文章编号:1674-6864(2021)01-0079-05 DOI: 10.16508/j.cnki.11-5866/n.2021.01.015

基于LSTM神经网络的股票价格预测研究

黄超斌,程希明
(北京信息科技大学 理学院,北京 100192)

摘 要:基于LSTM神经网络模型进行股票价格的预测研究。利用开市以来的七千多条上证综合指数数据,使用长短期记忆(LSTM)神经网络模型对上证综合指数进行预测分析,并将其预测结果与使用BP神经网络模型、CNN、RNN、GRU网络模型的预测结果进行对比。结果显示LSTM神经网络模型的预测效果最好,其评价指标中的平均绝对误差(MAE)为0.015 799,均方误差(MSE)为0.000 450,平均绝对百分比误差(MAPE)为0.019 867,预测误差低于其他模型;其预测值和真实值之间的相关系数为0.995 7,表明预测值和真实值的拟合程度较高。

关 键 词:股票;价格预测;神经网络;长短期记忆(LSTM)神经网络;时间序列
中图分类号:F 830.91 **文献标志码**:A

Research on stock price prediction based on LSTM neural network

HUANG Chaobin, CHENG Ximing
(School of Applied Science, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract: LSTM neural network model is used to predict stock prices. Based on more than 7 000 Shanghai Composite Index data since the opening of the market, long short-term memory (LSTM) neural network model is used to conduct the prediction analysis, and the prediction results are compared with those of the BP neural network model, CNN model, RNN model and GRU network model. The results show that the LSTM neural network model has the best prediction effect. The average absolute error (MAE) of the evaluation index is 0.015 799, the mean square error (MSE) is 0.000 450, the average absolute percentage error (MAPE) is 0.019 867, and the prediction error is lower than other models. The correlation coefficient between the predicted value and the true value is 0.995 7, indicating that the predicted value and true value have a high degree of fitting.

Keywords: stock; price prediction; neural network; long short-term memory (LSTM) neural network; time series

0 引言

股票价格的预测在商业和金融领域具有重要的意义,但由于股票价格受到众多因素的影响,导致股票价格预测十分困难。国内外许多相关工作者进行股票价格的预测研究,使用的方法大致可以分为3类:使用传统的时间序列模型、使用机器学习算法以及使用神经网络模型。

冯盼等[1]运用ARMA模型(自回归移动平均模型)对股票开盘价格进行预测,证实了ARMA模型在金融时间序列数据上预测的效果;王丽娜等[2]使用ARMA模型在经济非平稳时间序列上进行预测分析;田翔等[3]使用支持向量回归方法(support vector regression, SVR)构建股指短期预测模型,提出一种非线性时间序列预测模型;Wang Shuai等[4]以及Wang Jianzhou等[5]通过使用支持向量机(support vector machine, SVM)构建模型进行沪深300指数的趋势预测,验证了支持向量机在股票价格指数预测中的有效性。近年来,深度学习和神经网络的研究发展迅速,Dixon等[6]使用BP(back propagation)神

收稿日期:2020-10-07
第一作者简介:黄超斌,男,硕士研究生;通讯作者:程希明,男,教授。

经网络模型,预测商品期货和外汇期货在未来5 min的价格波动。

传统的时间序列模型在使用时要求时间序列数据必须是平稳的且无自相关性,在进行中长期的变化趋势预测时误差较大。随着研究的不断进行,许多学者将机器学习算法和神经网络模型应用到股票价格的预测之中。相比于传统的机器学习算法,使用神经网络模型时不需要构建复杂的、特定的特征工程即可构建有效的学习模型,因此在数据预处理阶段会节省大量的工作和时间。神经网络模型可以有效解决股票数据的不确定、非线性、非平稳等问题。许多学者的研究结果显示神经网络模型在预测股票价格上具有很好的效果。本文利用长短期记忆(long short-term memory, LSTM)神经网络模型对上证综合指数数据进行预测研究,并将预测结果与使用BP神经网络模型、卷积神经网络(convolutional neural network, CNN)模型、循环神经网络(recurrent neural network, RNN)模型、GRU(gated recurrent unit)网络模型的预测结果进行对比分析。

1 长短期记忆神经网络

由于神经网络模型更容易对复杂的非线性关系进行拟合,而股票指数、股票价格等金融时间序列数据因受到多种因素的影响呈现出非线性的特征,因此利用神经网络模型可以提高股指以及股票价格预测的精确度。但是传统的BP神经网络只在网络层与层之间建立了全连接,忽略了时间维度上的信息,由此RNN应运而生。RNN通过在神经网络的神经元之间建立连接,对隐藏层的中间结果进行循环利用,使得网络可以从时间维度上提取到有用的信息。RNN神经网络结构如图1所示。

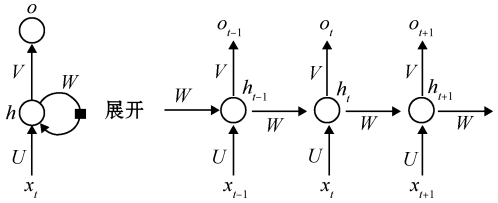

图1 RNN神经网络结构

其中x_{t-1}、x_t、x_{t+1}分别表示在$t-1$、t、$t+1$时刻神经元节点的输入,h_{t-1}、h_t、h_{t+1}分别表示对应的中间输出结果。在t时刻的输出结果o_t不仅仅与x_t有关,还和$t-1$时刻的输出结果h_{t-1}有关,这样就有效地保留了时间维度上的历史信息。因此RNN模型比BP神经网络模型更适合处理时间序列上的预测问题。其中t时刻的输出结果o_t为

$$o_t = g(\boldsymbol{V} \cdot h_t) \tag{1}$$

其中:

$$h_t = f(\boldsymbol{U} \cdot x_t, \boldsymbol{W} \cdot h_{t-1}) \tag{2}$$

神经网络中的误差梯度在反向传播过程中,随着时间序列长度的增加以及网络层数的加深,后层的梯度以连乘的方式叠加到前层。由于网络中使用的激活函数为Sigmoid函数,该函数具有饱和性,在输入达到一定值的情况下,输出不会发生明显变化。当后层梯度较小时,误差梯度传到前层时几乎衰减为0,导致RNN出现梯度消失的问题,无法对前层的网络参数进行学习,以至于预测能力无法提升。为了改进RNN的不足,Hochreiter和Schmidhuber提出了LSTM神经网络结构[7]。LSTM神经网络是一种特殊的循环神经网络,主要是为了解决长序列数据在训练过程中的梯度消失和梯度爆炸问题,相比于RNN,LSTM神经网络可以在更长的序列中取得较好的效果,能够对有价值的信息进行长期的记忆。LSTM神经网络的神经单元结构如图2所示,其中包含了输入门i_t、遗忘门f_t、输出门o_t和记忆单元c_t。输入门控制当前计算的新状态以多大程度更新到记忆单元中;遗忘门控制前一步记忆单元中的信息有多大程度被遗忘掉;输出门控制当前的输出在多大程度上取决于当前的记忆单元[8]。

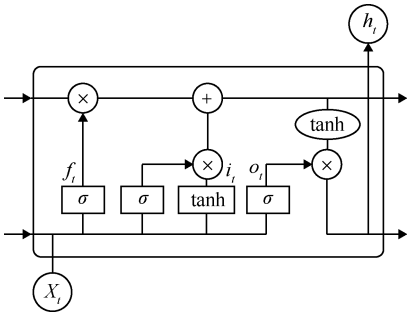

图2 LSTM神经网络的神经单元结构

LSTM神经网络在t时刻的输出$\boldsymbol{h}_t$为

$$\boldsymbol{h}_t = \boldsymbol{o}_t \cdot \tanh(\boldsymbol{c}_t) \tag{3}$$

其中:

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \tilde{\boldsymbol{c}}_t \tag{5}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{6}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{7}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1}) \tag{8}$$

式中:$\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\boldsymbol{o}_t$ 分别为 t 时刻的遗忘门、输入门和输出门的输出;$\tilde{\boldsymbol{c}}_t$ 和 $\boldsymbol{c}_t$ 分别为 t 时刻记忆单元中存储的内容;$\boldsymbol{x}_t$、$\boldsymbol{h}_t$ 分别为 t 时刻的输入向量和隐藏层的输出,σ 表示 Sigmoid 函数;$\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{b}$ 分别为计算时权重矩阵和偏置向量。

LSTM 神经网络的神经单元结构中的门控结构使用的是 Sigmoid 激活函数和 tanh 激活函数。其中 Sigmoid 函数为

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{9}$$

tanh 函数为

$$f(x) = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{\mathrm{e}^{x} + \mathrm{e}^{-x}} \tag{10}$$

两者的区别在于 Sigmoid 函数的值域为(0,1),而 tanh 函数的值域为(-1,1)。两个激活函数都是随着变量 x 的增大而增大,以实现门控结构的作用。

2　实验与分析

2.1　数据集介绍

本文使用的数据集为通过 tushare 库获取到的上证综合指数数据,该数据集一共包含 11 列数据,包括:指数代码(ts_code)、交易日期(trade_date)、收盘点位(close)、开盘点位(open)、最高点位(high)、最低点位(low)、昨日收盘点(pre_close)、涨跌值(change)、涨跌幅(pct_chg)、成交量(vol)、成交额(amount)。该数据集包含了 1990 年 12 月 19 日至 2019 年 12 月 31 日之间共计 7101 天的股票数据,其数据形式如表 1 所示。

表 1　数据集形式

指数代码	交易日期	收盘点位	开盘点位	最高点位	最低点位	昨日收盘点	涨跌值	涨跌幅	成交量	成交额
000001.SH	19901219	99.98	96.05	99.98	95.79	100.00	-0.02	-0.0200	1260.0	494.311
000001.SH	19901220	104.39	104.30	104.39	99.98	99.98	4.41	4.4109	197.0	84.992
000001.SH	19901221	109.13	109.07	109.13	103.73	104.39	4.74	4.5407	28.0	16.096

2.2　实验流程

本文进行上证综合指数预测的流程如图 3 所示。

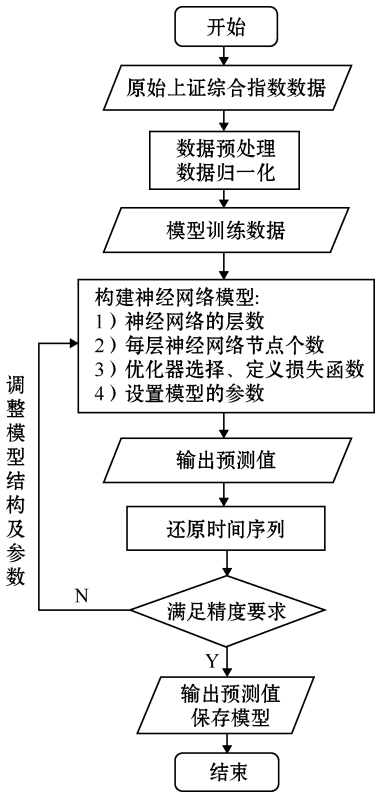

图 3　上证综合指数预测流程

在数据预处理阶段对原始数据进行简单处理,把数据集中第 2 天的收盘价作为当天的标签值,然后将数据集划分为训练集和测数集,其中训练数据占 70%,用于模型的训练,测试数据占 30%,用于测试模型泛化误差。

在数据输入到神经网络模型之前,需要对数据进行归一化处理。对数值类型的特征做归一化可以将所有的特征都统一到一个大致相同的数值区间内。数据的归一化可以消除量纲,从而避免数据对度量单位选择的依赖,并且有助于提高模型性能,有利于进行模型的训练,加快模型的收敛速度。本文采用的归一化方法为

$$x'_i = \frac{x_i - \bar{x}}{\max(x) - \min(x)} \tag{11}$$

式中:x_i 为第 i 个变量;$\bar{x}$ 为 x_i 的均值;$\max(x)$、$\min(x)$ 分别表示 x_i 的最大值和最小值。

通过使用不同的神经网络模型在训练集上训练模型并在测试集上进行测试,对比不同的神经网络模型在不同的参数设置下的模型性能。本文选取模型评价指标中的平均绝对误差(MAE)、均方误差(MSE)、平均绝对百分比误差(MAPE)以及预测数据和真实数据之间的相关系数(ρ)作为模型性能的评价指标。

平均绝对误差为

$$e_{\mathrm{MAE}} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

均方误差为

$$e_{\mathrm{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

平均百分比误差为

$$e_{\mathrm{MAPE}} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{14}$$

相关系数为

$$\rho = \frac{\mathrm{COV}(\boldsymbol{Y},\hat{\boldsymbol{Y}})}{\sqrt{\mathrm{VAR}(\boldsymbol{Y})\cdot\mathrm{VAR}(\hat{\boldsymbol{Y}})}} \tag{15}$$

式中：n 为测试数据集的数量；y_i 为第 i 个样本点的真实值；$\hat{y}_i$ 为第 i 样本点的模型预测值；$\boldsymbol{Y}$ 为样本真实值；$\hat{\boldsymbol{Y}}$ 为模型预测值；$\mathrm{COV}(\boldsymbol{Y},\hat{\boldsymbol{Y}})$ 为 $\boldsymbol{Y}$ 与 $\hat{\boldsymbol{Y}}$ 的协方差；$\mathrm{VAR}(\boldsymbol{Y})$ 为 $\boldsymbol{Y}$ 的方差；$\mathrm{VAR}(\hat{\boldsymbol{Y}})$ 为 $\hat{\boldsymbol{Y}}$ 的方差。

2.3 实验结果及分析

使用训练集训练不同的神经网络模型，然后在测试集上测试模型的性能，各模型在测试集上的预测结果如图4～9所示。

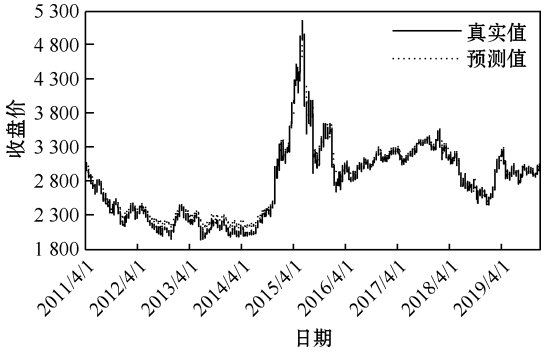

图4 LSTM 神经网络模型在测试集上的预测值

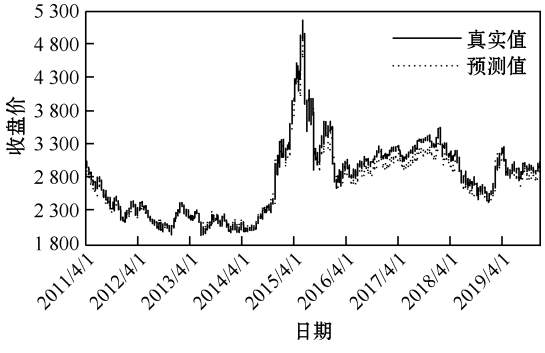

图5 BP 神经网络模型在测试集上的预测值

其中图4至图8是使用LSTM神经网络模型、BP神经网络模型、CNN模型、RNN模型、GRU神经网络模型在测试集上的预测值与真实值的对比图，图9是各个模型在测试集上的预测结果与真实值的对比图。从图中可以看出，LSTM神经网络模型和RNN模型在测试集上的预测值和真实值的曲线最为接近，模型的预测结果与真实值之间的误差较小，拟合程度较高；BP神经网络模型和GRU神经网络模型在测试集上的预测值和真实值之间的误差较大，预测值曲线与真实值曲线的拟合程度较低；CNN神经网络模型在测试集上的预测值和真实值之间的误差最大。

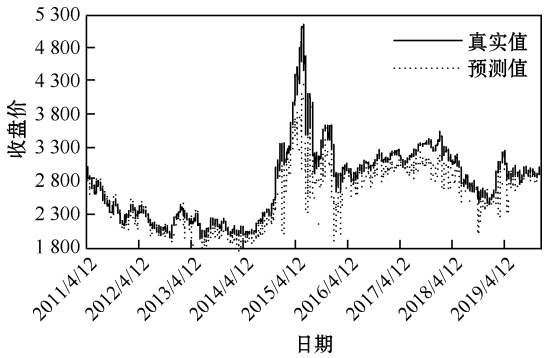

图6 CNN 模型在测试集上的预测值

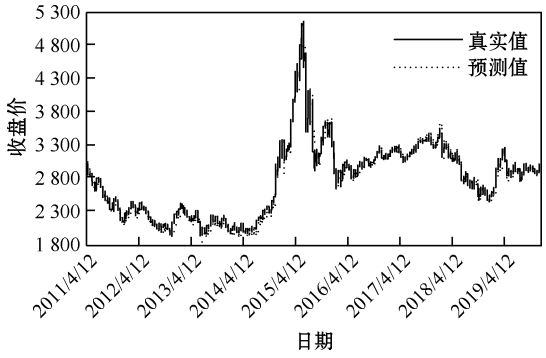

图7 RNN 模型在测试集上的预测值

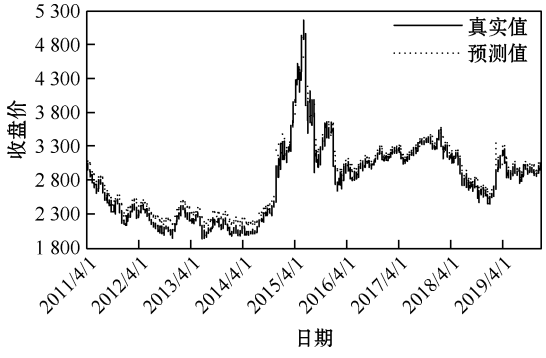

图8 GRU 神经网络模型在测试集上的预测值

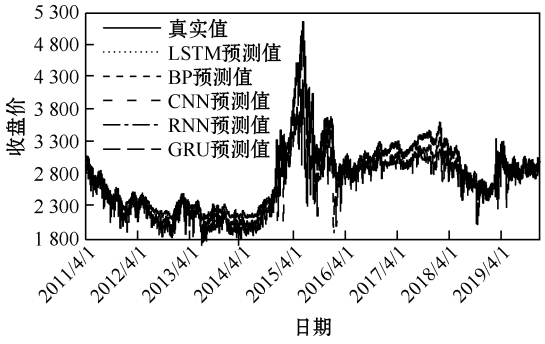

图 9　各神经网络模型在测试集上的预测值

各个模型在测试集上的模型评价指标值如表 2所示。其中 LSTM 神经网络模型的平均绝对误差为 0.015 799,均方误差为 0.000 450,平均绝对百分比误差为 0.019 867,在测试集上的预测值和真实值之间的相关系数为 0.995 7。从评价指标可以看出 LSTM 神经网络模型在测试集上的预测效果最好。

通过对以上各种模型在测试集上的预测结果和模型评价指标分析,在相同的样本数据下,LSTM 网络模型作为对 RNN 模型的改进,比 RNN 模型的预测效果更好。由于 RNN 模型考虑了时间序列上的先后关系,因此在预测效果上优于 BP 神经网络模型;而 CNN 模型表现最差,其原因在于 CNN 更适合处理图像识别、图像分类以及预测问题,在时间序列上的效果反而不佳。

表 2　不同模型在测试集上的评价指标

模型	模型结构	平均绝对误差	均方误差	平均绝对百分比误差	相关系数 ρ
1	LSTM 神经网络	0.015 799	0.000 450	0.019 867	0.995 7
2	BP 神经网络	0.028 388	0.001 256	0.030 203	0.993 2
3	CNN 网络	0.064 225	0.009 193	0.067 489	0.917 1
4	RNN 网络	0.019 350	0.000 744	0.022 844	0.990 9
5	GRU 神经网络	0.023 211	0.000 836	0.029 179	0.993 9

3　结束语

本文应用深度学习理论,基于金融时间序列数据的特性,使用 LSTM 神经网络模型进行上证综合指数的预测,将其预测结果和使用 BP 神经网络模型、CNN 模型、RNN 模型、GRU 神经网络模型的预测结果进行对比,实验结果显示 LSTM 神经网络模型在测试集上的表现效果最好。作为 RNN 模型的改进结构,LSTM 模型既继承了 RNN 模型适合处理时间序列数据问题的特点,又进一步解决了时间维度上长期依赖的问题,提高了预测的精确度,其预测效果优于 BP 神经网络、RNN、CNN、GRU 等神经网络模型。

下一步的工作将通过股票网站、论坛、贴吧等提取线上用户对股票、股价的相关评论、新闻报道、公司报表、重大事件等语言文字信息中的相关特征,将其加入到股票价格预测的模型当中,或者将 LSTM 神经网络模型与其他模型结合起来形成集成模型,以进一步提高模型的预测精确度。

参考文献:

[1] 冯盼,曹显兵. 基于 ARMA 模型的股价分析与预测的实证研究[J]. 数学的实践与认识,2011(22):84 - 90.

[2] 王丽娜,肖冬荣. 基于 ARMA 模型的经济非平稳时间序列的预测分析[J]. 武汉理工大学学报(交通科学与工程版),2004,28(1):133 - 136.

[3] 田翔,邓飞其. 精确在线支持向量回归在股指预测中的应用[J]. 计算机工程,2005,31(22):18 - 20.

[4] Wang Shuai, Shang Wei. Forecasting direction of china security index 300 movement with least squares support vector machine [J]. Procedia Computer Science, 2014, 31: 869 - 874.

[5] Wang J, Hou R, Wang C, et al. Improved v-support vector regression model based on variable selection and brain storm optimization for stock price forecasting [J]. Applied Soft Computing, 2016, 49: 164 - 178.

[6] Matthew Dixon, Diego Klabjan, Jin Hoon Bang. Classification-based financial markets prediction using deep neural networks [J]. Algorithmic Finance, 2017, 6(3 - 4): 67 - 77.

[7] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.

[8] 诸葛越. 百面机器学习[M]. 北京:人民邮电出版社,2018:236 - 250.