_i}^{(T-1)}+f_T(x_i))+\Omega(f_T)+C \tag{5}$$

其中, C 为常数. 将式 (5) 通过二阶泰勒展开可优化为:

$$O^T=\sum_{i=0}^{t}\left[l(y_i,\widehat{y_i}^{(T-1)})+g_if_T(x_i)+\frac{1}{2}h_if_T^2(x_i)\right]+\Omega(f_T) \tag{6}$$

其中, $g_i=\partial_{\widehat{y}^{(T-1)}}l(y_i,\widehat{y}^{(T-1)})$ 为预测误差对当前模型的一阶导数, $h_i=\partial^2_{\widehat{y}^{(T-1)}}l(y_i,\widehat{y}^{(T-1)})$ 为许多预测误差对当前模型的二阶导数.

由于在第 T 次迭代时, 第 $T-1$ 次的模型残差已知, 因此去掉常数项 $l(y_i,\widehat{y_i}^{(T-1)})$ 并对式 (6) 进行展开, 将目标函数转换成按叶节点累加的形式为:

$$O^T=\sum_{j=1}^{N}\left[Gw_j+\frac{1}{2}(H+\lambda)w_j^2\right]+\gamma N \tag{7a}$$

$$G=\sum_{i\in I}g_i \tag{7b}$$

$$H=\sum_{i\in I}h_i \tag{7c}$$

其中, I 表示每个叶子节点上样本的集合, $I_j=\{i|q(x_i)=j\}$; $q(x_i)$ 为树结构函数; w_j 表示每棵树叶子节点的输出分数; N 表示分裂树的叶子节点个数; λ 和 γ 为权重因子, 用来控制对应部分的比重.

创建好提升决策树后通过计算增益得到每个特征

的特征重要性. 与决策树中的信息增益及基尼指数类似, XGBoost 算法在每一次尝试对已有的叶子加入一个分割时, 都会计算选取特征的增益 *Gain*:

$$Gain=\frac{1}{2}\left[\frac{G_{\mathrm{L}}^2}{H_{\mathrm{L}}+\lambda}+\frac{G_{\mathrm{R}}^2}{H_{\mathrm{R}}+\lambda}-\frac{(G_{\mathrm{L}}+G_{\mathrm{R}})^2}{H_{\mathrm{L}}+H_{\mathrm{R}}+\lambda}\right]-\gamma \tag{8}$$

其中, 下标 L、R 分别表示左子树和右子树; $\frac{G_{\mathrm{L}}^2}{H_{\mathrm{L}}+\lambda}$表示左子树分数; $\frac{G_{\mathrm{R}}^2}{H_{\mathrm{R}}+\lambda}$表示右子树分数; $\frac{(G_{\mathrm{L}}+G_{\mathrm{R}})^2}{H_{\mathrm{L}}+H_{\mathrm{R}}+\lambda}$表示当前节点不分割的分数.

在单个提升树中通过每个特征分裂点的增益大小来计算特征重要性, 增益越大其权值越大. 特征被越多的提升树所选择, 该特征越重要. 最终将一个特征在所有提升树中的结果进行加权求和后取平均, 得到重要性分数. 按重要性分数从高到低排序后, 通过设置不同的阈值筛选出 m (m<n) 个影响棉花价格的重要特征.

2.2.2 基于 TCN-Attention 的棉花价格预测模型

时间卷积网络 (TCN) 是一种能够处理时间序列问题的网络结构, 其本质上是一种一维卷积网络的变形, 通过扩张因果卷积和残差连接来进行序列建模和预测. 棉花数据集$M=\{(x_i,y_i)\}(i=0,1,\cdots,t)$通过 XGBoost 算法从原始的 n 个特征中筛选出 m 个对棉花价格影响较大的特征后, 作为 TCN 网络的输入, 输入到网络中进行卷积计算提取时序特征, 得到输出向量h_i ($i=0,1,\cdots,t$), 其中$x_i=(x_0^i,x_1^i,\cdots,x_{m-1}^i,x_m^i)$ ($m<n$).

(1) 因果卷积

因果卷积 (causal convolution) 使得 TCN 具有严格的时间约束. 对于输入到网络中的棉花价格序列$\{x_0,x_1,\cdots,x_{t-1},x_t,\cdots\}$, t 时刻的输出h_t只能通过当前时刻的输入x_t和之前时刻的输入$x_0,x_1,\cdots,x_{t-1}$计算得到, 即当前时刻 t 的信息只能通过当前时刻 t 及其之前的信息得到. 为了确保输出张量与输入张量具有相同的长度, 采取在输入张量左侧进行零填充的策略, 这样也满足了因果卷积.

然而, 要想获得较长的历史信息甚至覆盖完整的历史记录则需要特别深的网络, 随着网络深度的增加就会出现梯度消失、计算复杂、拟合效果不好等问题, 所以引入了扩张卷积.

(2) 扩张卷积

扩张卷积能够在不增加参数和模型复杂度的情况下, 通过间隔采样指数级增大感受野, 捕获长期依赖关系.

如图 3 所示是 TCN 的网络结构, 和传统卷积不同的是, 扩张卷积允许卷积时的输入存在间隔采样, 采样率受扩张因子 d 控制. 最下面一层 d=1 表示输入时每个时间点都采样, 隐藏层 d=2 表示输入时每 2 个时间点采样一个作为输入. 对于输入的棉花价格序列 X=$(x_0,x_1,\cdots,x_{t-1},x_t)$, 滤波器$f:\{0,\cdots,k-1\}\to R$, t 时刻的卷积运算 F 为:

$$F(x_t)=\sum_{i=0}^{k-1}f(i)\cdot x_{t-d\cdot i} \tag{9}$$

其中, k 为卷积核大小, d 为扩张因子, $t-d\cdot i$表示历史数据. 此时 TCN 网络的感受野 w 为:

$$w=1+(k-1)\cdot\frac{b^n-1}{b-1} \tag{10}$$

其中, n 为层数、b 为扩张基数 (扩张因子$d=b^i,i=1,2,\cdots,n$).

可以看到当卷积核大小为 3、扩张因子为 [1, 2, 4] 时 t 时刻的输出h_t由输入$x_0,x_1,\cdots,x_{t-1},x_t$共同决定, 即感受野能够覆盖输入序列中的所有值.

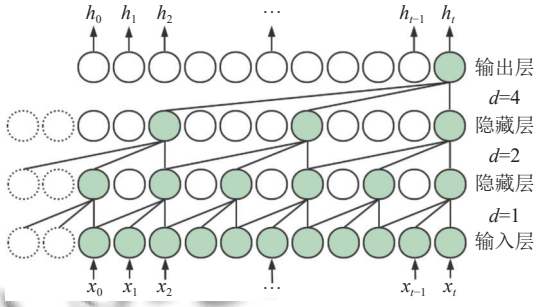


图 3 TCN 网络结构图

(3) 残差连接

残差连接使得浅层网络可以很容易扩展成深层网络. 本文构建了两个残差模块来代替两层卷积. 残差模块结构如图 4 所示, 一个残差模块由两层扩张因果卷积和非线性映射构成, 卷积核大小为 3.

首先对输入的特征向量进行扩张因果卷积, 之后通过 *WeightNorm* 归一化来加速网络训练, 然后使用 *ReLU* 激活函数进行非线性计算, 并加入 Dropout 防止过拟合. 最后将得到的结果与输入求和得到输出向量. 计算过程如下:

$$S_i=Conv(W_i\times F_j+b_i) \tag{11}$$

$$\{S_0,S_1,\cdots,S_{t-1},S_t\}=WeightNorm(\{S_0,S_1,\cdots,S_{t-1},S_t\}) \tag{12}$$

$$\{S_0,S_1,\cdots,S_{t-1},S_t\}=ReLU(\{S_0,S_1,\cdots,S_{t-1},S_t\}) \tag{13}$$

其中, S_i表示 i 时刻经过卷积计算得到的特征向量, W_i表示 i 时刻卷积计算的权重矩阵, F_j表示第 j 层的卷积核, b_i表示偏置向量; $WeightNorm(x)=\frac{\|w_x\|}{\|v\|}v$, $\|w_x\|$表示$ReLU(x)=\max(0,x)$权重 w 的大小, $\frac{v}{\|v\|}$表示与 w 同方向的单位向量; $\{S_0,S_1,\cdots,S_{t-1},S_t\}$表示经过第 j 层完整卷积之后得到的特征图.

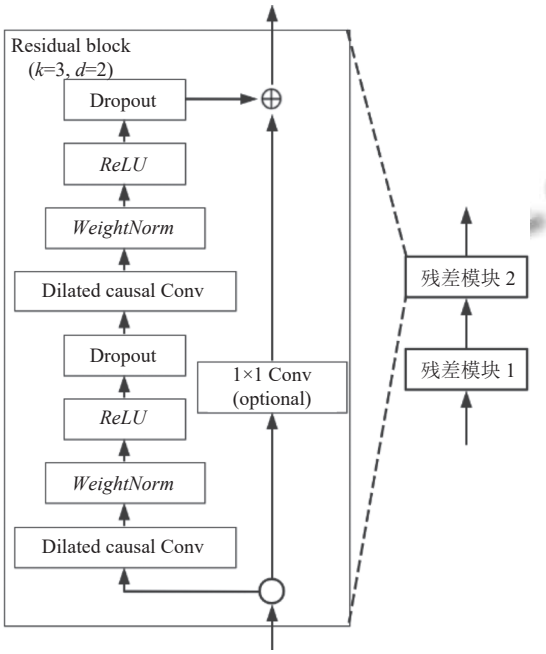


图 4 TCN 网络残差模块

(4) 注意力机制

为了更好地从棉花价格时间序列中学习到每个特征的重要程度, 进一步捕捉其中的时序关系, 在模型中加入注意力机制. 注意力机制对由 TCN 网络输入的特征向量进行加权求和, 通过 Softmax 函数计算得到注意力分数α, 将输入特征与注意力分数相乘生成权重矩阵. 注意力机制原理如图 5 所示.

其中$x_i\ (i=0,1,\cdots,t)$是 TCN 网络的输入, $h_i\ (i=0,1,\cdots,t)$是通过 TCN 网络输出得到的隐藏层向量, $\alpha_i\ (i=0,1,\cdots,t)$是通过对隐藏层输出计算得到的注意力分数, $y_i\ (i=0,1,\cdots,t)$是加入注意力机制后的输出向量. 计算公式为:

$$c_i=u\cdot\tanh(wh_i+b) \tag{14}$$

$$\alpha_i=\frac{\exp(c_i)}{\sum_{j=0}^{i}c_j} \tag{15}$$

$$y_i=\sum_{i=0}^{t}\alpha_i\cdot h_i \tag{16}$$

其中, u 和 w 为权重系数, b 为偏置系数, 表示第 i 时刻 TCN 网络输出的隐藏层向量 所确定的注意力权重.

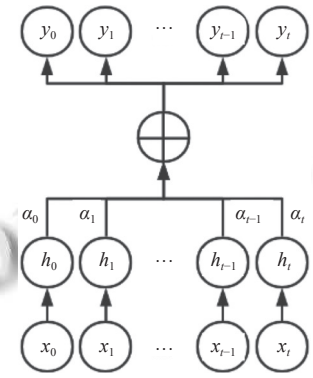


图 5 注意力机制原理

最后将输出向量输入全连接网络, 经过迭代训练得到最终的价格预测值.

3 特征工程

3.1 数据选取及来源

本文的预测对象为新疆棉花现货价格, 数据来源于 Wind 数据库, 时间周期从 2009 年 5 月–2022 年 5 月 (2009/5–2022/5), 价格走势如图 6 所示. 棉花价格的影响因素从供需关系、宏观经济、国际市场、产业链这 4 个角度综合选择了平均气温、平均风速、降水量、原油现货价格、汇率、棉粕现货价、棉粕现货平均价. 其中气候数据来源于美国国家信息中心, 其余数据均来源于万德数据库.

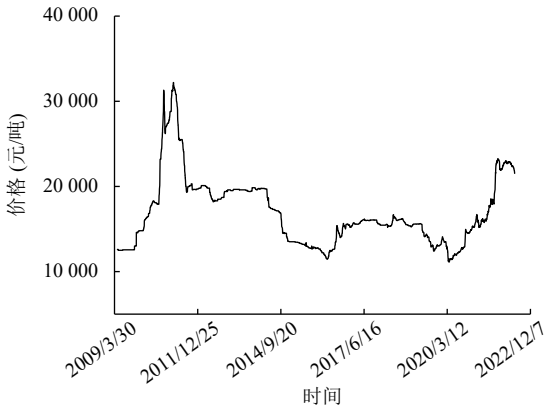


图 6 棉花价格走势

3.2 数据预处理

(1) 数据清洗

从万德数据库获取到的棉花日现货价格数据均为工作日交易数据, 整体价格数据是真实可靠, 不存在不完整、异常值等脏数据. 因此其他特征数据均按照棉花日现货价格日期进行对齐处理. 针对对齐后部分特征数据存在缺失值情况, 因为考虑到输入数据均为时序序列, 所以选择先以前向填充方法进行填补, 剩余缺失值再以均值填充法处理. 总共包含 10 个特征指标, 共 32 540 条数据.

(2) 数据变换

原始数据集中包含价格、原油、汇率、气候等不同方面的特征数据, 而所有特征数据之间的数量单位和数量级存在差异. 这不可避免会对模型的效果产生不利影响, 因此需要对数据集进行归一化处理, 将所有特征数据映射到 [0, 1] 之间, 消除不同单位、量级带来的影响. 本文采用 Min-Max 归一化进行处理, 公式如下所示:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{17}$$

其中, x、x'、$x_{\min}$和$x_{\max}$分别为输入数据、输出数据、输入数据的最小值和输入数据的最大值.

(3) 数据划分

数据集时间维度从 2009/5/1–2022/5/31. 在使用数据集训练模型之前, 按 7:1:2 的比重将其划分为训练集、验证集和测试集, 用以增强模型的泛化能力. 其中, 训练集用于更新网络中的权值参数, 寻找最佳参数组合模型. 验证集用于在训练过程中调整模型超参数. 测试集用于评估模型预测性能. 数据划分结果如表 1 所示.

表 1 数据集划分结果

数据	训练集	验证集	测试集
划分时间	2009/5–2018/6	2018/6–2019/10	2019/10–2022/5
数据维度	10	10	10
数据量	22 780	3 250	6 510

4 实验与分析

4.1 实验环境

本文的实验环境如表 2 所示.

4.2 评价指标

本文选择平均绝对误差 (MAE)、均方根误差 ($RMSE$)、平均绝对百分比误差 ($MAPE$)、决定系数 R 平方 (R^2) 和运行时间 (TIME) 这 5 项指标来评估模型的预测性能. 其中, MAE、$MAPE$、$RMSE$ 和R^2用以评估模型的预测性能. “TIME”反映模型的运行时间, 用以评估模型的复杂性和操作效率, 单位为 s. 计算公式如下所示:

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{18}$$

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{19}$$

$$MAPE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{20}$$

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{21}$$

其中, $\hat{y}_i$为预测值; y_i为真实值; $\bar{y}$为真实值的平均值; n 为总的样本个数.

表 2 实验环境

名称	型号及版本
操作系统	Windows 10
内存	16 GB DDR4
处理器 (CPU)	Intel core i7-10750H处理器
显卡 (GPU)	NVIDIA RTX2070 8 GB独立显卡
编程语言	Python 3.8
框架	TensorFlow 2.4
运算平台	cuda11

4.3 特征评估

将特征输入 XGBoost 模型中对特征进行训练, 通过 F score 指标来衡量各个特征对棉花价格的影响强度, 为特征选择提供参考. 得分如图 7 所示, 其中各个特征的符号及含义如表 3 所示.

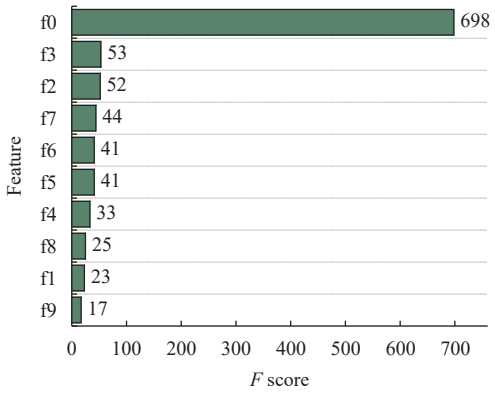


图 7 各个特征的评估得分

表 3 输入特征的符号和含义

变量	特征
f0	棉花现货价 (price)
f1	棉粕现货价 (MZ)
f2	棉籽现货价 (MP)
f3	Dtd布伦特原油现货价 (Dtd)
f4	WTI原油现货价 (WTI)
f5	英镑兑人民币汇率 (GBP)
f6	美元兑人民币汇率 (USD)
f7	平均温度 (TEMP)
f8	平均风速 (WDSP)
f9	平均降水量 (PRCP)

从图 7 中不同特征的 *F* score 值可以看出, 棉花历史价格对输出几乎有决定性的影响. 这可能是因为棉花价格数据属于时间序列, 未来值会受历史值的影响. 此外, 作为棉花种子的棉籽, 位于产业链的上游与棉花之间存在着纵向价格传递, 对棉花价格具有较强的影响. 而由棉籽加工生产的棉粕对棉花价格影响较小. 同时, 气象条件中温度和风速的变化也会对棉花价格造成影响, 但是风速对输出的影响较小. 此外, 国际原油价格和汇率也具有影响, 其价格波动会导致棉花价格产生变化. 随着特征重要性分数的降低, 特征提供的有效信息逐渐减少. 为了简化输入特征, 同时尽可能保留特征信息, 初步设置阈值大小为 40, 选用 *F* score 值大于 40 的前 6 个特征作为 TCN 网络的输入进行实验, 分别为棉花现货价、原油现货价、棉籽现货价、平均温度、英国汇率和美国汇率. 其中棉花现货价、原油现货价和棉籽现货价是棉花价格的主要决定因素之一, 与棉花的生产和加工密切相关; 平均温度是影响棉花生产的重要环境因素之一, 对棉花的生长和收成有很大的影响; 英国汇率和美国汇率则是影响棉花进出口贸易的关键因素之一, 因为棉花在国际贸易中往往以美元和英镑计价.

图 8 是各个特征之间的皮尔逊相关性热力图, 可以看出, 棉籽现货价、原油现货价和美元汇率与棉花现货价之间存在一定的正相关关系, 而英镑汇率与棉花现货价之间存在一定的负相关关系, 天气因素与棉花现货价的相关性较弱.

可以发现相关性分析得到的结果与 XGBoost 特征评估得到的结果存在一定差异, 这是因为相关系数体现的是特征间的线性相关关系, 而重要性分数 *F* score 衡量的是特征在模型中的贡献程度, 它可以帮助我们理解模型中每个特征的重要性, 找出对预测结果最重要的特征. 对于受多种因素影响的棉花价格预测问题, 线性相关仅是其中的一种影响因素, 而且不一定是最重要的影响因素. 这进一步说明了线性相关并不能很好地解决非线性问题. 在实际的数据分析和建模过程中, 我们需要综合运用相关系数和特征重要性评估方法, 以全面了解特征之间的关系和模型中各个特征的重要性, 从而更好地解决复杂的非线性问题.

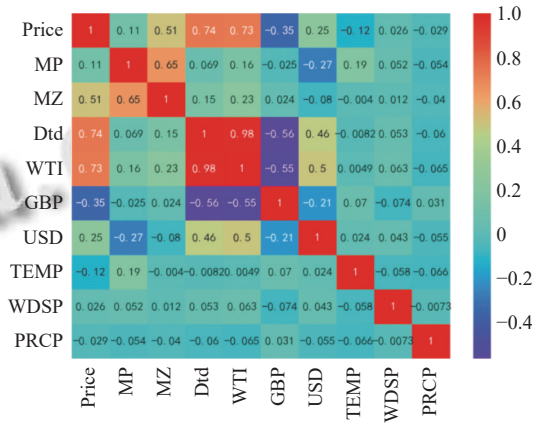


图 8 各特征间的相关性热力图

4.4 参数确定

时间卷积网络中的卷积核大小、卷积核个数、 舍弃率以及膨胀系数的设置对模型结果有至关重要的影响. 基于训练样本集对网络进行训练, 经过大量实验确定最佳超参数组合, 如表 4 所示.

表 4 TCN 参数

参数	值
残差模块个数	2
卷积核大小	3
卷积核个数	128
扩张系数	[1, 2]
丢弃率	0.2
初始学习率	0.000 5
批量大小	4

当时间卷积网络 (TCN) 的结构为 2 个隐藏层, 每个隐藏层节点数为 128, 卷积核大小为 3, 个数为 128, 丢弃率为 0.2, 扩张系数设置为, 损失函数设置为 *MAE*; 优化器为 Adam; 迭代次数为 100; 学习率初始为 0.000 5; 每迭代 30 次学习率会衰减为原来的 1/2 时, 模型取得了较为出色的预测效果.

4.5 特征选择

根据图 7 中基于 XGBoost 的特征评价结果, 选择

F score 指标前 6 个特征作为备选项, 分别为 f0 (棉花现货价)、f3 (布伦特原油现货价)、f2 (棉籽现货平均价)、f7 (TEMP 平均温度)、f6 (USD 美元兑人民币汇率)、f5 (GBP 英镑兑人民币汇率). 按照 *F* score 值的大小, 依次输入不同的特征组合进行实验, 评价结果如表 5 所示.

表 5 不同特征组合的评价结果

特征	*MAE*	*RMSE*	*MAPE*	R^2	TIME
f0, f3	83.16	111.69	0.00477	0.99920	61.17
f0, f3, f2	63.10	100.31	0.00445	0.99935	62.21
f0, f3, f2, f7	55.22	79.78	0.00341	0.99959	63.22
f0, f3, f2, f7, f6	41.47	58.76	0.00273	0.99978	64.33
f0, f3, f2, f7, f6, f5	81.35	108.87	0.00543	0.99924	65.91

当选取 f0, f3, f2, f7, f6 这 5 个特征作为输入时预测误差最小, 尽管模型运算时间比选取 4 个特征时要长, 但是预测问题更需要关注模型的预测误差. 所以当“TIME”指标差异不大时, 预测误差指标更为重要.

4.6 模型对比

(1) 消融实验

为了验证特征评估和注意力机制的有效性. 使用 TCN 和引入注意力机制的 TCN 模型作为预测模型设计了消融实验. 其中, “A+”代表将所有特征都输入模型进行预测, “X+”代表使用经过 XGBoost 排序筛选后的特征组合来进行预测. “+ATT”代表引入注意力机制的模型. 实验结果如图 9 和表 6 所示.

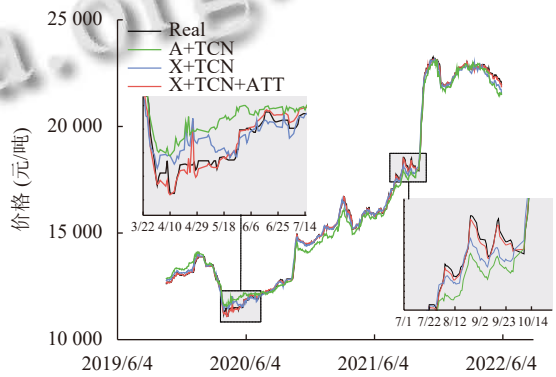
图 9 中给出消融实验中不同模型的预测曲线, 表 6 中展示了各模型预测指标. 可以看出:

利用 XGBoost 进行输入特征的筛选能够有效提高预测精度和模型运算效率. X+TCN 与 A+TCN 相比, *MAE*、*RMSE*、*MAPE* 的值分别下降了 50.1%、44%、51.8%. X+TCN+ATT 与 A+TCN+ATT 相比, *MAE*、*RMSE*、*MAPE* 的值分别下降了 77.57%、76.49%、75.18%. 运算时间均有所减少. 证明如果盲目引入其他因素建立预测模型, 会对整个模型的性能造成影响.

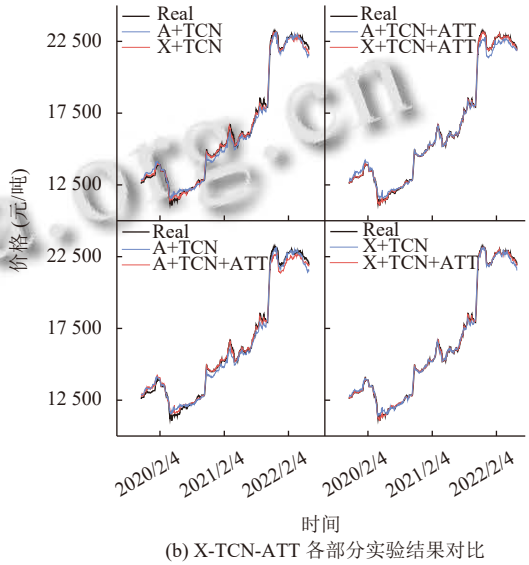
同时, 无论以哪种特征组合作为输入, 引入注意力机制的 TCN 模型的预测效果和模型运算效率都会更好. A+TCN 与 A+TCN+ATT 相比, *MAE*、*RMSE*、*MAPE* 的值分别下降了 27%、17.79%、32.76%. X+TCN 与 X+TCN+ATT 相比, *MAE*、*RMSE*、*MAPE* 的值分别下降了 67.2%、65.48%、65.4%. 运算时间均有所减少. 证明引入注意力机制的预测模型的有效性.

(2) 预测模型比较实验

为了验证本文提出的 XGBoost+TCN+ATT 模型的优越性, 分别建立加入 XGBoost 算法的 X+LSTM、X+GRU 和 X+TCN 这 3 种组合模型, 并对其进行了对比实验. 实验结果如图 10 和表 7 所示. 所有模型的输入数据均由特征选择时选取的 5 个特征组成, 各比较模型的网络参数如表 8 所示. 同时, 由于一层 TCN 网络由两个残差模块组成, 故 LSTM 和 GRU 设置两层网络以保证网络层数一致.



(a) X-TCN-ATT 消融实验结果



(b) X-TCN-ATT 各部分实验结果对比

图 9 消融实验预测结果

表 6 消融实验评价结果

组合	*MAE*	*RMSE*	*MAPE*	R^2	TIME
A+TCN	253.29	303.99	0.01636	0.99410	64.19
A+TCN+ATT	184.88	249.91	0.01100	0.99601	65.20
X+TCN	126.42	170.21	0.00789	0.99815	61.51
X+TCN+ATT	41.47	58.76	0.00273	0.99978	64.33

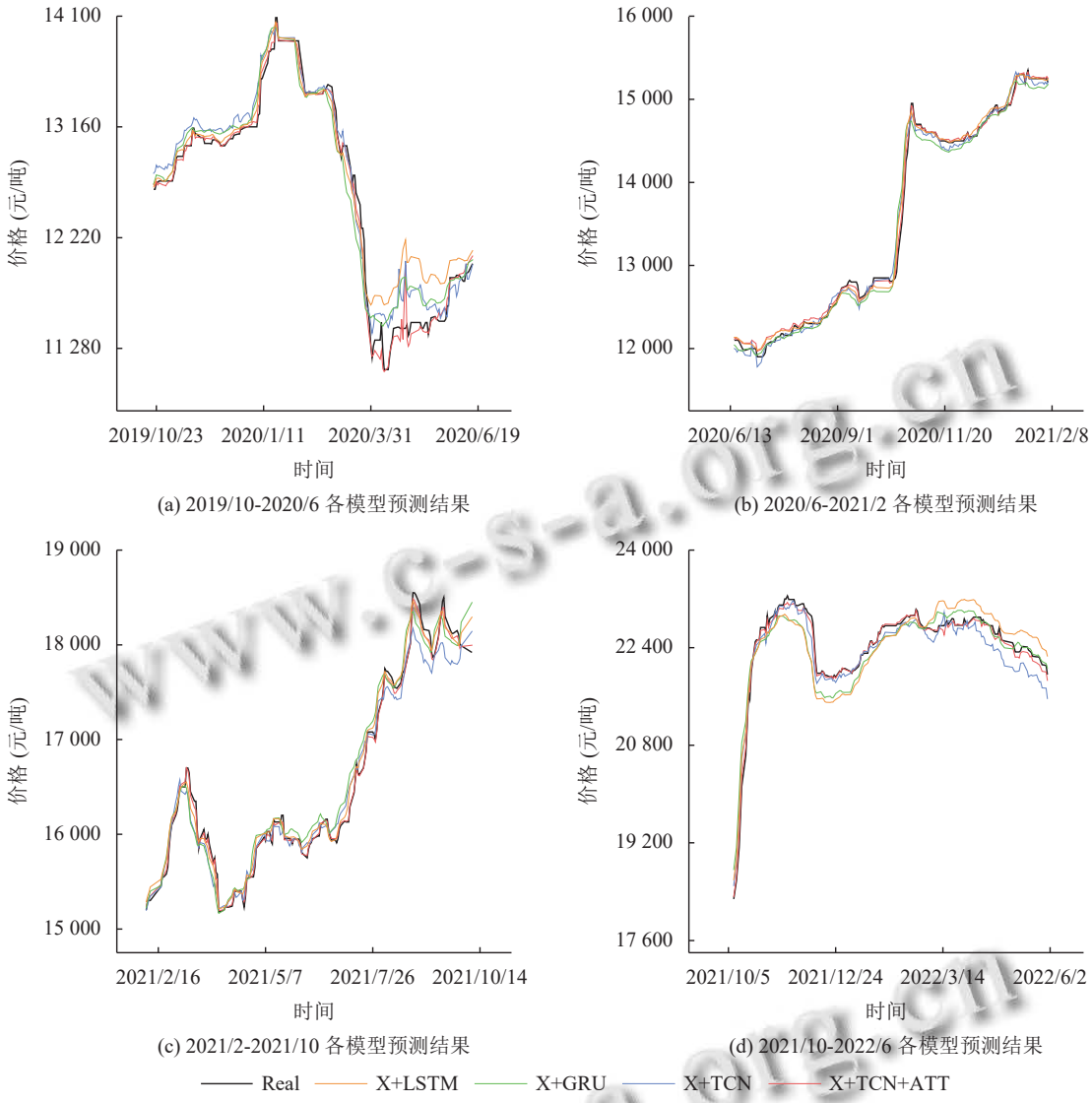


图 10 各模型预测结果对比

表 7 各模型预测误差

模型	*MAE*	*RMSE*	*MAPE*	R^2	TIME
X+GRU	149.01	202.18	0.009 25	0.997 39	150.51
X+LSTM	145.91	208.13	0.008 97	0.997 23	158.72
X+TCN	126.42	170.21	0.007 89	0.998 15	61.51
X+TCN+ATT	41.47	58.76	0.002 73	0.999 78	64.33

图 10 中给出不同模型的预测曲线. 各个模型的误差如表 7 所示. 通过实验结果可以看出: 在棉花价格真实值曲线整体波动较大的情况下, 基于极限梯度提升 (XGBoost) 和引入注意力机制的时间卷积网络 (TCN) 的模型预测结果最优, 预测值与真实结果拟合状况较好, *MAE* 为 41.47, *RMSE* 为 58.76, *MAPE* 为 0.002 73, 在棉花平均价格为 16 867 的基数下, 该误差非常小. 这是因为相比于循环神经网络 LSTM 和 GRU, TCN 不需要门控机制, 通过多层卷积和残差连接的方式来建模序列的长期依赖性, 允许并行计算的同时避免了梯度消失或梯度爆炸的问题, 从而可以更加高效地训练模型, 捕捉序列的长期信息. 同时, 在 TCN 基础上引入注意力机制, 通过学习注意力权重来关注对于预测重要的特征, 可以更加准确地捕捉序列中的重要信息, 从而提高了预测性能和预测结果的可解释性.

5 结论与展望

本文通过对棉花现货价格和其影响因素的分析, 构建了一个基于极限梯度提升 (XGBoost) 和引入注意力机制的时间卷积网络 (TCN) 的预测模型. 通过 XGBoost 算法对所选择的影响因素进行特征评估, 选出影响棉

花价格波动的主要因素. 之后将筛选的多特征数据依次组合输入到引入注意力机制的 TCN 网络模型中进行预测. 最后, 分别通过消融实验和对比实验证明基于 XGBoost 进行特征评估预测方法比未进行评估的预测方法效果更好, *MAE*、*RMSE*、*MAPE* 整体下降了 40% 以上. 而引入了注意力机制的 TCN 模型也要比未引入注意力机制的 TCN 模型、LSTM 模型、GRU 模型预测精度更高, *MAE*、*RMSE*、*MAPE* 整体下降了 50% 以上. 本文主要贡献如下.

(1) 在棉花价格影响因素选择时不仅综合考虑了供需关系、国际市场、宏观经济方面的影响, 同时引入现有的价格预测研究中没有考虑到的棉花产业链因素. 通过特征选择和预测结果证明棉籽、棉粕对棉花价格的影响.

(2) 通过 XGBoost 算法衡量各因素与棉花价格的影响强度, 来对输入影响因素数据进行选择, 通过实验验证特征筛选在价格预测中能够成功降低模型复杂度, 提高模型预测精度和效率.

(3) 将引入注意力机制的时间卷积网络 (TCN) 的预测模型和 TCN、LSTM、GRU 模型相比, 验证了无论以哪种特征组合作为输入, 引入注意力机制的 TCN 模型的预测效果和模型运算效率整体更优.

表 8 LSTM、GRU、TCN 网络参数

模型	参数	值
LSTM、GRU	网络层数	2
	神经元个数	128
	丢弃率	0.2
	初始学习率	0.000 5
	批量大小	4
TCN	残差模块个数	2
	卷积核大小	3
	卷积核个数	128
	扩张系数	[1, 2]
	丢弃率	0.2
	初始学习率	0.000 5
	批量大小	4

总体来说, 本文所构建的 XGBoost-TCN-Attention 模型能够选择最少但最有效的特征达到最好的预测效果, 最大提升预测时间, 在棉花现货价格预测中具备可行性和适用性. 同时, 也可以考虑将其应用到不同农产品的价格预测中, 为其他领域的时间序列数据预测提供一定的借鉴. 不过在预测过程中, 存在预测准确度下降点, 之后将对这一情况进行深入研究, 探明其出现的原因, 并采取合适办法解决该问题, 以此提高模型稳定性.

参考文献

1 Wang LY, Feng JY, Sui XJ, *et al*. Agricultural product price forecasting methods: Research advances and trend. British Food Journal, 2020, 122(7): 2121–2138. [doi: 10.1108/BFJ-09-2019-0683]

2 高翔, 喻晓玲. 后疫情时代新疆棉花价格走势分析及预测. 合作经济与科技, 2022, (8): 68–70. [doi: 10.3969/j.issn.1672-190X.2022.08.029]

3 国家发改委学术委员会办公室课题组. 新形势下我国棉花价格问题研究. 经济研究参考, 2013, (39): 3–38. [doi: 10.3969/j.issn.2095-3151.2013.39.001]

4 褚志磐, 刘荣俊, 张燕飞, 等. 基于 VAR 模型的中国棉花价格波动及影响因素分析. 湖南农业科学, 2022, (1): 96–99, 104. [doi: 10.16498/j.cnki.hnnykx.2022.001.025]

5 彭新宇, 樊海利. 国际原油价格对中国大宗农产品价格的影响研究. 宏观经济研究, 2019, (1): 99–109, 124. [doi: 10.16304/j.cnki.11-3952/f.2019.01.010]

6 Bodjongo MJM. Climate change, cotton prices and production in Cameroon. The European Journal of Development Research, 2022, 34(1): 22–50. [doi: 10.1057/s41287-020-00345-1]

7 许馨露, 顾光同. 指数投资对大宗农产品价格的影响. 合作经济与科技, 2019, (15): 78–82. [doi: 10.13665/j.cnki.hzjjykj.2019.15.028]

8 丁建国, 穆月英, 钱加荣. 目标价格政策下新疆棉花供给反应研究. 中国农业大学学报, 2020, 25(8): 184–193. [doi: 10.11841/j.issn.1007-4333.2020.08.18]

9 张兆同, 余潜. 灰色关联分析与 RBF 神经网络在我国棉花价格预测中的应用研究. 价格月刊, 2017, (9): 31–36. [doi: 10.14076/j.issn.1006-2025.2017.09.06]

10 郭少红, 董玲, 李达, 等. 市场情绪、棉花期货价格和现货价格的相关性研究——基于 MSVAR 模型的实证检验. 粮食与油脂, 2019, 32(6): 97–100. [doi: 10.3969/j.issn.1008-9578.2019.06.025]

11 胡雨. 我国棉花产业链价格传递机制的研究——基于 VAR 模型 [硕士学位论文]. 广州: 广东财经大学, 2018.

12 向雪燕, 张立杰. 基于马尔科夫链的棉花价格预测. 中国棉花, 2016, 43(10): 1–6. [doi: 10.11963/issn.1000-632X.2016 10001]

13 张立杰, 寇纪淞, 李敏强, 等. 基于自回归移动平均及支持向量机的中国棉花价格预测. 统计与决策, 2013, (6): 30–33. [doi: 10.13546/j.cnki.tjyjc.2013.06.037]

14 张永礼, 赵蕾, 董志良. 基于信息粒化和 PSO-SVR 模型的

棉花价格波动区间和变化趋势预测. 广东农业科学, 2015, 42(11): 180–185. [doi: 10.3969/j.issn.1004-874X.2015.11.032]
15 王伟国, 支小军. 基于 EMD-ARMA 模型的我国棉花价格预测方法研究. 新疆农垦经济, 2012, (11): 14–16. [doi: 10.3969/j.issn.1000-7652.2012.11.006]
16 李君华, 王志坚, 张立杰, 等. 基于小波理论及 ARIMA 模型的短期棉花价格预测. 中国棉花学会 2012 年年会暨第八次代表大会论文汇编. 运城: 《棉花学报》编辑部, 2012. 37–40.
17 张立杰, 朱新杰. 我国棉花价格长期走势与短期预测——基于差分自回归移动平均模型 (ARIMA) 的分析. 价格理论与实践, 2012, (6): 53–54. [doi: 10.19851/j.cnki.cn11-1010/f.2012.06.022]
18 高欣宇, 余国新. 对我国棉花期货价格预测的方法研究——基于 EGARCH-EWMA 模型与 ARIMA 模型比较. 价格理论与实践, 2014, (12): 85–87. [doi: 10.19851/j.cnki.cn11-1010/f.2014.12.035]
19 吴叶, 刘婷婷, 方少勇. 基于 MIV-GA-BP 神经网络的我国棉价预测研究. 棉纺织技术, 2018, 46(7): 77–80. [doi: 10.3969/j.issn.1001-7415.2018.07.020]
20 江知航, 王艳霞, 颜家均, 等. 基于 BiLSTM 的棉花价格预测建模与分析. 中国农机化学报, 2021, 42(8): 151–160. [doi: 10.13733/j.jcam.issn.2095-5553.2021.08.21]
21 Shin S, Lee M, Song SK. A prediction model for agricultural products price with LSTM network. The Korea Contents Association, 2018, 18(11): 416–429.
22 Gu YH, Jin D, Yin HL, *et al*. Forecasting agricultural commodity prices using dual input attention LSTM. Agriculture, 2022, 12(2): 256. [doi: 10.3390/agriculture12020256]
23 Li KQ, Shen N, Kang Y, *et al*. Livestock product price forecasting method based on heterogeneous GRU neural network and energy decomposition. IEEE Access, 2021, 9: 158322–158330. [doi: 10.1109/ACCESS.2021.3128960]
24 Bai SJ, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271, 2018.
25 Lara-Benítez P, Carranza-García M, Luna-Romera JM, *et al*. Temporal convolutional networks applied to energy-related time series forecasting. Applied Sciences, 2020, 10(7): 2322. [doi: 10.3390/app10072322]
26 Zha WT, Liu J, Li YL, *et al*. Ultra-short-term power forecast method for the wind farm based on feature selection and temporal convolution network. ISA Transactions, 2022, 129: 405–414. [doi: 10.1016/j.isatra.2022.01.024]
27 Luo S, Ni ZW, Zhu XH, *et al*. A novel methanol futures price prediction method based on multicycle CNN-GRU and attention mechanism. Arabian Journal for Science and Engineering, 2023, 48(2): 1487–1501. [doi: 10.1007/s13369-022-06902-6]
28 李占山, 刘兆赓. 基于 XGBoost 的特征选择算法. 通信学报, 2019, 40(10): 101–108. [doi: 10.11959/j.issn.1000-436x.2019154]

(校对责编: 牛欣悦)