基于小波去噪的ARIMA-LSTM混合模型及对股票价格指数的预测

娄磊，刘璐，刘先俊，施三支
（长春理工大学 理学院，长春 130022）

摘 要：由于传统ARIMA模型只对数据线性部分有非常好的拟合效果，而LSTM模型对非线性数据有很好的拟合效果，基于误差补偿的思想，给出了基于小波去噪的ARIMA-LSTM混合模型，并用此模型来对我国上证指数每日收盘价格进行预测，并将预测结果与单独使用ARIMA模型和LSTM模型的预测结果进行对比，结果表明使用ARIMA-LSTM混合模型可以有效地提高股指预测的精准度.

关键词：股票价格指数；小波去噪；ARIMA-LSTM混合模型
中图分类号： O212.1 **文献标志码**：A **文章编号**：1672-9870(2021)02-0119-05

ARIMA-LSTM Hybrid Model Based on Wavelet Denoising and Prediction of Stock Price Index

LOU Lei, LIU Lu, LIU Xian-jun, SHI San-zhi
(School of Science, Changchun University of Science and Technology, Changchun 130022)

Abstract: Because the traditional ARIMA model only has a very good fitting effect on the linear part of the data, while the LSTM model has a good fitting effect on the nonlinear data. Based on the idea of error compensation, the ARIMA-LSTM hybrid model based on wavelet denoising is given. And use this model to predict the daily closing price of China's Shanghai Composite Index, and compare the prediction results with the prediction results of ARIMA model and LSTM model alone. The results show that the ARIMA-LSTM hybrid model can effectively improve the accuracy of stock index prediction.

Key words: stock price index; wavelet denoising; ARIMA-LSTM hybrid model

ARIMA模型由Box与Jenkins于上世纪七十年代提出[1]，广泛应用于时间序列的预测，并取得了非常好的效果，因此ARIMA模型也被应用在了股票价格指数的预测上。我国经济学者陈守东、孟庆顺、杨兴武等人[2]最早于1998年对中国股票市场的有效性进行检验分析，并且应用ARIMA模型验证了上海、深圳股市的同步性，得出我国股票市场还在发展初期，但他们并没有对股指进行预测。赵志峰[3]于2003年建立ARIMA模型对我国股票价格指数进行预测，并得出干预模型更好的结果，同时也说明了我国市场是一个政策市场。吴玉霞、温欣[4]在2016年使用“华泰证券”250期的股票收盘价格作为时间序列分析数据，建立ARIMA模型进行实证检验，证

收稿日期：2019-12-06
基金项目：国家自然科学基金（11601039）；吉林省自然科学基金（20140101199JC）
作者简介：娄磊（1995-），男，硕士研究生，E-mail：2425031862@qq.com
通讯作者：施三支（1968-）女，博士，教授，E-mail：shisz@cust.edu.cn

明了ARIMA模型在短期动态、静态预测效果好的特点。

早在二十世纪八、九十年代，许多专家学者开始探讨循环神经网络（RNN），并在二十一世纪将其发展成为深度学习算法之一[5]。由于循环神经网络无法克服梯度消失和梯度爆炸的缺点，Sepp Hochreiter和Jurgen Schmidhuber于1997年提出长短期记忆神经网络（LSTM）来进行长期预测[6]，主要应用领域是语言模型以及文本生成[7-8]、机器翻译[9-11]、语音识别[12]、图像生成等。近十多年随着机器学习的兴起，越来越多的学者用神经网络对股票价格指数进行预测。林春燕、朱东华[13]于2006年利用了Elman神经网络对股票价格进行预测，并得到了不错的效果。随着时间的推移，我国经济学家开始使用混合模型进行预测。回旋[14]认为股票市场是一个极其复杂的非线性动力学系统，具有高噪声、非线性和投资者的盲目任意性等因素，造成其价格的波动往往表现出较强的非线性特征，针对这些问题提出了TS模糊规则与神经网络结合的模型，并对绿景地产等进行实证检验。彭燕、刘宇红[15]结合LSTM神经网络的特点，先对数据进行插值、小波去噪预处理，再通过调整LSTM层数与隐藏神经元的个数提高预测精准度。于水玲[16]于2018年基于深度学习对金融市场的波动率进行了预测，也取得了不错的效果。

综上，对股票价格指数的研究大致可以分为三个阶段。第一阶段使用ARIMA模型对股票价格指数进行预测；第二阶段使用神经网络对股票价格指数进行预测；第三阶段使用各种混合模型对股票价格指数进行预测，旨在不断提高预测精准度。

为了提高预测精度，提出了使用小波去噪后的数据，采用ARIMA-LSTM混合模型的方法进行预测。采集2009年7月至2018年12月共计114个月上证指数每日收盘价格，经小波去噪法预处理后做为训练集，对2019年1月至2019年6月共计6个月的上证指数进行实证预测，预测结果与单独使用ARIMA模型与LSTM神经网络模型相比，ARIMA-LSTM混合模型效果较好。

1 相关的模型理论简介

1.1 ARIMA模型

假设有时间序列输入为$x_1,x_2,\cdots,x_T,\{\varepsilon_t\}$是高斯白噪声，则：

$$\varphi(B)\nabla^d X_t = \theta(B)\varepsilon_t \tag{1}$$

为求和自回归滑动平均模型，记为ARIMA(p,d,q)，其中d为差分的阶数，p，q为自回归与滑动平均的阶数，且：

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p,$$

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q.$$

式中，B表示后移算子。根据平稳性的要求，多项式$\varphi(z)$和$\theta(z)$的根在单位圆外，即$|z| > 1$。

1.2 LSTM模型

当序列长度很大时，RNN模型会发生爆炸/消失梯度，因此很难捕获序列数据中的长期相关性，为了克服RNN模型缺点，Sepp Hochreiter和Jurgen Schmidhuber提出了LSTM模型[6]。

假设有时间序列输入为$x_1,x_2,\cdots,x_T$，LSTM模型有如下形式：

$$\begin{pmatrix} i_t \\ f_t \\ O_t \\ \tilde{C}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \\ 1 \end{pmatrix} \tag{2}$$

其中，W是具有适当维数的大权重矩阵，且：

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t$$

$$h_t = O_t \tanh(C_t)$$

根据LSTM的结构，单元状态向量C_t表示携带序列信息，遗忘门f_t决定C_{t-1}的值在时间t内保留多少，输入门i_t控制单元状态的更新量，输出门O_t给出C_t向h_t显示多少，b表示偏差权重向量（例如，b_i是输入门的偏差权重向量），σ为sigmoid或relu函数。LSTM具有类似RNN的链式结构，如图1所示，但是与RNN相比较而言，LSTM的结构相对复杂。

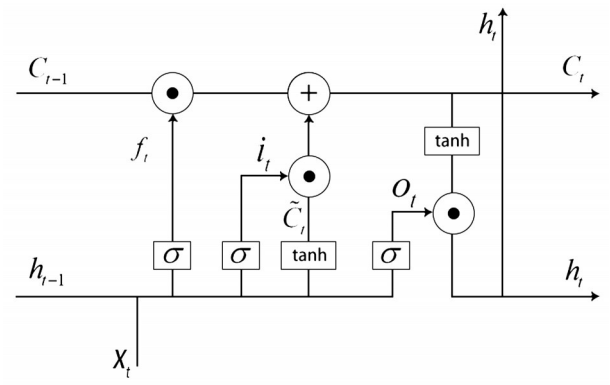


图1　LSTM结构图

LSTM包含一个特殊的单元，即递归隐藏层中的内存块。这些内存块包含具有自连接的内存单元，这些内存单元存储网络的时间状态，此外还包含特殊乘法单元“门”，用于控制信息流。每个内存块都包含一个输入门和一个输出门和遗忘门。

假设包含一个LSTM层的三层长短期记忆网络输入向量序列为 $x = (x_1,\cdots x_T)$，输出向量序列 $h = (h_1,\cdots h_T)$，则LSTM的向前传播流程可以用如下具体表示：

$$\tilde{C}_t = \tanh\left(W_C\left[h_{t-1},x_t\right] + x_C\right)$$

$$i_t = \sigma\left(W_i\left[h_{t-1},x_t\right] + b_i\right)$$

$$f_t = \sigma\left(W_f\left[h_{t-1},x_t\right] + b_f\right)$$

$$C_t = i_t \cdot z_t + f_t \cdot C_{t-1}$$

$$O_t = \sigma\left(W_O\left[h_{t-1},x_t\right] + b_o\right), h_t = O_t \tanh\left(C_t\right)$$

其中，$t = 1,\cdots,T$。内模块输出 h_t 受到 $f_t, i_t, \tilde{C}_t, O_t$，四个输入的加权影响，并且在学习时还要将拼接的权重 W 进行分割，即：

$$W_c = W_{cx} + W_{ch}, W_i = W_{ix} + W_{ih}$$

$$W_f = W_{fx} + W_{fh}, W_o = W_{ox} + W_{oh}$$

当输出层的输入为：$y_t^i = W_{yi}h_t$，则输出 $y_t^o = \sigma\left(y_t^i\right)$

1.3　ARIMA-LSTM混合模型

假设有时间序列输入为 $x_1, x_2, \cdots, x_T$，ARIMA-LSTM混合模型有如下形式：

$$\varphi(B)\nabla^d X_t = \theta(B)\varepsilon_t$$

$$\begin{pmatrix} i_t \\ f_t \\ O_t \\ \tilde{C}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \\ 1 \end{pmatrix} \quad (3)$$

ARIMA模型对数据线性部分有很好的拟合效果，而LSTM模型对非线性数据有好的预测效果，而上证指数的收盘价也同时具有线性和非线性成分，基于误差的补偿思想给出ARIMA-LSTM混合模型可对上证指数进行预测。由于股票数据的高波动性，因此在预测前对数据进行小波去噪预处理，来剔除异常数据可以提高预测准确率，这里用平均绝对误差（MAE）和均方根误差（RMSE）：

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\left(f_i - y_i\right)\right|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - f_i\right)^2}$$

作为评价模型预测效果，预测步骤如下：第一步使用小波去噪法对数据进行预处理；第二步建立ARIMA预测模型；第三步进行模型检验；第四步使用ARIMA模型预测；第五步建立LSTM模型预测残差序列；第六步加和得到ARIMA-LSTM混合模型的预测结果；第七步进行误差分析。预测的流程图如图2所示。

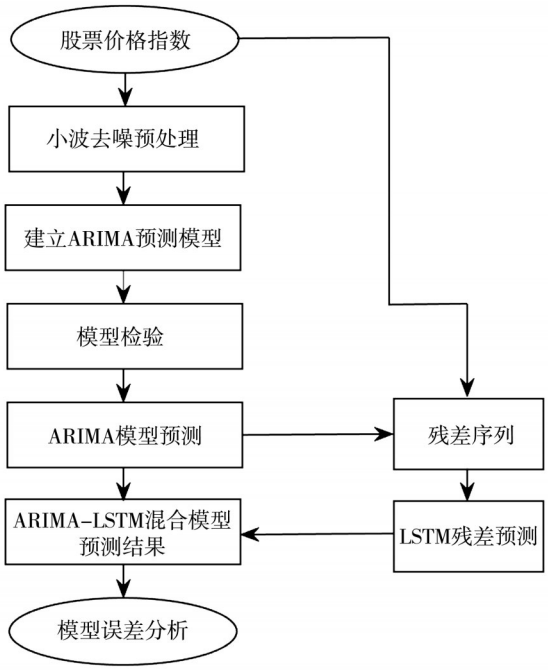


图2　基于小波去噪的ARIMA-LSTM预测模型流程图

2 实证分析

以原始的和经小波去噪后的2009年7月至2018年12月共计114个月的上证指数每日收盘价格作为训练集，建立ARIMA模型处理数据的线性部分，再利用LSTM模型拟合非线性数据效果好的优点，对ARIMA模型无法拟合的残差进行拟合，用ARIMA-LSTM混合模型对2019年1月至2019年6月共计6个月的上证指数每日收盘价格进行预测。

由Occam's Razor法则知，误差相同时模型越小其效果就会越好。因此常用AIC信息准则、BIC信息准则来对ARIMA模型进行定阶。由于股票价格指数高波动性等因素，有时通过理论知识选出的最优模型不一定是预测效果最好的，所以根据不同的训练集，分别建立几组不同的模型，对模型检验后，用平均绝对误差（MAE）和均方根误差（RMSE）作为评价模型预测效果的标准进行预测误差分析，选出不同训练集下的误差最小的模型。

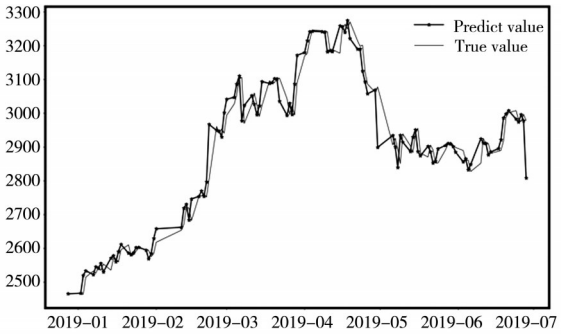
在建立LSTM模型预测时，通过增加训练次数来提高预测精度。在实证分析中参数Dropout设为0.2，采用3层神经网络，时间步长设置为20，每批训练60个样本数据，学习率设为0.000 6，输入层到隐藏层之间的激活函数使用tanh函数，预测时将训练次数从800次增加到2 000次，每次训练都增加100次并记录结果。

建立不同模型在不同训练集下的最优模型误差对比如表1所示。

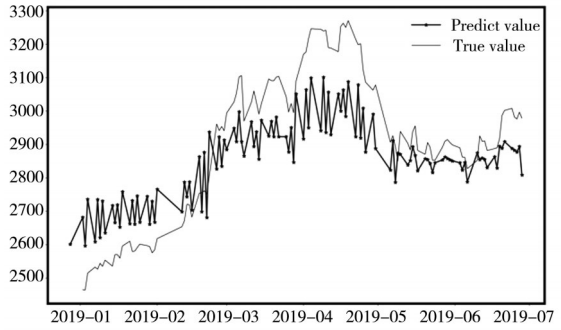
表1 不同训练集下的最优模型误差对比

模型	平均绝对误差（MAE）	均方根误差（RMSE）
原始最优ARIMA(2,2,2)模型	15.038 7	18.025 7
去噪最优ARIMA(1,2,1)模型	2.045 8	2.801 0
原始LSTM模型	32.978 6	43.175 4
去噪LSTM模型	28.451 3	37.258 7
原始ARIMA-LSTM混合模型	1.972 1	2.799 4
去噪ARIMA-LSTM混合模型	1.897 6	2.765 3

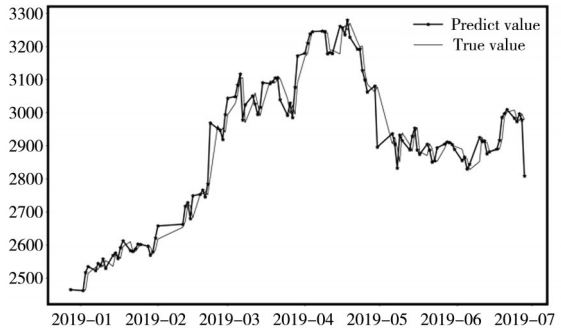
根据表1的误差对比可以看出：训练集经小波去噪预处理后，再进行建模预测得到的预测结果要好于用未经预处理作为训练集的预测结果，去噪后的混合模型和单模型预测效果图、误差对比分别如图3、表2所示。



（a）去噪最优ARIMA(1,2,1)模型



（b）去噪LSTM模型



（c）去噪ARIMA-LSTM混合模型

图3 单模型和去噪后的混合模型预测效果图

表2 三种去噪后的模型预测误差对比

模型	平均绝对误差（MAE）	均方根误差（RMSE）
去噪最优ARIMA(1,2,1)模型	2.045 8	2.801 0
去噪LSTM模型	28.451 3	37.258 7
去噪ARIMA-LSTM混合模型	1.897 6	2.765 3

根据表2的误差对比可以看出：第一，去噪后的ARIMA模型的MAE和RMSE均远远小于去噪后的LSTM模型，由此说明比起LSTM模型，ARIMA模型可以更好的拟合我国股票价格指

数，而ARIMA模型主要用来拟合时间序列中的线性部分，所以股票价格指数中大部分数据是线性的，但是股票价格具有高波动性等不可控因素，因此对非线性因素预测也至关重要；第二，去噪后的ARIMA-LSTM混合模型的MAE和RMSE均稍微小于单独建立去噪后的ARIMA模型的MAE和RMSE，因此建立LSTM模型对残差进行拟合，将结果与单独建立ARIMA模型预测的结果相加后得到结果更加接近真实值，所以去噪后的ARIMA-LSTM混合模型可以提高预测精度。

3　结论

通过实证分析对ARIMA模型、LSTM模型以及ARIMA-LSTM混合模型的预测效果进行了对比，结果表明基于小波去噪的ARIMA-LSTM混合模型对上证指数收盘价格进行预测效果最好，其次是ARIMA模型，最后是LSTM模型。这是因为股票价格指数中大部分数据是呈线性趋势的，与LSTM模型相比ARIMA模型可以更好的拟合线性数据，所以单独建立ARIMA模型预测的效果要比单独建立LSTM模型的预测效果好；由于股票价格指数具有高波动性等因素，线性数据中也会参杂非线性数据，因此给出基于误差补偿思想的去噪ARIMA-LSTM混合模型，在建立ARIMA模型对线性数据拟合后，再建立LSTM模型对非线性残差数据进行拟合，进一步提高了预测的精准度；该模型只适用于对我国上证指数的预测，并不适用于其它个股以及创业板指数的预测。

参考文献

[1] Ho S L, Xie M, Goh T N. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction[J]. Computers & Industrial Engineering, 2002, 42(2): 371-375.

[2] 陈守东，孟庆顺，杨兴武. 中国股票市场的有效性检验与分析[J]. 吉林大学社会科学学报，1998，7(2)：69-74.

[3] 赵志峰. 对建立中国股票价格指数时间序列模型的探讨[J]. 统计与信息论坛，2003，18(1)：66-69.

[4] 吴玉霞，温欣. 基于ARIMA模型的短期股票价格预测[J]. 统计与决策，2016，10(23)：83-86.

[5] Schmidhuber, Jürgen. Deep learning in neural networks: an overview[J].Neural Netw, 2015, 61(2): 85-117.

[6] Hochreiter, Sepp, Schmidhuber Jürgen. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[7] Graves, Alex. Generating sequences with recurrent neural networks[J]. Computer Science, 2013, 6(12): 1-43.

[8] Sak, Haşim, Senior, et al. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [J]. Computer Science, 2014, 7(10): 338-342.

[9] Cho, Kyunghyun, Van Merrienboer, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. Computer Science, 2014, 7(11): 1724-1734.

[10] Cinar, Yagmur G, Mirisaee, et al. Position-based content attention for time series forecasting with sequence-to-sequence RNNs [C]. International Conference on Neural Information Processing, 2017.

[11] Yang T H, Tseng T H, Chen C P. Recurrent neural network-based language models with variation in net topology, language, and granularity [C]. International Conference on Hsian Language Processing, 2017.

[12] Graves, Alex, Mohamed, et al. Speech recognition with deep recurrent neural networks [J]. International Conference, 2013, 2(13): 6645-6649.

[13] 林春燕，朱东华. 基于Elman神经网络的股票价格预测研究[J]. 计算机应用，2006，26(2)：476-477.

[14] 回旋. 模糊神经网络在股票预测中的应用研究[D]. 蚌埠：安徽财经大学，2012.

[15] 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，2019，12(11)：209-212.

[16] 于水玲. 基于深度学习的金融市场波动率预测和风险值计算[D]. 吉林：长春理工大学，2018.