

基于 ARIMA-LSTM 模型的 卷烟制丝质量预测研究

李志敏[†], 叶春明, 闫文凯

(上海理工大学, 上海 200093)

摘 要:为了应对卷烟制丝工艺中的不确定性所带来的质量风险,提出了一种将差分移动平均自回归 (ARIMA) 模型与深度学习中的长短期记忆单元 (LSTM) 模型相结合的制丝质量数据预测方法,并用该方法对某卷烟企业制丝工艺中的加热处理 (HT) 出口温度和叶丝增温增湿的入口水分进行预测,以均方根误差 (RMSE)、平均绝对误差 (MAE) 和平均绝对百分比误差 (MAPE) 为指标来评估各模型的预测精度。结果表明,相较于单一 ARIMA 和 LSTM 模型,所提出的 ARIMA-LSTM 组合模型的预测值更精准,在对 HT 出口温度进行预测时,组合模型的 RMSE、MAE、MAPE 值分别至少降低了 60.1%、63.1% 和 63%;在对叶丝增温增湿的入口水分进行预测时,组合模型的 RMSE、MAE、MAPE 值分别至少降低了 49.5%、49.4% 和 49.3%,有望为卷烟企业及时制定或调整生产方案提供合理的参考依据。

关键词:ARIMA 模型;长短期记忆网络;残差;时序预测;卷烟制丝质量

中图分类号:TP183

文献标识码:A

Research on Quality Prediction of Tobacco Primary Processing Based on ARIMA-LSTM Model

LI Zhimin[†], YE Chunming, YAN Wenkai

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

Abstract: In order to address quality risks associated with the uncertainty in the tobacco primary process, a tobacco forming quality data prediction method that combines the autoregressive integrated moving average (ARIMA) model with the long short-term memory (LSTM) model in deep learning is proposed, and the method is used to predict the outlet temperature of heat treatment (HT) and the inlet moisture of leaf filament warming and humidification in the tobacco primary process of a tobacco enterprise. It evaluates the prediction accuracy of each model using metrics such as root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The results show that the combined ARIMA-LSTM model provides more accurate predictions compared to the individual ARIMA and LSTM models. When predicting the HT export temperature, the combined model reduces the RMSE, MAE, and MAPE values by at least 60.1%, 63.1%, and 63% respectively. When predicting the inlet moisture content of the leaf conditioning process, the combined model reduces the RMSE, MAE, and MAPE values by at least 49.5%, 49.4%, and 49.3% respectively. This method has the potential to provide a reasonable reference for tobacco companies to timely formulate or adjust their production plans.

Key words: ARIMA model; long short-term memory networks; residuals; time series forecasting; tobacco primary quality

烟草行业是我国重点国营产业之一,对财政贡献巨大^[1]。随着竞争加剧以及消费者对产品质量要求的不断提高,烟草公司必须提升自己的加工水平^[2]。制丝是决定卷烟品质但又充满不确定性和复杂性的关键环节之一^[3],但由于监控体系不完整,滞后的生产调整方案无疑增添了大量的成本,探寻质量变化动向从而及时制定更有效的生产方案,对于卷烟企业意义重大。在大数据发展背景下,时间序列数据日益丰富^[4],卷烟企业可以充分利用历史数据对未来的生产进行有效的预测,从而帮助制定合理的决策和规划^[5]。

时间序列数据预测的研究方法经历了不断的演变,传统计量经济模型和人工智能模型针对不同的时间序列数据预测效果各有优势^[6]。以 ARIMA、GARCH 为例的传统计量经济模型对于平稳、低噪声数据的预测效果较佳^[7],例如对新冠疫情患病率预测^[8],新冠疫情对中国股市崩盘风险的预测等^[9]。以 RNN 为代表的人工神经网络智能模型在预测非线性、高噪声的时间序列数据时拟合效果更好^[10]。长短期记忆神经网络 LSTM 在 RNN 的基础上引入了门控制^[11],可以有效地解决长时间序列过程中梯度消失和梯度爆炸的问题^[12],且能够自适应地学习输入序列之间的时序关系。尚文利等设计了一种以 LSTM 为基础的异常探测模型,有效地提升了工艺数据中的异常探测精度^[13]。但卷烟企业的实际生产中的时间序列数据通常既有线性的特征,又有非线性的特征,单用一种模型难以达到精准的预测效果^[14]。近年来,国内外诸多学者通过结合 ARIMA-LSTM 在机械^[15]、能源^[16]、金融^[17]等领域的预测方面取得了不少研究成果,但很少有学者将其应用到烟草质量的预测领域^[18],且多数研究的做法是将单个模型预测结果赋予一定的权重直接求和得到组合模型的预测值,未考虑到数据线性和非线性的双重关系。

鉴此,本文选取了国内 X 卷烟企业制丝过程中的质量数据,首先运用 ARIMA 模型对其线性部分进行预测,把预测值与实际值的差值作为 LSTM 模型的输入,对非线性部分预测,再进行叠加得到组合模型预测值。此外,本文分别用 ARIMA 模型、LSTM 模型、ARIMA-LSTM 组合模型对 X 企业制丝过程中的两组卷烟物理指标数据依次进行预测并对比预测效果,以验证本文所提出的预测模型的精准性和普适性。

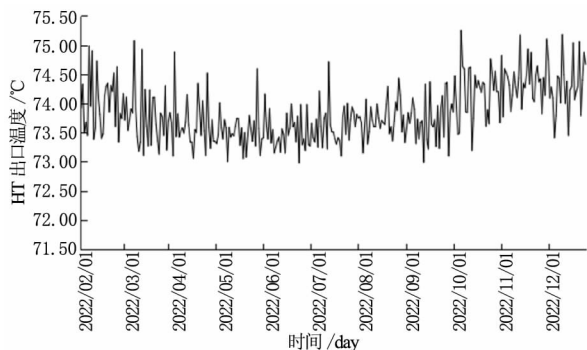
1 数据与方法

1.1 数据

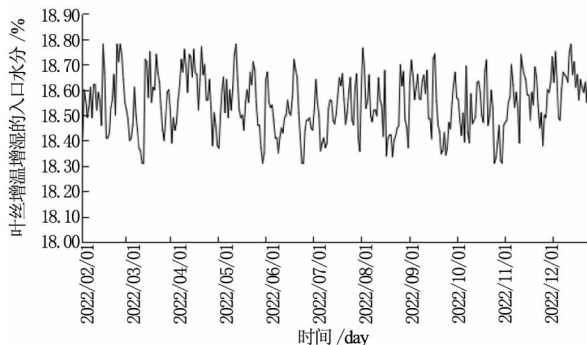
使用的实验数据均来自 X 卷烟企业制丝过程中的质量数据,卷烟的生产牌号为利群(新版),批次号为 FL2207LL1510,包括两个时间序列数据集:HT 出口温度($^{\circ}\text{C}$)数据集、叶丝增温增湿的入口水分($\%$)数据集,所取数据均为一天内的均值。

HT 是加热处理,卷烟 HT 出口温度是指在卷烟生产过程中,经过高温处理后,烟丝从卷烟生产线的热风道口出来时的温度。这个温度对烟丝的质量有一定的影响。通过合理预测并控制温度,可以保持烟丝的适宜含水率,保留烟草中的香气成分,同时保持叶片的颜色和气味,从而提高卷烟产品的品质和口感,并提高生产效率。叶丝增温增湿是卷烟生产过程中的一个关键步骤,它指的是通过加热和增湿,使烟叶中的水分适当增加,以便更好地进行卷烟生产。入口水分是指烟叶在进行叶丝增温增湿之前的初始水分含量。通过合理预测并调节入口水分,可以保持烟叶的柔软度和口感,保留烟叶中的香气成分,同时确保卷烟的燃烧性能和质量稳定,从而达到生产效率和卷烟品质的最佳平衡。

两个数据集如图 1 所示,时间覆盖范围是 2022 年 2 月 1 日至 2022 年 12 月 22 日。



(a) HT 出口温度数据集



b. 叶丝增温增湿的入口水分数据集

图 1 两个时间序列数据集

1.2 模型构建

1.2.1 数据预处理

首先对数据进行平稳性检验,若为非平稳时间序列,则需要差分处理。这里通过 ADF 检验判断原始时间序列是否为非平稳序列,设显著性水平为 0.05。对于 HT 出口温度数据集,在差分为 0 阶时, p 值为 0.381,即时间序列为非平稳的,进行一阶差分处理,得到 p 值远小于 0.05,即该一阶差分后的序列为平稳的时间序列;对于叶丝增温增湿的入口水分数据集,差分为 0 阶时, p 值远小于 0.05,故该原始序列为平稳的时间序列。其次,对数据进行白噪声检验,若序列是非白噪声的,则可以建立预测模型。实验得出一阶差分后的 HT 出口温度时间序列和原始的叶丝增温增湿的入口水分时间序列白噪声检验的 p 值均远小于 0.05,说明两个序列都是非白噪声序列。

1.2.2 数据集构建

两个时间序列数据集的时间覆盖范围均是 2022 年 2 月 1 日至 2022 年 12 月 22 日,其中 2 月 1 日至 10 月 18 日的数据用作训练集,其余用作测试集。

1.2.3 模型建立与训练

在处理时间序列数据预测方面问题时,不论是传统的 ARIMA 模型还是 LSTM 神经网络模型,都有自身的优势和局限。ARIMA 模型能够很好地挖掘线性信息^[19],LSTM 更善于处理非线性关系^[20]。但在实际应用中,许多数据很难判断到底是完全的线性关系还是非线性关系,所以本文结合 ARIMA 模型和 LSTM 模型设计出基于 ARIMA-LSTM 的组合预测模型,如图 2 所示。将原始时间序列数据进行分解得到线性部分和非线性部分,再用 ARIMA 预测线性部分,非线性部分使用 LSTM 来预测,最后叠加结果得到预测值,用公式表示为:

$$\hat{L}_{pre} = \hat{L}_{arima} + \hat{L}_{lstm} \quad (1)$$

式中, $\hat{L}_{pre}$ 代表组合模型预测序列值, $\hat{L}_{arima}$ 为 ARIMA 模型预测序列值, $\hat{L}_{lstm}$ 是 LSTM 模型的残差预测序列值。

结合图 2,建立 ARIMA-LSTM 模型可以分为以下几个步骤:

(1) 获取样本数据。记原始时间序列为 X_t ,线性部分为 L_t ,非线性部分为 N_t ,有:

$$X_t = L_t + N_t \quad (2)$$

(2) 平稳性检验。采用 ADF 检验来判断原始时间序列是否为非平稳序列,若为非平稳,则需要

进行差分运算,直到数据达到平稳, d 阶差分公式如下:

$$\nabla^d X_t = (1 - B)^d X_t \quad (3)$$

式中, ∇^d 为 d 阶差分算子。

(3) 模型定阶。通过 d 阶差分使数据达到平稳状态,接着观察序列的自相关图、偏自相关图选取合适的 p 和 q ,根据 AIC 判别准则来判定最终阶数,即选择使 AIC 达到最小的 p 和 q 值,AIC 的计算公式为:

$$AIC = n \ln \hat{\sigma}_e^2 + 2(p + q + 1) \quad (4)$$

式中, n 为样本容量, $\hat{\sigma}_e^2$ 与 p 、 q 有关。若当 $p = p'$, $q = q'$ 时,AIC 值最达到小,则认为模型的阶数为 p' 和 q' 。

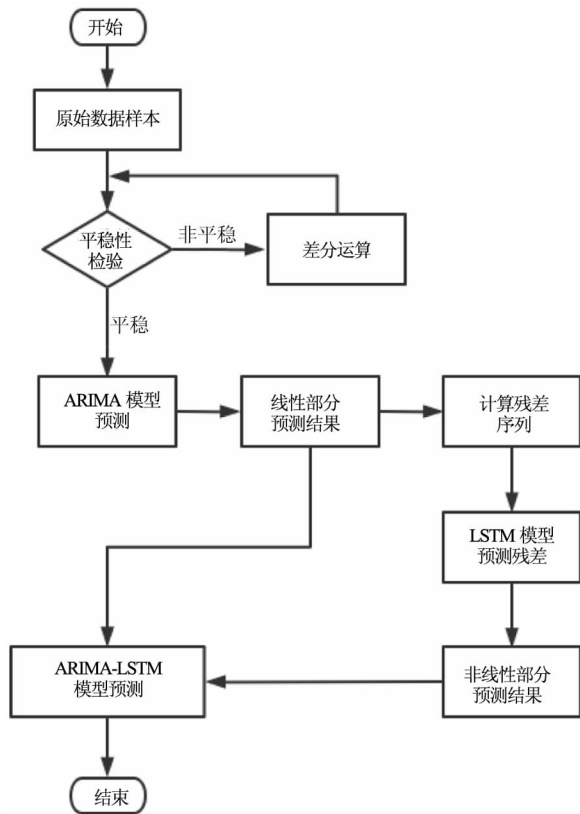


图 2 ARIMA-LSTM 建模流程图

(4) ARIMA 模型的线性部分预测。采用 ARIMA 模型来处理原始时间序列线性部分 L_t ,经过 ARIMA 模型得到拟合值 $\hat{L}_t$ 以及预测序列 $\hat{L}_{arima}$ 。

(5) 计算残差序列。用真实值序列 X_t 减去拟合值序列 $\hat{L}_t$ 得到残差序列 e_t ,公式如下:

$$e_t = X_t - \hat{L}_t \quad (5)$$

(6) LSTM 模型的非线性部分预测。在残差序列中,隐藏着原始序列的非线性特性,因此,对残差序列进行拟合就等于是对原始时间序列的非线

性部分拟合,用 LSTM 拟合残差序列得到的预测值 $\hat{L}_{lstm}$ 公式可表示为:

$$\hat{L}_{lstm} = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t \tag{6}$$

(7)最后,将线性部分与非线性部分预测值进行叠加得出最终组合模型预测值 $\hat{L}_{pre}$ 。

根据以上步骤可以将 ARIMA 模型和 LSTM 模型的各自优点有机结合,修正单一模型的误差,有效提高预测精度。

在对 LSTM 模型进行训练时,激活函数选择 ReLU,输入变量和输出变量个数均为 1,选择双隐藏层结构,每个隐藏层的神经元个数为 32,训练块大小 batch_size 设为 20,迭代次数 epoche 设置为 500 次。

1.3 模型评估

为了对不同模型的预测性能有一个客观的评价,选取三个常用的评价指标:均方根误差(RMSE)、平均绝对误差(MAE)和平均绝对百分比误差(MAPE),公式如下:

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \tag{7}$$

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \tag{8}$$

$$MAPE(y, \hat{y}) = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \tag{9}$$

式中, y_t 表示真实值, $\hat{y}_t$ 表示预测值, n 为时间序列数。通常, RMSE、MAE 和 MAPE 越小说明预测精度越高。

1.4 模型运行环境

开发工具选择 PyCharm,语言是 Python 3.7.0,深度学习框架为 Pytorch1.10.2。图像处理单元(Graphic Processing Unit, GPU)采用 NVIDIA GeForce GTX 1050(美国 NVIDIA 公司)。

2 结果与分析

2.1 模型训练结果

2.1.1 ARIMA 模型的训练结果

经过平稳性检验、白噪声检验和差分处理后,接着进行模型定阶,通过式(4)计算最小的 AIC 值,这里为了提高计算速度,采用 Python 语言编程实现,最终选择 ARIMA(0,1,2)和 ARIMA(1,0,0)作为 HT 出口温度和叶丝增温增湿入口水分的 ARIMA 模型。用相应的 ARIMA 模型分别对两个序列预测,结果如图 3 所示。

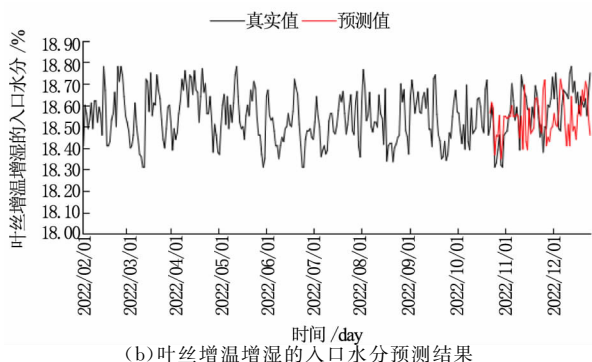
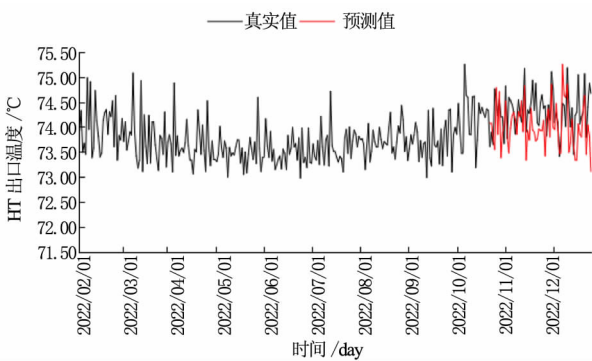


图 3 ARIMA 模型预测结果

2.1.2 LSTM 模型的训练结果

用训练好的 LSTM 模型分别对 HT 出口温度和叶丝增温增湿的入口水分两个序列残差进行预测,结果如图 4 所示。

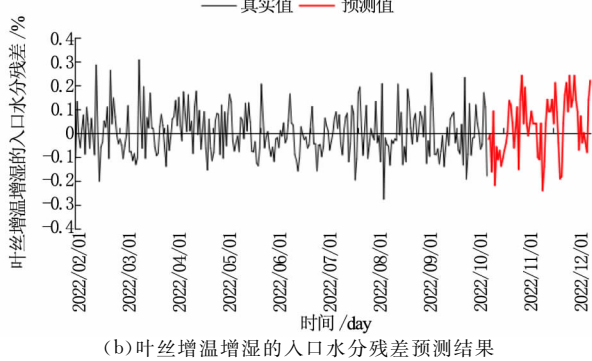
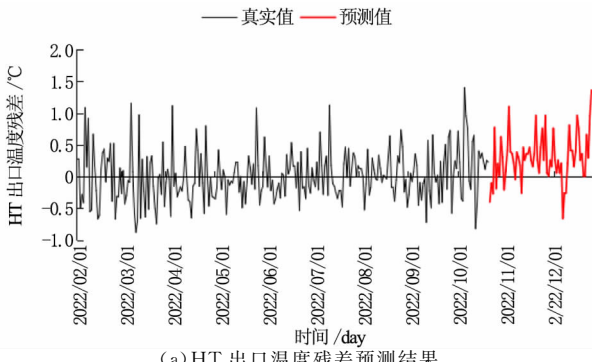


图 4 LSTM 模型预测结果

2.1.3 ARIMA-LSTM 组合模型的训练结果

利用式(7)构造 ARIMA-LSTM 组合模型,并分别对 HT 出口温度和叶丝增温增湿的入口水分

预测,组合模型和单一模型的预测结果对比如图 5 所示,计算各模型的预测相对误差,如图 6 所示。

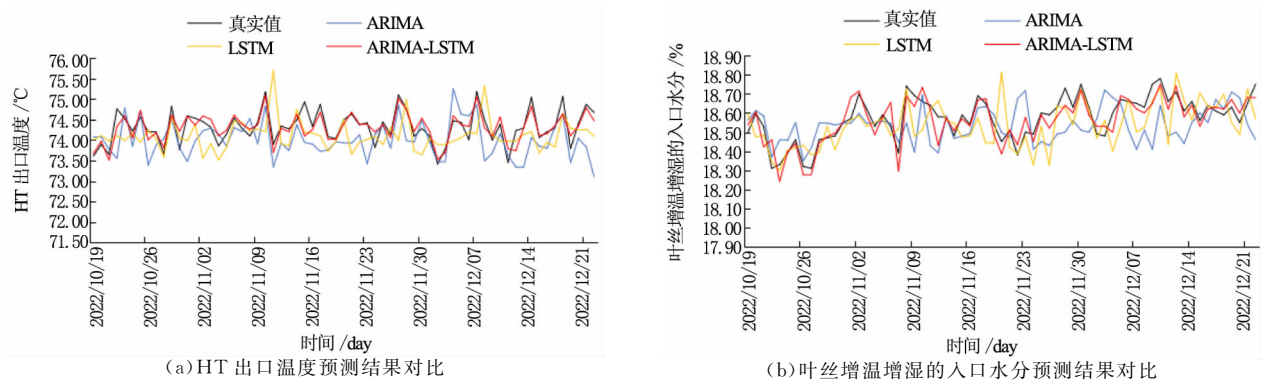


图 5 不同模型预测结果对比

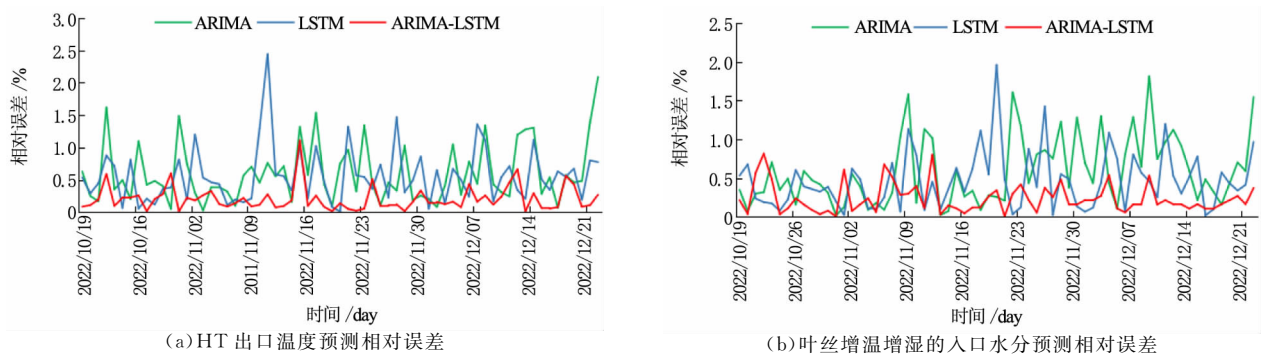


图 6 不同模型预测相对误差

2.2 模型效果评估

从图 5 和图 6 可以看出,组合模型对 HT 出口温度和叶丝增温增湿的入口水分的预测更精准。在对 RMSE、MAE、MAPE 三个指标进行对比之后,得到了组合模型和单一模型的预测性能,如表 1 所示。从表中可以看出,与单独的 ARIMA 模型、LSTM 模型相比,组合模型具有更好的预测效

果。在对 HT 出口温度进行预测时,与单一模型相比,组合模型的 RMSE、MAE、MAPE 值分别至少降低了 60.1%、63.1%和 63%;在对叶丝增温增湿的入口水分进行预测时,与单一模型相比,组合模型的 RMSE、MAE、MAPE 值分别至少降低了 49.5%、49.4%和 49.3%。

表 1 组合模型和单一模型预测性能对比

预测参量	预测模型	评价指标		
		RMSE	MAE	MAPE
HT 出口温度	ARIMA	0.5648	0.4494	0.6037%
	LSTM	0.5243	0.4168	0.5600%
	ARIMA-LSTM	0.2091	0.1540	0.2072%
叶丝增温增湿的入口水分	ARIMA	0.1367	0.1092	0.5870%
	LSTM	0.1104	0.0861	0.4631%
	ARIMA-LSTM	0.0558	0.0436	0.2350%

3 结 论

在卷烟企业的实际生产中,尤其是制丝工艺中的各项物理指标,其时间序列数据通常既有线性的特征,又有非线性的特征,单用一种模型难以达到精准的预测效果,为此,本文提出了 ARIMA-LSTM 组合模型来对 X 卷烟企业制丝工艺中的两项质量数据进行预测,选择 ARIMA(0,1,2)和 ARIMA(1,0,0)分别预测 HT 出口温度和叶丝增温增湿入口水分的线性部分,使用 LSTM 模型对 ARIMA 模型预测结果进行修正,得到非线性部分预测,结合两模型预测结果计算出最终的组合模型预测值。经计算,在对 HT 出口温度进行预测时,组合模型的 RMSE、MAE、MAPE 值分别至少降低了 60.1%、63.1%和 63%;在对叶丝增温增湿的入口水分进行预测时,组合模型的 RMSE、MAE、MAPE 值分别至少降低了 49.5%、49.4%和 49.3%,这说明与单一的 ARIMA 模型和 LSTM 模型相比,组合模型的预测值更精准,可以为卷烟企业在制丝工艺中的质量预测提供参考。

在将来,还可以进一步拓展其应用范围,完善技术水平。例如:该方法可用于除烟叶产业外的其他企业,对其生产流程进行质量预报;或者加入某些相关的特性,作为模型的输入,寻找与不同的序列数据相对应的最优网络结构。

参考文献

[1] 陈超英. 变革与挑战: 新型烟草制品发展展望[J]. 中国烟草学报, 2017, 23(3): 14—18.

[2] 窦玉青, 沈轶, 杨举田, 等. 新型烟草制品发展现状及展望[J]. 中国烟草科学, 2016, 37(5): 92—97.

[3] 鲁宇童, 张艺洁, 丁丹阳, 等. 基于正交设计的制丝工艺参数对叶丝质量的影响[J]. 湖南农业大学学报(自然科学版), 2018, 44(2): 140—144.

[4] 张美英, 何杰. 时间序列预测模型研究综述[J]. 数学的实践与认识, 2011, 41(18): 189—195.

[5] 钟宇, 董浩, 邢军, 等. 烟草行业质量数据利用现状及应用

需求分析[J]. 烟草科技, 2023, 56(2): 104—112.

[6] ABBASIMEHR H, SHABANI M, YOUSEFI M. An optimized model using LSTM network for demand forecasting[J]. Computers & Industrial Engineering, 2020, 143: 106435.

[7] 徐映梅, 陈尧. 季节 ARIMA 模型与 LSTM 神经网络预测的比较[J]. 统计与决策, 2021, 37(2): 46—50.

[8] CEYLAN Z. Estimation of Covid-19 prevalence in Italy, Spain, and France[J]. Science of The Total Environment, 2020, 729: 138817.

[9] LIU Z, HUYNH T L, DAI P F. The impact of covid-19 on the stock market crash risk in China[J]. Research in International Business and Finance, 2021, 57: 101419.

[10] LARA-BENÍTEZ P, CARRANZA-GARCÍA M, RIQUELME J C. An experimental review on Deep Learning Architectures for time series forecasting[J]. International Journal of Neural Systems, 2021, 31(3): 2130001.

[11] 刘博, 王明烁, 李永, 等. 深度学习在时空序列预测中的应用综述[J]. 北京工业大学学报, 2021, 47(8): 925—941.

[12] 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2): 1—6+26.

[13] 尚文利, 石贺, 赵剑明, 等. 基于 SAE-LSTM 的工艺数据异常检测方法[J]. 电子学报, 2021, 49(8): 1561—1568.

[14] 王英伟, 马树才. 基于 ARIMA 和 LSTM 混合模型的时间序列预测[J]. 计算机应用与软件, 2021, 38(2): 291—298.

[15] 崔建国, 李鹏程, 崔霄, 等. 基于 ARIMA-LSTM 的飞机液压泵性能趋势预测方法[J]. 振动. 测试与诊断, 2021, 41(4): 735—740+832.

[16] FAN D, SUN H, YAO J, et al. Well production forecasting based on ARIMA-LSTM model considering manual operations[J]. Energy, 2021, 220: 119708.

[17] BUKHARI A H, RAJA M A Z, SULAIMAN M, et al. Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting[J]. IEEE Access, 2020, 8: 71326—71338.

[18] 刘颖, 吴甦, 黄国梁, 等. 基于扎根理论的制造企业质量管理关键要素研究[J]. 工业工程与管理, 2019, 24(1): 182—188.

[19] BABU C N, REDDY B E. A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data[J]. Applied Soft Computing, 2014, 23: 27—38.

[20] YU Y, SI X S, HU C H, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural Computation, 2019, 31(7): 1235—1270.